

# Analyse numérique et optimisation

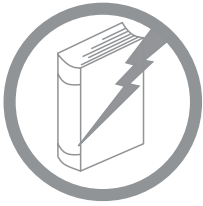
Une introduction à  
la modélisation mathématique  
et à la simulation numérique

Grégoire Allaire

DEUXIÈME  
ÉDITION



**DANGER**



**LE PHOTOCOPILLAGE  
TUE LE LIVRE**

Ce logo a pour objet d'alerter le lecteur sur la menace que représente pour l'avenir de l'écrit, tout particulièrement dans le domaine universitaire, le développement massif du « photocopillage ».

Cette pratique qui s'est généralisée, notamment dans les établissements d'enseignement, provoque une baisse brutale des achats de livres, au point que la possibilité même pour les auteurs de créer des œuvres nouvelles et de les faire éditer correctement est aujourd'hui menacée. Nous rappelons donc que la production et la vente sans autorisation, ainsi que le recel, sont passibles de poursuites.

Les demandes d'autorisation de photocopier doivent être adressées à l'éditeur ou au Centre français d'exploitation du droit de copie : 20, rue des Grands-Augustins, 75006 Paris. Tél. : 01 44 07 47 70.

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

# Table des matières

<b>1</b>	<b>INTRODUCTION A LA MODÉLISATION MATHÉMATIQUE ET A LA SIMULATION NUMÉRIQUE</b>	<b>1</b>
1.1	Introduction générale . . . . .	1
1.2	Un exemple de modélisation . . . . .	2
1.3	Quelques modèles classiques . . . . .	9
1.3.1	Équation de la chaleur . . . . .	9
1.3.2	Équation des ondes . . . . .	10
1.3.3	Le Laplacien . . . . .	12
1.3.4	Équation de Schrödinger . . . . .	12
1.3.5	Système de Lamé . . . . .	13
1.3.6	Système de Stokes . . . . .	14
1.3.7	Équations des plaques . . . . .	14
1.4	Calcul numérique par différences finies . . . . .	15
1.4.1	Principes de la méthode . . . . .	15
1.4.2	Résultats numériques pour l'équation de la chaleur . . . . .	18
1.4.3	Résultats numériques pour l'équation d'advection . . . . .	22
1.5	Remarques sur les modèles mathématiques . . . . .	26
1.5.1	Notion de problème bien posé . . . . .	27
1.5.2	Classification des équations aux dérivées partielles . . . . .	29
<b>2</b>	<b>MÉTHODE DES DIFFÉRENCES FINIES</b>	<b>31</b>
2.1	Introduction . . . . .	31
2.2	Différences finies pour l'équation de la chaleur . . . . .	32
2.2.1	Divers exemples de schémas . . . . .	32
2.2.2	Consistance et précision . . . . .	35
2.2.3	Stabilité et analyse de Fourier . . . . .	37
2.2.4	Convergence des schémas . . . . .	42
2.2.5	Schémas multiniveaux . . . . .	44
2.2.6	Le cas multidimensionnel . . . . .	46
2.3	Autres modèles . . . . .	51
2.3.1	Équation d'advection . . . . .	51
2.3.2	Équation des ondes . . . . .	59

<b>3 FORMULATION VARIATIONNELLE DES PROBLÈMES ELLIPTIQUES</b>	<b>65</b>
3.1 Généralités . . . . .	65
3.1.1 Introduction . . . . .	65
3.1.2 Formulation classique . . . . .	66
3.1.3 Le cas de la dimension un d'espace . . . . .	67
3.2 Approche variationnelle . . . . .	68
3.2.1 Formules de Green . . . . .	68
3.2.2 Formulation variationnelle . . . . .	71
3.3 Théorie de Lax-Milgram . . . . .	74
3.3.1 Cadre abstrait . . . . .	74
3.3.2 Application au Laplacien . . . . .	77
<b>4 ESPACES DE SOBOLEV</b>	<b>81</b>
4.1 Introduction et avertissement . . . . .	81
4.2 Fonctions de carré sommable et dérivation faible . . . . .	82
4.2.1 Quelques rappels d'intégration . . . . .	82
4.2.2 Dérivation faible . . . . .	83
4.3 Définition et principales propriétés . . . . .	86
4.3.1 Espace $H^1(\Omega)$ . . . . .	86
4.3.2 Espace $H_0^1(\Omega)$ . . . . .	90
4.3.3 Traces et formules de Green . . . . .	92
4.3.4 Un résultat de compacité . . . . .	97
4.3.5 Espaces $H^m(\Omega)$ . . . . .	98
4.4 Quelques compléments utiles . . . . .	101
4.4.1 Démonstration du Théorème 4.3.5 de densité . . . . .	101
4.4.2 Espace $H(div)$ . . . . .	103
4.4.3 Espaces $W^{m,p}(\Omega)$ . . . . .	105
4.4.4 Dualité . . . . .	106
4.5 Lien avec les distributions . . . . .	107
<b>5 ÉTUDE MATHÉMATIQUE DES PROBLÈMES ELLIPTIQUES</b>	<b>111</b>
5.1 Introduction . . . . .	111
5.2 Étude du Laplacien . . . . .	112
5.2.1 Conditions aux limites de Dirichlet . . . . .	112
5.2.2 Conditions aux limites de Neumann . . . . .	118
5.2.3 Coefficients variables . . . . .	125
5.2.4 Propriétés qualitatives . . . . .	129
5.3 Résolution d'autres modèles . . . . .	138
5.3.1 Système de l'élasticité linéarisée . . . . .	138
5.3.2 Équations de Stokes . . . . .	147

<b>6</b>	<b>MÉTHODE DES ÉLÉMENTS FINIS</b>	<b>151</b>
6.1	Approximation variationnelle	151
6.1.1	Introduction	151
6.1.2	Approximation interne générale	152
6.1.3	Méthode de Galerkin	155
6.1.4	Méthode des éléments finis (principes généraux)	155
6.2	Éléments finis en dimension $N = 1$	156
6.2.1	Éléments finis $\mathbb{P}_1$	156
6.2.2	Convergence et estimation d'erreur	161
6.2.3	Éléments finis $\mathbb{P}_2$	166
6.2.4	Propriétés qualitatives	168
6.2.5	Éléments finis d'Hermite	172
6.3	Éléments finis en dimension $N \geq 2$	174
6.3.1	Éléments finis triangulaires	174
6.3.2	Convergence et estimation d'erreur	186
6.3.3	Éléments finis rectangulaires	194
6.3.4	Éléments finis pour Stokes	199
6.3.5	Visualisation des résultats numériques	205
<b>7</b>	<b>PROBLÈMES AUX VALEURS PROPRES</b>	<b>209</b>
7.1	Motivation et exemples	209
7.1.1	Introduction	209
7.1.2	Résolution des problèmes instationnaires	210
7.2	Théorie spectrale	213
7.2.1	Généralités	213
7.2.2	Décomposition spectrale d'un opérateur compact	215
7.3	Valeurs propres d'un problème elliptique	217
7.3.1	Problème variationnel	217
7.3.2	Valeurs propres du Laplacien	222
7.3.3	Autres modèles	226
7.4	Méthodes numériques	229
7.4.1	Discrétisation par éléments finis	229
7.4.2	Convergence et estimations d'erreur	232
<b>8</b>	<b>PROBLÈMES D'ÉVOLUTION</b>	<b>235</b>
8.1	Motivation et exemples	235
8.1.1	Introduction	235
8.1.2	Modélisation et exemples d'équations paraboliques	236
8.1.3	Modélisation et exemples d'équations hyperboliques	237
8.2	Existence et unicité dans le cas parabolique	238
8.2.1	Formulation variationnelle	238
8.2.2	Un résultat général	240
8.2.3	Applications	245
8.3	Existence et unicité dans le cas hyperbolique	250

8.3.1	Formulation variationnelle . . . . .	250
8.3.2	Un résultat général . . . . .	251
8.3.3	Applications . . . . .	254
8.4	Propriétés qualitatives dans le cas parabolique . . . . .	257
8.4.1	Comportement asymptotique . . . . .	257
8.4.2	Principe du maximum . . . . .	259
8.4.3	Propagation à vitesse infinie . . . . .	260
8.4.4	Régularité et effet régularisant . . . . .	261
8.4.5	Équation de la chaleur dans tout l'espace . . . . .	263
8.5	Propriétés qualitatives dans le cas hyperbolique . . . . .	265
8.5.1	Réversibilité en temps . . . . .	265
8.5.2	Comportement asymptotique et équipartition de l'énergie . . . . .	266
8.5.3	Vitesse de propagation finie . . . . .	267
8.6	Méthodes numériques dans le cas parabolique . . . . .	271
8.6.1	Semi-discrétisation en espace . . . . .	271
8.6.2	Discrétisation totale en espace-temps . . . . .	272
8.7	Méthodes numériques dans le cas hyperbolique . . . . .	276
8.7.1	Semi-discrétisation en espace . . . . .	277
8.7.2	Discrétisation totale en espace-temps . . . . .	278
<b>9</b>	<b>INTRODUCTION À L'OPTIMISATION</b>	<b>283</b>
9.1	Motivation et exemples . . . . .	283
9.1.1	Introduction . . . . .	283
9.1.2	Exemples . . . . .	284
9.1.3	Définitions et notations . . . . .	290
9.1.4	Optimisation en dimension finie . . . . .	291
9.2	Existence d'un minimum en dimension infinie . . . . .	293
9.2.1	Exemples de non-existence . . . . .	293
9.2.2	Analyse convexe . . . . .	296
9.2.3	Résultats d'existence . . . . .	299
<b>10</b>	<b>CONDITIONS D'OPTIMALITÉ ET ALGORITHMES</b>	<b>303</b>
10.1	Généralités . . . . .	303
10.1.1	Introduction . . . . .	303
10.1.2	Différentiabilité . . . . .	304
10.2	Conditions d'optimalité . . . . .	309
10.2.1	Inéquations d'Euler et contraintes convexes . . . . .	309
10.2.2	Multiplicateurs de Lagrange . . . . .	312
10.3	Point-selle, théorème de Kuhn et Tucker, dualité . . . . .	324
10.3.1	Point-selle . . . . .	324
10.3.2	Théorème de Kuhn et Tucker . . . . .	325
10.3.3	Dualité . . . . .	327
10.4	Applications . . . . .	330
10.4.1	Énergie duale ou complémentaire . . . . .	330

10.4.2	Commande optimale . . . . .	332
10.4.3	Optimisation des systèmes distribués . . . . .	337
10.5	Algorithmes numériques . . . . .	339
10.5.1	Introduction . . . . .	339
10.5.2	Algorithmes de type gradient (cas sans contraintes) . . . . .	340
10.5.3	Algorithmes de type gradient (cas avec contraintes) . . . . .	343
10.5.4	Méthode de Newton . . . . .	349
<b>11</b>	<b>MÉTHODES DE LA RECHERCHE OPÉRATIONNELLE</b>	
	<b>(Rédigé en collaboration avec Stéphane Gaubert)</b>	<b>353</b>
11.1	Introduction . . . . .	353
11.2	Programmation linéaire . . . . .	354
11.2.1	Définitions et propriétés . . . . .	354
11.2.2	Algorithme du simplexe . . . . .	359
11.2.3	Algorithmes de points intérieurs . . . . .	364
11.2.4	Dualité . . . . .	364
11.3	Polyèdres entiers . . . . .	368
11.3.1	Points extrémaux de compacts convexes . . . . .	368
11.3.2	Matrices totalement unimodulaires . . . . .	371
11.3.3	Problèmes de flots . . . . .	374
11.4	Programmation dynamique . . . . .	378
11.4.1	Principe d'optimalité de Bellman . . . . .	378
11.4.2	Problème en horizon fini . . . . .	379
11.4.3	Problème du chemin de coût minimum, ou d'arrêt optimal . . . . .	382
11.5	Algorithmes gloutons . . . . .	387
11.5.1	Généralités sur les méthodes gloutonnes . . . . .	387
11.5.2	Algorithme de Kruskal pour le problème de l'arbre couvrant de coût minimum . . . . .	387
11.6	Séparation et relaxation . . . . .	390
11.6.1	Séparation et évaluation (branch and bound) . . . . .	390
11.6.2	Relaxation de problèmes combinatoires . . . . .	395
	<b>ANNEXE : ESPACES DE HILBERT</b>	<b>405</b>
	<b>ANNEXE : ANALYSE NUMÉRIQUE MATRICIELLE</b>	<b>411</b>
13.1	Résolution des systèmes linéaires . . . . .	411
13.1.1	Rappels sur les normes matricielles . . . . .	412
13.1.2	Conditionnement et stabilité . . . . .	415
13.1.3	Méthodes directes . . . . .	417
13.1.4	Méthodes itératives . . . . .	430
13.1.5	Méthode du gradient conjugué . . . . .	434
13.2	Calcul de valeurs et vecteurs propres . . . . .	442
13.2.1	Méthode de la puissance . . . . .	442
13.2.2	Méthode de Givens-Householder . . . . .	445

13.2.3 Méthode de Lanczos . . . . .	448
<b>Bibliographie</b>	<b>453</b>
<b>Index</b>	<b>456</b>
<b>Index des applications</b>	<b>460</b>
<b>Index des notations</b>	<b>461</b>



*A la mémoire de Jacques-Louis LIONS (1928-2001)*  
*Professeur à l'Ecole Polytechnique de 1966 à 1986*

## Introduction

Ce cours traite de deux sujets essentiels, mais parmi tant d'autres, en mathématiques appliquées : l'analyse 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 est l'art (ou la science, selon le point de vue) de représenter (ou de transformer) une réalité physique en des modèles abstraits 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.

Mais, tout d'abord, que sont les mathématiques appliquées ? Dire qu'il s'agit des mathématiques tournées vers les applications serait une tautologie et une fausse caractérisation. En effet, de tout temps les mathématiciens ont été inspirés pas des problèmes pratiques qu'ils ont essayé de résoudre, et cependant l'émergence des mathématiques appliquées comme discipline indépendante est relativement récente. En fait, tout a changé avec l'apparition des premiers ordinateurs au lendemain de la seconde guerre mondiale. Plus que pour tout autre discipline l'ordinateur a été une révolution pour les mathématiques : il a en effet ouvert un champ nouveau, celui de la modélisation et de la simulation. L'ordinateur a fait des mathématiques 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 la simulation numérique. Ces progrès ont aussi permis aux mathématiques de s'attaquer à des problèmes beaucoup plus complexes et concrets, issus de motivations immédiates industrielles ou scientifiques, auxquels on peut apporter des réponses à la fois qualitatives mais aussi quantitatives : c'est la modélisation mathématique.

On peut donc caractériser les mathématiques appliquées comme les mathématiques de la modélisation et de la simulation numérique. De ce point de vue, les mathématiques appliquées se situent à l'intersection de plusieurs disciplines scientifiques : mathématiques, calcul informatique, sciences physiques, chimiques, mécaniques, biologiques, économiques, et sciences de l'ingénieur (sous ce dernier vocable on regroupe usuellement les différents domaines d'applications industriels comme l'aéronautique, la production d'énergie, la finance, etc.). Le mathématicien américain Joseph Keller affirmait sous forme de boutade que les mathématiques appliquées sont "la science dont les mathématiques pures sont juste une branche". Il voulait mettre ainsi en relief le caractère pluridisciplinaire des mathématiques appliquées (mais il n'est pas exclu qu'il ait voulu aussi rendre la monnaie de leur pièce à certains mathématiciens "purs"

qui affectent de mépriser les mathématiques appliquées).

En paraphrasant le titre d'un film célèbre, mon collègue Pierre-Louis Lions prétend que les mathématiques appliquées sont caractérisées par trois choses : *Sex, Lies, and Videotapes*. Les cassettes vidéo sont bien sûr le symbole de la simulation numérique (et des jolis films qu'elle produit), les mensonges correspondent aux modèles (pas toujours fidèles à la réalité), et le sexe c'est évidemment l'analyse mathématique (moteur inépuisable des passions humaines et source de tant de plaisirs)...

Après ce (long) détour nous pouvons maintenant revenir au titre de ce cours. L'analyse numérique est donc la discipline qui conçoit et analyse les méthodes ou algorithmes de calcul numérique. Par ailleurs l'optimisation est la théorie des méthodes qui permettent d'améliorer le fonctionnement, le rendement, ou la réponse d'un système en maximisant ou minimisant des fonctions associées. C'est donc un outil essentiel pour la modélisation.

Les **objectifs de ce cours** sont de familiariser le lecteur avec les principaux modèles (qui sont souvent des équations aux dérivées partielles), leurs méthodes de résolution numérique et leur optimisation. Bien sûr, l'ambition de ce cours est 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. En effet, de nombreuses décisions industrielles ou politiques se prennent désormais sur la foi de calculs ou de simulations 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.

Le plan de ce cours est le suivant. Après un premier chapitre d'introduction aux principaux modèles "classiques" et à leur résolution numérique, le Chapitre 2 est consacré à l'étude de la méthode numérique des **différences finies**. Ces deux premiers chapitres permettent d'aller très vite vers des questions numériques essentielles qui motivent les développements théoriques qui suivront. Les Chapitres 3, 4, et 5 sont consacrés à la résolution théorique par **l'approche variationnelle** de modèles stationnaires (indépendants du temps). Ils posent aussi les bases d'une méthode numérique très importante, dite des **éléments finis**, qui est présentée en détail au Chapitre 6. La méthode des éléments finis est à la base de nombreux logiciels de calculs industriels ou académiques. Les Chapitres 7 et 8 portent sur la résolution de **problèmes instationnaires** (ou d'évolution en temps), tant du point de vue théorique que numérique. Si les 8 premiers chapitres sont dédiés à l'analyse numérique, les 3 derniers traitent **d'optimisation**. Le Chapitre 9 présente une série d'exemples concrets de problèmes d'optimisation et donne une théorie d'existence de solutions à ces problèmes. Le Chapitre 10 dérive 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. Finalement, le Chapitre 11 est une introduction à la **recherche opérationnelle**. Après avoir étudié la programmation

linéaire, nous donnons un aperçu des méthodes de l'optimisation combinatoire (c'est-à-dire de l'optimisation en variables discrètes) qui est essentielle pour la planification optimale des ressources et des tâches dans toutes les grandes entreprises. Chaque chapitre commence par une introduction qui en donne le plan et les idées principales.

L'épaisseur de ce cours ne doit pas inquiéter le lecteur : en plus des points essentiels qui seront traités dans le cours oral, le cours écrit contient de nombreux développements complémentaires qui permettent au lecteur curieux "d'aller un peu plus loin" et de faire le lien avec d'autres ouvrages ou d'autres disciplines. Il s'agit donc plus d'un ouvrage de référence que de la transcription exacte du contenu des cours magistraux.

Pour terminer cette introduction nous donnons quelques renseignements d'ordre pratique. Dans la mesure du possible ce cours s'est voulu "auto-contenu" pour éviter de trop fréquents renvois à d'autres ouvrages. Cela est particulièrement sensible pour de nombreux résultats d'analyse qui ne sont ici que des outils techniques utiles, mais pas essentiels. Les énoncer sans démonstration reviendrait à les utiliser en "boîte noire" ce qui leur donne un aspect "recette de cuisine" trop artificiel. Dans la mesure du possible, nous avons donc inclus leur démonstration, mais plus à titre d'information et pour les "démystifier" que pour l'intérêt théorique des arguments mathématiques. Afin de les distinguer nous employons pour tous ces passages difficiles, ou d'intérêt complémentaire, des caractères plus petits comme ceux-ci. Le lecteur pourra donc considérer ces passages en petits caractères comme "hors programme". *Les énoncés de résultats ou de définitions sont en caractères italiques comme ceux-ci.* Les exercices sont en caractères sans sérif comme ceux-ci. La fin d'une démonstration est indiquée par le caractère  $\square$ , tandis que la fin d'une remarque ou d'un exemple est indiquée par le caractère  $\bullet$ . Un index est disponible à la fin de l'ouvrage.

Les corrigés des exercices seront prochainement publiés. La plupart des programmes informatiques qui mettent en oeuvre les méthodes numériques étudiées, et qui ont permis de réaliser les figures de cet ouvrage, sont disponibles sur le site web [http://www.cmap.polytechnique.fr/~allaire/cours\\_X\\_annee2.html](http://www.cmap.polytechnique.fr/~allaire/cours_X_annee2.html)

où le lecteur pourra les télécharger librement. 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 l'INRIA et l'ENPC, disponible gratuitement sur le site web

<http://www.scilab.org>

tandis que les résultats de la méthode des éléments finis en dimension deux ont été obtenus à l'aide du logiciel FreeFem++ développé par F. Hecht et O. Pironneau et aussi disponible gratuitement sur le site web

<http://www.freefem.org>

Par ailleurs, la plupart des figures bidimensionnelles et la totalité des figures tridimensionnelles ont été tracées à l'aide du logiciel graphique xd3d développé par François Jouve à l'École Polytechnique et aussi disponible gratuitement sur le site web

<http://www.cmap.polytechnique.fr/~jouve/xd3d>

Indiquons une autre adresse web pour le lecteur curieux d'en savoir plus sur l'histoire des mathématiques ou la vie de certains mathématiciens cités dans ce cours

<http://www-history.mcs.st-and.ac.uk/history>

Le lecteur qui voudrait se tenir au courant des progrès et des avancées des mathématiques appliquées peut consulter avec bénéfice le site de la Société de Mathématiques Appliquées et Industrielles

<http://smai.emath.fr>

ou celui de sa consœur américaine, the Society for Industrial and Applied Mathematics

<http://www.siam.org>

Le niveau de ce cours est introductif et il n'exige aucun autre prérequis que le niveau de connaissances acquis en classes préparatoires ou en premier cycle universitaire. Reconnaissons qu'il est difficile de faire preuve de beaucoup d'originalité sur ce sujet déjà bien classique dans la littérature. En particulier, notre cours doit beaucoup à ces prédécesseurs et notamment aux cours de B. Larrouturnou, P.-L. Lions, et P.-A. Raviart auxquels il fait parfois de larges emprunts. L'auteur remercie tous ceux qui ont relu certaines parties du manuscrit, notamment Frédéric Bonnans, Bruno Després et Bertrand Maury. Une mention spéciale est due à Stéphane Gaubert, qui a participé à la rédaction du Chapitre 11, ainsi qu'à Olivier Pantz, qui a relu l'intégralité du manuscrit avec beaucoup d'attention et qui a vérifié les exercices et rédigé leur corrigé. L'auteur remercie à l'avance tous ceux qui voudront bien lui signaler les inévitables erreurs ou imperfections de cette édition, par exemple par courrier électronique à l'adresse

[gregoire.allaire@polytechnique.fr](mailto:gregoire.allaire@polytechnique.fr)

G. Allaire

Paris, le 7 Juillet 2005

La seconde édition de ce cours a permis de corriger de multiples fautes de frappe, incorrections ou petites erreurs (merci aux nombreux étudiants ou collègues qui me les ont signalées). Elle contient aussi un résultat supplémentaire sur l'équation des ondes (Proposition 8.5.3) au Chapitre 8.

G. Allaire

Paris, le 7 juillet 2012

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

## 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).

Pour l'instant nous laissons de côté un troisième aspect fondamental des mathématiques appliquées, à savoir l'analyse mathématique des modèles, sur lequel nous reviendrons un peu plus longuement dans les chapitres suivants. En pratiquant de la sorte, nous voulons en quelque sorte, motiver et justifier cette nécessaire intrusion de l'analyse mathématique. 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 mathé-

matiques appliquées sont fondamentalement **non-linéaires** (voir par exemple [14], [30]), 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 resterons souvent un peu flou dans l'argumentaire mathématique pour ne pas alourdir inutilement l'exposé. Que le lecteur rigoureux se rassure : nous reprendrons tous les concepts introduits de manière plus précise au prochain chapitre.

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  $\Omega$  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  $\Omega$ , et  $t$  la variable de temps. Dans  $\Omega$  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  $\theta$  et vaut  $c\theta$  où  $c$  est une constante physique (qui dépend du type de matériau) appelée chaleur spécifique. Pour déterminer la température  $\theta$ , nous écrivons la **loi de conservation de l'énergie** ou de la quantité de chaleur. Dans un volume élémentaire  $V$  inclus dans  $\Omega$ , 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,

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

où  $\partial V$  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

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

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

$$c \frac{\partial \theta}{\partial t} + \operatorname{div} q = f \quad (1.2)$$

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

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

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

$$q = -k \nabla \theta, \quad (1.3)$$

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

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

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

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

où  $\Delta = \operatorname{div} \nabla$  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  $\Omega$ , une relation, appelée **condition aux limites**, qui indique ce qui se passe à la frontière ou au

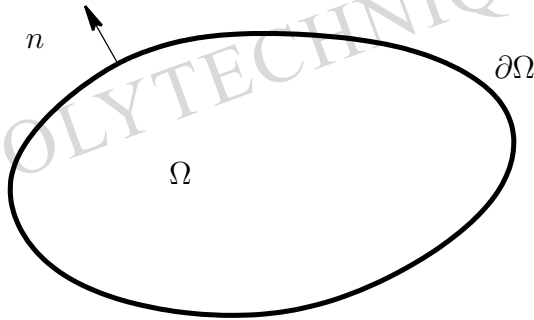


FIGURE 1.1 – Vecteur normal unité orienté vers l'extérieur.

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**

$$\theta(t = 0, x) = \theta_0(x), \quad (1.4)$$

où  $\theta_0$  est la fonction de distribution initiale de température dans le domaine  $\Omega$ . 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, quitte à modifier l'échelle des températures, la température vérifie la condition aux limites de Dirichlet

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

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

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

où  $n$  est la normale extérieure unité de  $\Omega$  (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

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

où  $\alpha$  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 satis-

faites par la température, on obtient l'équation de la chaleur

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

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** Dans ce modèle de propagation de la chaleur, il nous faut préciser les unités ou dimensions physiques : la température  $\theta$  s'exprime en Kelvin ( $K$ ), la chaleur spécifique  $c$  en Joule par kilogramme par Kelvin ( $J/(kg \times K)$ ), la conductivité thermique (par unité de masse)  $k$  en Joule mètre carré par kilogramme par Kelvin par seconde ( $Jm^2/(kg \times K \times s)$ ). D'un point de vue mathématique, nous allons très souvent oublier ces unités, et même ces constantes, en supposant que  $c$  et  $k$  valent 1 (cela revient à adimensionner les grandeurs physiques). •

**Remarque 1.2.2** Nous avons mentionné trois types de conditions aux limites, Dirichlet, Neumann, Fourier (mais il en existe d'autres) qui ont lieu sur l'intégralité de la frontière  $\partial\Omega$ . Bien sûr, on peut aisément imaginer des situations où les conditions aux limites sont mélangées : Dirichlet sur  $\partial\Omega_D$ , Neumann sur  $\partial\Omega_N$ , et Fourier sur  $\partial\Omega_F$ , avec  $\partial\Omega_D, \partial\Omega_N, \partial\Omega_F$  formant une partition de la frontière  $\partial\Omega$ . •

**Remarque 1.2.3** L'équation de la chaleur (1.8) est **linéaire** au sens où sa solution  $\theta$  dépend linéairement des données  $(f, \theta_0)$ . En physique cette propriété de linéarité est souvent traduite sous la forme d'un principe de superposition : une combinaison linéaire des données  $(f, \theta_0)$  conduit à une solution  $\theta$  qui est la même combinaison linéaire des solutions correspondant à chaque terme de la décomposition des données. D'un point de vue physique, la linéarité n'est qu'une hypothèse parmi d'autres. En effet, pour les problèmes à forte variation de température, la loi de Fourier est fautive, et il faut la corriger en supposant que la conductivité thermique  $k$  dépend en fait de la température  $\theta$  et de son gradient  $\nabla\theta$  (ce qui rend le problème non-linéaire). Encore pire, pour des phénomènes extrêmement rapides (explosion, par exemple) il est nécessaire d'abandonner le principe même de la loi de Fourier qui suppose la proportionnalité du flux de chaleur  $q$  avec le gradient de température  $\nabla\theta$ . En effet, cette hypothèse ("naturelle" à première vue) entraîne une propriété paradoxale : la chaleur se propage à une vitesse infinie dans le domaine  $\Omega$ . Nous verrons plus loin (voir la Remarque 1.2.9) comment établir ce paradoxe. Retenons pour l'instant que modéliser c'est faire des hypothèses et préciser leur domaine de validité... •

**Remarque 1.2.4** 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  $\Omega$  (imaginer un polluant diffusant dans l'atmosphère, ou bien une espèce chimique migrant dans un substrat). Dans ce cas,  $\theta$  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. •

**Remarque 1.2.5** Le problème (1.8) intervient aussi en finance où il porte le nom de **modèle de Black et Scholes**. Une variante de (1.8) permet de trouver le prix de l'option d'achat (ou call) d'une action qui vaut initialement  $x$  et qu'on pourra acheter au prix  $k$  dans un temps ultérieur  $T$ . Ce prix est la solution  $u$  de

$$\begin{cases} \frac{\partial u}{\partial t} - ru + 1/2rx \frac{\partial u}{\partial x} + 1/2\sigma^2 x^2 \frac{\partial^2 u}{\partial x^2} = 0 & \text{pour } (x, t) \in \mathbb{R} \times (0, T) \\ u(t = T, x) = \max(x - k, 0) & \text{pour } x \in \mathbb{R} \end{cases} \quad (1.9)$$

Plus précisément,  $u(0, x)$  est le prix au temps  $t = 0$  de l'option d'achat de prix d'exercice  $k$  à l'échéance  $T > 0$ , et d'actif  $x$  en  $t = 0$ . On note  $\sigma$  la volatilité de l'action et  $r$  le taux d'intérêt. Remarquons que (1.9) est un problème avec condition finale et non pas initiale, mais que le signe de la dérivée seconde en espace est opposé à celui dans (1.8). Par conséquent, après inversion du temps (1.9) est bien une équation parabolique. •

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**

$$\begin{cases} c \frac{\partial \theta}{\partial t} + cV \cdot \nabla \theta - k \Delta \theta = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ \theta = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{dans } \Omega \end{cases} \quad (1.10)$$

La différence entre (1.8) et (1.10) 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

$$\text{Pe} = \frac{cVL}{k}, \quad (1.11)$$

où  $L$  est une longueur caractéristique du problème (par exemple le diamètre du domaine  $\Omega$ ). 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 (1.10) est plus réaliste que (1.8). Par contre, si le nombre de Péclet est très grand, on peut simplifier (1.10) en supprimant le terme de diffusion. On obtient alors l'équation dite **d'advection**

$$\begin{cases} c \frac{\partial \theta}{\partial t} + cV \cdot \nabla \theta = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ \theta(t, x) = 0 & \text{pour } (x, t) \in \partial\Omega \times \mathbb{R}_*^+ \text{ si } V(x) \cdot n(x) < 0 \\ \theta(t = 0, x) = \theta_0(x) & \text{dans } \Omega \end{cases} \quad (1.12)$$

Remarquons la différence dans la condition aux limites de (1.12) par rapport à celle de (1.10) : on n'impose plus à la température  $\theta$  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), (1.10), (1.12), 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, (1.10) devient

$$\begin{cases} \frac{\partial \theta}{\partial t} + V \frac{\partial \theta}{\partial x} - \nu \frac{\partial^2 \theta}{\partial x^2} = 0 & \text{pour } (x, t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{pour } x \in \mathbb{R} \end{cases} \quad (1.13)$$

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

$$\theta(t, x) = \frac{1}{\sqrt{4\pi\nu t}} \int_{-\infty}^{+\infty} \theta_0(y) \exp\left(-\frac{(x - Vt - y)^2}{4\nu t}\right) dy. \quad (1.14)$$

Une solution de (1.8) est facilement obtenue en faisant  $V = 0$  dans l'expression (1.14).

**Exercice 1.2.1** On suppose que la donnée initiale  $\theta_0$  est continue et uniformément bornée sur  $\mathbb{R}$ . Vérifier que (1.14) est bien une solution de (1.13).

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

$$\begin{cases} \frac{\partial \theta}{\partial t} + V \frac{\partial \theta}{\partial x} = 0 & \text{pour } (x, t) \in \mathbb{R} \times \mathbb{R}_*^+ \\ \theta(t = 0, x) = \theta_0(x) & \text{pour } x \in \mathbb{R} \end{cases} \quad (1.15)$$

On vérifie que

$$\theta(t, x) = \theta_0(x - Vt) \quad (1.16)$$

est une solution de l'équation (1.15).

**Exercice 1.2.2** On suppose que la donnée initiale  $\theta_0$  est dérivable et uniformément bornée sur  $\mathbb{R}$ . Vérifier que (1.16) est bien une solution de (1.15). Montrer que (1.16) est la limite de (1.14) lorsque le paramètre  $\nu$  tend vers zéro.

**Remarque 1.2.6** Si l'on résolvait l'équation de la chaleur (1.8) sur un intervalle borné (et non dans tout l'espace), on pourrait aussi calculer une solution explicite en utilisant l'analyse de Fourier (voir le cours de mathématiques [7]). Cette solution serait un peu moins "explicite" que (1.14) car définie comme la somme d'une série infinie. Remarquons que c'est précisément pour résoudre l'équation de la chaleur que Fourier a inventé l'analyse qui porte son nom. •

**Remarque 1.2.7** Le rôle du temps est fondamentalement différent dans les équations (1.8) et (1.12). 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 (1.12) 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 : (1.16) est invariant par changement de signe de  $t$  et  $V$ , alors que (1.14) (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 1.2.8** Une autre différence fondamentale entre les équations (1.8) et (1.12) porte sur les propriétés **d'invariance par changement d'échelle**. Supposons que le terme source est nul,  $f = 0$ . Il est facile de voir que si  $\theta(x, t)$  est une solution de l'équation de la chaleur (1.8), alors, pour tout  $\lambda > 0$ ,  $\theta(\frac{x}{\lambda}, \frac{t}{\lambda^2})$  est aussi solution de la même équation (pour une donnée initiale différente). De même, en supposant que la vitesse  $V$  est constante, si  $\theta(x, t)$  est une solution de l'équation d'advection (1.12), alors  $\theta(\frac{x}{\lambda}, \frac{t}{\lambda})$  est aussi solution. On voit bien que la mise à l'échelle de la variable de temps n'est pas la même dans les deux cas. Remarquons aussi que, dans les deux cas, l'équation est invariante par translation en espace et en temps. •

**Remarque 1.2.9** 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 (1.14)). En particulier, dans le cas de (1.13), 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 (1.15) la donnée initiale est convectée à la vitesse  $V$  (voir la formule (1.16)) : il y a donc **propagation à vitesse finie**. •

**Remarque 1.2.10** Grâce aux formules explicites (1.14) et (1.16), on vérifie aisément que les solutions de l'équation de convection-diffusion (1.13) et de l'équation d'advection (1.15) 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 (1.10) et de l'équation d'advection (1.12). Nous l'étudierons précisément dans la suite. •

## 1.3 Quelques modèles classiques

Dans cette section nous donnons rapidement la forme de quelques modèles classiques. Le but de cette énumération est de dégager dès maintenant les principales classes d'équations aux dérivées partielles que nous étudierons par la suite, et de montrer que ces équations jouent un rôle important dans des domaines scientifiques très divers. Désormais nous adimensionnons toutes les variables, ce qui permet de fixer toutes les constantes des modèles égales à 1.

### 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

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

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 (voir plus loin la Sous-section 1.5.2). 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 (1.14)) de la solution de l'équation de la chaleur (1.17) 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 (1.17). 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

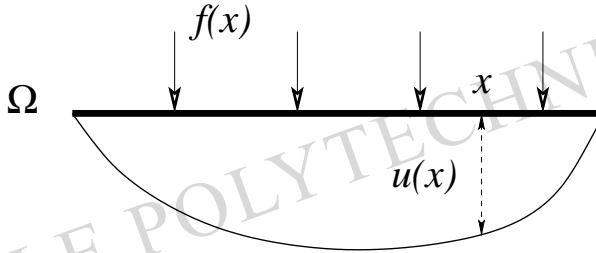


FIGURE 1.2 – Déplacement d'une corde élastique.

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  $\Omega$ . 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

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0 & \text{dans } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1 & \text{dans } \Omega \end{cases} \quad (1.18)$$

Remarquons qu'il s'agit d'une équation du deuxième ordre en temps et qu'il faut donc deux conditions initiales pour  $u$ . On dira que cette équation est hyperbolique (voir plus loin la Sous-section 1.5.2).



**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

$$u(t, x) = \frac{1}{2}(u_0(x+t) + u_0(x-t)) + \frac{1}{2}(U_1(x+t) - U_1(x-t)), \quad (1.19)$$

est la solution unique de (1.18) dans la classe des fonctions régulières.

L'équation des ondes partage avec l'équation d'advection (1.12) 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 (1.14), 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 (1.19) 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 (1.18) dans  $\Omega \times \mathbb{R}_*^-$  quitte à changer le signe de la vitesse initiale  $u_1(x)$ .

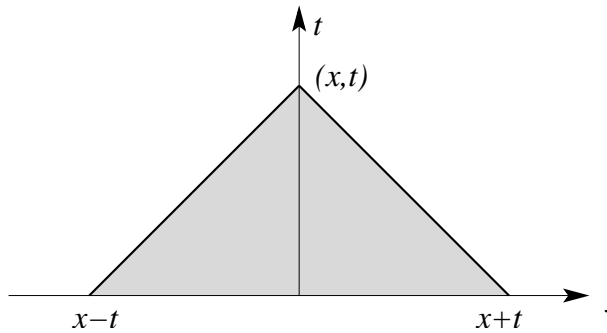


FIGURE 1.3 – Domaine ou cône de dépendance de l'équation des ondes.

**Exercice 1.3.4** On se propose de démontrer un principe de conservation de l'énergie pour l'équation des ondes (1.18) sans utiliser la formule explicite (1.19). En une dimension d'espace, on pose  $\Omega = (0, 1)$  et on suppose  $f = 0$ . Soit  $u(t, x)$  une solution régulière de

(1.18). 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 (1.17) 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

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (1.20)$$

que l'on appelle Laplacien ou équation de Laplace. On dira que cette équation est elliptique (voir plus loin la Sous-section 1.5.2). Remarquons que le Laplacien est aussi la version stationnaire de l'équation des ondes (1.19). Le Laplacien intervient aussi dans de très nombreux domaines des sciences de l'ingénieur. Par exemple, (1.20) 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

$$\begin{cases} i \frac{\partial u}{\partial t} + \Delta u - V u = 0 & \text{dans } \mathbb{R}^N \times \mathbb{R}_*^+ \\ u(t = 0) = u_0 & \text{dans } \mathbb{R}^N \end{cases} \quad (1.21)$$

Il n'y a pas de condition aux limites apparentes dans (1.21) puisque l'équation a lieu dans tout l'espace (qui n'a pas de bord). Néanmoins, nous verrons qu'un choix "raisonnable" d'espace fonctionnel dans lequel nous chercherons la solution entraîne de facto une condition de décroissance à l'infini de  $u$  qui peut s'interpréter comme une sorte de condition aux limites à l'infini.

**Exercice 1.3.5** On se propose de démontrer des principes de conservation de l'énergie pour l'équation de Schrödinger (1.21). Soit  $u(t, x)$  une solution régulière de (1.21) en une

dimension d'espace qui décroît vers zéro (ainsi que  $\frac{\partial u}{\partial x}$ ) lorsque  $|x| \rightarrow +\infty$ . Montrer que pour toute fonction dérivable  $v(t)$  on a

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

où  $\mathcal{R}$  désigne la partie réelle et  $\bar{v}$  le complexe conjugué de  $v$ . En multipliant l'équation par  $\bar{u}$  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  $\frac{\partial \bar{u}}{\partial t}$ , 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 (voir la Sous-section 5.3.1 pour plus de détails sur la modélisation). 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  $\Omega$  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  $\Omega$  dans  $\mathbb{R}^N$ , comme le déplacement  $u(x)$ . Ce dernier est solution de

$$\begin{cases} -\mu \Delta u - (\mu + \lambda) \nabla(\operatorname{div} u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (1.22)$$

où  $\lambda$  et  $\mu$  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 (1.22) 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  $\mathbb{R}^N$ , (1.22) est équivalent à

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

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

$$\begin{cases} \nabla p - \mu \Delta u = f & \text{dans } \Omega \\ \operatorname{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (1.23)$$

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  $\operatorname{div} u = 0$  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  $\Omega$  la surface moyenne de la plaque, et  $f(x)$  (une fonction de  $\Omega$  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)

$$\begin{cases} \Delta(\Delta u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega \end{cases} \quad (1.24)$$

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).

Remarquons qu'il est possible de justifier l'équation des plaques (1.24) par un raisonnement asymptotique à partir du système de Lamé (1.22) dans lequel on fait tendre l'épaisseur de la plaque vers zéro. Il s'agit là d'un exemple de modélisation mathématique.

## 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).

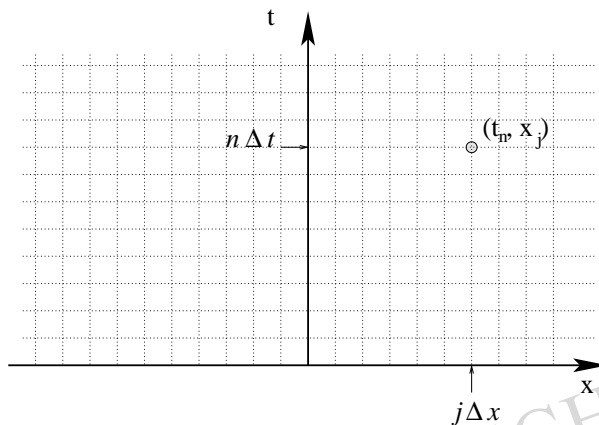


FIGURE 1.4 – Maillage en différences finies.

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

$$-\frac{\partial^2 u}{\partial x^2}(t_n, x_j) \approx \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} \quad (1.25)$$

où l'on reconnaît la formule de Taylor

$$\begin{aligned} -u(t, x - \Delta x) + 2u(t, x) - u(t, x + \Delta x) = & -(\Delta x)^2 \frac{\partial^2 u}{\partial x^2}(t, x) \\ & - \frac{(\Delta x)^4}{12} \frac{\partial^4 u}{\partial x^4}(t, x) + \mathcal{O}((\Delta x)^6) \end{aligned} \quad (1.26)$$

Si  $\Delta x$  est "petit", la formule (1.25) est une "bonne" approximation (elle est naturelle mais pas unique). La formule (1.25) est dite **centrée** car elle est symétrique en  $j$ .

Pour discrétiser l'équation de convection-diffusion

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0 \quad (1.27)$$

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**)

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

Aussi “naturel” et évident soit-il, **ce schéma est incapable de calculer des solutions approchées** de l’équation de convection-diffusion (1.27) (voir l’exemple numérique de la Figure 1.5) ! Nous justifierons cette incapacité du schéma à approcher la solution exacte dans le Lemme 2.2.23. 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

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

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**)

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

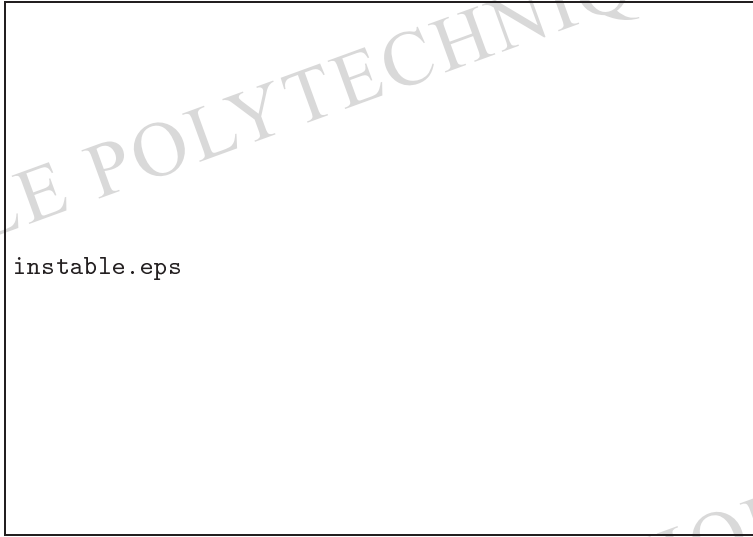
conduit au schéma

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} = 0. \quad (1.30)$$

La différence principale entre ces deux derniers schémas est que (1.29) 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 (1.30) 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 (1.29) et (1.30) n’est évidemment qu’apparent puisqu’on peut réécrire de manière équivalente (1.30) 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 (1.27). Remarquons que pour le “mauvais” schéma centré (1.28) 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).

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

### 1.4.2 Résultats numériques pour l'équation de la chaleur

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

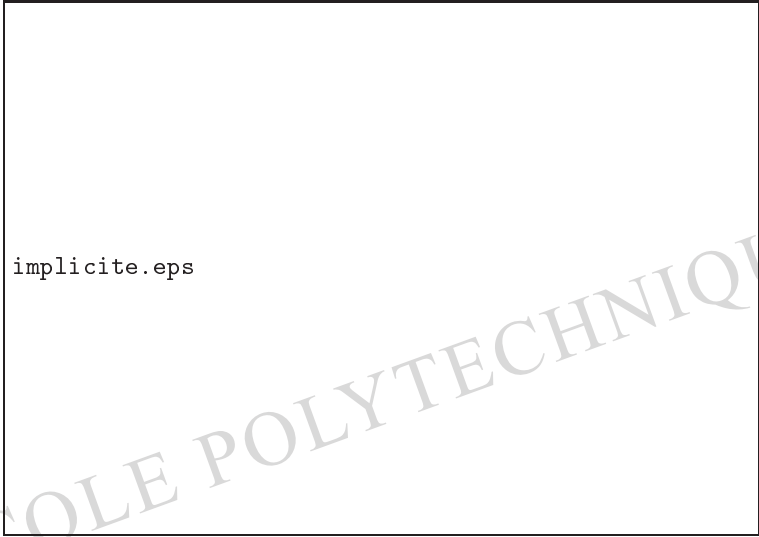
$$u_0(x) = \max(1 - x^2, 0).$$

Pour pouvoir comparer les solutions numériques approchées avec la solution exacte (1.14), 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 (1.14), et la solution approchée numérique considérée.

Réglons tout de suite le sort du schéma centré (1.28) : 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  $\Delta t$ , 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  $\Delta x$  et  $\Delta t$ . 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  $\Delta t$  et  $\Delta x$ , 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 2.2.23).



implicite.eps

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

A l'opposé du précédent schéma, le schéma implicite (1.29) calcule de “bonnes” solutions approchées de l'équation de la chaleur **quel que soit** le pas de temps  $\Delta t$  (voir la Figure 1.6). En particulier, on n'observe jamais d'oscillations numériques quel que soit le choix des pas  $\Delta t$  et  $\Delta x$ . On dit que le schéma implicite est **inconditionnellement stable**.

Considérons maintenant le schéma explicite (1.30) : des expériences numériques montrent facilement que selon les valeurs du pas de temps  $\Delta t$  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  $\Delta t$  et  $\Delta x$  qui **vérifient** la condition

$$2\nu\Delta t \leq (\Delta x)^2 \quad (1.31)$$

le schéma est stable, tandis que si (1.31) n'est pas vérifiée, alors le schéma est instable. On dit que le schéma explicite est **conditionnellement stable**. La condition de stabilité (1.31) est **une des remarques les plus simples et les plus profondes de l'analyse numérique**. Elle fut découverte en 1928 (avant l'apparition des pre-

miers ordinateurs!) par Courant, Friedrichs, et Lewy. Elle porte depuis le nom de **condition CFL ou condition de Courant, Friedrichs, et Lewy**.

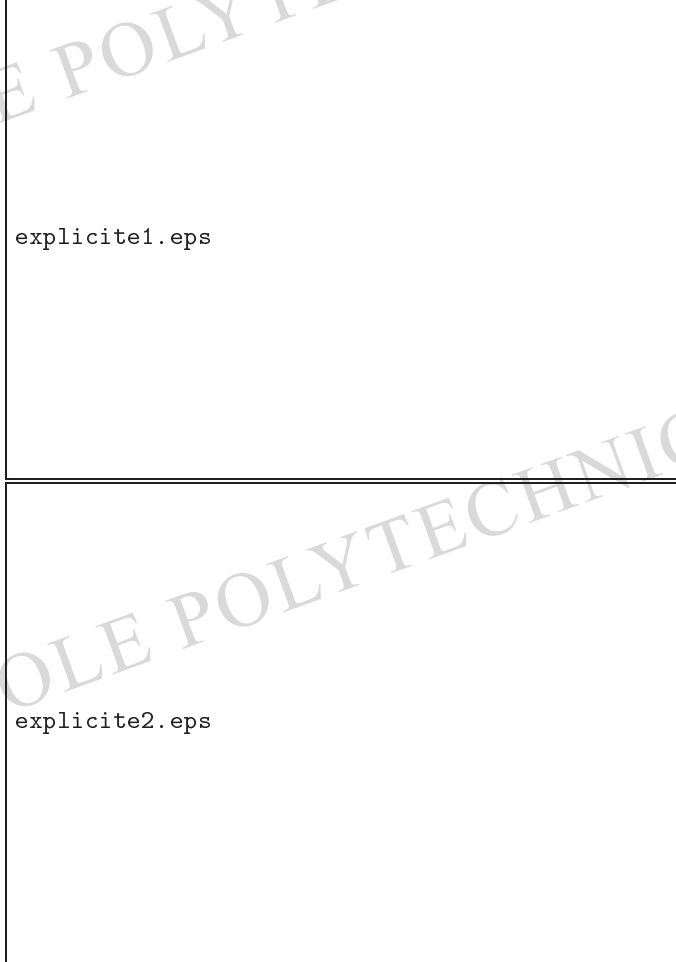


FIGURE 1.7 – Schéma explicite avec  $\nu\Delta t = 0.4(\Delta x)^2$  (haut) et  $\nu\Delta t = 0.51(\Delta x)^2$  (bas).

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

$$u_j^{n+1} = \frac{\nu\Delta t}{(\Delta x)^2} u_{j-1}^n + \left(1 - 2\frac{\nu\Delta t}{(\Delta x)^2}\right) u_j^n + \frac{\nu\Delta t}{(\Delta x)^2} u_{j+1}^n. \quad (1.32)$$

Si la condition CFL est vérifiée, alors (1.32) 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 (1.32) 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

$$m \leq u_j^n \leq M \text{ pour tout } j \in \mathbb{Z} \text{ et pour tout } n \geq 0. \quad (1.33)$$

La propriété (1.33) empêche le schéma d'osciller de manière non bornée : il est donc stable sous la condition CFL. La propriété (1.33) 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 1.2.10.

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.

**Exercice 1.4.1** Le but de cet exercice est de montrer que le schéma implicite (1.29), 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 (1.29) 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  $\Delta t$  et  $\Delta x$ ).

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

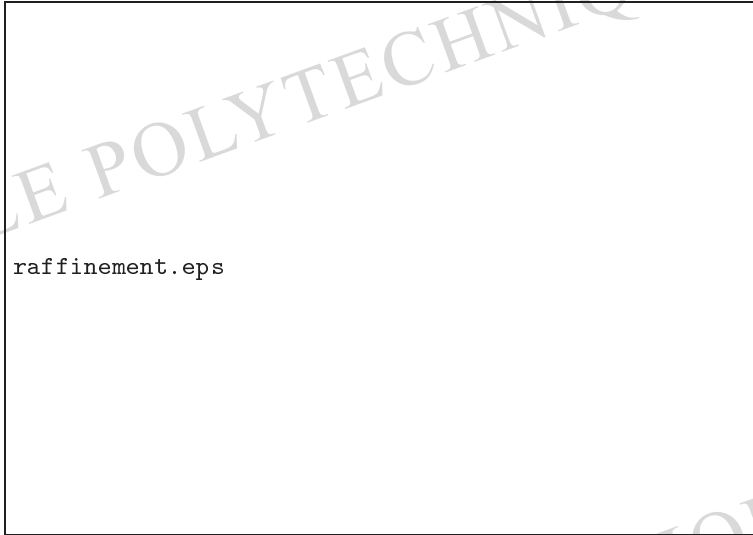


FIGURE 1.8 – Schéma explicite avec  $\nu\Delta t = 0.4(\Delta x)^2$  pour diverses valeurs de  $\Delta x$ .

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  $\Delta x$  (qui prend les valeurs 0.5, 0.1, et 0.05) ainsi que le pas de temps  $\Delta t$  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  $\Delta t$  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** (1.27) 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  $\nu$  (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  $\nu$  est petit, plus le terme convectif est prédominant sur le terme

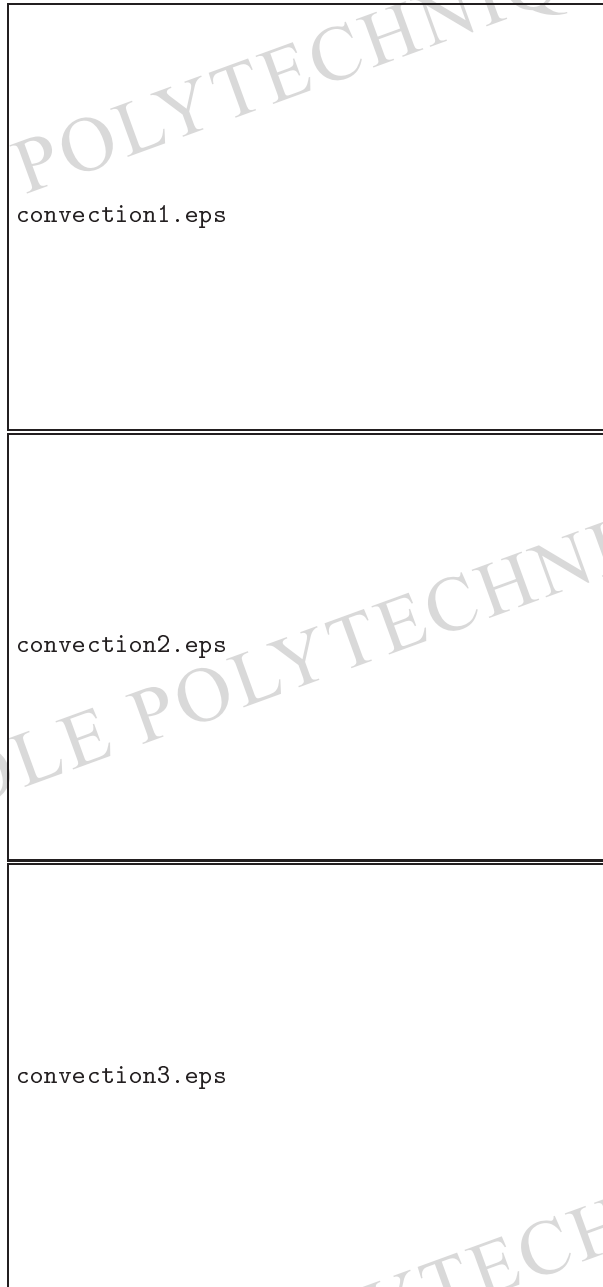


FIGURE 1.9 – Schéma explicite pour l'équation de convection-diffusion avec  $\nu\Delta t = 0.4(\Delta x)^2$  et  $V = 1$ . En haut,  $\nu = 1$ , au milieu  $\nu = 0.1$ , et en bas  $\nu = 0.01$ .

diffusif. Par conséquent, la condition CFL (1.31), obtenue lorsque la vitesse  $V$  est nulle, est de moins en moins valable au fur et à mesure que  $\nu$  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 (1.31) est automatiquement satisfaite si  $\nu = 0$  (quel que soit  $\Delta t$  et  $\Delta x$ ), ce qui semble contradictoire avec le résultat expérimental du bas de la Figure 1.9.

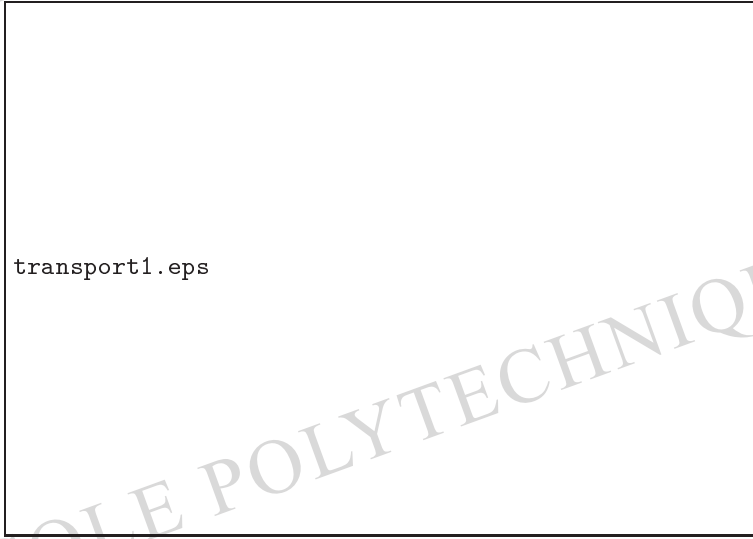


FIGURE 1.10 – Schéma explicite centré pour l'équation d'advection avec  $\Delta t = 0.9\Delta x$ ,  $V = 1$ ,  $\nu = 0$ .

Pour l'équation d'advection (c'est-à-dire (1.27) avec  $\nu = 0$ ), le schéma explicite (1.30) peut se réécrire

$$u_j^{n+1} = \frac{V\Delta t}{2\Delta x} u_{j-1}^n + u_j^n - \frac{V\Delta t}{2\Delta x} u_{j+1}^n. \quad (1.34)$$

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  $\Delta t$ ) 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 (1.34), nous avons choisi de traiter le terme convectif de manière centrée. 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”

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

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

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} = 0 \quad (1.36)$$

qui donne les résultats de la Figure 1.11. On vérifie aisément que le schéma (1.36) est

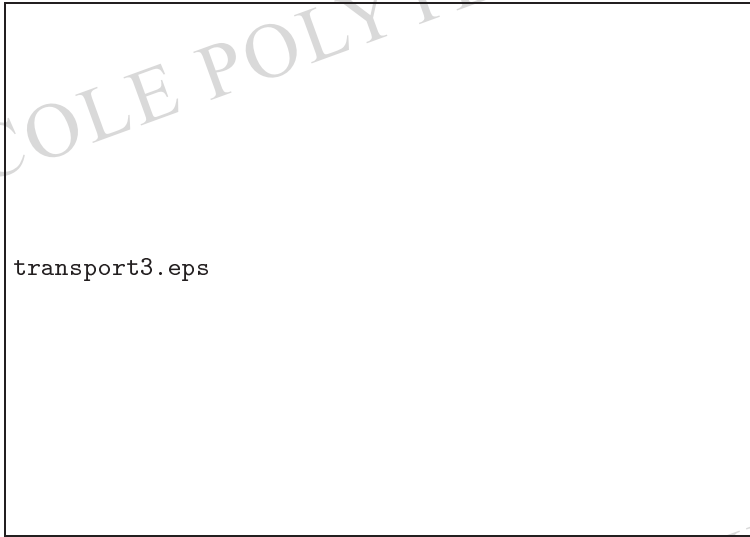


FIGURE 1.11 – Schéma explicite décentré amont pour l’équation d’advection avec  $\Delta t = 0.9\Delta x$ ,  $V = 1$ .

stable sous une nouvelle condition CFL (différente de la précédente condition CFL (1.31))

$$|V|\Delta t \leq \Delta x. \quad (1.37)$$

En effet, on peut réécrire (1.36) 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 (1.37) 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 (1.36) 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 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  $\nu$ , il faut absolument décentrer vers l'amont le terme convectif et suivre la condition CFL (1.37) plutôt que celle (1.31). A ce prix on peut améliorer les résultats de la Figure 1.9.

**Exercice 1.4.2** Montrer que, si la condition CFL (1.37) n'est pas satisfaite, le schéma décentré amont (1.36) 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 (1.18) 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 Remarques sur les modèles mathématiques

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.



### 1.5.1 Notion de problème bien posé

**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 (1.20) est un problème aux limites. A contrario, l'équation différentielle ordinaire

$$\begin{cases} \frac{dy}{dt} = f(t, y) \text{ pour } 0 < t < T \\ y(t=0) = y_0 \end{cases} \quad (1.38)$$

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 (1.38) est un problème de Cauchy, mais pas le Laplacien (1.20) (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

$$\mathcal{A}(u) = f \quad (1.39)$$

**Définition 1.5.3** On dit que le problème (1.39) 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 un point plus délicat car, s'il est clair que, lorsque en météorologie on prévoit le temps qu'il va faire demain, il vaut mieux pouvoir prédire "soleil" ou "pluie" (avec un "ou" exclusif) mais pas les deux avec des chances égales, il existe d'autres problèmes qui admettent "raisonnablement" plusieurs ou une infinité de solutions. Par exemple, les problèmes de plus court chemin admettent souvent plusieurs solutions : pour aller du pôle sud au pôle nord tout méridien convient, et de même, pour aller en avion de Paris à New York, votre agence de voyage vous fait passer tantôt par Bruxelles ou Londres, plutôt qu'un trajet direct, car plus court chemin veut dire ici plus économique. Hadamard exclut de sa définition ce type de problème car la multiplicité des solutions cache une indétermination du modèle : pour choisir finalement un chemin parmi tous ceux qui sont les plus courts, on utilise un autre critère (qu'on avait "oublié" jusque là) comme par exemple, le trajet le plus pratique ou confortable. C'est une situation courante en mathématiques appliquées : quand un modèle admet trop de solutions, il faut lui ajouter un critère de sélection de la "bonne" solution (voir l'exemple typique de la dynamique des gaz [19]). 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 (1.39) revient à perturber les données (qui de continues deviennent discrètes) et à résoudre (1.39) 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. Remarquons que cette condition est aussi très importante d'un point de vue physique car les appareils de mesure physique des données n'ont qu'une précision relative : si l'on est incapable de distinguer deux données très proches mais conduisant à des phénomènes très différents, le modèle représenté par (1.39) n'a aucune valeur prédictive, et donc un intérêt pratique à peu près nul.

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

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

Vérifier que  $u(x, y) = \frac{e^{-\sqrt{n}}}{n} \sin(ny) \operatorname{sh}(nx)$  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.

### 1.5.2 Classification des équations aux dérivées partielles

**Définition 1.5.4** On appelle **ordre** d'une équation aux dérivées partielles l'ordre de la plus grande dérivée présente dans l'équation.

Par exemple, le Laplacien (1.20) est une équation du deuxième ordre, tandis que l'équation des plaques (1.24) est une équation du quatrième ordre. On distingue souvent l'ordre par rapport à la variable de temps  $t$  et par rapport à la variable d'espace  $x$ . Ainsi, on dira que l'équation de la chaleur (1.8) est du premier ordre en temps et du deuxième ordre en espace, alors que l'équation des ondes (1.18) est du deuxième ordre en espace-temps.

Pour comprendre le vocabulaire souvent employé d'équation aux dérivées partielles **soit elliptique, soit parabolique, soit hyperbolique**, nous allons brièvement classifier les équations aux dérivées partielles linéaires du deuxième ordre portant sur des fonctions réelles de deux variables réelles  $u(x, y)$  (nous ne cherchons absolument pas à effectuer une classification systématique de toutes les e.d.p.). Une telle équation s'écrit

$$a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial x \partial y} + c \frac{\partial^2 u}{\partial y^2} + d \frac{\partial u}{\partial x} + e \frac{\partial u}{\partial y} + fu = g. \quad (1.40)$$

Pour simplifier nous supposons que les coefficients  $a, b, c, d, e, f$  sont constants.

**Définition 1.5.5** On dit que l'équation (1.40) est **elliptique** si  $b^2 - 4ac < 0$ , **parabolique** si  $b^2 - 4ac = 0$ , et **hyperbolique** si  $b^2 - 4ac > 0$ .

L'origine de ce vocabulaire est bien sûr la classification des coniques du plan, sur laquelle la Définition 1.5.5 est calquée. En effet, il est bien connu que l'équation du deuxième degré

$$ax^2 + bxy + cy^2 + dx + ey + f = 0$$

définit une courbe plane qui est (sauf certains cas dégénérés) une ellipse si  $b^2 - 4ac < 0$ , une parabole si  $b^2 - 4ac = 0$ , et une hyperbole si  $b^2 - 4ac > 0$ .

Si on applique la Définition 1.5.5 aux divers modèles du deuxième ordre que nous avons évoqué dans ce chapitre (en remplaçant le couple  $(x, y)$  par les variables  $(t, x)$  en une dimension d'espace), nous concluons que **l'équation de la chaleur est parabolique** (de même que l'équation de convection-diffusion), que **le Laplacien est elliptique**, et que **l'équation des ondes est hyperbolique**. Une généralisation adéquate de cette définition permet

d'affirmer que l'équation d'advection est hyperbolique, et que les équations de Stokes, de l'élasticité, ou des plaques sont elliptiques. En règle générale les problèmes stationnaires (indépendants du temps) sont modélisés par des e.d.p. elliptiques, tandis que les problèmes d'évolution sont modélisés par des e.d.p. paraboliques ou hyperboliques.

Nous verrons plus loin que les problèmes aux limites sont bien posés pour les équations aux dérivées partielles elliptiques, tandis que les problèmes de Cauchy en temps et aux limites en espace sont bien posés pour les équations aux dérivées partielles paraboliques ou hyperboliques. Il y a donc une différence importante de comportement entre ces types d'équations.

**Remarque 1.5.6** Le caractère elliptique, parabolique, ou hyperbolique de l'équation (1.40) n'est pas modifié par un changement de variables. Soit  $(x, y) \rightarrow (X, Y)$  un tel changement de variables non singulier, c'est-à-dire tel que son Jacobien  $J = X_x Y_y - X_y Y_x$  ne s'annule pas (on note  $Z_z$  la dérivée de  $Z$  par rapport à  $z$ ). Un calcul, simple dans le principe mais long dans le détail, montre que (1.40) devient

$$A \frac{\partial^2 u}{\partial X^2} + B \frac{\partial^2 u}{\partial X \partial Y} + C \frac{\partial^2 u}{\partial Y^2} + D \frac{\partial u}{\partial X} + E \frac{\partial u}{\partial Y} + Fu = G,$$

avec notamment  $A = aX_x^2 + bX_x X_y + cX_y^2$ ,  $B = 2aX_x Y_x + b(X_x Y_y + X_y Y_x) + 2cX_y Y_y$ ,  $C = aY_x^2 + bY_x Y_y + cY_y^2$ , et on vérifie que  $B^2 - 4AC = J^2(b^2 - 4ac)$ . En particulier, un changement de variables adéquat permet de simplifier l'équation aux dérivées partielles (1.40) pour la ramener sous sa forme "canonique". Ainsi, toute équation elliptique peut se ramener au Laplacien  $\frac{\partial^2}{\partial X^2} + \frac{\partial^2}{\partial Y^2}$ , toute équation parabolique à l'équation de la chaleur  $\frac{\partial}{\partial X} - \frac{\partial^2}{\partial Y^2}$ , et toute équation hyperbolique à l'équation des ondes  $\frac{\partial^2}{\partial X^2} - \frac{\partial^2}{\partial Y^2}$ . •

**Remarque 1.5.7** On sait bien que l'équation générale des coniques du plan admet un certain nombre de cas dégénérés où elle ne décrit plus une conique mais un ensemble de droites, voire un seul point. La même situation de dégénérescence peut avoir lieu avec l'équation aux dérivées partielles (1.40). Par exemple, l'équation  $\frac{\partial^2 u}{\partial x^2} = 1$  avec  $a = 1$  et  $b = c = d = e = f = 0$  n'est pas parabolique en dimension deux (bien que  $b^2 - 4ac = 0$ ) mais elliptique en dimension un (la variable  $y$  ne joue ici aucun rôle). Il faut donc faire un peu attention avant de conclure sur le type de ces équations "dégénérées". •

## 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.

Terminons cette introduction en disant que la méthode des différences finies est une des plus anciennes méthodes de simulation numérique qui est encore utilisée pour certaines applications, comme la propagation d'ondes (sismiques ou électromagnétiques) ou la mécanique des fluides compressibles. Pour d'autres applications, comme la mécanique du solide ou celle des fluides incompressibles, on lui préfère souvent la méthode des éléments finis. Néanmoins, de nombreux concepts en différences finies se retrouvent dans toutes les autres méthodes numériques. Ainsi, les schémas numériques du Chapitre 8 combineront des éléments finis pour la discrétisation spatiale et des différences finies pour la discrétisation temporelle. La généralité et la simplicité de la méthode des différences finies motive donc son exposition détaillée en début de cet ouvrage.

## 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  $(0, 1)$

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

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

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

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

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

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**

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

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

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

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  $N$

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

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  **$\theta$ -schéma**

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

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

$$\begin{aligned} & \frac{u_{j+1}^{n+1} - u_{j+1}^n}{12\Delta t} + \frac{5(u_j^{n+1} - u_j^n)}{6\Delta t} + \frac{u_{j-1}^{n+1} - u_{j-1}^n}{12\Delta t} \\ & + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{2(\Delta x)^2} + \nu \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{2(\Delta x)^2} = 0. \end{aligned} \quad (2.6)$$

**Exercice 2.2.1** Montrer que le schéma (2.6) n'est rien d'autre que le  $\theta$ -schéma avec  $\theta = 1/2 - (\Delta x)^2/12\nu\Delta t$ .

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 multiniveaux : 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**

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} + \nu \frac{-u_{j-1}^n + u_j^{n+1} + u_j^{n-1} - u_{j+1}^n}{(\Delta x)^2} = 0, \quad (2.7)$$

le **schéma de Gear**

$$\frac{3u_j^{n+1} - 4u_j^n + u_j^{n-1}}{2\Delta t} + \nu \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} = 0. \quad (2.8)$$

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 faut 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

$$F_{\Delta t, \Delta x} \left( \{u_{j+k}^{n+m}\}_{m^- \leq m \leq m^+, k^- \leq k \leq k^+} \right) = 0 \quad (2.9)$$

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 (2.9) 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*

$$F_{\Delta t, \Delta x} (\{u(t + m\Delta t, x + k\Delta x)\}_{m^- \leq m \leq m^+, k^- \leq k \leq k^+}), \quad (2.10)$$

*tend vers zéro, uniformément par rapport à  $(t, x)$ , lorsque  $\Delta t$  et  $\Delta x$  tendent vers zéro indépendamment.*

*De plus, on dit que le schéma est précis à l'ordre  $p$  en espace et à l'ordre  $q$  en temps si l'erreur de troncature (2.10) tend vers zéro comme  $\mathcal{O}((\Delta x)^p + (\Delta t)^q)$  lorsque  $\Delta t$  et  $\Delta x$  tendent vers zéro.*

**Remarque 2.2.5** Il faut prendre garde dans la formule (2.9) 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  $\Delta t$  et  $\Delta x$  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 (2.9) 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  $\Delta t$  et  $\Delta x$  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  $\Delta t$  et  $\Delta x$ . •

**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)

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

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  $\Delta t$  et en  $(\Delta x)^2$  se simplifient car  $v_{tt} = \nu v_{txx} = \nu^2 v_{xxxx}$ . □

Schéma	Erreur de troncature	Stabilité
Explicite (2.2)	$\mathcal{O}\left(\Delta t + (\Delta x)^2\right)$	stable $L^2$ et $L^\infty$ si condition CFL $2\nu\Delta t \leq (\Delta x)^2$
Implicite (2.3)	$\mathcal{O}\left(\Delta t + (\Delta x)^2\right)$	stable $L^2$ et $L^\infty$
Crank-Nicolson (2.5) (avec $\theta = 1/2$ )	$\mathcal{O}\left((\Delta t)^2 + (\Delta x)^2\right)$	stable $L^2$
$\theta$ -schéma (2.5) (avec $\theta \neq 1/2$ )	$\mathcal{O}\left(\Delta t + (\Delta x)^2\right)$	stable $L^2$ si condition CFL $2(1 - 2\theta)\nu\Delta t \leq (\Delta x)^2$
Schéma à 6 points (2.6)	$\mathcal{O}\left((\Delta t)^2 + (\Delta x)^4\right)$	stable $L^2$
DuFort-Frankel (2.7)	$\mathcal{O}\left((\frac{\Delta t}{\Delta x})^2 + (\Delta x)^2\right)$	stable $L^2$ si condition CFL $\Delta t/(\Delta x)^2$ borné
Gear (2.8)	$\mathcal{O}\left((\Delta t)^2 + (\Delta x)^2\right)$	stable $L^2$

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

**Exercice 2.2.2** 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 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  $\Delta x$  :

$$\|u^n\|_p = \left( \sum_{j=1}^N \Delta x |u_j^n|^p \right)^{1/p} \quad \text{pour } 1 \leq p \leq +\infty, \quad (2.11)$$

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  $\Delta x$  à travers la pondération mais aussi à travers l'entier  $N$  car  $\Delta x = 1/(N+1)$ . Grâce à la pondération par  $\Delta x$ , 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  $\|\cdot\|$ , définie par (2.11), s'il existe une constante  $K > 0$  indépendante de  $\Delta t$  et  $\Delta x$  (lorsque ces valeurs tendent vers zéro) telle que*

$$\|u^n\| \leq K \|u^0\| \quad \text{pour tout } n \geq 0, \quad (2.12)$$

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

*Si (2.12) n'a lieu que pour des pas  $\Delta t$  et  $\Delta x$  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  $\Delta x$  alors même que les normes définies par (2.11) dépendent de  $\Delta x$ .

**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

$$u^{n+1} = Au^n, \quad (2.13)$$

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

$$\begin{pmatrix} 1-2c & c & & & 0 \\ c & 1-2c & c & & \\ & \ddots & \ddots & \ddots & \\ & & c & 1-2c & c \\ 0 & & & c & 1-2c \end{pmatrix} \quad \text{avec } c = \frac{\nu \Delta t}{(\Delta x)^2}, \quad (2.14)$$

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 13.1.1)

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

la stabilité du schéma est équivalente à

$$\|A^n\| \leq K \quad \forall n \geq 0, \quad (2.15)$$

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

### Stabilité en norme $L^\infty$ .

La stabilité en norme  $L^\infty$  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 (1.33) 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  $L^\infty$  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  $L^\infty$  quelque soit les pas de temps  $\Delta t$  et d'espace  $\Delta x$  (on dit qu'il est inconditionnellement stable).*

**Exercice 2.2.3** Montrer que le schéma de Crank-Nicolson (2.5) (avec  $\theta = 1/2$ ) est stable en norme  $L^\infty$  si  $\nu\Delta t \leq (\Delta x)^2$ , et que le schéma de DuFort-Frankel (2.7) est stable en norme  $L^\infty$  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  $L^\infty$ . 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 le théorème 4.5.13 de [7]). Plus précisément on a

$$u^n(x) = \sum_{k \in \mathbb{Z}} \hat{u}^n(k) \exp(2i\pi kx), \quad (2.16)$$

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

$$\int_0^1 |u^n(x)|^2 dx = \sum_{k \in \mathbb{Z}} |\hat{u}^n(k)|^2. \quad (2.17)$$

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

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

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

$$2\nu\Delta t(\sin(\pi k\Delta x))^2 \leq (\Delta x)^2. \quad (2.18)$$

Si la condition CFL (1.31), i.e.  $2\nu\Delta t \leq (\Delta x)^2$ , est satisfaite, alors l'inégalité (2.18) 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  $\Delta x$  (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  $\pi$ ) 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  $L^2$ .*

**Remarque 2.2.16** Pour les schémas explicite (2.2) et implicite (2.3) la condition de stabilité  $L^2$  est la même que celle de stabilité  $L^\infty$ . Cela n'est pas toujours le cas pour d'autres schémas. •

**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

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

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.  $\square$

**Remarque 2.2.17** L'analyse de Fourier repose sur le choix des conditions aux limites de périodicité. On peut aussi l'effectuer si l'équation aux dérivées partielles a lieu sur tout  $\mathbb{R}$  au lieu de  $[0, 1]$  (on a alors affaire à une intégrale de Fourier plutôt qu'à une série de Fourier). Néanmoins, il n'est pas très réaliste de parler de schéma numérique sur tout  $\mathbb{R}$  puisque cela implique un nombre infini de valeurs  $u_j^n$  à chaque pas de temps  $n$  alors qu'un ordinateur ne peut que traiter un nombre fini de valeurs.

La stabilité  $L^2$  peut aussi se démontrer dans le cas de conditions aux limites de Dirichlet. Il faut alors adapter les idées de l'analyse de Fourier. Par exemple, ce qui remplace la transformée de Fourier dans ce cas est la décomposition sur une base de vecteurs propres de la matrice d'itération (2.13) qui permet de passer du vecteur  $u^n$  au vecteur  $u^{n+1}$ . •

**Remarque 2.2.18 (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

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

et on en déduit la valeur du facteur d'amplification  $A(k)$ . Rappelons que, pour l'instant, nous nous sommes limité 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

$$|A(k)| \leq 1 \text{ pour tout mode } k \in \mathbb{Z}. \quad (2.19)$$

Si la condition de stabilité de Von Neumann est satisfaite (avec éventuellement des restrictions sur  $\Delta t$  et  $\Delta x$ ), 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 2.2.4** Montrer que le  $\theta$ -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  $0 \leq \theta < 1/2$ .

**Exercice 2.2.5** Montrer que le schéma à 6 points (2.6) est inconditionnellement stable en norme  $L^2$ .

**Remarque 2.2.19** Certains auteurs utilisent une autre définition de la stabilité, moins restrictive que la Définition 2.2.8 mais plus complexe. Dans cette définition le schéma est dit stable pour la norme  $\|\cdot\|$  si pour tout temps  $T > 0$  il existe une constante  $K(T) > 0$  indépendante de  $\Delta t$  et  $\Delta x$  telle que

$$\|u^n\| \leq K(T)\|u^0\| \text{ pour tout } 0 \leq n \leq T/\Delta t,$$

quelle que soit la donnée initiale  $u^0$ . Cette nouvelle définition permet à la solution de croître avec le temps comme c'est le cas, par exemple, pour la solution de l'équation

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = cu \text{ pour } (t, x) \in \mathbb{R}^+ \times \mathbb{R},$$

qui, par le changement d'inconnue  $v(t, x) = e^{-ct}u(t, x)$ , se ramène à l'équation de la chaleur (donc pour  $c > 0$  suffisamment grand, la solution  $u$  croît exponentiellement en temps). Avec une telle définition de la stabilité, la condition de stabilité de Von Neumann devient l'inégalité

$$|A(k)| \leq 1 + C\Delta t \text{ pour tout mode } k \in \mathbb{Z}.$$

Par souci de simplicité nous préférons nous en tenir à la Définition 2.2.8 de la stabilité. •

## 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 2.2.20 (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  $\|\cdot\|$ . Alors le schéma est convergent au sens où

$$\forall T > 0, \quad \lim_{\Delta t, \Delta x \rightarrow 0} \left( \sup_{t_n \leq T} \|e^n\| \right) = 0, \quad (2.20)$$

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



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

$$\sup_{t_n \leq T} \|\epsilon^n\| \leq C_T \left( (\Delta x)^p + (\Delta t)^q \right). \quad (2.21)$$

**Remarque 2.2.21** Nous n'avons pas encore 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). Pour l'instant nous faisons donc l'hypothèse de l'existence et l'unicité d'une telle solution (ainsi que de sa régularité), mais nous verrons au Chapitre 8 que ce résultat est vrai de manière générale. •

**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 (2.13), 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  $\epsilon^n$  tel que

$$\tilde{u}^{n+1} = A\tilde{u}^n + \Delta t \epsilon^n \text{ avec } \lim_{\Delta t, \Delta x \rightarrow 0} \|\epsilon^n\| = 0, \quad (2.22)$$

et la convergence de  $\epsilon^n$  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 (2.22) à (2.13)

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

d'où par récurrence

$$e^n = A^n e^0 - \Delta t \sum_{k=1}^n A^{n-k} \epsilon^{k-1}. \quad (2.23)$$

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 (2.23) donne

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

ce qui donne l'inégalité (2.21) avec la constante  $C_T = TKC$ . La démonstration de (2.20) est similaire. □

**Remarque 2.2.22** Le Théorème de Lax 2.2.20 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 (2.21) est exactement la précision du schéma. Enfin, il est bon de noter que cette estimation (2.21) 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 (2.13). Par contre, si on pose

$$U^n = \begin{pmatrix} u^n \\ u^{n-1} \end{pmatrix}, \quad (2.24)$$

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

$$U^{n+1} = A U^n = \begin{pmatrix} A_1 & A_2 \\ \text{Id} & 0 \end{pmatrix} U^n, \quad (2.25)$$

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 (2.24). En guise d'exemple, nous démontrons un résultat pressenti au Chapitre 1.

**Lemme 2.2.23** *Le schéma centré (1.28) est instable en norme  $L^2$ .*

**Démonstration.** Avec les notations usuelles le schéma (1.28) 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

$$\|A(k)^n\|_2 = \sup_{U \in \mathbb{R}^2, U \neq 0} \frac{\|A(k)^n U\|_2}{\|U\|_2} \leq K \quad \forall n \geq 1, \quad (2.26)$$

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 (2.26) est vraie quelque soit le mode de Fourier  $k \in \mathbb{Z}$ , par la formule de Plancherel on en déduit

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

c'est-à-dire la stabilité  $L^2$  du schéma. À 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 13.1.6). Donc, l'inégalité (2.26) 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  $L^2$ .  $\square$

**Remarque 2.2.24** La méthode de l'analyse de Fourier que nous venons d'utiliser dans la démonstration du Lemme 2.2.23 est un peu plus compliquée dans le cas des schémas multiniveaux que dans le cas à deux niveaux (voir la Remarque 2.2.18). 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

$$\rho(A(k)) \leq 1 \text{ pour tout mode } k \in \mathbb{Z}, \quad (2.27)$$

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 13.1.6), donc la condition de Von Neumann (2.27) est nécessaire et suffisante (nous avons eu la “chance” lors de la démonstration du Lemme 2.2.23 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). •

**Remarque 2.2.25** Le Théorème de Lax 2.2.20 se généralise sans difficulté aux schémas multiniveaux si on choisit la norme  $L^2$ . La méthode de démonstration est inchangée : elle utilise l'analyse de Fourier et la notation vectorielle (2.24). •

**Remarque 2.2.26** Tout ce que nous venons de dire sur la stabilité et la convergence des schémas multiniveaux se généralise immédiatement aux schémas pour des systèmes d'équations. Dans ce dernier cas, on a aussi affaire à une écriture vectorielle de la relation de récurrence (2.25) et à une matrice d'amplification (au lieu d'un facteur scalaire). •

**Exercice 2.2.6** Montrer que le schéma de Gear (2.8) est inconditionnellement stable et donc convergent en norme  $L^2$ .

**Exercice 2.2.7** Montrer que le schéma de DuFort-Frankel (2.7) est stable en norme  $L^2$ , et donc convergent, si le rapport  $\Delta t/(\Delta x)^2$  reste borné lorsqu'on fait tendre  $\Delta t$  et  $\Delta x$  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

$$\begin{cases} \frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} - \nu \frac{\partial^2 u}{\partial y^2} = 0 \text{ pour } (x, y, t) \in \Omega \times \mathbb{R}_*^+ \\ u(t = 0, x, y) = u_0(x, y) \text{ pour } (x, y) \in \Omega \\ u(t, x, y) = 0 \text{ pour } t \in \mathbb{R}_*^+, (x, y) \in \partial\Omega. \end{cases} \quad (2.28)$$

Pour discrétiser le domaine  $\Omega$ , 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 (voir la Figure 2.1)

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

On note  $u_{j,k}^n$  la valeur d'une solution discrète approchée au point  $(t_n, x_j, y_k)$ , et  $u(t, x, y)$  la solution exacte de (2.28).

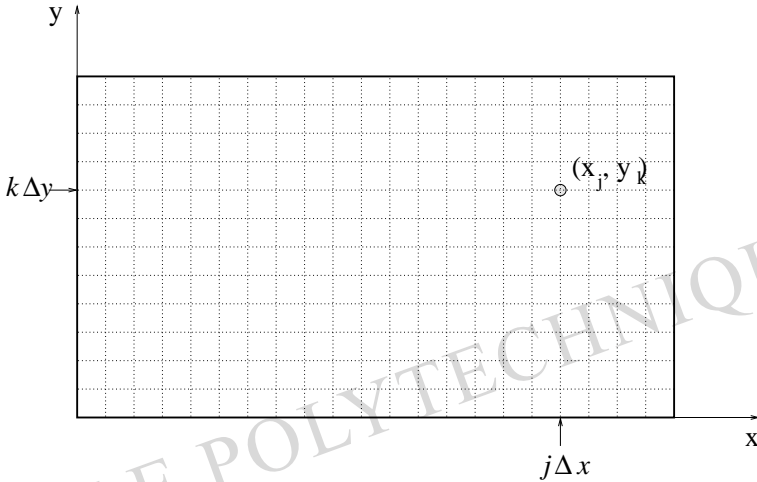


FIGURE 2.1 – Maillage d'un rectangle en différences finies.

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

$$u_{0,k}^n = u_{N_x+1,k}^n = 0, \forall k, \text{ et } u_{j,0}^n = u_{j,N_y+1}^n = 0, \forall j.$$

La donnée initiale est discrétisée par

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

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

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

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

**Exercice 2.2.8** Montrer que le schéma explicite (2.29) est stable en norme  $L^\infty$  (et même qu'il vérifie le principe du maximum) sous la condition CFL

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

Nous illustrons le schéma explicite (2.29) (auquel nous ajoutons un terme de convection) par la Figure 2.2 qui représente la convection-diffusion d'une "bosse" (le coefficient de diffusion vaut 0.01 et la vitesse (1., 0.)).

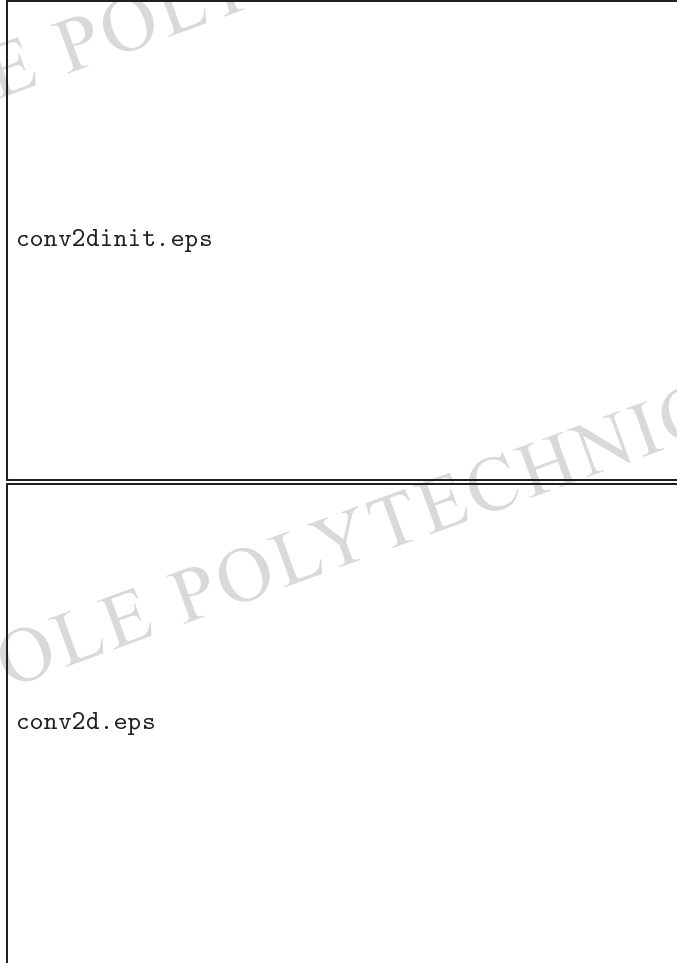


FIGURE 2.2 – Schéma explicite pour l'équation de convection-diffusion en deux dimensions : donnée initiale (haut) et solution (bas).

De même, on a le **schéma implicite**

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

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  $u_{j,k}^n$  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  $u_{j,k}^n$  est d'écrire

$$u^n = (u_{1,1}^n, \dots, u_{1,N_y}^n, u_{2,1}^n, \dots, u_{2,N_y}^n, \dots, u_{N_x,1}^n, \dots, u_{N_x,N_y}^n).$$

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 (2.30) 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 13.1.21 et 13.1.41). La situation serait encore pire en trois dimensions.

**Exercice 2.2.9** Montrer que le schéma de Peaceman-Rachford

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

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 (2.28), 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 (2.28). Par exemple, en utilisant dans chaque direction un schéma de Crank-Nicolson pour un demi pas de temps  $\Delta/2$ , on obtient un **schéma de directions alternées**

$$\begin{aligned} \frac{u_{j,k}^{n+1/2} - u_{j,k}^n}{\Delta t} + \nu \frac{-u_{j-1,k}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j+1,k}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^n + 2u_{j,k}^n - u_{j,k+1}^n}{2(\Delta y)^2} &= 0 \\ \frac{u_{j,k}^{n+1} - u_{j,k}^{n+1/2}}{\Delta t} + \nu \frac{-u_{j,k-1}^{n+1/2} + 2u_{j,k}^{n+1/2} - u_{j,k+1}^{n+1/2}}{2(\Delta x)^2} + \nu \frac{-u_{j,k-1}^{n+1} + 2u_{j,k}^{n+1} - u_{j,k+1}^{n+1}}{2(\Delta y)^2} &= 0 \end{aligned} \quad (2.31)$$

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 (2.28).

**Exercice 2.2.10** Montrer que le schéma de directions alternées (2.31) 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).

Concluons cette section par quelques considérations pratiques à propos de la méthode des différences finies. Son avantage principal est sa simplicité aussi bien conceptuelle que de mise en oeuvre informatique. Elle présente cependant un certain nombre de défauts qui, pour de nombreux problèmes complexes, lui font préférer d'autres méthodes comme celle des éléments finis (voir les Chapitres 6 et 8). Une des



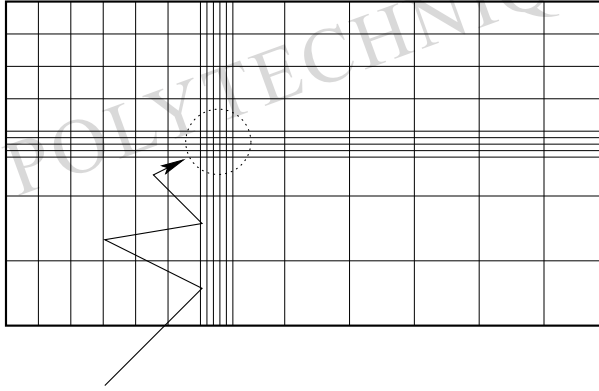


FIGURE 2.3 – Raffinement de maillage en différences finies : la zone entourée est celle où l'on veut plus de précision.

principales limitations de la méthode est qu'elle ne fonctionne que pour des maillages réguliers, dits **rectangulaires**. Il n'est pas toujours facile de paver un domaine quelconque de l'espace par des mailles rectangulaires ! Par ailleurs, il n'est pas possible de raffiner localement le maillage pour avoir une meilleure précision en un endroit précis du domaine de calcul. Il est possible de faire varier dans chaque direction le pas d'espace mais cette variation est uniforme dans les directions perpendiculaires ( $\Delta x$  et  $\Delta y$  peuvent changer le long des axes  $x$  et  $y$ , respectivement, mais cette variation est uniforme dans les directions orthogonales ; voir la Figure 2.3). Un raffinement de maillage en différences finies produit donc des mailles élancées loin de la zone d'intérêt. Par ailleurs, la théorie comme la pratique des différences finies se compliquent singulièrement lorsque les coefficients dans les équations aux dérivées partielles sont variables et lorsque les problèmes sont non-linéaires.

## 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

$$\begin{cases} \frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} = 0 \text{ pour } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(t, x+1) = u(t, x) \text{ pour } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(0, x) = u_0(x) \text{ pour } x \in (0, 1). \end{cases} \quad (2.32)$$

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  $n \geq 0, j \in \{0, 1, \dots, N+1\}$ ,

$u_j^n$  la valeur d'une solution discrète approchée au point  $(t_n, x_j)$ , et  $u(t, x)$  la solution exacte de (2.32). 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 (2.32). Au Chapitre 1 nous avons déjà constaté le mauvais comportement numérique du **schéma explicite centré**

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

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

**Lemme 2.3.1** *Le schéma explicite centré (2.33) est consistant avec l'équation d'advection (2.32), 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.  $\square$

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

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

**Exercice 2.3.1** Montrer que le schéma implicite centré (2.34) est consistant avec l'équation d'advection (2.32), 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**

$$\frac{2u_j^{n+1} - u_{j+1}^n - u_{j-1}^n}{2\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0 \quad (2.35)$$

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

**Lemme 2.3.2** *Le schéma de Lax-Friedrichs (2.35) 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  $\Delta t$  et  $\Delta x$  tendent vers zéro, il est consistant avec l'équation d'advection (2.32) 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  $u$  :

$$\begin{aligned} & \frac{2u(t_{n+1}, x_j) - u(t_n, x_{j+1}) - u(t_n, x_{j-1}))}{2\Delta t} + V \frac{u(t_n, x_{j+1}) - u(t_n, x_{j-1}))}{2\Delta x} = \\ & (u_t + V u_x)(t_n, x_j) - \frac{(\Delta x)^2}{2\Delta t} \left( 1 - \frac{(V \Delta t)^2}{(\Delta x)^2} \right) u_{xx}(t_n, x_j) + \mathcal{O}\left((\Delta x)^2 + \frac{(\Delta x)^4}{\Delta t}\right). \end{aligned} \quad (2.36)$$

Comme l'erreur de troncature contient un terme en  $\mathcal{O}\left((\Delta x)^2/\Delta t\right)$ , le schéma n'est pas consistant si  $\Delta t$  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 2.2.20. L'erreur  $e^n$  est toujours majorée par l'erreur de troncature, et donc ici

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

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

**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  $\Delta t$  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.  $\bullet$

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

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} - \left( \frac{V^2 \Delta t}{2} \right) \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} = 0. \quad (2.37)$$

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

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

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

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

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

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

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 (2.37) 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  $\alpha, \beta, \gamma$  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 (2.32) est le **décentrement amont**. Nous donnons la forme générale du **schéma décentré amont**

$$\begin{aligned} \frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_j^n - u_{j-1}^n}{\Delta x} &= 0 \quad \text{si } V > 0 \\ \frac{u_j^{n+1} - u_j^n}{\Delta t} + V \frac{u_{j+1}^n - u_j^n}{\Delta x} &= 0 \quad \text{si } V < 0. \end{aligned} \quad (2.38)$$

On a déjà vu au Chapitre 1 que le schéma décentré amont est stable en norme  $L^\infty$  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  $L^\infty$  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 (2.38) est consistant avec l'équation d'advection (2.32), 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.

**Remarque 2.3.4** Pour des problèmes non-linéaires (où la vitesse  $V$  dépend elle-même de l'inconnue  $u$ ), et notamment pour des modèles d'écoulements de fluides, le schéma décentré amont est vraiment supérieur aux autres. Il est à la source de nombreuses généralisations, beaucoup plus complexes que l'original (voir à ce sujet le cours [19]). En particulier, bien que le schéma original soit seulement d'ordre 1, il admet des variantes d'ordre 2. •

Schéma	Stabilité	Erreur de troncature
Explicite centré (2.33)	instable	$\mathcal{O}(\Delta t + (\Delta x)^2)$
Implicite centré (2.34)	stable $L^2$	$\mathcal{O}(\Delta t + (\Delta x)^2)$
Lax-Friedrichs (2.35)	stable $L^2$ et $L^\infty$ si condition CFL $ V \Delta t \leq \Delta x$	$\mathcal{O}(\Delta t + \frac{(\Delta x)^2}{\Delta t})$
Lax-Wendroff (2.37)	stable $L^2$ si condition CFL $ V \Delta t \leq \Delta x$	$\mathcal{O}((\Delta t)^2 + (\Delta x)^2)$
Décentré amont (2.38)	stable $L^2$ et $L^\infty$ si condition CFL $ V \Delta t \leq \Delta x$	$\mathcal{O}(\Delta t + \Delta x)$

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

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

**Définition 2.3.5** 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.

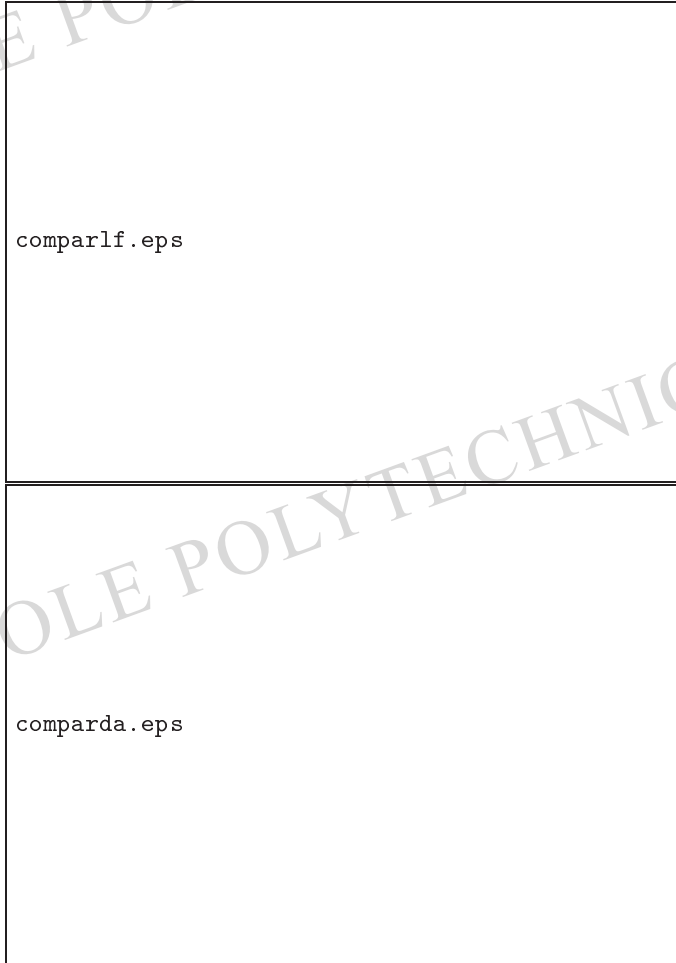


FIGURE 2.4 – Influence de la CFL sur la diffusion numérique du schéma de Lax-Friedrichs (haut) et du schéma décentré amont (bas).

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'équa-

tion d'origine. Prenons l'exemple du schéma de Lax-Friedrichs (2.35) pour l'équation d'advection : d'après (2.36), 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

$$\frac{\partial u}{\partial t} + V \frac{\partial u}{\partial x} - \nu \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{avec} \quad \nu = \frac{(\Delta x)^2}{2\Delta t} \left(1 - \frac{(V\Delta t)^2}{(\Delta x)^2}\right). \quad (2.39)$$

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 (2.39) où le coefficient de diffusion  $\nu$  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  $\nu$  peut être très grand et le schéma mauvais car trop porté à la diffusion (voir la Figure 2.4). Le coefficient de diffusion  $\nu$  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 (2.38) est

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

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  $\Delta x$  fixé), ce qui est une nette amélioration par rapport au schéma de Lax-Friedrichs (voir la Figure 2.4). Cet effet de diffusion numérique est illustré par la Figure 2.4 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  $\Delta t = 0.45\Delta x$ .

**Exercice 2.3.7** Montrer que l'équation équivalente du schéma de Lax-Wendroff (2.37) 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 2.5). 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).

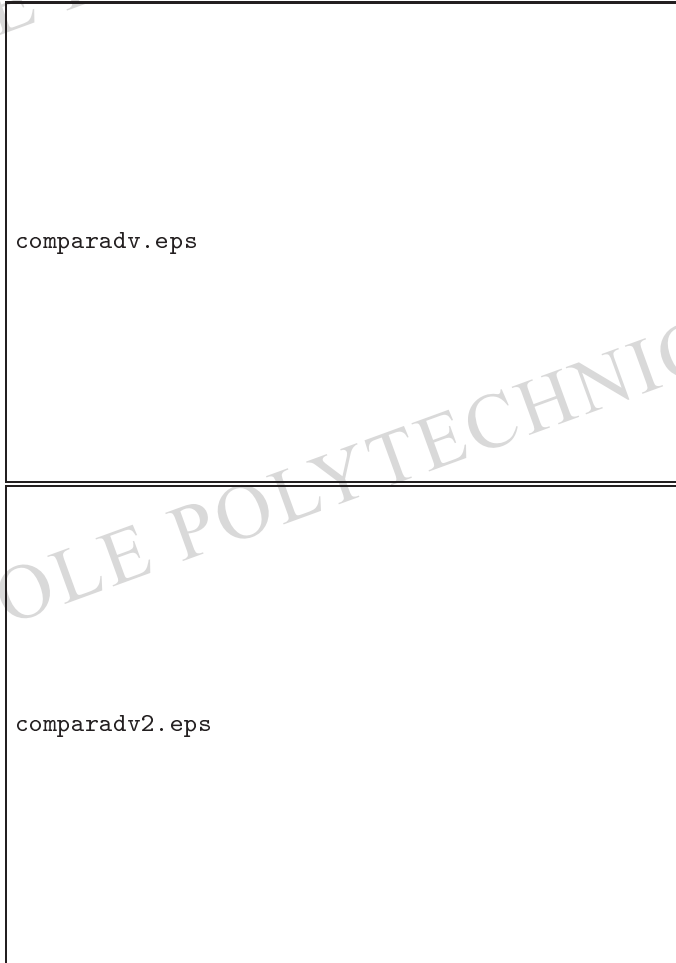


FIGURE 2.5 – Comparaison des schémas de Lax-Friedrichs, de Lax-Wendroff, et décentré amont pour une donnée initiale sinusoïdale (haut) ou en créneau (bas).

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 2.5). 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

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

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

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

(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

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \frac{\partial^2 u}{\partial x^2} = 0 \text{ pour } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(t, x+1) = u(t, x) \text{ pour } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ u(t=0, x) = u_0(x) \text{ pour } x \in (0, 1) \\ \frac{\partial u}{\partial t}(t=0, x) = u_1(x) \text{ pour } x \in (0, 1). \end{cases} \quad (2.40)$$

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 (2.40) est  $u(t, x) = Ct$ . Pour éliminer cet effet, on fait donc l'hypothèse que la vitesse initiale est de moyenne nulle

$$\int_0^1 u_1(x) dx = 0. \quad (2.41)$$

Pour l'équation des ondes (2.40) le schéma habituel est le  $\theta$ -schéma centré : pour  $n \geq 1$  et  $j \in \{0, \dots, N\}$ ,

$$\begin{aligned} & \frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{(\Delta t)^2} + \theta \frac{-u_{j-1}^{n+1} + 2u_j^{n+1} - u_{j+1}^{n+1}}{(\Delta x)^2} \\ & + (1 - 2\theta) \frac{-u_{j-1}^n + 2u_j^n - u_{j+1}^n}{(\Delta x)^2} + \theta \frac{-u_{j-1}^{n-1} + 2u_j^{n-1} - u_{j+1}^{n-1}}{(\Delta x)^2} = 0 \end{aligned} \quad (2.42)$$

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 (2.41). Comme chacune des différences finies centrées qui approchent les dérivées secondes dans (2.42) est d'ordre 2, le  $\theta$ -schéma centré (2.42) 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 2.3.6** *Si  $1/4 \leq \theta \leq 1/2$ , le  $\theta$ -schéma centré (2.42) 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

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

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  $(\lambda_1, \lambda_2)$  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 2.2.24), 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  $\lambda_1$  et  $\lambda_2$  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.  $\square$

**Exercice 2.3.11** Finir la démonstration du Lemme 2.3.6 en calculant  $A(k)^n$ , et montrer la stabilité du schéma sous condition CFL grâce à (2.41).

**Exercice 2.3.12** On considère le cas limite du Lemme 2.3.6, c'est-à-dire  $\Delta t/\Delta x = 1/\sqrt{1-4\theta}$  avec  $0 \leq \theta < 1/4$ . Montrer que le  $\theta$ -schéma centré (2.42) 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).

Nous illustrons ces schémas par la Figure 2.6 sur laquelle sont représentées les résultats obtenus avec le schéma centré explicite et le  $\theta$ -schéma implicite (pour  $\theta = 0.25$ ). Les calculs sont 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  $\Delta t = 0.9 * \Delta x$ , et un temps final  $T = 5$ . La conditions initiale  $u_0$  est un sinus, tandis que  $u_1$  est nulle.

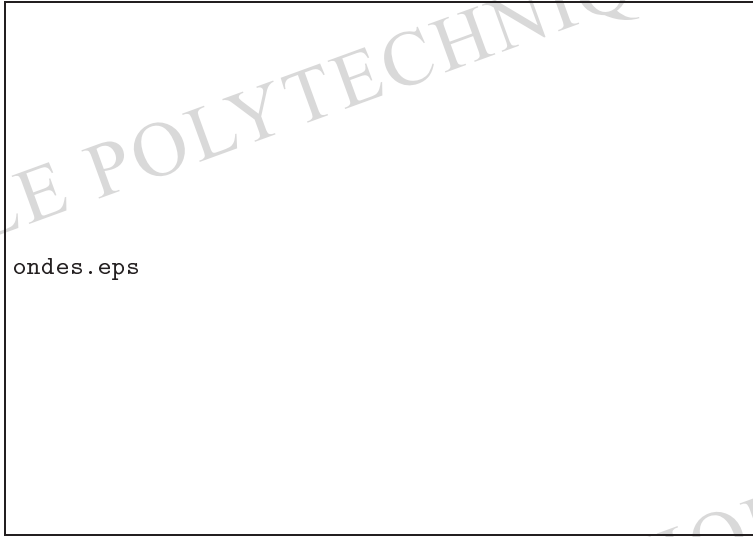


FIGURE 2.6 – Schémas pour l'équation des ondes.

Nous avons vu lors de l'Exercice 1.3.4 que l'équation des ondes (2.40) vérifie une propriété de conservation de l'énergie, c'est-à-dire que, pour tout  $t > 0$ ,

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

Il est souvent désirable qu'un schéma numérique vérifie (exactement ou approximativement) une version discrète de cette conservation de l'énergie. Pour le  $\theta$ -schéma on introduit **l'énergie discrète**

$$E^{n+1} = \sum_{j=0}^N \left( \frac{u_j^{n+1} - u_j^n}{\Delta t} \right)^2 + a_{\Delta x}(u^{n+1}, u^n) + \theta a_{\Delta x}(u^{n+1} - u^n, u^{n+1} - u^n)$$

avec

$$a_{\Delta x}(u, v) = \sum_{j=0}^N \left( \frac{u_{j+1} - u_j}{\Delta x} \right) \left( \frac{v_{j+1} - v_j}{\Delta x} \right).$$

Clairement,  $E^{n+1}$  est une approximation, à  $\mathcal{O}(\Delta x + \Delta t)$  près, de l'énergie exacte  $E(t_{n+1})$ . Nous laissons au lecteur le soin de démontrer une propriété de conservation de l'énergie discrète.

**Exercice 2.3.13** Montrer que le  $\theta$ -schéma centré (2.42) conserve l'énergie discrète, c'est-à-dire que  $E^n = E^0$  pour tout  $n \geq 0$ .

Une autre façon de définir des schémas pour l'équation des ondes est de commencer par réécrire (2.40) comme un système d'équations du premier ordre. Introduisant  $v = \frac{\partial u}{\partial t}$  et  $w = \frac{\partial u}{\partial x}$ , (2.40) est équivalent à

$$\begin{cases} \frac{\partial}{\partial t} \begin{pmatrix} v \\ w \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \frac{\partial}{\partial x} \begin{pmatrix} v \\ w \end{pmatrix} \text{ pour } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ v(t, x+1) = v(t, x), w(t, x+1) = w(t, x) \text{ pour } (x, t) \in (0, 1) \times \mathbb{R}_*^+ \\ w(t=0, x) = \frac{\partial u_0}{\partial x}(x) \text{ pour } x \in (0, 1) \\ v(t=0, x) = u_1(x) \text{ pour } x \in (0, 1). \end{cases} \quad (2.43)$$

On peut donner une interprétation physique ou mécanique de ces nouvelles variables. Si  $u$  modélise un déplacement (celui de la corde vibrante, par exemple), alors  $v$  est une vitesse et  $w$  est une déformation. Le système de deux équations (2.43) apparaît comme une généralisation de l'équation d'advection. On peut ainsi définir un schéma de type **Lax-Friedrichs**

$$\frac{1}{2\Delta t} \begin{pmatrix} 2v_j^{n+1} - v_{j+1}^n - v_{j-1}^n \\ 2w_j^{n+1} - w_{j+1}^n - w_{j-1}^n \end{pmatrix} - \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_{j+1}^n - v_{j-1}^n \\ w_{j+1}^n - w_{j-1}^n \end{pmatrix} = 0, \quad (2.44)$$

ou bien un schéma de type **Lax-Wendroff**

$$\begin{aligned} \frac{1}{\Delta t} \begin{pmatrix} v_j^{n+1} - v_j^n \\ w_j^{n+1} - w_j^n \end{pmatrix} - \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_{j+1}^n - v_{j-1}^n \\ w_{j+1}^n - w_{j-1}^n \end{pmatrix} \\ + \frac{\Delta t}{2(\Delta x)^2} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}^2 \begin{pmatrix} -v_{j-1}^n + 2v_j^n - v_{j+1}^n \\ -w_{j-1}^n + 2w_j^n - w_{j+1}^n \end{pmatrix} = 0. \end{aligned} \quad (2.45)$$

**Exercice 2.3.14** Montrer que le schéma de Lax-Friedrichs (2.44) est stable en norme  $L^2$  sous la condition CFL  $\Delta t \leq \Delta x$ , et qu'il est précis à l'ordre 1 en espace et temps si le rapport  $\Delta t/\Delta x$  est gardé constant lorsque  $\Delta t$  et  $\Delta x$  tendent vers zéro.

**Exercice 2.3.15** Montrer que le schéma de Lax-Wendroff (2.45) est stable en norme  $L^2$  sous la condition CFL  $\Delta t \leq \Delta x$ , et qu'il est précis à l'ordre 2 en espace et temps.

Comme pour l'équation d'advection, une idée fondamentale pour obtenir de "bons" schémas est le **décentrement amont**. Cependant, ici il s'agit d'un système de deux équations et il n'est pas clair de savoir quelle est la vitesse qui permet de savoir dans quel sens décentrer. En fait, il suffit de diagonaliser la matrice

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

pour obtenir deux équations d'advection découplées qui sont décentrées indépendamment et différemment l'une de l'autre. Ce sont donc les valeurs propres de la matrice

$J$  (1 et  $-1$  en l'occurrence) qui jouent le rôle de la vitesse, et on décentre composante par composante dans la décomposition sur une base de vecteurs propres de cette matrice. Ce type de schéma est systématiquement utilisé pour les systèmes hyperboliques et, en particulier, pour la dynamique des gaz (voir [19] auquel nous renvoyons pour plus de détails).

## Chapitre 3

# FORMULATION VARIATIONNELLE DES PROBLÈMES ELLIPTIQUES

### 3.1 Généralités

#### 3.1.1 Introduction

Dans ce chapitre nous nous intéressons à l'analyse mathématique des **équations aux dérivées partielles de type elliptique** (voir la Définition 1.5.5). En règle générale ces équations elliptiques correspondent à des modèles physiques stationnaires, c'est-à-dire indépendants du temps. Nous allons montrer que les problèmes aux limites sont bien posés pour ces e.d.p. elliptiques, c'est-à-dire qu'elles admettent une solution, unique, et dépendant continûment des données. L'approche que nous allons suivre est appelée **approche variationnelle**. Disons tout de suite que l'intérêt de cette approche dépasse, et de loin, le cadre des e.d.p. elliptiques et même le cadre d'analyse mathématique "pure" auquel nous nous restreignons pour l'instant. En effet, nous reprendrons cette approche variationnelle pour les problèmes d'évolution en temps (e.d.p. de type parabolique ou hyperbolique), et elle sera cruciale pour comprendre la méthode numérique des éléments finis que nous développerons au Chapitre 6. Par ailleurs, cette approche admet une interprétation physique ou mécanique très naturelle. Autant dire que le lecteur ne peut pas faire l'économie de la présentation qui suit de cette approche variationnelle!

Au cours de ce chapitre et des suivants, l'exemple prototype d'équation aux dérivées partielles de type elliptique sera le Laplacien pour lequel nous étudierons le problème aux limites suivant

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

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),  $\Omega$  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. Bien sûr, nous donnerons au Chapitre 5 de nombreux autres exemples d'équations aux dérivées partielles de type elliptique qui peuvent s'étudier grâce à l'approche variationnelle.

Le plan de ce chapitre est le suivant. Dans la Section 3.2 nous rappelons quelques formules d'intégration par parties, dites **formules de Green**, puis nous définissons ce qu'est une **formulation variationnelle**. La Section 3.3 est consacrée au **théorème de Lax-Milgram** qui sera l'outil essentiel permettant de démontrer des résultats d'existence et d'unicité de solutions de formulation variationnelle. Nous verrons que pour pouvoir appliquer ce théorème il est inéluctable de devoir abandonner l'espace  $C^1(\overline{\Omega})$  des fonctions continûment différentiables au profit de sa "généralisation", l'espace de Sobolev  $H^1(\Omega)$ .

Concluons cette introduction en mentionnant l'existence d'autres méthodes de résolution des équations aux dérivées partielles mais qui sont moins puissantes ou plus compliquées que l'approche variationnelle (nous renvoyons à l'encyclopédie [18] le lecteur curieux... et courageux).

### 3.1.2 Formulation classique

La formulation "classique" de (3.1), qui pourrait paraître "naturelle" à première vue, est de supposer suffisamment de régularité pour la solution  $u$  afin que les équations de (3.1) aient un sens en tout point de  $\Omega$  ou de  $\partial\Omega$ . Rappelons tout d'abord quelques notations d'espaces de fonctions régulières.

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

Une **solution classique** (on parle aussi de **solution forte**) de (3.1) est une solution  $u \in C^2(\Omega) \cap C(\overline{\Omega})$ , ce qui implique que le second membre  $f$  doit appartenir à  $C(\Omega)$ . Cette formulation classique pose malheureusement un certain nombre de problèmes! Sans rentrer dans le détail, signalons que, sous la seule hypothèse  $f \in C(\overline{\Omega})$ , il n'existe en général pas de solution de classe  $C^2$  pour (3.1) si la dimension d'espace est plus grande que deux ( $N \geq 2$ ). En fait, il existe bien une solution, comme nous le verrons plus loin, mais elle n'est pas de classe  $C^2$  (elle est un peu moins régulière sauf si la donnée  $f$  est plus régulière que  $C(\overline{\Omega})$ ). Le cas de la dimension un d'espace ( $N = 1$ ) est particulier puisqu'il est facile de trouver des solutions classiques (voir l'Exercice 3.1.1), mais nous verrons néanmoins que, même dans ce cas favorable, cette formulation classique est peu commode.

Dans la suite, pour étudier (3.1), nous remplacerons sa formulation classique par une formulation, dite variationnelle, beaucoup plus avantageuse.



### 3.1.3 Le cas de la dimension un d'espace

En une dimension d'espace ( $N = 1$ ), si  $\Omega = (0, 1)$ , le problème aux limites (3.1) devient

$$\begin{cases} -\frac{d^2u}{dx^2} = f & \text{pour } 0 < x < 1 \\ u(0) = u(1) = 0. \end{cases} \quad (3.2)$$

Ce problème est tellement simple qu'il admet une solution explicite !

**Exercice 3.1.1** Si  $f$  est une fonction continue sur  $[0, 1]$ , montrer que (3.2) a une solution unique dans  $C^2([0, 1])$  donnée par la formule

$$u(x) = x \int_0^1 f(s)(1-s)ds - \int_0^x f(s)(x-s)ds \text{ pour } x \in [0, 1]. \quad (3.3)$$

Pour le reste de cette sous-section nous allons oublier la formule explicite (3.3) qui n'a pas toujours d'équivalent pour des problèmes plus compliqués.

En une dimension d'espace, l'appellation "équation aux dérivées partielles" perd de sa justesse puisque, comme il n'y a plus qu'une seule variable, on peut plus simplement parler "d'équation différentielle ordinaire". Cependant, l'équation (3.2) n'est pas une équation différentielle "usuelle" au sens où la solution doit satisfaire des conditions "aux deux bouts" plutôt qu'une condition initiale en une seule extrémité de l'intervalle  $[0, 1]$ . C'est là précisément la différence entre un problème aux limites (avec des conditions "aux deux bouts") et un problème de Cauchy (avec une condition initiale en "un seul bout").

Il est intéressant cependant de voir pourquoi, même en dimension un, les méthodes classiques d'équations différentielles ordinaires ne sont pas très commodées pour étudier (3.2) (et sont totalement inopérantes en dimension supérieure). Pour un paramètre  $m \in \mathbb{R}$ , on considère le problème de Cauchy pour le Laplacien avec donnée initiale en 0

$$\begin{cases} -\frac{d^2u}{dx^2} = f \text{ pour } 0 < x < 1 \\ u(0) = 0, \quad \frac{du}{dx}(0) = m. \end{cases} \quad (3.4)$$

De façon évidente il existe une unique solution de (3.4) : il suffit d'intégrer cette équation linéaire (ou plus généralement d'utiliser le théorème d'existence de Cauchy-Lipschitz). Il n'est pas du tout clair, par contre, que la solution de (3.4) coïncide avec celle de (3.2) (si elle existe). La question qui se pose est de savoir s'il existe un paramètre  $m$  tel que la solution de (3.4) vérifie aussi  $u(1) = 0$  et soit donc une solution de (3.2). C'est le principe de la **méthode du tir** qui permet de résoudre, aussi bien d'un point de vue théorique que numérique, le problème aux limites (3.2). Itérativement, on prédit une valeur de  $m$  (on tire du point 0), on intègre le problème de Cauchy (3.4) (on calcule la trajectoire du tir), puis selon le résultat  $u(1)$  on corrige la valeur de  $m$ . En pratique c'est une méthode peu efficace qui a l'inconvénient majeur de ne pas se généraliser en dimension supérieure.

La conclusion est qu'il faut des méthodes propres aux problèmes aux limites qui n'ont rien à voir avec celles attachées aux problèmes de Cauchy.

## 3.2 Approche variationnelle

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.

### 3.2.1 Formules de Green

Dans toute cette sous-section  $\Omega$  est un ouvert de l'espace  $\mathbb{R}^N$  (borné ou non), dont le bord (ou la frontière) est noté  $\partial\Omega$ . Nous supposons aussi que  $\Omega$  est un ouvert **régulier** de classe  $\mathcal{C}^1$ . La définition précise d'un ouvert régulier est donné plus bas dans la Définition 3.2.5, mais sa connaissance n'est absolument pas nécessaire pour la bonne compréhension de la suite de ce cours. Il suffit juste de savoir qu'un ouvert régulier est *grosso modo* un ouvert dont le bord est une hypersurface (une variété de dimension  $N - 1$ ) régulière, et que cet ouvert est localement situé d'un seul coté 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  $\Omega$  et pointant vers l'extérieur de  $\Omega$  (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 théorème suivant que nous admettrons (voir le théorème 7.6.2 dans [28] pour une démonstration qui fait appel à des arguments de géométrie différentielle, ou le théorème 5.4.9 dans [7]).

**Théorème 3.2.1 (Formule de Green)** *Soit  $\Omega$  un ouvert régulier de classe  $\mathcal{C}^1$ . Soit  $w$  une fonction de  $\mathcal{C}^1(\overline{\Omega})$  à support borné dans le fermé  $\overline{\Omega}$ . Alors elle vérifie la formule de Green*

$$\int_{\Omega} \frac{\partial w}{\partial x_i}(x) dx = \int_{\partial\Omega} w(x) n_i(x) ds, \quad (3.5)$$

où  $n_i$  est la  $i$ -ème composante de la normale extérieure unité de  $\Omega$ .

**Remarque 3.2.2** Dire qu'une fonction régulière  $w$  a son support borné dans le fermé  $\overline{\Omega}$  veut dire qu'elle s'annule à l'infini si le fermé n'est pas borné. On dit aussi que la fonction  $w$  a un support compact dans  $\overline{\Omega}$  (attention : cela n'implique pas que  $w$  s'annule sur le bord  $\partial\Omega$ ). En particulier, l'hypothèse du Théorème 3.2.1 à propos du support borné de la fonction  $w$  dans  $\overline{\Omega}$  est inutile si l'ouvert  $\Omega$  est borné. Si  $\Omega$  n'est pas borné, cette hypothèse assure que les intégrales dans (3.5) sont finies. •

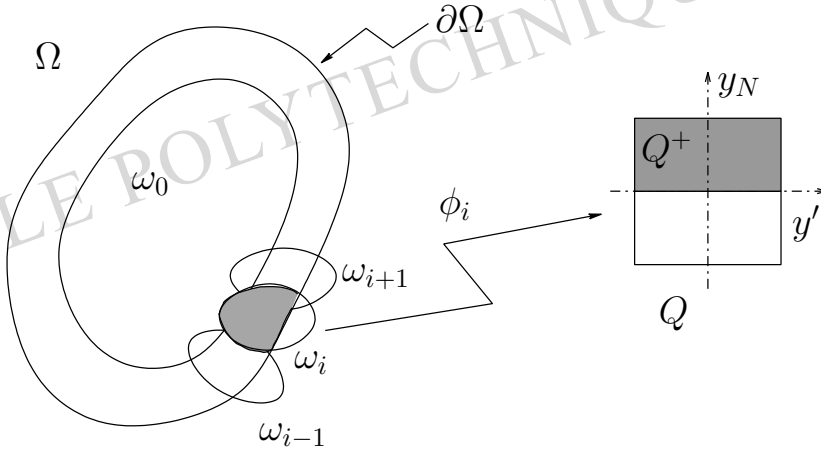


FIGURE 3.1 – Définition de la régularité d'un ouvert.

Le Théorème 3.2.1 a de nombreux corollaires qui sont tous des conséquences immédiates de la formule de Green (3.5). Le lecteur qui voudra économiser sa mémoire ne retiendra donc que la formule de Green (3.5) !

**Corollaire 3.2.3 (Formule d'intégration par parties)** Soit  $\Omega$  un ouvert régulier de classe  $C^1$ . Soit  $u$  et  $v$  deux fonctions de  $C^1(\overline{\Omega})$  à support borné dans le fermé  $\overline{\Omega}$ . Alors elles vérifient la formule d'intégration par parties

$$\int_{\Omega} u(x) \frac{\partial v}{\partial x_i}(x) dx = - \int_{\Omega} v(x) \frac{\partial u}{\partial x_i}(x) dx + \int_{\partial\Omega} u(x) v(x) n_i(x) ds. \quad (3.6)$$

**Démonstration.** Il suffit de prendre  $w = uv$  dans le Théorème 3.2.1.  $\square$

**Corollaire 3.2.4** Soit  $\Omega$  un ouvert régulier de classe  $C^1$ . Soit  $u$  une fonction de  $C^2(\overline{\Omega})$  et  $v$  une fonction de  $C^1(\overline{\Omega})$ , toutes deux à support borné dans le fermé  $\overline{\Omega}$ . Alors elles vérifient la formule d'intégration par parties

$$\int_{\Omega} \Delta u(x) v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) ds, \quad (3.7)$$

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

**Démonstration.** On applique le Corollaire 3.2.3 à  $v$  et  $\frac{\partial u}{\partial x_i}$  et on somme en  $i$ .  $\square$

**Définition 3.2.5** On dit qu'un ouvert  $\Omega$  de  $\mathbb{R}^N$  est régulier de classe  $C^k$  (avec un entier  $k \geq 1$ ) s'il existe un nombre fini d'ouverts  $(\omega_i)_{0 \leq i \leq I}$  tels que

$$\overline{\omega_0} \subset \Omega, \quad \overline{\Omega} \subset \bigcup_{i=0}^I \omega_i, \quad \partial\Omega \subset \bigcup_{i=1}^I \omega_i,$$

et que, pour chaque  $i \in \{1, \dots, I\}$  (voir la Figure 3.1), il existe une application bijective  $\phi_i$  de classe  $C^k$ , de  $\omega_i$  dans l'ensemble

$$Q = \{y = (y', y_N) \in \mathbb{R}^{N-1} \times \mathbb{R}, |y'| < 1, |y_N| < 1\},$$

dont l'inverse est aussi de classe  $C^k$ , et telle que

$$\begin{aligned}\phi_i(\omega_i \cap \Omega) &= Q \cap \{y = (y', y_N) \in \mathbb{R}^{N-1} \times \mathbb{R}, y_N > 0\} = Q^+, \\ \phi_i(\omega_i \cap \partial\Omega) &= Q \cap \{y = (y', y_N) \in \mathbb{R}^{N-1} \times \mathbb{R}, y_N = 0\}.\end{aligned}$$

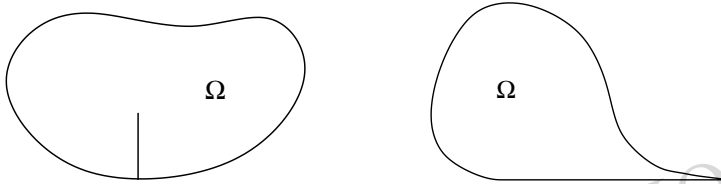


FIGURE 3.2 – Deux exemples d'ouvert non régulier : ouvert fissuré à gauche, ouvert avec un point de rebroussement à droite.

**Remarque 3.2.6** Bien que la Figure 3.1 représente un ouvert régulier qui est borné, la Définition 3.2.5 s'applique aussi à des ouverts non bornés. La Définition 3.2.5 n'exclut pas seulement les ouverts dont le bord n'est pas une surface régulière, mais elle exclut aussi les ouverts qui ne sont pas localement situés d'un seul côté de leur frontière. La Figure 3.2 contient deux exemples typiques d'ouvert non régulier qui présentent une singularité irrémédiable, soit le long de la fissure, soit en un point de rebroussement. Ces exemples ne sont pas des “inventions mathématiques” : l'ouvert fissuré est typiquement utilisé pour étudier les problèmes de fissures en mécanique des structures. On peut néanmoins généraliser un peu la classe des ouverts réguliers aux ouverts “réguliers par morceaux”, à condition que ces morceaux de frontières se “recollent” en formant des angles différents de 0 (cas d'un point de rebroussement) ou de  $2\pi$  (cas d'une fissure). Tous ces détails dépassent largement le cadre de ce cours, et nous renvoyons le lecteur à la Remarque 4.3.7 pour une autre explication sur ces problèmes de régularité. •

**Exercice 3.2.1** Dédurre de la formule de Green (3.5) la formule de Stokes

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

où  $\phi$  est une fonction scalaire de  $C^1(\overline{\Omega})$  et  $\sigma$  une fonction à valeurs vectorielles de  $C^1(\overline{\Omega})$ , à supports bornés dans le fermé  $\overline{\Omega}$ .

**Exercice 3.2.2** En dimension  $N = 3$  on définit le rotationnel d'une fonction de  $\Omega$  dans  $\mathbb{R}^3$ ,  $\phi = (\phi_1, \phi_2, \phi_3)$ , comme la fonction de  $\Omega$  dans  $\mathbb{R}^3$  définie par

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

Pour  $\phi$  et  $\psi$ , fonctions à valeurs vectorielles de  $C^1(\overline{\Omega})$ , à supports bornés dans le fermé  $\overline{\Omega}$ , déduire de la formule de Green (3.5)

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

### 3.2.2 Formulation variationnelle

Pour simplifier la présentation, nous supposons que l'ouvert  $\Omega$  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 3.2.7** Soit  $u$  une fonction de  $C^2(\overline{\Omega})$ . Soit  $X$  l'espace défini par

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

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

$$\int_{\Omega} \nabla u(x) \cdot \nabla v(x) \, dx = \int_{\Omega} f(x)v(x) \, dx \text{ pour toute fonction } v \in X. \quad (3.8)$$

L'égalité (3.8) est appelée la **formulation variationnelle** du problème aux limites (3.1).

**Remarque 3.2.8** Un intérêt immédiat de la formulation variationnelle (3.8) 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 (3.8) que (3.1) puisqu'on est moins exigeant sur la régularité de la solution.

Dans la formulation variationnelle (3.8), 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 (3.8), 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  $\Omega$  (le terme de gauche de (3.8)) et une énergie potentielle associée à  $f$  (le terme de droite de (3.8)).

**Démonstration.** Si  $u$  est solution du problème aux limites (3.1), on multiplie l'équation par  $v \in X$  et on utilise la formule d'intégration par parties du Corollaire 3.2.4

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

Or  $v = 0$  sur  $\partial\Omega$  puisque  $v \in X$ , 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 (3.8). Réciproquement, si  $u \in X$  vérifie (3.8), en utilisant "à l'envers" la formule d'intégration par parties précédente on obtient

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

Comme  $(\Delta u + f)$  est une fonction continue, grâce au Lemme 3.2.9 on conclut que  $-\Delta u(x) = f(x)$  pour tout  $x \in \Omega$ . Par ailleurs, comme  $u \in X$ , 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).  $\square$

**Lemme 3.2.9** Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$ . Soit  $g(x)$  une fonction continue dans  $\Omega$ . Si pour toute fonction  $\phi$  de  $C^\infty(\Omega)$  à support compact dans  $\Omega$ , on a

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

alors la fonction  $g$  est nulle dans  $\Omega$ .

**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,  $\phi$  à support inclus dans  $\omega$ . 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  $x \in \Omega$ .  $\square$

**Remarque 3.2.10** En notation compacte on peut réécrire la formulation variationnelle (3.8) sous la forme : trouver  $u \in X$  tel que

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

avec

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx$$

et

$$L(v) = \int_{\Omega} f(x)v(x) dx,$$

où  $a(\cdot, \cdot)$  est une forme bilinéaire sur  $X$  et  $L(\cdot)$  est une forme linéaire sur  $X$ . C'est sous cette forme abstraite que nous résoudrons (avec quelques hypothèses) la formulation variationnelle dans la prochaine section. •

L'idée principale de l'**approche variationnelle** est de montrer l'existence et l'unicité de la solution de la formulation variationnelle (3.8), ce qui entraînera le même résultat pour l'équation (3.1) à cause de la Proposition 3.2.7. En effet, nous allons voir qu'il existe une théorie à la fois simple et puissante pour analyser les formulations variationnelles. Néanmoins cette théorie ne fonctionne que si l'espace dans lequel on cherche la solution et dans lequel on prend les fonctions tests (dans les notations précédentes, l'espace  $X$ ) est un espace de Hilbert, ce qui n'est pas le cas pour  $X = \{v \in C^1(\overline{\Omega}), v = 0 \text{ sur } \partial\Omega\}$  muni du produit scalaire "naturel" pour ce problème. La principale difficulté dans l'application de l'approche variationnelle sera donc qu'il faudra utiliser un autre espace que  $X$ , à savoir l'espace de Sobolev  $H_0^1(\Omega)$  qui est bien un espace de Hilbert (voir le Chapitre 4).

**Exercice 3.2.3** On considère le Laplacien avec condition aux limites de Neumann

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega. \end{cases} \quad (3.9)$$

Soit  $u$  une fonction de  $C^2(\overline{\Omega})$ . Montrer que  $u$  est une solution du problème aux limites (3.9) si et seulement si  $u$  appartient à  $C^1(\overline{\Omega})$  et vérifie l'égalité

$$\int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx = \int_{\Omega} f(x)v(x) dx \text{ pour toute fonction } v \in C^1(\overline{\Omega}). \quad (3.10)$$

En déduire qu'une condition nécessaire d'existence d'une solution dans  $C^2(\overline{\Omega})$  de (3.9) est que  $\int_{\Omega} f(x)dx = 0$ .

**Exercice 3.2.4** On considère l'équation des plaques

$$\begin{cases} \Delta(\Delta u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega \end{cases} \quad (3.11)$$

On note  $X$  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 (3.11) si et seulement si  $u$  appartient à  $X$  et vérifie l'égalité

$$\int_{\Omega} \Delta u(x) \Delta v(x) dx = \int_{\Omega} f(x)v(x) dx \text{ pour toute fonction } v \in X. \quad (3.12)$$

### 3.3 Théorie de Lax-Milgram

#### 3.3.1 Cadre abstrait

Nous décrivons une théorie abstraite pour obtenir l'existence et l'unicité de la solution d'une formulation variationnelle dans un espace de Hilbert. On note  $V$  un espace de Hilbert réel de produit scalaire  $\langle \cdot, \cdot \rangle$  et de norme  $\| \cdot \|$ . Suivant la Remarque 3.2.10 nous considérons une formulation variationnelle du type :

$$\text{trouver } u \in V \text{ tel que } a(u, v) = L(v) \text{ pour toute fonction } v \in V. \quad (3.13)$$

Les hypothèses sur  $a$  et  $L$  sont

1.  $L(\cdot)$  est une forme linéaire continue sur  $V$ , c'est-à-dire que  $v \rightarrow L(v)$  est linéaire de  $V$  dans  $\mathbb{R}$  et il existe  $C > 0$  tel que

$$|L(v)| \leq C\|v\| \text{ pour tout } v \in V;$$

2.  $a(\cdot, \cdot)$  est une forme bilinéaire sur  $V$ , c'est-à-dire que  $w \rightarrow a(w, v)$  est une forme linéaire de  $V$  dans  $\mathbb{R}$  pour tout  $v \in V$ , et  $v \rightarrow a(w, v)$  est une forme linéaire de  $V$  dans  $\mathbb{R}$  pour tout  $w \in V$ ;
3.  $a(\cdot, \cdot)$  est continue, c'est-à-dire qu'il existe  $M > 0$  tel que

$$|a(w, v)| \leq M\|w\|\|v\| \text{ pour tout } w, v \in V; \quad (3.14)$$

4.  $a(\cdot, \cdot)$  est **coercive** (ou elliptique), c'est-à-dire qu'il existe  $\nu > 0$  tel que

$$a(v, v) \geq \nu\|v\|^2 \text{ pour tout } v \in V. \quad (3.15)$$

Comme nous le verrons au cours de cette sous-section, toutes les hypothèses ci-dessus sont nécessaires pour pouvoir résoudre (3.13). En particulier, la coercivité de  $a(\cdot, \cdot)$  est essentielle.

**Théorème 3.3.1 (Lax-Milgram)** *Soit  $V$  un espace de Hilbert réel,  $L(\cdot)$  une forme linéaire continue sur  $V$ ,  $a(\cdot, \cdot)$  une forme bilinéaire continue coercive sur  $V$ . Alors la formulation variationnelle (3.13) admet une unique solution. De plus cette solution dépend continûment de la forme linéaire  $L$ .*

**Démonstration.** Pour tout  $w \in V$ , l'application  $v \rightarrow a(w, v)$  est une forme linéaire continue sur  $V$  : par conséquent, le Théorème 12.1.18 de représentation de Riesz entraîne qu'il existe un élément de  $V$ , noté  $A(w)$ , tel que

$$a(w, v) = \langle A(w), v \rangle \text{ pour tout } v \in V.$$

Par ailleurs, la bilinéarité de  $a(w, v)$  implique évidemment la linéarité de l'application  $w \rightarrow A(w)$ . De plus, en prenant  $v = A(w)$ , la continuité (3.14) de  $a(w, v)$  montre que

$$\|A(w)\|^2 = a(w, A(w)) \leq M\|w\|\|A(w)\|,$$



c'est-à-dire que  $\|A(w)\| \leq M\|w\|$  et donc  $w \rightarrow A(w)$  est continue. Une autre application du Théorème 12.1.18 de représentation de Riesz implique qu'il existe un élément de  $V$ , noté  $f$ , tel que  $\|f\|_V = \|L\|_{V'}$  et

$$L(v) = \langle f, v \rangle \text{ pour tout } v \in V.$$

Finalement, le problème variationnel (3.13) est équivalent à :

$$\text{trouver } u \in V \text{ tel que } A(u) = f. \quad (3.16)$$

Pour démontrer le théorème il nous faut donc montrer que l'opérateur  $A$  est bijectif de  $V$  dans  $V$  (ce qui implique l'existence et l'unicité de  $u$ ) et que son inverse est continu (ce qui prouve la dépendance continue de  $u$  par rapport à  $L$ ).

La coercivité (3.15) de  $a(w, v)$  montre que

$$\nu\|w\|^2 \leq a(w, w) = \langle A(w), w \rangle \leq \|A(w)\|\|w\|,$$

ce qui donne

$$\nu\|w\| \leq \|A(w)\| \text{ pour tout } w \in V, \quad (3.17)$$

c'est-à-dire que  $A$  est injectif. Pour montrer que  $A$  est surjectif, c'est-à-dire que  $\text{Im}(A) = V$  (ce qui n'est pas évident si  $V$  est de dimension infinie), il suffit de montrer que  $\text{Im}(A)$  est fermé dans  $V$  et que  $\text{Im}(A)^\perp = \{0\}$ . En effet, dans ce cas on voit que  $V = \{0\}^\perp = (\text{Im}(A)^\perp)^\perp = \overline{\text{Im}(A)} = \text{Im}(A)$ , ce qui prouve bien que  $A$  est surjectif. Soit  $A(w_n)$  une suite dans  $\text{Im}(A)$  qui converge vers  $b$  dans  $V$ . En vertu de (3.17) on a

$$\nu\|w_n - w_p\| \leq \|A(w_n) - A(w_p)\|$$

qui tend vers zéro quand  $n$  et  $p$  tendent vers l'infini. Donc  $w_n$  est une suite de Cauchy dans l'espace de Hilbert  $V$ , c'est-à-dire qu'elle converge vers une limite  $w \in V$ . Alors, par continuité de  $A$  on en déduit que  $A(w_n)$  converge vers  $A(w) = b$ , c'est-à-dire que  $b \in \text{Im}(A)$  et  $\text{Im}(A)$  est donc fermé. D'autre part, soit  $v \in \text{Im}(A)^\perp$ ; la coercivité (3.15) de  $a(w, v)$  implique que

$$\nu\|v\|^2 \leq a(v, v) = \langle A(v), v \rangle = 0,$$

c'est-à-dire que  $v = 0$  et  $\text{Im}(A)^\perp = \{0\}$ , ce qui prouve que  $A$  est bijectif. Soit  $A^{-1}$  son inverse : l'inégalité (3.17) avec  $w = A^{-1}(v)$  prouve que  $A^{-1}$  est continu, donc la solution  $u$  dépend continûment de  $f$ .  $\square$

**Remarque 3.3.2** Si l'espace de Hilbert  $V$  est de dimension finie (ce qui n'est cependant jamais le cas pour les applications que nous visons), la démonstration du Théorème 3.3.1 de Lax-Milgram se simplifie considérablement. En effet, en dimension finie toutes les applications linéaires sont continues et l'injectivité (3.17) de  $A$  est équivalent à son inversibilité. On voit bien dans ce cas (comme dans le cas général) que l'hypothèse de coercivité de la forme bilinéaire  $a(w, v)$  est essentielle puisque c'est elle qui donne l'injectivité de  $A$ . Remarquons pour finir que, si  $V = \mathbb{R}^N$ , une formulation variationnelle n'est que l'écriture,  $\langle Au, v \rangle = \langle f, v \rangle$  pour tout  $v \in \mathbb{R}^N$ , d'un simple système linéaire  $Au = f$ .  $\bullet$

**Remarque 3.3.3** Une autre démonstration (un peu moins technique mais qui camoufle un peu les arguments essentiels) du Théorème 3.3.1 de Lax-Milgram est la suivante. On démarre comme précédemment jusqu'à la formulation (3.16) du problème. Pour montrer l'existence et l'unicité de la solution  $u$  de (3.16), on introduit une application affine  $T$  de  $V$  dans  $V$ , définie par

$$T(w) = w - \mu(A(w) - f) \text{ avec } \mu = \frac{\nu}{M^2},$$

dont on va montrer qu'elle est strictement contractante, ce qui prouve l'existence et l'unicité de  $u \in V$  tel que  $T(u) = u$  (d'où le résultat). En effet, on a

$$\begin{aligned} \|T(v) - T(w)\|^2 &= \|v - w - \mu A(v - w)\|^2 \\ &= \|v - w\|^2 - 2\mu \langle A(v - w), v - w \rangle + \mu^2 \|A(v - w)\|^2 \\ &= \|v - w\|^2 - 2\mu a(v - w, v - w) + \mu^2 \|A(v - w)\|^2 \\ &\leq (1 - 2\mu\nu + \mu^2 M^2) \|v - w\|^2 \\ &\leq (1 - \nu^2/M^2) \|v - w\|^2. \end{aligned}$$

•

Une formulation variationnelle possède souvent une interprétation physique, en particulier si la forme bilinéaire est symétrique. En effet dans ce cas, la solution de la formulation variationnelle (3.13) réalise le **minimum d'une énergie** (très naturelle en physique ou en mécanique).

**Proposition 3.3.4** *On se place sous les hypothèses du Théorème 3.3.1 de Lax-Milgram. On suppose en plus que la forme bilinéaire est symétrique  $a(w, v) = a(v, w)$  pour tout  $v, w \in V$ . Soit  $J(v)$  l'énergie définie pour  $v \in V$  par*

$$J(v) = \frac{1}{2}a(v, v) - L(v). \quad (3.18)$$

*Soit  $u \in V$  la solution unique de la formulation variationnelle (3.13). Alors  $u$  est aussi l'unique point de minimum de l'énergie, c'est-à-dire que*

$$J(u) = \min_{v \in V} J(v).$$

*Réciproquement, si  $u \in V$  est un point de minimum de l'énergie  $J(v)$ , alors  $u$  est la solution unique de la formulation variationnelle (3.13).*

**Démonstration.** Si  $u$  est solution de la formulation variationnelle (3.13), on développe (grâce à la symétrie de  $a$ )

$$J(u + v) = J(u) + \frac{1}{2}a(v, v) + a(u, v) - L(v) = J(u) + \frac{1}{2}a(v, v) \geq J(u).$$

Comme  $u + v$  est quelconque dans  $V$ ,  $u$  minimise bien l'énergie  $J$  dans  $V$ . Réciproquement, soit  $u \in V$  tel que

$$J(u) = \min_{v \in V} J(v).$$

Pour  $v \in V$  on définit une fonction  $j(t) = J(u + tv)$  de  $\mathbb{R}$  dans  $\mathbb{R}$  (il s'agit d'un polynôme du deuxième degré en  $t$ ). Comme  $t = 0$  est un minimum de  $j$ , on en déduit que  $j'(0) = 0$  qui, par un calcul simple, est exactement la formulation variationnelle (3.13).  $\square$

**Remarque 3.3.5** Nous verrons plus loin au Chapitre 9 que, lorsque la forme bilinéaire  $a$  est symétrique, il existe un autre argument que le Théorème 3.3.1 de Lax-Milgram pour conclure à l'existence et l'unicité d'une solution de (3.13). En effet, on démontrera directement l'existence d'un unique point de minimum de l'énergie  $J(v)$ . En vertu de la Proposition 3.3.4, cela démontre l'existence et l'unicité de la solution de la formulation variationnelle.  $\bullet$

### 3.3.2 Application au Laplacien

Essayons d'appliquer le Théorème 3.3.1 de Lax-Milgram à la formulation variationnelle (3.8) du Laplacien avec conditions aux limites de Dirichlet. Celle-ci s'écrit bien sous la forme (3.13) avec

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx$$

et

$$L(v) = \int_{\Omega} f(x)v(x) dx,$$

où clairement  $a(\cdot, \cdot)$  est une forme bilinéaire, et  $L(\cdot)$  une forme linéaire. L'espace  $V$  (noté précédemment  $X$ ) est

$$V = \{v \in C^1(\overline{\Omega}), v = 0 \text{ sur } \partial\Omega\}. \quad (3.19)$$

Comme produit scalaire sur  $V$  nous choisissons

$$\langle w, v \rangle = \int_{\Omega} \nabla w(x) \cdot \nabla v(x) dx, \quad (3.20)$$

qui a pour norme associée

$$\|v\| = \left( \int_{\Omega} |\nabla v(x)|^2 dx \right)^{1/2}.$$

On vérifie aisément que (3.20) définit un produit scalaire sur  $V$  : le seul point qui mérite de s'y attarder est la propriété  $\|v\| = 0 \Rightarrow v = 0$ . En effet, de l'égalité

$$\int_{\Omega} |\nabla v(x)|^2 dx = 0$$

on déduit que  $v$  est une constante dans  $\Omega$ , et comme  $v = 0$  sur  $\partial\Omega$  on a bien  $v = 0$ . La motivation du choix de (3.20) comme produit scalaire est bien sûr le fait que la forme bilinéaire  $a(\cdot, \cdot)$  est **automatiquement coercive** pour (3.20). On vérifie par ailleurs aisément que  $a$  est continue. Pour montrer que  $L$  est continue, il faut faire appel à l'inégalité de Poincaré du Lemme 3.3.6 : on a alors

$$\left| \int_{\Omega} f(x)v(x) dx \right| \leq \left( \int_{\Omega} |f(x)|^2 dx \right)^{1/2} \left( \int_{\Omega} |v(x)|^2 dx \right)^{1/2} \leq C \|v\|,$$

où  $C$  est une constante qui dépend de  $f$  mais pas de  $v$ . Donc  $L$  est continue sur  $V$ . Toutes les hypothèses du Théorème 3.3.1 de Lax-Milgram semblent vérifiées, et pourtant il en manque une qui empêche son application : l'espace  $V$  n'est pas un espace de Hilbert car il n'est pas complet pour la norme induite par (3.20) ! L'obstruction ne vient pas tant du choix du produit scalaire que de l'exigence de régularité  $C^1$  des fonctions de l'espace  $V$ . Une façon immédiate, quoique peu explicite, de résoudre la difficulté est de remplacer  $V$  par  $\overline{V}$ , sa fermeture pour le produit scalaire (3.20). Évidemment, on n'a fait que déplacer la difficulté : à quoi peut bien ressembler l'espace  $\overline{V}$  ? La réponse sera apportée au Chapitre 4 :  $\overline{V}$  est l'espace de Sobolev  $H_0^1(\Omega)$  dont les éléments ne sont plus des fonctions régulières mais seulement mesurables. Une autre difficulté sera de voir en quel sens la Proposition 3.2.7 (qui exprime l'équivalence entre le problème aux limites (3.1) et sa formulation variationnelle (3.8)) reste vrai lorsque on remplace l'espace  $V$  par  $\overline{V}$ .

Nous espérons avoir ainsi convaincu le lecteur du **caractère naturel et inéluctable des espaces de Sobolev dans la résolution des formulations variationnelles** d'équations aux dérivées partielles elliptiques. Terminons ce chapitre par un lemme technique, appelé inégalité de Poincaré, que nous avons utilisé un peu plus haut.

**Lemme 3.3.6** *Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$  borné dans au moins une direction de l'espace. Il existe une constante  $C > 0$  telle que, pour toute fonction  $v \in C^1(\overline{\Omega})$  qui s'annule sur le bord  $\partial\Omega$ ,*

$$\int_{\Omega} |v(x)|^2 dx \leq C \int_{\Omega} |\nabla v(x)|^2 dx.$$

**Démonstration.** L'hypothèse sur le caractère borné de  $\Omega$  dit (après une éventuelle rotation) que pour tout  $x \in \Omega$  la première composante  $x_1$  est bornée,  $-\infty < a \leq x_1 \leq b < +\infty$ . Soit  $v$  une fonction de  $C^1(\overline{\Omega})$  qui est nulle sur  $\partial\Omega$ . On peut l'étendre par continuité par zéro en dehors de  $\Omega$  ( $v$  est alors une fonction continue de classe  $C^1$  par morceaux dans  $\mathbb{R}^N$ ) et écrire, pour  $x \in \Omega$ ,

$$v(x) = \int_a^{x_1} \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) dt,$$

d'où l'on déduit par l'inégalité de Cauchy-Schwarz

$$|v(x)|^2 \leq (x_1 - a) \int_a^{x_1} \left| \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) \right|^2 dt \leq (b - a) \int_a^b \left| \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) \right|^2 dt.$$

Intégrant sur  $\Omega$  on obtient

$$\int_{\Omega} |v(x)|^2 dx \leq (b-a) \int_{\Omega} \int_a^b \left| \frac{\partial v}{\partial x_1}(t, x_2, \dots, x_N) \right|^2 dt dx,$$

et permutant les deux intégrations par rapport à  $t$  et  $x_1$ , on conclut

$$\int_{\Omega} |v(x)|^2 dx \leq (b-a)^2 \int_{\Omega} \left| \frac{\partial v}{\partial x_1}(x) \right|^2 dx \leq (b-a)^2 \int_{\Omega} |\nabla v(x)|^2 dx.$$

□

**Exercice 3.3.1** Le but de cet exercice est de montrer que l'espace  $V$ , défini par (3.19) et muni du produit scalaire (3.20), n'est pas complet. Soit  $\Omega$  la boule unité ouverte de  $\mathbb{R}^N$ . Si  $N = 1$ , on définit la suite

$$u_n(x) = \begin{cases} -x - 1 & \text{si } -1 < x < -n^{-1}, \\ (n/2)x^2 - 1 + 1/(2n) & \text{si } -n^{-1} \leq x \leq n^{-1}, \\ x - 1 & \text{si } n^{-1} < x < 1. \end{cases}$$

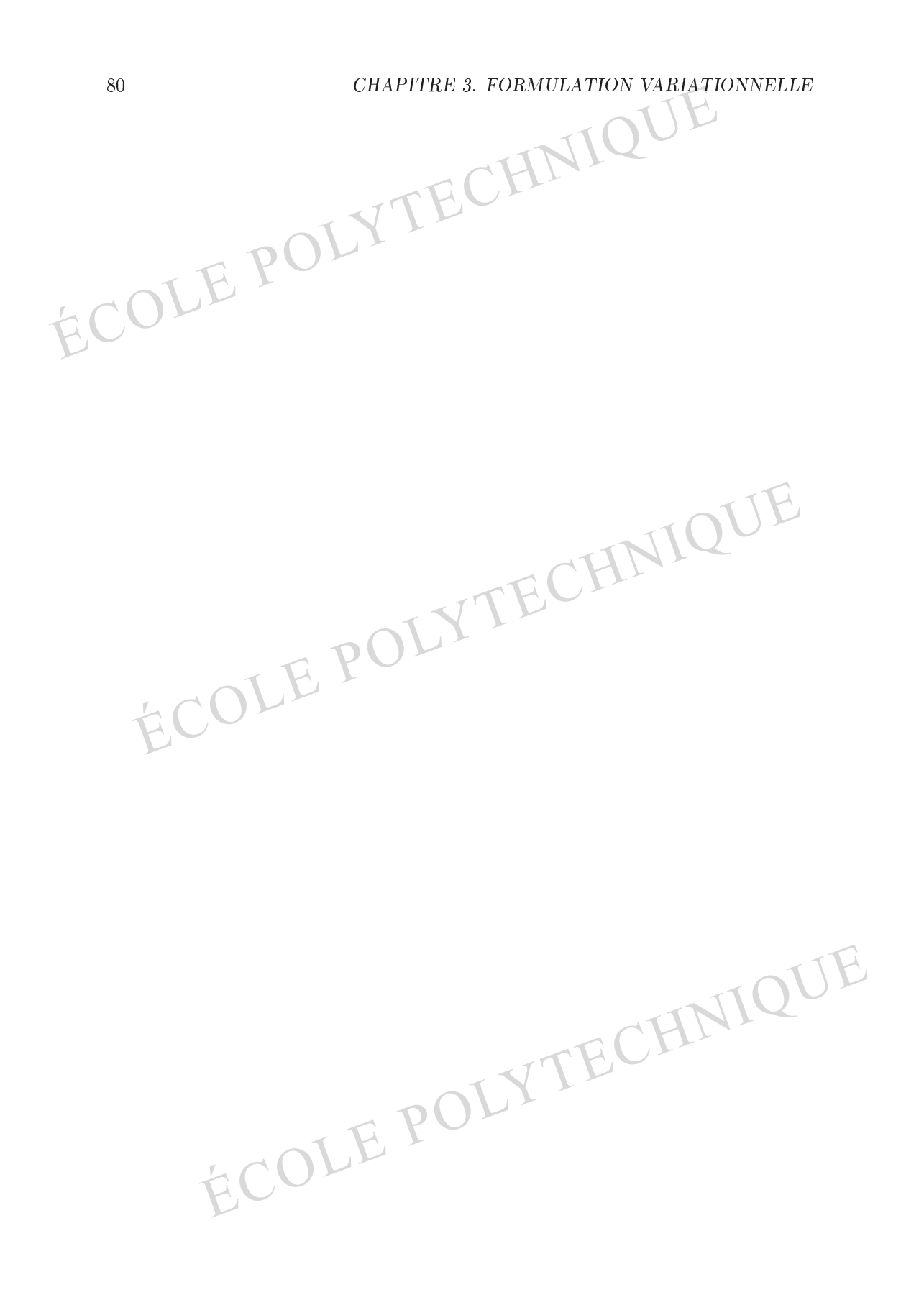
Si  $N = 2$ , pour  $0 < \alpha < 1/2$ , on définit la suite

$$u_n(x) = |\log(|x|^2 + n^{-1})|^{\alpha/2} - |\log(1 + n^{-1})|^{\alpha/2}.$$

Si  $N \geq 3$ , pour  $0 < \beta < (N-2)/2$ , on définit la suite

$$u_n(x) = \frac{1}{(|x|^2 + n^{-1})^{\beta/2}} - \frac{1}{(1 + n^{-1})^{\beta/2}}.$$

Montrer que la suite  $u_n$  est de Cauchy dans  $V$  mais qu'elle ne converge pas dans  $V$  lorsque  $n$  tend vers l'infini.



## Chapitre 4

# ESPACES DE SOBOLEV

### 4.1 Introduction et avertissement

Dans ce chapitre nous définissons les espaces de Sobolev qui sont **les espaces “naturels” de fonctions permettant de résoudre les formulations variationnelles d’équations aux dérivées partielles**. Physiquement, les espaces de Sobolev s’interprètent comme des espaces de **fonctions d’énergie finie**. Ce chapitre est le plus “technique” de cet ouvrage et relève en partie d’un cours de mathématiques “pures”. Néanmoins, il est nécessaire de bien connaître les résultats ci-dessous pour aborder la suite du cours, y compris dans ses aspects les plus numériques. En règle générale, il n’est pas nécessaire de connaître les démonstrations de ces résultats (sauf pour les plus simples et les plus utiles). Cependant, pour la commodité du lecteur et afin d’éviter de trop fréquents renvois à d’autres ouvrages, nous avons inclus la plupart des démonstrations. Le lecteur intéressé, ou tout simplement curieux, y trouvera les idées et les arguments clés qui lui permettront de bien comprendre la structure et l’intérêt des espaces de Sobolev. **Insistons encore pour dire que c’est l’esprit des résultats plus que la lettre des démonstrations qui est important ici.**

Le plan de ce chapitre est le suivant. Comme les espaces de Sobolev se construisent à partir de la notion de fonction mesurable et de l’espace  $L^2$  des fonctions de carrés sommables, la Section 4.2 donne quelques rappels à ce sujet. On y introduit aussi la notion de **dérivation faible**. La Section 4.3 contient toutes les définitions et les résultats qu’il faut absolument connaître sur les espaces de Sobolev pour suivre le reste du cours. La Section 4.4 donne quelques résultats complémentaires pour le lecteur curieux. Enfin, la Section 4.5 permet au lecteur qui connaît la théorie des distributions (qui n’est pas nécessaire ici), de faire le lien entre espaces de Sobolev et distributions. A la fin du chapitre le **Tableau 4.1 récapitule tous les résultats nécessaires pour la suite.**

## 4.2 Fonctions de carré sommable et dérivation faible

### 4.2.1 Quelques rappels d'intégration

Tous les résultats de cette sous-section sont détaillés dans le cours de mathématiques [7]. Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$  muni de la mesure de Lebesgue. On définit l'espace  $L^2(\Omega)$  des fonctions mesurables de carré sommable dans  $\Omega$ . Muni du produit scalaire

$$\langle f, g \rangle = \int_{\Omega} f(x)g(x) dx,$$

$L^2(\Omega)$  est un espace de Hilbert (voir le théorème 3.3.2 de [7]). On note

$$\|f\|_{L^2(\Omega)} = \left( \int_{\Omega} |f(x)|^2 dx \right)^{1/2}$$

la norme correspondante. Rappelons que les fonctions mesurables dans  $\Omega$  sont définies **presque partout** dans  $\Omega$  : si on change les valeurs d'une fonction mesurable  $f$  sur un sous-ensemble de  $\Omega$  de mesure nulle, on ne change pas la fonction mesurable  $f$ . Autrement dit, deux fonctions mesurables  $f$  et  $g$  seront dites égales si  $f(x) = g(x)$  presque partout dans  $\Omega$ , c'est-à-dire s'il existe  $E \subset \Omega$  tel que la mesure de Lebesgue de  $E$  est nulle et  $f(x) = g(x)$  pour tout  $x \in (\Omega \setminus E)$ .

On note  $C_c^\infty(\Omega)$  (ou  $\mathcal{D}(\Omega)$ ) l'espace des fonctions de classe  $C^\infty$  à support compact dans  $\Omega$ . Remarquons que l'espace  $C_c^\infty(\Omega)$  n'est pas réduit à la seule fonction nulle partout (ce qui n'est pas évident ! Voir le corollaire 3.2.6 dans [7]). Notons aussi que les fonctions de  $C_c^\infty(\Omega)$  s'annulent, ainsi que toutes leurs dérivées, sur le bord de  $\Omega$ . Nous rappelons le résultat de densité suivant (voir le théorème 3.4.3 de [7])

**Théorème 4.2.1** *L'espace  $C_c^\infty(\Omega)$  est dense dans  $L^2(\Omega)$ , c'est-à-dire que pour tout  $f \in L^2(\Omega)$  il existe une suite  $f_n \in C_c^\infty(\Omega)$  telle que*

$$\lim_{n \rightarrow +\infty} \|f - f_n\|_{L^2(\Omega)} = 0.$$

La propriété suivante généralise le Lemme 3.2.9.

**Corollaire 4.2.2** *Soit  $f \in L^2(\Omega)$ . Si pour toute fonction  $\phi \in C_c^\infty(\Omega)$  on a*

$$\int_{\Omega} f(x)\phi(x) dx = 0,$$

*alors  $f(x) = 0$  presque partout dans  $\Omega$ .*

**Démonstration.** Soit  $f_n \in C_c^\infty(\Omega)$  la suite de fonctions régulières qui converge vers  $f$  dans  $L^2(\Omega)$  en vertu du Théorème 4.2.1. On a

$$0 = \lim_{n \rightarrow +\infty} \int_{\Omega} f(x)f_n(x) dx = \int_{\Omega} |f(x)|^2 dx,$$



d'où l'on déduit que  $f(x) = 0$  presque partout dans  $\Omega$ .  $\square$

Plus généralement, on peut définir les espaces  $L^p(\Omega)$  avec  $1 \leq p \leq +\infty$ . Pour  $1 \leq p < +\infty$ ,  $L^p(\Omega)$  est l'espace des fonctions mesurables de puissance  $p$ -ème intégrable sur  $\Omega$ . Muni de la norme

$$\|f\|_{L^p(\Omega)} = \left( \int_{\Omega} |f(x)|^p dx \right)^{1/p}, \quad (4.1)$$

$L^p(\Omega)$  est un espace de Banach, c'est-à-dire un espace vectoriel normé complet. Pour  $p = +\infty$ ,  $L^\infty(\Omega)$  est l'espace des fonctions mesurables  $f$  essentiellement bornées sur  $\Omega$ , c'est-à-dire qu'il existe une constante  $C > 0$  telle que  $|f(x)| \leq C$  presque partout dans  $\Omega$ . Muni de la norme

$$\|f\|_{L^\infty(\Omega)} = \inf \{ C \in \mathbb{R}^+ \text{ tel que } |f(x)| \leq C \text{ p.p. dans } \Omega \}, \quad (4.2)$$

$L^\infty(\Omega)$  est un espace de Banach. Rappelons que, si  $\Omega$  est un ouvert borné, alors  $L^p(\Omega) \subset L^q(\Omega)$  pour  $1 \leq q \leq p \leq +\infty$ .

### 4.2.2 Dérivation faible

On définit tout d'abord le concept de dérivée faible dans  $L^2(\Omega)$ . Cette notion généralise la dérivation usuelle (parfois appelée, par opposition, dérivation forte) et est un cas particulier de la dérivation au sens des distributions (voir [7] ou la Section 4.5 pour un bref résumé).

**Définition 4.2.3** Soit  $v$  une fonction de  $L^2(\Omega)$ . On dit que  $v$  est dérivable au sens faible dans  $L^2(\Omega)$  s'il existe des fonctions  $w_i \in L^2(\Omega)$ , pour  $i \in \{1, \dots, N\}$ , telles que, pour toute fonction  $\phi \in C_c^\infty(\Omega)$ , on a

$$\int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_{\Omega} w_i(x) \phi(x) dx.$$

Chaque  $w_i$  est appelée la  $i$ -ème dérivée partielle faible de  $v$  et notée désormais  $\frac{\partial v}{\partial x_i}$ .

La Définition 4.2.3 a bien un sens : en particulier, la notation  $w_i = \frac{\partial v}{\partial x_i}$  est univoque car, en vertu du Corollaire 4.2.2, les fonctions  $w_i$  sont uniques (si elles existent). Bien sûr, si  $v$  est dérivable au sens usuel et que ses dérivées partielles appartiennent à  $L^2(\Omega)$ , alors les dérivées usuelle et faible de  $v$  coïncident (utiliser le Corollaire 3.2.3). Donnons tout de suite un critère simple et pratique pour déterminer si une fonction est dérivable au sens faible.

**Lemme 4.2.4** Soit  $v$  une fonction de  $L^2(\Omega)$ . S'il existe une constante  $C > 0$  telle que, pour toute fonction  $\phi \in C_c^\infty(\Omega)$  et pour tout indice  $i \in \{1, \dots, N\}$ , on a

$$\left| \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx \right| \leq C \|\phi\|_{L^2(\Omega)}, \quad (4.3)$$

alors  $v$  est dérivable au sens faible.

**Démonstration.** Soit  $L$  la forme linéaire définie par

$$L(\phi) = \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx.$$

A priori  $L(\phi)$  n'est définie que pour  $\phi \in C_c^\infty(\Omega)$ , mais grâce à l'inégalité (4.3), on peut étendre  $L$  par continuité à toutes les fonctions de  $L^2(\Omega)$  car  $C_c^\infty(\Omega)$  est dense dans  $L^2(\Omega)$  d'après le Théorème 4.2.1. En fait, l'inégalité (4.3) prouve que la forme linéaire  $L$  est continue sur  $L^2(\Omega)$ . En vertu du Théorème 12.1.18 de représentation de Riesz, il existe une fonction  $(-w_i) \in L^2(\Omega)$  telle que

$$L(\phi) = - \int_{\Omega} w_i(x) \phi(x) dx,$$

ce qui prouve que  $v$  est dérivable au sens faible dans  $L^2(\Omega)$ .  $\square$

**Exercice 4.2.1** Soit  $\Omega = (0, 1)$ . Montrer que la fonction  $x^\alpha$  est dérivable au sens faible dans  $L^2(\Omega)$  si et seulement si  $\alpha > 1/2$ .

**Exercice 4.2.2** Soit  $\Omega$  un ouvert borné. Montrer qu'une fonction continue sur  $\overline{\Omega}$ , et  $C^1$  par morceaux est dérivable au sens faible dans  $L^2(\Omega)$ .

**Exercice 4.2.3** Soit  $\Omega$  un ouvert borné. Montrer qu'une fonction  $C^1$  par morceaux mais pas continue n'est pas dérivable au sens faible dans  $L^2(\Omega)$ .

On retrouve un résultat bien connu pour la dérivée usuelle.

**Proposition 4.2.5** Soit  $v$  une fonction de  $L^2(\Omega)$  dérivable au sens faible et telle que toutes ses dérivées partielles faibles  $\frac{\partial v}{\partial x_i}$ , pour  $1 \leq i \leq N$ , sont nulles. Alors, pour chaque composante connexe de  $\Omega$ , il existe une constante  $C$  telle que  $v(x) = C$  presque partout dans cette composante connexe.

**Démonstration.** Pour tout  $\psi \in C_c^\infty(\Omega)$ , on a donc

$$\int_{\Omega} v(x) \frac{\partial \psi}{\partial x_i}(x) dx = 0. \quad (4.4)$$

Soit  $Q = ]-\ell, +\ell[^N$  un cube ouvert inclus dans  $\Omega$  (avec  $\ell > 0$ ), et soit  $\theta(t) \in C_c^\infty(-\ell, +\ell)$  telle que

$$\int_{-\ell}^{+\ell} \theta(t) dt = 1.$$

Pour toute fonction  $\phi \in C_c^\infty(Q)$  on définit

$$\psi(x', x_i) = \int_{-\ell}^{x_i} \left( \theta(t) \int_{-\ell}^{+\ell} \phi(x', s) ds - \phi(x', t) \right) dt,$$

avec la notation  $x = (x', x_i)$  pour  $x' \in \mathbb{R}^{N-1}$  et  $x_i \in \mathbb{R}$ . On vérifie facilement que  $\psi$  appartient aussi à  $C_c^\infty(Q)$  et que

$$\frac{\partial \psi}{\partial x_i}(x', x_i) = \theta(x_i) \int_{-\ell}^{+\ell} \phi(x', s) ds - \phi(x', x_i).$$

Avec une telle fonction  $\psi$  l'équation (4.4) devient

$$\begin{aligned} \int_Q v(x) \phi(x) dx &= \int_Q v(x) \theta(x_i) \left( \int_{-\ell}^{+\ell} \phi(x', s) ds \right) dx' dx_i \\ &= \int_Q \phi(x', s) \left( \int_{-\ell}^{+\ell} v(x', x_i) \theta(x_i) dx_i \right) dx' ds \end{aligned}$$

grâce au théorème de Fubini. Comme  $\phi$  est quelconque, par application du Corollaire 4.2.2 on en déduit

$$v(x) = \int_{-\ell}^{+\ell} v(x', s) \theta(s) ds,$$

c'est-à-dire que  $v$  ne dépend pas de  $x_i$  dans  $Q$ . En répétant cet argument pour toutes les composantes  $x_i$ , on obtient ainsi que  $v(x)$  est constant dans  $Q$ . Comme tout couple de points dans une même composante connexe de  $\Omega$  peuvent être reliés par une chaîne de tels cubes (de taille variable) qui se recoupent, on peut en conclure que  $v(x)$  est constant dans chaque composante connexe de  $\Omega$ .  $\square$

On peut facilement généraliser la Définition 4.2.3 de la dérivée faible à certains opérateurs différentiels qui ne font intervenir que certaines combinaisons de dérivées partielles (et non pas toutes). C'est par exemple le cas de la divergence d'une fonction à valeurs vectorielles qui nous sera utile par la suite.

**Définition 4.2.6** Soit  $\sigma$  une fonction de  $\Omega$  dans  $\mathbb{R}^N$  dont toutes les composantes appartiennent à  $L^2(\Omega)$  (on note  $\sigma \in L^2(\Omega)^N$ ). On dit que  $\sigma$  admet une divergence au sens faible dans  $L^2(\Omega)$  s'il existe une fonction  $w \in L^2(\Omega)$  telle que, pour toute fonction  $\phi \in C_c^\infty(\Omega)$ , on a

$$\int_\Omega \sigma(x) \cdot \nabla \phi(x) dx = - \int_\Omega w(x) \phi(x) dx.$$

La fonction  $w$  est appelée la divergence faible de  $\sigma$  et notée désormais  $\operatorname{div} \sigma$ .

La justification de la Définition 4.2.6 est que, si  $\sigma$  est une fonction régulière, alors une simple intégration par parties (voir le Corollaire 3.2.3) montre que l'on a bien  $w = \operatorname{div} \sigma$ . Une généralisation facile du critère de dérivation faible du Lemme 4.2.4 est donnée par le résultat suivant (dont nous laissons la démonstration au lecteur en guise d'exercice).

**Lemme 4.2.7** Soit  $\sigma$  une fonction de  $L^2(\Omega)^N$ . S'il existe une constante  $C > 0$  telle que, pour toute fonction  $\phi \in C_c^\infty(\Omega)$ , on a

$$\left| \int_{\Omega} \sigma(x) \cdot \nabla \phi(x) dx \right| \leq C \|\phi\|_{L^2(\Omega)},$$

alors  $\sigma$  admet une divergence au sens faible.

**Exercice 4.2.4** Soit  $\Omega$  un ouvert borné constitué de deux ouverts  $\Omega_1$  et  $\Omega_2$  séparés par une surface  $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$ . Montrer qu'une fonction vectorielle de classe  $C^1$  sur chaque morceau  $\Omega_1$  et  $\Omega_2$  admet une divergence faible dans  $L^2(\Omega)$  si et seulement si sa composante normale est continue à travers la surface  $\Gamma$ .

**Remarque 4.2.8** La notion de dérivation faible et tous les résultats de cette sous-section s'étendent aux espaces  $L^p(\Omega)$  pour  $1 \leq p \leq +\infty$ . Comme pour  $p \neq 2$ , l'espace  $L^p(\Omega)$  n'est pas un espace de Hilbert, le critère de dérivation faible (4.3) doit être remplacé par

$$\left| \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_i}(x) dx \right| \leq C \|\phi\|_{L^{p'}(\Omega)} \text{ avec } \frac{1}{p} + \frac{1}{p'} = 1 \text{ et } 1 < p \leq +\infty,$$

et le Théorème 12.1.18 de représentation de Riesz est remplacé par l'utilisation du dual  $L^p(\Omega)$  de  $L^{p'}(\Omega)$  (voir [8]). •

## 4.3 Définition et principales propriétés

### 4.3.1 Espace $H^1(\Omega)$

**Définition 4.3.1** Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$ . L'espace de Sobolev  $H^1(\Omega)$  est défini par

$$H^1(\Omega) = \left\{ v \in L^2(\Omega) \text{ tel que } \forall i \in \{1, \dots, N\} \frac{\partial v}{\partial x_i} \in L^2(\Omega) \right\}, \quad (4.5)$$

où  $\frac{\partial v}{\partial x_i}$  est la dérivée partielle faible de  $v$  au sens de la Définition 4.2.3.

En physique ou en mécanique l'espace de Sobolev est souvent appelé **espace d'énergie** au sens où il est constitué des fonctions d'énergie finie (c'est-à-dire de norme  $\|u\|_{H^1(\Omega)}$  finie). Les fonctions d'énergie finie peuvent éventuellement être "singulières" ce qui a un sens physique possible (concentration ou explosion locale de certaines grandeurs). On consultera avec intérêt les exemples explicites de l'Exercice 4.3.2 et du Lemme 5.2.33.

**Proposition 4.3.2** Muni du produit scalaire

$$\langle u, v \rangle = \int_{\Omega} (u(x)v(x) + \nabla u(x) \cdot \nabla v(x)) dx \quad (4.6)$$

et de la norme

$$\|u\|_{H^1(\Omega)} = \left( \int_{\Omega} (|u(x)|^2 + |\nabla u(x)|^2) dx \right)^{1/2}$$

l'espace de Sobolev  $H^1(\Omega)$  est un espace de Hilbert.

**Démonstration.** Il est évident que (4.6) est bien un produit scalaire dans  $H^1(\Omega)$ . Il reste donc à montrer que  $H^1(\Omega)$  est complet pour la norme associée. Soit  $(u_n)_{n \geq 1}$  une suite de Cauchy dans  $H^1(\Omega)$ . Par définition de la norme de  $H^1(\Omega)$ ,  $(u_n)_{n \geq 1}$  ainsi que  $(\frac{\partial u_n}{\partial x_i})_{n \geq 1}$  pour  $i \in \{1, \dots, N\}$  sont des suites de Cauchy dans  $L^2(\Omega)$ . Comme  $L^2(\Omega)$  est complet, il existe des limites  $u$  et  $w_i$  telles que  $u_n$  converge vers  $u$  et  $\frac{\partial u_n}{\partial x_i}$  converge vers  $w_i$  dans  $L^2(\Omega)$ . Or, par définition de la dérivée faible de  $u_n$ , pour toute fonction  $\phi \in C_c^\infty(\Omega)$ , on a

$$\int_{\Omega} u_n(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_{\Omega} \frac{\partial u_n}{\partial x_i}(x) \phi(x) dx. \quad (4.7)$$

Passant à la limite  $n \rightarrow +\infty$  dans (4.7), on obtient

$$\int_{\Omega} u(x) \frac{\partial \phi}{\partial x_i}(x) dx = - \int_{\Omega} w_i(x) \phi(x) dx,$$

ce qui prouve que  $u$  est dérivable au sens faible et que  $w_i$  est la  $i$ -ème dérivée partielle faible de  $u$ ,  $\frac{\partial u}{\partial x_i}$ . Donc,  $u$  appartient bien à  $H^1(\Omega)$  et  $(u_n)_{n \geq 1}$  converge vers  $u$  dans  $H^1(\Omega)$ .  $\square$

**Exercice 4.3.1** Montrer que les fonctions continues,  $C^1$  par morceaux et à support borné dans  $\overline{\Omega}$ , appartiennent à  $H^1(\Omega)$ .

En dimension  $N \geq 2$ , les fonctions de  $H^1(\Omega)$  ne sont en général **ni continues ni bornées**, comme le montre le contre-exemple suivant.

**Exercice 4.3.2** Soit  $B$  la boule unité ouverte de  $\mathbb{R}^N$ . Si  $N = 2$ , montrer que la fonction  $u(x) = |\log(|x|)|^\alpha$  appartient à  $H^1(B)$  pour  $0 < \alpha < 1/2$ , mais n'est pas bornée au voisinage de l'origine. Si  $N \geq 3$ , montrer que la fonction  $u(x) = |x|^{-\beta}$  appartient à  $H^1(B)$  pour  $0 < \beta < (N-2)/2$ , mais n'est pas bornée au voisinage de l'origine.

La dimension d'espace  $N = 1$  fait "exception" à la non-continuité des fonctions de  $H^1(\Omega)$  comme l'affirme le lemme suivant où, sans perte de généralité, on prend  $\Omega = (0, 1)$ .

**Lemme 4.3.3** Pour toute fonction  $v \in H^1(0, 1)$  et pour tout  $x, y \in [0, 1]$ , on a

$$v(y) = v(x) + \int_x^y v'(s) ds. \quad (4.8)$$

Plus généralement, pour tout  $x \in [0, 1]$ , l'application  $v \rightarrow v(x)$ , définie de  $H^1(0, 1)$  dans  $\mathbb{R}$ , est une forme linéaire continue sur  $H^1(0, 1)$ . En particulier, toute fonction  $v \in H^1(0, 1)$  est continue sur  $[0, 1]$ .

**Démonstration.** Soit  $v \in H^1(0, 1)$ . On définit une fonction  $w(x)$  sur  $[0, 1]$  par

$$w(x) = \int_0^x v'(s) ds.$$

Cette définition a un sens car, par l'inégalité de Cauchy-Schwarz,

$$\left| \int_0^x v'(s) ds \right| \leq \sqrt{x} \sqrt{\int_0^x |v'(s)|^2 ds} \leq \sqrt{\int_0^1 |v'(s)|^2 ds} < +\infty.$$

En fait, le même raisonnement montre que la fonction  $w$  est continue sur  $[0, 1]$

$$|w(x) - w(y)| = \left| \int_y^x v'(s) ds \right| \leq \sqrt{|x - y|} \sqrt{\int_y^x |v'(s)|^2 ds} \leq \sqrt{|x - y|} \sqrt{\int_0^1 |v'(s)|^2 ds}.$$

Montrons que  $w$  est dérivable au sens faible et que  $w' = v'$ . Soit  $\phi \in C_c^\infty(0, 1)$ . En notant  $T$  le triangle  $T = \{(x, s) \in \mathbb{R}^2, 0 \leq s \leq x \leq 1\}$ , on a

$$\int_0^1 w(x) \phi'(x) dx = \int_0^1 \left( \int_0^x v'(s) ds \right) \phi'(x) dx = \int_T v'(s) \phi'(x) ds dx.$$

Par application du théorème de Fubini, on a

$$\int_T v'(s) \phi'(x) ds dx = \int_0^1 \left( \int_s^1 \phi'(x) dx \right) v'(s) ds = - \int_0^1 \phi(s) v'(s) ds,$$

et par Cauchy-Schwarz on en déduit

$$\left| \int_0^1 w(x) \phi'(x) dx \right| \leq \|v'\|_{L^2(0,1)} \|\phi\|_{L^2(0,1)}.$$

Donc,  $w$  est bien dérivable au sens faible et par définition de  $w'$  on a

$$- \int_0^1 w'(x) \phi(x) dx = \int_0^1 w(x) \phi'(x) dx = - \int_0^1 \phi(s) v'(s) ds,$$

pour tout  $\phi \in C_c^\infty(0, 1)$ , ce qui implique que  $w' = v'$ . Le Lemme 4.2.5 nous dit alors que  $w - v$  est égale à une constante presque partout dans  $(0, 1)$ , ce qui établit (4.8).

À partir de (4.8) et en utilisant Cauchy-Schwarz on obtient

$$|v(x)| \leq |v(y)| + \sqrt{|y - x|} \sqrt{\int_x^y |v'(s)|^2 ds} \leq |v(y)| + \sqrt{\int_0^1 |v'(s)|^2 ds},$$

et en intégrant par rapport à  $y$

$$\begin{aligned} |v(x)| &\leq \int_0^1 |v(y)| dy + \sqrt{\int_0^1 |v'(s)|^2 ds} \\ &\leq \sqrt{\int_0^1 |v(y)|^2 dy} + \sqrt{\int_0^1 |v'(s)|^2 ds} \leq \sqrt{2} \|v\|_{H^1(0,1)}, \end{aligned}$$

ce qui prouve que  $v \rightarrow v(x)$  est une forme linéaire continue sur  $H^1(0, 1)$ .  $\square$

**Remarque 4.3.4** L'affirmation que toute fonction de  $H^1(0, 1)$  est continue peut sembler à première vue contradictoire avec le fait que les fonctions de  $H^1(0, 1)$ , comme toutes les fonctions mesurables, ne sont définies que presque partout (autrement dit, on peut changer certaines valeurs ponctuelles de  $v$  sans changer de fonction dans  $H^1(0, 1)$  mais en détruisant la continuité de  $v$ ). Pour résoudre cet apparent paradoxe, il faut se rappeler qu'une fonction de  $H^1(0, 1)$  est en fait une classe de fonctions puisqu'on décide d'identifier deux fonctions égales presque partout (autrement dit, deux représentants d'une même classe de fonctions sont égaux presque partout). Dans ces conditions, le résultat du Lemme 4.3.3 doit se comprendre dans le sens qu'il existe un représentant de la classe de fonctions  $v \in H^1(0, 1)$  qui est continu.

Il est très important en pratique de savoir si **les fonctions régulières sont denses dans l'espace de Sobolev**  $H^1(\Omega)$ . Cela justifie en partie la notion d'espace de Sobolev qui apparaît ainsi très simplement comme l'ensemble des fonctions régulières complété par les limites de suites de fonctions régulières dans la norme de l'énergie  $\|u\|_{H^1(\Omega)}$ . Cela permet de démontrer facilement de nombreuses propriétés en les établissant d'abord sur les fonctions régulières puis en utilisant un argument de "densité" (voir par exemple les démonstrations des Théorèmes 4.3.13 et 4.3.15 ci-dessous).

**Théorème 4.3.5 (de densité)** *Si  $\Omega$  est un ouvert borné régulier de classe  $C^1$ , ou bien si  $\Omega = \mathbb{R}_+^N$ , ou encore si  $\Omega = \mathbb{R}^N$ , alors  $C_c^\infty(\overline{\Omega})$  est dense dans  $H^1(\Omega)$ .*

La démonstration du Théorème 4.3.5 (qu'on peut admettre en première lecture) se trouve à la Section 4.4. Rappelons que la notation  $\mathbb{R}_+^N$  désigne le demi-espace  $\{x \in \mathbb{R}^N \text{ tel que } x_N > 0\}$ .

**Remarque 4.3.6** L'espace  $C_c^\infty(\overline{\Omega})$  qui est dense dans  $H^1(\Omega)$  est constitué des fonctions régulières de classe  $C^\infty$  à support borné (ou compact) dans le fermé  $\overline{\Omega}$ . En particulier, si  $\Omega$  est borné, toutes les fonctions de  $C^\infty(\overline{\Omega})$  ont nécessairement un support borné, et donc  $C_c^\infty(\overline{\Omega}) = C^\infty(\overline{\Omega})$ . Précisons que les fonctions de  $C_c^\infty(\overline{\Omega})$  ne s'annulent pas nécessairement sur le bord de l'ouvert  $\Omega$ , ce qui différencie cet espace de  $C_c^\infty(\Omega)$  (voir la Remarque 3.2.2). Par contre, si  $\Omega$  n'est pas borné, les fonctions de  $C_c^\infty(\overline{\Omega})$  s'annulent "à l'infini".

**Remarque 4.3.7** La notion de régularité d'un ouvert a été introduite dans la Définition 3.2.5. Il n'est pas nécessaire de connaître précisément cette définition de la régularité d'un ouvert. Il suffit de savoir *grosso modo* que l'on demande que le bord de l'ouvert soit une surface régulière et que l'on exclut certaines "pathologies" (voir la Remarque 3.2.6). Lorsque nous énonçons un résultat sous une hypothèse de régularité de l'ouvert, cette régularité est toujours nécessaire (le résultat tombe en défaut pour certains ouverts non réguliers, voir le contre-exemple de l'Exercice 4.3.3). Néanmoins, l'hypothèse de régularité suivant la Définition 3.2.5 peut souvent être affaiblie : l'appartenance à la classe des fonctions  $C^1$  peut être remplacée par l'appartenance à la classe des fonctions lipschitziennes (voir [18]). Bien que ces détails dépassent largement le cadre de ce cours, nous faisons cette remarque afin que le lecteur pointilleux

ne s'insurge pas lorsque nous utiliserons de tels résultats (où l'hypothèse de régularité est nécessaire) dans le cas d'ouverts "avec des coins" qui apparaissent naturellement dans tous les calculs numériques (voir les différentes images de maillages qui illustrent ce cours). •

### 4.3.2 Espace $H_0^1(\Omega)$

Définissons maintenant un autre espace de Sobolev qui est un sous-espace de  $H^1(\Omega)$  et qui nous sera très utile pour les problèmes avec conditions aux limites de Dirichlet.

**Définition 4.3.8** Soit  $C_c^\infty(\Omega)$  l'espace des fonctions de classe  $C^\infty$  à support compact dans  $\Omega$ . L'espace de Sobolev  $H_0^1(\Omega)$  est défini comme l'adhérence de  $C_c^\infty(\Omega)$  dans  $H^1(\Omega)$ .

On verra un peu plus loin (voir le Corollaire 4.3.16) que  $H_0^1(\Omega)$  est en fait le sous-espace de  $H^1(\Omega)$  constitué des **fonctions qui s'annulent sur le bord**  $\partial\Omega$  puisque tel est le cas des fonctions de  $C_c^\infty(\Omega)$ . En général,  $H_0^1(\Omega)$  est **strictement plus petit** que  $H^1(\Omega)$  car  $C_c^\infty(\Omega)$  est un sous-espace **strict** de  $C_c^\infty(\overline{\Omega})$  (voir le Théorème 4.3.5 et la Remarque 4.3.6). Une exception importante est le cas où  $\Omega = \mathbb{R}^N$  : en effet, dans ce cas  $\overline{\Omega} = \mathbb{R}^N = \Omega$  et le Théorème 4.3.5 affirme que  $C_c^\infty(\mathbb{R}^N)$  est dense dans  $H^1(\mathbb{R}^N)$ , donc on a  $H_0^1(\mathbb{R}^N) = H^1(\mathbb{R}^N)$ . Cette exception se comprend aisément puisque l'espace entier  $\mathbb{R}^N$  n'a pas de bord.

**Proposition 4.3.9** Muni du produit scalaire (4.6) de  $H^1(\Omega)$ , l'espace de Sobolev  $H_0^1(\Omega)$  est un espace de Hilbert.

**Démonstration.** Par définition  $H_0^1(\Omega)$  est un sous-espace fermé de  $H^1(\Omega)$  (qui est un espace de Hilbert), donc c'est aussi un espace de Hilbert.  $\square$

Un résultat essentiel pour les applications du prochain chapitre est l'inégalité suivante.

**Proposition 4.3.10 (Inégalité de Poincaré)** Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$  borné dans au moins une direction de l'espace. Il existe une constante  $C > 0$  telle que, pour toute fonction  $v \in H_0^1(\Omega)$ ,

$$\int_{\Omega} |v(x)|^2 dx \leq C \int_{\Omega} |\nabla v(x)|^2 dx. \quad (4.9)$$

**Démonstration.** Pour les fonctions  $v \in C_c^\infty(\Omega)$  on a déjà démontré l'inégalité de Poincaré (4.9) dans le Lemme 3.3.6. Par un argument de densité le résultat reste vrai pour toute fonction  $v \in H_0^1(\Omega)$ . En effet, comme  $C_c^\infty(\Omega)$  est dense dans  $H_0^1(\Omega)$  (par sa Définition 4.3.8), il existe une suite  $v_n \in C_c^\infty(\Omega)$  telle que

$$\lim_{n \rightarrow +\infty} \|v - v_n\|_{H^1(\Omega)}^2 = \lim_{n \rightarrow +\infty} \int_{\Omega} (|v - v_n|^2 + |\nabla(v - v_n)|^2) dx = 0.$$



En particulier, on en déduit que

$$\lim_{n \rightarrow +\infty} \int_{\Omega} |v_n|^2 dx = \int_{\Omega} |v|^2 dx \text{ et } \lim_{n \rightarrow +\infty} \int_{\Omega} |\nabla v_n|^2 dx = \int_{\Omega} |\nabla v|^2 dx.$$

Par application du Lemme 3.3.6, on a

$$\int_{\Omega} |v_n(x)|^2 dx \leq C \int_{\Omega} |\nabla v_n(x)|^2 dx. \quad (4.10)$$

On passe alors à la limite  $n \rightarrow +\infty$  dans chacun des deux termes de l'inégalité (4.10) pour obtenir le résultat recherché. Ce type d'argument "par densité" sera très souvent repris par la suite.  $\square$

**Remarque 4.3.11** L'inégalité de Poincaré (4.9) n'est pas vraie pour les fonctions de  $H^1(\Omega)$ . En effet, les fonctions constantes (non nulles) annulent le terme de droite dans (4.9) mais pas le terme de gauche. L'hypothèse sous-jacente essentielle dans l'inégalité de Poincaré est que les fonctions de  $H_0^1(\Omega)$  s'annulent sur le bord  $\partial\Omega$  de l'ouvert  $\Omega$  (voir la Remarque 4.3.18 pour des variantes de cette hypothèse). •

Un corollaire important de l'inégalité de Poincaré est le résultat suivant qui fournit une norme équivalente plus simple dans  $H_0^1(\Omega)$ .

**Corollaire 4.3.12** *Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$  borné dans au moins une direction de l'espace. Alors la semi-norme*

$$|v|_{H_0^1(\Omega)} = \left( \int_{\Omega} |\nabla v(x)|^2 dx \right)^{1/2}$$

*est une norme sur  $H_0^1(\Omega)$  équivalente à la norme usuelle induite par celle de  $H^1(\Omega)$ .*

**Démonstration.** Soit  $v \in H_0^1(\Omega)$ . La première inégalité

$$|v|_{H_0^1(\Omega)} \leq \|v\|_{H^1(\Omega)} = \left( \int_{\Omega} (|v|^2 + |\nabla v|^2) dx \right)^{1/2}$$

est évidente. D'autre part, l'inégalité de Poincaré du Lemme 3.3.6 conduit à

$$\|v\|_{H^1(\Omega)}^2 \leq (C+1) \int_{\Omega} |\nabla v|^2 dx = (C+1) |v|_{H_0^1(\Omega)}^2,$$

ce qui prouve que  $|v|_{H_0^1(\Omega)}$  est une norme équivalente à  $\|v\|_{H^1(\Omega)}$ .  $\square$

### 4.3.3 Traces et formules de Green

Nous avons vu qu'en dimension  $N \geq 2$  les fonctions de  $H^1(\Omega)$  ne sont pas continues (voir le contre-exemple de l'Exercice 4.3.2). Comme pour toute fonction mesurable, on ne peut donc parler de la valeur ponctuelle d'une fonction  $v \in H^1(\Omega)$  que "presque partout" dans  $\Omega$ . En particulier, il n'est pas clair de savoir si on peut définir la "valeur au bord", ou "trace" de  $v$  sur le bord  $\partial\Omega$  car  $\partial\Omega$  est un ensemble négligeable ou de mesure nulle. Fort heureusement pour les problèmes aux limites que nous étudions, il y a tout de même un moyen pour définir la trace  $v|_{\partial\Omega}$  d'une fonction de  $H^1(\Omega)$ . Ce résultat essentiel, appelé théorème de trace, est le suivant.

**Théorème 4.3.13 (de trace)** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$ , ou bien  $\Omega = \mathbb{R}_+^N$ . On définit l'application trace  $\gamma_0$*

$$\begin{aligned} H^1(\Omega) \cap C(\overline{\Omega}) &\rightarrow L^2(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ v &\rightarrow \gamma_0(v) = v|_{\partial\Omega}. \end{aligned} \quad (4.11)$$

*Cette application  $\gamma_0$  se prolonge par continuité en une application linéaire continue de  $H^1(\Omega)$  dans  $L^2(\partial\Omega)$ , notée encore  $\gamma_0$ . En particulier, il existe une constante  $C > 0$  telle que, pour toute fonction  $v \in H^1(\Omega)$ , on a*

$$\|v\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^1(\Omega)}. \quad (4.12)$$

**Remarque 4.3.14** Grâce au Théorème de trace 4.3.13 on peut donc parler de la valeur d'une fonction de  $H^1(\Omega)$  sur le bord  $\partial\Omega$ . Ce résultat est remarquable car il n'est pas vrai pour une fonction de  $L^2(\Omega)$  (voir en particulier l'Exercice 4.3.4). •

**Démonstration.** Démontrons le résultat pour le demi-espace  $\Omega = \mathbb{R}_+^N = \{x \in \mathbb{R}^N, x_N > 0\}$ . Soit  $v \in C_c^\infty(\overline{\mathbb{R}_+^N})$ . Avec la notation  $x = (x', x_N)$ , on a

$$|v(x', 0)|^2 = -2 \int_0^{+\infty} v(x', x_N) \frac{\partial v}{\partial x_N}(x', x_N) dx_N,$$

et, en utilisant l'inégalité  $2ab \leq a^2 + b^2$ ,

$$|v(x', 0)|^2 \leq \int_0^{+\infty} \left( |v(x', x_N)|^2 + \left| \frac{\partial v}{\partial x_N}(x', x_N) \right|^2 \right) dx_N.$$

Par intégration en  $x'$ , on en déduit

$$\int_{\mathbb{R}^{N-1}} |v(x', 0)|^2 dx' \leq \int_{\mathbb{R}_+^N} \left( |v(x)|^2 + \left| \frac{\partial v}{\partial x_N}(x) \right|^2 \right) dx,$$

c'est-à-dire  $\|v\|_{L^2(\partial\mathbb{R}_+^N)} \leq \|v\|_{H^1(\mathbb{R}_+^N)}$ . Par densité de  $C_c^\infty(\overline{\mathbb{R}_+^N})$  dans  $H^1(\mathbb{R}_+^N)$ , on obtient ainsi le résultat.

Pour un ouvert borné régulier de classe  $\mathcal{C}^1$ , on utilise un argument de cartes locales du bord qui permet de se ramener au cas de  $\Omega = \mathbb{R}_+^N$ . Nous ne détaillons pas cet argument de cartes locales (assez technique) qui est le même que celui utilisé dans la démonstration de la Proposition 4.4.2 ci-dessous.  $\square$

Le Théorème de trace 4.3.13 permet de généraliser aux fonctions de  $H^1(\Omega)$  la formule de Green précédemment établie pour des fonctions de classe  $C^1$  au Corollaire 3.2.3.

**Théorème 4.3.15 (Formule de Green)** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$ . Si  $u$  et  $v$  sont des fonctions de  $H^1(\Omega)$ , elles vérifient*

$$\int_{\Omega} u(x) \frac{\partial v}{\partial x_i}(x) dx = - \int_{\Omega} v(x) \frac{\partial u}{\partial x_i}(x) dx + \int_{\partial\Omega} u(x) v(x) n_i(x) ds, \quad (4.13)$$

où  $n = (n_i)_{1 \leq i \leq N}$  est la normale unité extérieure à  $\partial\Omega$ .

**Démonstration.** Rappelons que la formule (4.13) a été établie pour des fonctions de classe  $C^1$  dans le Corollaire 3.2.3. On utilise à nouveau un argument de densité. Par densité de  $C_c^\infty(\bar{\Omega})$  dans  $H^1(\Omega)$  (voir le Théorème 4.3.5), il existe des suites  $(u_n)_{n \geq 1}$  et  $(v_n)_{n \geq 1}$  dans  $C_c^\infty(\bar{\Omega})$  qui convergent dans  $H^1(\Omega)$  vers  $u$  et  $v$ , respectivement. En vertu du Corollaire 3.2.3 on a

$$\int_{\Omega} u_n \frac{\partial v_n}{\partial x_i} dx = - \int_{\Omega} v_n \frac{\partial u_n}{\partial x_i} dx + \int_{\partial\Omega} u_n v_n n_i ds. \quad (4.14)$$

On peut passer à la limite  $n \rightarrow +\infty$  dans les deux premiers termes de (4.14) car  $u_n$  et  $\frac{\partial u_n}{\partial x_i}$  (respectivement,  $v_n$  et  $\frac{\partial v_n}{\partial x_i}$ ) convergent vers  $u$  et  $\frac{\partial u}{\partial x_i}$  (respectivement,  $v$  et  $\frac{\partial v}{\partial x_i}$ ) dans  $L^2(\Omega)$ . Pour passer à la limite dans la dernière intégrale de (4.14), on utilise la continuité de l'application trace  $\gamma_0$ , c'est-à-dire l'inégalité (4.12), qui permet d'affirmer que  $\gamma_0(u_n)$  (respectivement,  $\gamma_0(v_n)$ ) converge vers  $\gamma_0(u)$  (respectivement,  $\gamma_0(v)$ ) dans  $L^2(\partial\Omega)$ . On obtient ainsi la formule (4.13) pour des fonctions  $u$  et  $v$  de  $H^1(\Omega)$ .  $\square$

Comme conséquence du Théorème de trace 4.3.13 on obtient une caractérisation très simple de l'espace  $H_0^1(\Omega)$ .

**Corollaire 4.3.16** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$ . L'espace  $H_0^1(\Omega)$  coïncide avec le sous-espace de  $H^1(\Omega)$  constitué des fonctions qui s'annulent sur le bord  $\partial\Omega$ .*

**Démonstration.** Comme toute fonction de  $H_0^1(\Omega)$  est limite d'une suite de fonctions appartenant à  $C_c^\infty(\Omega)$  qui ont bien sûr une trace nulle, la continuité de l'application trace  $\gamma_0$  implique que la trace de la limite est aussi nulle. On en déduit que  $H_0^1(\Omega)$  est contenu dans le sous-espace de  $H^1(\Omega)$  des fonctions qui s'annulent sur le bord  $\partial\Omega$ . La réciproque est plus technique et découle d'un double procédé de cartes locales (voire la preuve de la Proposition 4.4.2) puis de régularisation et de translation (similaire à la preuve du Théorème 4.4.1). Nous renvoyons à [8], [36] pour plus de détails.  $\square$

**Remarque 4.3.17** Le Corollaire 4.3.16 affirme que le noyau de l'application trace  $\gamma_0$  est précisément  $H_0^1(\Omega)$ . Une question naturelle, mais beaucoup plus délicate, est de caractériser l'image de  $\gamma_0$ . Contentons nous de dire que cette image  $\text{Im}(\gamma_0)$  n'est pas  $L^2(\partial\Omega)$ , mais un sous-espace strict, dense dans  $L^2(\partial\Omega)$ , constitué de fonctions "plus régulières", et noté  $H^{1/2}(\partial\Omega)$ . Pour plus de détails, nous renvoyons à [31]. •

Grâce au Corollaire 4.3.16 nous pouvons donner une autre démonstration de la Proposition 4.3.10 à propos de l'inégalité de Poincaré. Cette nouvelle démonstration n'est plus "constructive" mais est basée sur un argument de contradiction qui possède le mérite de se généraliser très facilement. En effet, il existe de nombreuses variantes de l'inégalité de Poincaré, adaptées aux différents modèles d'équations aux dérivées partielles. Au vu de l'importance de cette inégalité pour la suite, il n'est donc pas inutile d'en donner une démonstration aisément adaptable à tous les cas de figure.

**Autre démonstration de la Proposition 4.3.10.** On procède par contradiction. S'il n'existe pas de constante  $C > 0$  telle que, pour toute fonction  $v \in H_0^1(\Omega)$ ,

$$\int_{\Omega} |v(x)|^2 dx \leq C \int_{\Omega} |\nabla v(x)|^2 dx,$$

cela veut dire qu'il existe une suite  $v_n \in H_0^1(\Omega)$  telle que

$$1 = \int_{\Omega} |v_n(x)|^2 dx > n \int_{\Omega} |\nabla v_n(x)|^2 dx. \quad (4.15)$$

En particulier, (4.15) implique que la suite  $v_n$  est bornée dans  $H_0^1(\Omega)$ . Par application du Théorème de Rellich 4.3.21 ci-dessous, il existe une sous-suite  $v_{n'}$  qui converge dans  $L^2(\Omega)$ . De plus, (4.15) montre que la suite  $\nabla v_{n'}$  converge vers zéro dans  $L^2(\Omega)$  (composante par composante). Par conséquent,  $v_{n'}$  est une suite de Cauchy dans  $H_0^1(\Omega)$ , qui est un espace de Hilbert, donc elle converge dans  $H_0^1(\Omega)$  vers une limite  $v$ . Comme on a

$$\int_{\Omega} |\nabla v(x)|^2 dx = \lim_{n \rightarrow +\infty} \int_{\Omega} |\nabla v_n(x)|^2 dx \leq \lim_{n \rightarrow +\infty} \frac{1}{n} = 0,$$

on en déduit, en vertu du Lemme 4.2.5, que  $v$  est une constante dans chaque composante connexe de  $\Omega$ . Mais comme  $v$  est nulle sur le bord  $\partial\Omega$  (en vertu du Corollaire 4.3.16),  $v$  est identiquement nulle dans tout  $\Omega$ . Par ailleurs,

$$\int_{\Omega} |v(x)|^2 dx = \lim_{n \rightarrow +\infty} \int_{\Omega} |v_n(x)|^2 dx = 1,$$

ce qui est une contradiction avec le fait que  $v = 0$ . □

**Remarque 4.3.18** La démonstration par contradiction de la Proposition 4.3.10 se généralise facilement. Prenons par exemple, le cas d'un ouvert  $\Omega$ , borné connexe et régulier de classe  $C^1$ , dont le bord  $\partial\Omega$  se décompose en deux parties disjointes régulières

$\partial\Omega_N$  et  $\partial\Omega_D$  dont les mesures superficielles sont non nulles (voir la Figure 4.1). On définit un espace  $V$  par

$$V = \{v \in H^1(\Omega) \text{ tel que } v = 0 \text{ sur } \partial\Omega_D\}.$$

Par application du Théorème de trace 4.3.13, il est facile de voir que  $V$  est un sous-espace fermé de  $H^1(\Omega)$ , donc est un espace de Hilbert pour le produit scalaire de  $H^1(\Omega)$ . Comme pour  $H_0^1(\Omega)$ , l'argument de contradiction permet de démontrer l'existence d'une constante  $C > 0$  telle que toute fonction  $v \in V$  vérifie l'inégalité de Poincaré (4.9). •

Par application de la formule de Green du Théorème 4.3.15, nous pouvons construire une famille d'exemples de fonctions appartenant à  $H^1(\Omega)$ . Cette famille d'exemples nous sera très utile par la suite pour construire des sous-espaces de dimension finie de  $H^1(\Omega)$ .

**Lemme 4.3.19** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$ . Soit  $(\omega_i)_{1 \leq i \leq I}$  une partition régulière de  $\Omega$ , c'est-à-dire que chaque  $\omega_i$  est un ouvert régulier de classe  $C^1$ ,  $\omega_i \cap \omega_j = \emptyset$  si  $i \neq j$ , et  $\bar{\Omega} = \cup_{i=1}^I \bar{\omega}_i$ . Soit  $v$  une fonction dont la restriction à chaque  $\omega_i$ ,  $v_i = v|_{\omega_i}$ , appartient à  $H^1(\omega_i)$ . Si  $v$  est continue sur  $\bar{\Omega}$ , alors  $v$  appartient à  $H^1(\Omega)$ .*

**Démonstration.** Calculons la dérivée faible de  $v$  : pour  $\phi \in C_c^\infty(\Omega)$ , par application de la formule de Green dans chaque  $\omega_i$  on a

$$\begin{aligned} \int_{\Omega} v(x) \frac{\partial \phi}{\partial x_j}(x) dx &= \sum_{i=1}^I \int_{\omega_i} v_i(x) \frac{\partial \phi}{\partial x_j}(x) dx \\ &= - \sum_{i=1}^I \int_{\omega_i} \frac{\partial v_i}{\partial x_j}(x) \phi(x) dx + \sum_{i=1}^I \int_{\partial\omega_i} v_i(x) \phi(x) n_j^i(x) ds \\ &= - \sum_{i=1}^I \int_{\omega_i} \frac{\partial v_i}{\partial x_j}(x) \phi(x) dx, \end{aligned}$$

car les intégrales de bord s'annulent deux à deux. En effet, sur la partie  $\Gamma = \partial\omega_i \cap \partial\omega_k$  du bord commune aux deux ouverts  $\omega_i$  et  $\omega_k$  on a  $n_j^i(x) = -n_j^k(x)$  et donc, par continuité de  $v$  et  $\phi$ ,

$$\int_{\Gamma} v_i(x) \phi(x) n_j^i(x) ds + \int_{\Gamma} v_k(x) \phi(x) n_j^k(x) ds = 0.$$

On en déduit donc que  $v$  est dérivable au sens faible et que

$$\left. \frac{\partial v}{\partial x_j} \right|_{\omega_i} = \frac{\partial v_i}{\partial x_j}.$$

En particulier, ceci implique que  $v$  appartient à  $H^1(\Omega)$ . □

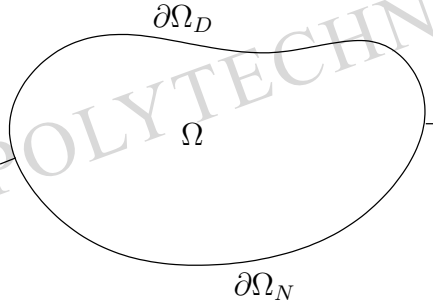


FIGURE 4.1 – Partition en deux parties disjointes du bord d'un ouvert.

**Remarque 4.3.20** Il n'est pas forcément facile de décomposer un ouvert régulier  $\Omega$  en une partition d'ouverts réguliers  $(\omega_i)_{1 \leq i \leq I}$ . Heureusement, le Lemme 4.3.19 reste valable si les ouverts  $\omega_i$  sont seulement réguliers "par morceaux". Nous utiliserons parfois cette généralisation, très commode, du Lemme 4.3.19. •

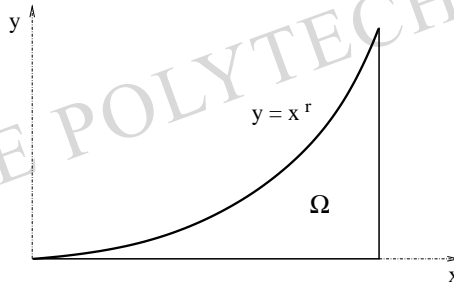


FIGURE 4.2 – Exemple d'un ouvert non régulier.

**Exercice 4.3.3** Le but de cet exercice est de montrer que le Théorème de trace 4.3.13 n'est pas vrai si l'ouvert  $\Omega$  n'est pas régulier. Soit l'ouvert  $\Omega \subset \mathbb{R}^2$  défini par  $0 < x < 1$  et  $0 < y < x^r$  avec  $r > 2$  (voir la Figure 4.2). Soit la fonction  $v(x) = x^\alpha$ . Montrer que  $v \in H^1(\Omega)$  si et seulement si  $2\alpha + r > 1$ , tandis que  $v \in L^2(\partial\Omega)$  si et seulement si  $2\alpha > -1$ . Conclure. (On peut aussi montrer avec ce même exemple que le Théorème 4.3.5 de densité et la Proposition 4.4.2 de prolongement ne sont pas vrais pour un tel ouvert.)

**Exercice 4.3.4** Le but de cet exercice est de montrer qu'il ne peut pas y avoir de notion de trace pour des fonctions de  $L^2(\Omega)$ , c'est-à-dire qu'il n'existe pas de constante  $C > 0$  telle que, pour toute fonction  $v \in L^2(\Omega)$ , on a

$$\|v|_{\partial\Omega}\|_{L^2(\partial\Omega)} \leq C\|v\|_{L^2(\Omega)}.$$

Pour simplifier, on choisit comme ouvert  $\Omega$  la boule unité. Construire une suite de fonctions régulières dans  $\bar{\Omega}$  égales à 1 sur  $\partial\Omega$  et dont la norme dans  $L^2(\Omega)$  tend vers zéro. Conclure.

### 4.3.4 Un résultat de compacité

Nous consacrons cette sous-section à l'étude d'une propriété de compacité connue sous le nom de théorème de Rellich qui jouera un rôle essentiel dans la théorie spectrale des problèmes aux limites (voir le Chapitre 7) que nous utiliserons pour résoudre les problèmes d'évolution en temps. Rappelons tout d'abord que, dans un espace de Hilbert de dimension infinie, il n'est pas vrai que, de toute suite bornée, on puisse extraire une sous-suite convergente (au contraire de ce qui se passe en dimension finie, voir l'Exercice 4.3.5).

**Théorème 4.3.21 (de Rellich)** *Si  $\Omega$  est un ouvert borné régulier de classe  $C^1$ , alors de toute suite bornée de  $H^1(\Omega)$  on peut extraire une sous-suite convergente dans  $L^2(\Omega)$  (on dit que l'injection canonique de  $H^1(\Omega)$  dans  $L^2(\Omega)$  est compacte).*

Le Théorème 4.3.21 peut être faux si l'ouvert  $\Omega$  n'est pas borné. Par exemple, si  $\Omega = \mathbb{R}^N$ , l'injection canonique de  $H^1(\mathbb{R}^N)$  dans  $L^2(\mathbb{R}^N)$  n'est pas compacte. Pour s'en convaincre il suffit de considérer la suite  $u_n(x) = u(x + ne)$  où  $e$  est un vecteur non nul et  $u$  une fonction de  $H^1(\mathbb{R}^N)$  (on translate  $u$  dans la direction  $e$ ). Il est clair qu'aucune sous-suite de  $u_n$  ne converge dans  $L^2(\mathbb{R}^N)$ .

**Remarque 4.3.22** Si l'on remplace  $H^1(\Omega)$  par  $H_0^1(\Omega)$ , alors non seulement le Théorème 4.3.21 de Rellich reste vrai, mais en plus il n'est pas nécessaire de supposer que l'ouvert  $\Omega$  est régulier. •

La démonstration du Théorème 4.3.21 est longue et délicate (elle fait appel soit à la transformée de Fourier, soit au théorème de compacité d'Ascoli; voir, par exemple, [8], [36]). Nous allons nous contenter de donner une démonstration plus simple et plus "parlante" en dimension  $N = 1$ .

**Démonstration.** On se place en une dimension d'espace  $N = 1$  et, sans perte de généralité, on suppose que  $\Omega = (0, 1)$ . Soit  $(u_n)_{n \geq 1}$  une suite bornée de  $H^1(0, 1)$ , c'est-à-dire qu'il existe  $K > 0$  tel que

$$\|u_n\|_{H^1(0,1)} \leq K \quad \forall n \geq 1.$$

D'après le Lemme 4.3.3, pour tout  $x, y \in [0, 1]$  on a

$$|u_n(x)| \leq CK \text{ et } |u_n(x) - u_n(y)| \leq CK\sqrt{|x - y|}. \quad (4.16)$$

(On a appliqué l'inégalité de Cauchy-Schwarz à (4.8).) Soit une suite (dénombrable)  $(x_p)_{p \geq 1}$  de points de  $[0, 1]$  qui est dense dans cet intervalle (par exemple, les points de  $\mathbb{Q} \cap [0, 1]$ ). Pour  $p$  fixé, la suite  $u_n(x_p)$  est bornée dans  $\mathbb{R}$  à cause de la première inégalité de (4.16). On peut donc en extraire une sous-suite qui converge dans  $\mathbb{R}$ . On

applique d'abord ce procédé à la suite  $u_n(x_1)$  et on obtient une sous-suite  $u_{n_1}(x_1)$  qui converge dans  $\mathbb{R}$ . Puis on extrait de cette sous-suite, indicée par  $n_1$ , une nouvelle sous-suite, indicée par  $n_2$ , telle que  $u_{n_2}(x_2)$  converge dans  $\mathbb{R}$  (mais, bien sûr, on a aussi  $u_{n_2}(x_1)$  qui converge). En extrayant successivement une sous-suite de la précédente, par récurrence, on construit ainsi une sous-suite, indicée par  $n_p$ , telle que  $u_{n_p}(x_p)$  converge, mais aussi  $u_{n_p}(x_k)$  pour  $1 \leq k \leq p$ . Évidemment, les sous-suites  $u_{n_p}$  sont de plus en plus “maigres” à mesure que  $p$  est grand. Pour éviter qu'il n'en reste plus rien “à la limite”, on utilise un argument d'extraction de suite diagonale. Autrement dit, on extrait une dernière sous-suite (dite diagonale) de cet ensemble de sous-suites, c'est-à-dire qu'on choisit le premier élément de la première sous-suite  $u_{n_1}$ , puis le deuxième élément de la deuxième  $u_{n_2}$ , et ainsi de suite le  $p$ -ème élément de la  $p$ -ème  $u_{n_p}$ . La suite ainsi obtenue est notée  $u_m$  et on voit que, pour tout  $p \geq 1$ , la suite  $u_m(x_p)$  converge dans  $\mathbb{R}$ .

Soit maintenant  $x \in [0, 1]$  et  $\epsilon > 0$  (un petit paramètre). Comme la suite  $(x_p)_{p \geq 1}$  est dense dans  $[0, 1]$ , il existe un point  $x_p$  tel que  $|x - x_p| \leq \epsilon$ . Par ailleurs, comme  $u_m(x_p)$  converge dans  $\mathbb{R}$  quand  $m$  tend vers l'infini, il existe  $m_0$  tel que, pour tout  $m, m' \geq m_0$ , on a  $|u_m(x_p) - u_{m'}(x_p)| \leq \epsilon$ . Par conséquent, on obtient

$$\begin{aligned} |u_m(x) - u_{m'}(x)| &\leq |u_m(x_p) - u_m(x)| + |u_m(x_p) - u_{m'}(x_p)| + |u_{m'}(x_p) - u_{m'}(x)| \\ &\leq \epsilon + 2CK\sqrt{\epsilon} \end{aligned}$$

ce qui prouve que la suite  $u_m(x)$  est de Cauchy dans  $\mathbb{R}$ , et donc converge pour tout  $x \in [0, 1]$ . Par application du théorème de convergence dominée de Lebesgue, on conclut que la suite  $u_m$  converge dans  $L^2(0, 1)$ .  $\square$

**Exercice 4.3.5** Soit  $\Omega = (0, 1)$  et  $u_n(x) = \sin(2\pi nx)$ . Montrer que la suite  $u_n$  est uniformément bornée dans  $L^2(\Omega)$ , mais qu'il n'existe aucune sous-suite convergente. Pour cela on montrera, grâce à une intégration par parties, que, pour toute fonction  $\phi \in C_c^\infty(\Omega)$ , on a

$$\lim_{n \rightarrow +\infty} \int_0^1 u_n(x) \phi(x) dx = 0,$$

et on en déduira une contradiction si une sous-suite de  $u_n$  converge dans  $L^2(\Omega)$ . Généraliser ce contre exemple à  $H^1(\Omega)$  en considérant une primitive de  $u_n$ .

### 4.3.5 Espaces $H^m(\Omega)$

On peut aisément généraliser la Définition 4.3.1 de l'espace de Sobolev  $H^1(\Omega)$  aux fonctions qui sont  $m \geq 0$  fois dérivables au sens faible. Commençons par donner une convention d'écriture bien utile. Soit  $\alpha = (\alpha_1, \dots, \alpha_N)$  un **multi-indice**, c'est-à-dire un vecteur à  $N$  composantes entières positives  $\alpha_i \geq 0$ . On note  $|\alpha| = \sum_{i=1}^N \alpha_i$  et, pour une fonction  $v$ ,

$$\partial^{\alpha} v(x) = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \dots \partial x_N^{\alpha_N}}(x).$$



A partir de la Définition 4.2.3 de la dérivée première faible, on définit par récurrence sur  $m$  la dérivée d'ordre  $m$  faible : on dit qu'une fonction  $v \in L^2(\Omega)$  est  $m$  fois dérivable au sens faible si toutes ses dérivées partielles faibles d'ordre  $m - 1$  sont dérivables faiblement au sens de la Définition 4.2.3. Remarquons que, dans la définition d'une dérivée croisée, l'ordre de dérivation n'est pas important, à cause du théorème de Schwarz  $\frac{\partial^2 v}{\partial x_i \partial x_j} = \frac{\partial^2 v}{\partial x_j \partial x_i}$ , ce qui justifie la notation  $\partial^\alpha v$  où l'ordre de dérivation n'est pas indiqué.

**Définition 4.3.23** Pour un entier  $m \geq 0$ , l'espace de Sobolev  $H^m(\Omega)$  est défini par

$$H^m(\Omega) = \{v \in L^2(\Omega) \text{ tel que, } \forall \alpha \text{ avec } |\alpha| \leq m, \partial^\alpha v \in L^2(\Omega)\}, \quad (4.17)$$

où la dérivée partielle  $\partial^\alpha v$  est à prendre au sens faible.

Nous laissons au lecteur le soin de vérifier le résultat facile suivant.

**Proposition 4.3.24** Muni du produit scalaire

$$\langle u, v \rangle = \int_{\Omega} \sum_{|\alpha| \leq m} \partial^\alpha u(x) \partial^\alpha v(x) dx \quad (4.18)$$

et de la norme  $\|u\|_{H^m(\Omega)} = \sqrt{\langle u, u \rangle}$ , l'espace de Sobolev  $H^m(\Omega)$  est un espace de Hilbert.

Les fonctions de  $H^m(\Omega)$  ne sont pas toujours continues ou régulières (cela dépend de  $m$  et de la dimension  $N$ ), mais si  $m$  est suffisamment grand alors toute fonction de  $H^m(\Omega)$  est continue. Rappelons qu'en vertu du Lemme 4.3.3, en dimension d'espace  $N = 1$ , les fonctions de  $H^1(\Omega)$  sont continues. Nous admettrons le résultat suivant qui généralise le Lemme 4.3.3 aux dimensions supérieures (voir le Lemme 4.4.9 pour la démonstration, plus simple, d'un résultat similaire).

**Théorème 4.3.25** Si  $\Omega$  est un ouvert borné régulier de classe  $C^1$ , et si  $m > N/2$ , alors  $H^m(\Omega)$  est un sous-espace de l'ensemble  $C(\overline{\Omega})$  des fonctions continues sur  $\overline{\Omega}$ .

**Remarque 4.3.26** Par application réitérée du Théorème 4.3.25 à une fonction et à ses dérivées, on peut en fait améliorer sa conclusion. S'il existe un entier  $k \geq 0$  tel que  $m - N/2 > k$ , alors  $H^m(\Omega)$  est un sous-espace de l'ensemble  $C^k(\overline{\Omega})$  des fonctions  $k$  fois différentiables sur  $\overline{\Omega}$ . •

La "morale" du Théorème 4.3.25 est que plus  $m$  est grand, plus les fonctions de  $H^m(\Omega)$  sont régulières, c'est-à-dire dérivables au sens usuel (il suffit d'appliquer successivement le Théorème 4.3.25 à une fonction  $v \in H^m(\Omega)$  et à ses dérivées  $\partial^\alpha v \in H^{m-|\alpha|}(\Omega)$ ).

Comme pour  $H^1(\Omega)$ , les fonctions régulières sont denses dans  $H^m(\Omega)$  (si du moins l'ouvert  $\Omega$  est régulier ; voir la Définition 3.2.5). La démonstration du Théorème de densité 4.3.5 se généralise très facilement à  $H^m(\Omega)$ . Nous ne la répéterons pas et nous énonçons seulement le résultat de densité suivant.

**Théorème 4.3.27** *Si  $\Omega$  est un ouvert borné régulier de classe  $C^m$ , ou bien si  $\Omega = \mathbb{R}_+^N$ , alors  $C_c^\infty(\overline{\Omega})$  est dense dans  $H^m(\Omega)$ .*

On peut aussi obtenir des résultats de trace et des formules de Green d'ordre plus élevés pour l'espace  $H^m(\Omega)$ . Par souci de simplicité, nous nous contentons de traiter le cas  $m = 2$  (qui est le seul que nous utiliserons par la suite).

**Théorème 4.3.28** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$ . On définit l'application trace  $\gamma_1$*

$$\begin{aligned} H^2(\Omega) \cap C^1(\overline{\Omega}) &\rightarrow L^2(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ v &\rightarrow \gamma_1(v) = \frac{\partial v}{\partial n} \Big|_{\partial\Omega}, \end{aligned} \quad (4.19)$$

avec  $\frac{\partial v}{\partial n} = \nabla u \cdot n$ . Cette application  $\gamma_1$  se prolonge par continuité en une application linéaire continue de  $H^2(\Omega)$  dans  $L^2(\partial\Omega)$ . En particulier, il existe une constante  $C > 0$  telle que, pour toute fonction  $v \in H^2(\Omega)$ , on a

$$\left\| \frac{\partial v}{\partial n} \right\|_{L^2(\partial\Omega)} \leq C \|v\|_{H^2(\Omega)}. \quad (4.20)$$

**Démonstration.** L'existence de l'application trace  $\gamma_1$  (et ses propriétés) est une simple conséquence du précédent Théorème de trace 4.3.13 pour les fonctions de  $H^1(\Omega)$ . En effet, si  $v \in H^2(\Omega)$ , alors  $\nabla v \in H^1(\Omega)^N$  et on peut donc définir la trace de  $\nabla v$  sur  $\partial\Omega$  comme une fonction de  $L^2(\partial\Omega)^N$ . Comme la normale est une fonction continue bornée sur  $\partial\Omega$ , on en déduit bien que  $\frac{\partial v}{\partial n} \in L^2(\partial\Omega)$ .  $\square$

**Remarque 4.3.29** Si  $\Omega$  est un ouvert borné régulier de classe  $C^2$ , on peut améliorer le précédent Théorème de trace 4.3.13. On redéfinit l'application trace  $\gamma_0$

$$\begin{aligned} H^2(\Omega) \cap C(\overline{\Omega}) &\rightarrow H^1(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ v &\rightarrow \gamma_0(v) = v|_{\partial\Omega}, \end{aligned} \quad (4.21)$$

qui se prolonge par continuité en une application linéaire continue de  $H^2(\Omega)$  dans  $H^1(\partial\Omega)$ . Autrement dit, la trace  $\gamma_0(v)$  admet des dérivées tangentielles. Si  $\Omega = \mathbb{R}_+^N$ , ce résultat est assez facile à concevoir et à démontrer. Dans le cas général, il faut savoir définir un espace de Sobolev sur une variété (ici  $\partial\Omega$ ) ce qui dépasse très largement le cadre de ce cours. Néanmoins ceci n'est pas sans intérêt puisque cela permet d'étudier des modèles d'équations aux dérivées partielles "vivant" à la fois dans un domaine et sur son bord (par exemple, des modèles de diffusion volumique et surfacique, ou bien un modèle d'élasticité volumique couplé avec un modèle de coque surfacique).  $\bullet$

Le Théorème de trace 4.3.28 permet de généraliser aux fonctions de  $H^2(\Omega)$  une formule de Green précédemment établie pour des fonctions de classe  $C^2$  au Corollaire 3.2.4.

**Théorème 4.3.30** Soit  $\Omega$  un ouvert borné régulier de classe  $C^2$ . Si  $u \in H^2(\Omega)$  et  $v \in H^1(\Omega)$ , on a

$$\int_{\Omega} \Delta u(x) v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) ds. \quad (4.22)$$

**Démonstration.** Comme (4.22) est vraie pour des fonctions de classe  $C^2$  et que les fonctions régulières sont denses dans  $H^2(\Omega)$  et  $H^1(\Omega)$ , on utilise un argument de densité. Nous renvoyons à la démonstration du Théorème 4.3.15 pour plus de détails. Le seul argument nouveau ici est qu'il faut utiliser la continuité de l'application trace  $\gamma_1$ , c'est-à-dire l'inégalité (4.20).  $\square$

## 4.4 Quelques compléments utiles

Cette section peut être omise en première lecture.

### 4.4.1 Démonstration du Théorème 4.3.5 de densité

Nous commençons par le cas  $\Omega = \mathbb{R}^N$  qui est le plus simple.

**Théorème 4.4.1** L'espace  $C_c^\infty(\mathbb{R}^N)$  des fonctions de classe  $C^\infty$  à support compact dans  $\mathbb{R}^N$  est dense dans  $H^1(\mathbb{R}^N)$ .

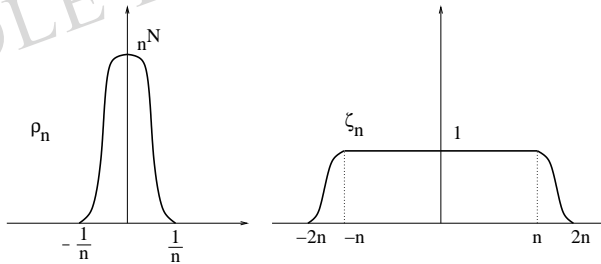


FIGURE 4.3 – Fonctions  $\rho_n$  de régularisation (à gauche), et  $\zeta_n$  de troncature (à droite).

**Démonstration.** La démonstration s'effectue par régularisation et troncature. Soit  $\rho \in C_c^\infty(B)$  (avec  $B$  la boule unité) telle que  $\rho \geq 0$  et  $\int_B \rho(x) dx = 1$ . On définit une suite "régularisante"  $\rho_n(x) = n^N \rho(nx)$  dont le support est contenu dans la boule de rayon  $1/n$  (voir la Figure 4.3). On régularise par convolution une fonction  $v \in H^1(\mathbb{R}^N)$  en définissant

$$v_n(x) = v \star \rho_n(x) = \int_{\mathbb{R}^N} \rho_n(x-y) v(y) dy,$$

qui est de classe  $C^\infty$  et telle que  $\nabla v_n = (\nabla v) \star \rho_n$ . On vérifie facilement que  $v_n$  (respectivement  $\nabla v_n$ ) converge vers  $v$  (respectivement  $\nabla v$ ) dans  $L^2(\mathbb{R}^N)$  : c'est évident si  $v$  (respectivement  $\nabla v$ ) est continu, et par densité des fonctions régulières dans  $L^2(\mathbb{R}^N)$  (voir le Théorème

4.2.1) le résultat s'étend à toute fonction de  $L^2(\mathbb{R}^N)$ . Il ne reste plus qu'à tronquer la suite  $v_n$  afin de lui donner un support compact. Soit  $\zeta \in C_c^\infty(\mathbb{R}^N)$  telle que  $0 \leq \zeta \leq 1$  et  $\zeta(x) = 1$  si  $|x| \leq 1$ ,  $\zeta(x) = 0$  si  $|x| \geq 2$ . On pose  $\zeta_n(x) = \zeta(\frac{x}{n})$  (voir la Figure 4.3) et on tronque  $v_n$  en définissant  $\tilde{v}_n(x) = v_n(x)\zeta_n(x)$ . On vérifie facilement aussi que  $\tilde{v}_n$  (respectivement  $\nabla \tilde{v}_n$ ) converge vers  $v$  (respectivement  $\nabla v$ ) dans  $L^2(\mathbb{R}^N)$ .  $\square$

Le Théorème 4.3.5 de densité pour un ouvert régulier ou pour le demi-espace est une conséquence immédiate du résultat suivant combiné avec le Théorème 4.4.1 de densité dans l'espace entier  $\mathbb{R}^N$ .

**Proposition 4.4.2** *Si  $\Omega$  est un ouvert borné régulier de classe  $C^1$ , ou bien si  $\Omega = \mathbb{R}_+^N$ , alors il existe un opérateur de prolongement  $P$  de  $H^1(\Omega)$  dans  $H^1(\mathbb{R}^N)$  qui est une application linéaire continue telle que, pour tout  $v \in H^1(\Omega)$ ,*

1.  $Pv|_\Omega = v$
2.  $\|Pv\|_{L^2(\mathbb{R}^N)} \leq C\|v\|_{L^2(\Omega)}$
3.  $\|Pv\|_{H^1(\mathbb{R}^N)} \leq C\|v\|_{H^1(\Omega)}$

où la constante  $C > 0$  dépend seulement de  $\Omega$ .

**Démonstration.** Tout d'abord, on démontre le résultat pour  $\Omega = \mathbb{R}_+^N$ . On note  $x = (x', x_N)$  avec  $x' = (x_1, \dots, x_{N-1})$ . Soit  $v \in H^1(\mathbb{R}_+^N)$ . On définit

$$Pv(x) = \begin{cases} v(x', x_N) & \text{si } x_N > 0 \\ v(x', -x_N) & \text{si } x_N < 0. \end{cases}$$

On vérifie alors que, pour  $1 \leq i \leq N-1$ ,

$$\frac{\partial Pv}{\partial x_i}(x) = \begin{cases} \frac{\partial v}{\partial x_i}(x', x_N) & \text{si } x_N > 0 \\ \frac{\partial v}{\partial x_i}(x', -x_N) & \text{si } x_N < 0, \end{cases}$$

et que

$$\frac{\partial Pv}{\partial x_N}(x) = \begin{cases} \frac{\partial v}{\partial x_N}(x', x_N) & \text{si } x_N > 0 \\ -\frac{\partial v}{\partial x_N}(x', -x_N) & \text{si } x_N < 0. \end{cases}$$

Ces égalités sont évidentes si  $v$  est une fonction régulière, mais nécessitent une justification lorsque  $v$  n'est dérivable que faiblement. Nous ne détaillons pas ici les arguments (faciles) qui justifient ces égalités (il faut notamment utiliser la symétrie par réflexion de  $Pv$ ; voir [36] pour les détails). On en déduit donc les propriétés désirées pour l'opérateur de prolongement  $P$  avec la constante  $C = \sqrt{2}$  (dans le cas  $\Omega = \mathbb{R}_+^N$ ).

Si  $\Omega$  est un ouvert borné régulier de classe  $C^1$ , on utilise un argument de "cartes locales" pour se ramener au cas  $\Omega = \mathbb{R}_+^N$ . En reprenant les notations de la Définition 3.2.5 d'un ouvert régulier, il existe un recouvrement fini de  $\Omega$  par des ouverts  $(\omega_i)_{0 \leq i \leq I}$ . On introduit alors une "partition de l'unité" associée à ce recouvrement, c'est-à-dire des fonctions  $(\theta_i)_{0 \leq i \leq I}$  de  $C_c^\infty(\mathbb{R}^N)$  telles que

$$\theta_i \in C_c^\infty(\omega_i), \quad 0 \leq \theta_i(x) \leq 1, \quad \sum_{i=0}^I \theta_i(x) = 1 \text{ dans } \overline{\Omega}.$$

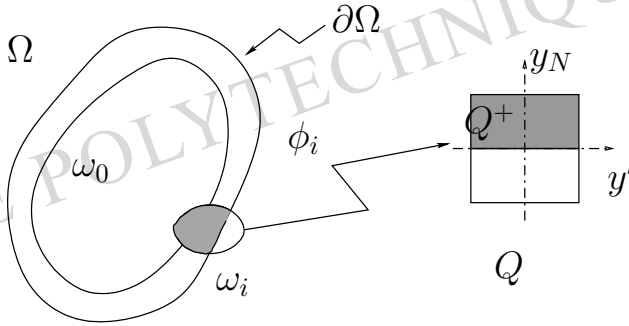


FIGURE 4.4 – Cartes locales d'un ouvert régulier.

(L'existence d'une telle partition de l'unité est classique : voir le théorème 3.2.9 dans [7].) On va définir  $Pv$  sous la forme

$$Pv = \sum_{i=0}^I P_i(\theta_i v),$$

où chaque opérateur  $P_i$  est défini localement dans  $\omega_i$ . Comme  $\theta_0 v$  est à support compact dans  $\Omega$ , on définit  $P_0(\theta_0 v)$  comme l'extension de  $\theta_0 v$  par zéro en dehors de  $\Omega$ . Pour chaque  $i \in \{1, \dots, I\}$ , notant  $\phi_i$  l'application qui transforme  $\omega_i$  en un domaine de référence  $Q$  (voir la Définition 3.2.5 et la Figure 4.4), on pose

$$w_i = (\theta_i v) \circ (\phi_i^{-1}|_{Q^+}) \text{ avec } Q^+ = Q \cap \mathbb{R}_+^N.$$

Cette fonction  $w_i$  appartient à  $H^1(Q^+)$  et est nulle dans un voisinage de  $\partial Q^+ \cap \mathbb{R}_+^N$ . Si on l'étend par 0 dans  $\mathbb{R}_+^N \setminus Q^+$ , on obtient une fonction  $\tilde{w}_i \in H^1(\mathbb{R}_+^N)$ . On peut alors prolonger par réflexion  $\tilde{w}_i$  pour obtenir une fonction  $P\tilde{w}_i \in H^1(\mathbb{R}^N)$  (on utilise l'opérateur de prolongement  $P$  qu'on vient de construire pour  $\mathbb{R}_+^N$ ). On revient dans  $\omega_i$  et on pose

$$P_i(\theta_i v) = (P\tilde{w}_i) \circ \phi_i.$$

Par la régularité  $C^1$  de  $\phi_i$  et de son inverse on obtient les propriétés désirées pour  $P_i$  et donc  $P$ .  $\square$

**Remarque 4.4.3** L'argument de “cartes locales” utilisé ci-dessus est très classique et d'utilisation courante dans de nombreuses démonstrations. Comme il est tout de même assez technique, nous l'invoquerons parfois sans donner plus de détails.  $\bullet$

#### 4.4.2 Espace $H(\text{div})$

Nous introduisons un autre espace, intermédiaire entre  $L^2(\Omega)$  et  $H^1(\Omega)$ , pour les fonctions à valeurs vectorielles. Cet espace est très utile dans certaines applications (voir par exemple la Remarque 5.2.15).

**Définition 4.4.4** L'espace  $H(\text{div})$  est défini par

$$H(\text{div}) = \left\{ \sigma \in L^2(\Omega)^N \text{ tel que } \text{div} \sigma \in L^2(\Omega) \right\}, \quad (4.23)$$

où  $\operatorname{div} \sigma$  est la divergence faible de  $\sigma$  au sens de la Définition 4.2.6.

On vérifie facilement que c'est un espace de Hilbert (la démonstration est laissée au lecteur en guise d'exercice).

**Proposition 4.4.5** *Muni du produit scalaire*

$$\langle \sigma, \tau \rangle = \int_{\Omega} (\sigma(x) \cdot \tau(x) + \operatorname{div} \sigma(x) \operatorname{div} \tau(x)) dx \quad (4.24)$$

et de la norme  $\|\sigma\|_{H(\operatorname{div})} = \sqrt{\langle \sigma, \sigma \rangle}$ , l'espace  $H(\operatorname{div})$  est un espace de Hilbert.

Comme pour les espaces de Sobolev, on peut démontrer un résultat de densité des fonctions régulières (nous omettons la démonstration qui est complètement similaire à celle du Théorème 4.3.5).

**Théorème 4.4.6** *Si  $\Omega$  est un ouvert borné régulier de classe  $C^1$ , ou bien si  $\Omega = \mathbb{R}_+^N$ , alors  $C_c^\infty(\overline{\Omega})^N$  est dense dans  $H(\operatorname{div})$ .*

Un des intérêts de l'espace  $H(\operatorname{div})$  est qu'il permet de démontrer un théorème de trace et une formule de Green avec encore moins de régularité que dans l'espace de Sobolev  $H^1(\Omega)$ . En effet, si  $\sigma$  appartient à  $H(\operatorname{div})$ , on ne "contrôle" qu'une seule combinaison de ses dérivées partielles (et non pas toutes comme dans  $H^1(\Omega)$ ), mais on peut néanmoins donner un sens à la **trace normale**  $\sigma \cdot n$  sur  $\partial\Omega$ .

Commençons par rappeler que  $\gamma_0$  désigne l'application trace de  $H^1(\Omega)$  dans  $L^2(\partial\Omega)$  (voir le Théorème de trace 4.3.13) et que  $\operatorname{Im}(\gamma_0) = H^{1/2}(\partial\Omega)$  qui est un sous-espace dense dans  $L^2(\partial\Omega)$  (voir la Remarque 4.3.17). On peut munir  $H^{1/2}(\partial\Omega)$  de la norme suivante

$$\|v\|_{H^{1/2}(\partial\Omega)} = \inf \{ \|\phi\|_{H^1(\Omega)} \text{ tel que } \gamma_0(\phi) = v \}$$

qui en fait un espace de Banach (et même un espace de Hilbert). On définit alors  $H^{-1/2}(\partial\Omega)$  comme le dual de  $H^{1/2}(\partial\Omega)$ .

**Théorème 4.4.7 (Formule de la divergence)** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$ . On définit l'application "trace normale"  $\gamma_n$*

$$\begin{aligned} H(\operatorname{div}) \cap C(\overline{\Omega}) &\rightarrow H^{-1/2}(\partial\Omega) \cap C(\overline{\partial\Omega}) \\ \sigma = (\sigma_i)_{1 \leq i \leq N} &\rightarrow \gamma_n(\sigma) = (\sigma \cdot n)|_{\partial\Omega} \end{aligned}$$

où  $n = (n_i)_{1 \leq i \leq N}$  est la normale unité extérieure à  $\partial\Omega$ . Cette application  $\gamma_n$  se prolonge par continuité en une application linéaire continue de  $H(\operatorname{div})$  dans  $H^{-1/2}(\partial\Omega)$ . De plus, si  $\sigma \in H(\operatorname{div})$  et  $\phi \in H^1(\Omega)$ , on a

$$\int_{\Omega} \operatorname{div} \sigma \phi dx + \int_{\Omega} \sigma \cdot \nabla \phi dx = \langle \sigma \cdot n, \gamma_0(\phi) \rangle_{H^{-1/2}, H^{1/2}(\partial\Omega)}. \quad (4.25)$$

**Démonstration.** Si  $\Omega$  est un ouvert régulier de classe  $C^1$ , l'Exercice 3.2.1 fournit la formule d'intégration, dite de la divergence, par parties suivante

$$\int_{\Omega} \operatorname{div} \sigma \phi dx + \int_{\Omega} \sigma \cdot \nabla \phi dx = \int_{\partial\Omega} \sigma \cdot n \phi ds, \quad (4.26)$$

pour des fonctions régulières  $\sigma$  et  $\phi$ . On remarque agréablement que le “vilain” terme de droite dans (4.25) n’est autre que la “gentille” intégrale de bord usuelle dans (4.26). On voit aisément que les deux termes de gauche de (4.26) ont bien un sens pour  $\phi \in H^1(\Omega)$  et  $\sigma \in H(\text{div})$ . Alors, par densité des fonctions régulières dans  $H^1(\Omega)$  et  $H(\text{div})$ , les termes de gauche de (4.26) sont étendus par continuité et le terme de droite apparaît comme une forme linéaire continue sur l’image de l’application trace  $\text{Im}(\gamma_0)$ , noté  $H^{1/2}(\partial\Omega)$ . Une telle forme linéaire s’écrit exactement comme dans la formule (4.25) et la trace normale  $\gamma_n(\sigma)$  est donc bien définie comme un élément de  $H^{-1/2}(\partial\Omega)$ .  $\square$

On peut bien sûr définir aussi l’équivalent de  $H_0^1(\Omega)$  pour l’espace  $H(\text{div})$ . On définit le sous-espace  $H_0(\text{div})$  de  $H(\text{div})$  comme l’adhérence de  $C_c^\infty(\Omega)$  dans  $H(\text{div})$ . C’est encore un espace de Hilbert qui s’interprète (si l’ouvert est régulier) comme le sous-espace des fonctions de  $H(\text{div})$  dont la trace normale est nulle.

#### 4.4.3 Espaces $W^{m,p}(\Omega)$

Plus généralement, on peut définir des espaces  $W^{m,p}(\Omega)$  pour un entier  $m \geq 0$  et pour un réel  $1 \leq p \leq +\infty$ . Ces espaces sont construits sur l’espace de Banach  $L^p(\Omega)$  (voir (4.1) et (4.2)). Comme nous l’avons dit à la Remarque 4.2.8, la notion de dérivée faible s’étend à  $L^p(\Omega)$ . Nous pouvons donc donner la définition suivante.

**Définition 4.4.8** *Pour tout entier  $m \geq 0$ , l’espace de Sobolev  $W^{m,p}(\Omega)$  est défini par*

$$W^{m,p}(\Omega) = \{v \in L^p(\Omega) \text{ tel que, } \forall \alpha \text{ avec } |\alpha| \leq m, \partial^\alpha v \in L^p(\Omega)\}, \quad (4.27)$$

où la dérivée partielle  $\partial^\alpha v$  est à prendre au sens faible.

Muni de la norme

$$\|u\|_{W^{m,p}(\Omega)} = \left( \sum_{|\alpha| \leq m} \|\partial^\alpha u\|^p \right)^{1/p}$$

on vérifie que  $W^{m,p}(\Omega)$  est un espace de Banach. Ces espaces sont particulièrement importants pour les problèmes non-linéaires (que nous n’aborderons pas ici, voir par exemple [30]), mais aussi dans les problèmes linéaires en raison des célèbres **inégalités de Sobolev**. Nous les énonçons sans démonstration. Si  $\Omega$  est un ouvert régulier, ou bien si  $\Omega = \mathbb{R}^N$  ou  $\Omega = \mathbb{R}_+^N$ , alors

$$\begin{cases} \text{si } p < N & W^{1,p}(\Omega) \subset L^q(\Omega) \quad \forall q \in [1, p^*] \text{ avec } 1/p^* = 1/p - 1/N \\ \text{si } p = N & W^{1,p}(\Omega) \subset L^q(\Omega) \quad \forall q \in [1, +\infty[ \\ \text{si } p > N & W^{1,p}(\Omega) \subset C(\overline{\Omega}), \end{cases} \quad (4.28)$$

avec injection continue, c’est-à-dire que  $W^{1,p}(\Omega) \subset E$  veut dire qu’il existe une constante  $C$  telle que, pour tout  $u \in W^{1,p}(\Omega)$ ,

$$\|u\|_E \leq C \|u\|_{W^{1,p}(\Omega)}.$$

Le cas particulier  $p = 1$  et  $m = N$  est remarquable car on peut démontrer très simplement une inégalité de type Sobolev.

**Lemme 4.4.9** *L'espace  $W^{N,1}(\mathbb{R}^N)$  s'injecte continûment dans l'espace des fonctions continues bornées sur  $\mathbb{R}^N$ , noté  $C_b(\mathbb{R}^N)$ , et pour tout  $u \in W^{N,1}(\mathbb{R}^N)$  on a*

$$\|u\|_{L^\infty(\mathbb{R}^N)} \leq \|u\|_{W^{N,1}(\mathbb{R}^N)}. \quad (4.29)$$

**Démonstration.** Soit  $u \in C_c^\infty(\mathbb{R}^N)$ . Pour  $x = (x_1, \dots, x_N)$ , on a

$$u(x) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_N} \frac{\partial^N u}{\partial x_1 \cdots \partial x_N}(y) dy_1 \dots dy_N,$$

d'où l'on déduit

$$\|u\|_{L^\infty(\mathbb{R}^N)} \leq \left\| \frac{\partial^N u}{\partial x_1 \cdots \partial x_N} \right\|_{L^1(\mathbb{R}^N)} \leq \|u\|_{W^{N,1}(\mathbb{R}^N)}.$$

Or  $C_c^\infty(\mathbb{R}^N)$  est dense dans  $W^{N,1}(\mathbb{R}^N)$  (cela se démontre comme le Théorème 4.4.1 de densité pour  $H^1(\mathbb{R}^N)$ ). Donc, par densité on obtient l'inégalité (4.29) pour tout  $u \in W^{N,1}(\mathbb{R}^N)$ . Par ailleurs, la fermeture de  $C_c^\infty(\mathbb{R}^N)$  pour la norme de  $L^\infty(\mathbb{R}^N)$  est exactement  $C_b(\mathbb{R}^N)$  (en fait, ces deux espaces ont la même norme). Donc, (4.29) implique que les fonctions de  $W^{N,1}(\mathbb{R}^N)$  sont continues et bornées.  $\square$

#### 4.4.4 Dualité

Rappelons que le dual  $V'$  d'un espace de Hilbert  $V$  est l'ensemble des formes linéaires continues sur  $V$ . Par application du Théorème 12.1.18 de représentation de Riesz, le dual de  $L^2(\Omega)$  est identifié à  $L^2(\Omega)$  lui-même. On peut aussi définir le dual d'un espace de Sobolev. En l'occurrence le dual de  $H_0^1(\Omega)$  joue un rôle particulier dans la suite.

**Définition 4.4.10** *Le dual de l'espace de Sobolev  $H_0^1(\Omega)$  est appelé  $H^{-1}(\Omega)$ . On note  $\langle L, \phi \rangle_{H^{-1}, H_0^1(\Omega)} = L(\phi)$  le produit de dualité entre  $H_0^1(\Omega)$  et son dual pour tout forme linéaire continue  $L \in H^{-1}(\Omega)$  et toute fonction  $\phi \in H_0^1(\Omega)$ .*

On peut caractériser ce dual  $H^{-1}(\Omega)$ . On admettra le résultat suivant (voir [8]).

**Proposition 4.4.11** *L'espace  $H^{-1}(\Omega)$  est caractérisé par*

$$H^{-1}(\Omega) = \left\{ f = v_0 + \sum_{i=1}^N \frac{\partial v_i}{\partial x_i} \text{ avec } v_0, v_1, \dots, v_N \in L^2(\Omega) \right\}.$$

Autrement dit, toute forme linéaire continue sur  $H_0^1(\Omega)$ , notée  $L \in H^{-1}(\Omega)$ , s'écrit pour tout  $\phi \in H_0^1(\Omega)$

$$L(\phi) = \int_{\Omega} \left( v_0 \phi - \sum_{i=1}^N v_i \frac{\partial \phi}{\partial x_i} \right) dx$$

avec  $v_0, v_1, \dots, v_N \in L^2(\Omega)$ .

Grâce à l'espace  $H^{-1}(\Omega)$  on peut définir une nouvelle notion de dérivation pour les fonctions de  $L^2(\Omega)$  (plus faible encore que la dérivée faible de la Définition 4.2.3). Devant cet afflux de notions de dérivation, rassurons le lecteur inquiet en disant qu'elles sont toutes des avatars de la dérivation au sens des distributions (c'est justement un des intérêts de la théorie des distributions d'unifier ces divers types de dérivation).



**Lemme 4.4.12** Soit  $v \in L^2(\Omega)$ . Pour  $1 \leq i \leq N$ , on peut définir une forme linéaire continue  $\frac{\partial v}{\partial x_i}$  dans  $H^{-1}(\Omega)$  par la formule

$$\left\langle \frac{\partial v}{\partial x_i}, \phi \right\rangle_{H^{-1}, H_0^1(\Omega)} = - \int_{\Omega} v \frac{\partial \phi}{\partial x_i} dx \quad \forall \phi \in H_0^1(\Omega), \quad (4.30)$$

qui vérifie

$$\left\| \frac{\partial v}{\partial x_i} \right\|_{H^{-1}(\Omega)} \leq \|v\|_{L^2(\Omega)}.$$

Si  $v \in H^1(\Omega)$ , alors la forme linéaire continue  $\frac{\partial v}{\partial x_i}$  coïncide avec la dérivée faible dans  $L^2(\Omega)$  de  $v$ .

**Démonstration.** On vérifie facilement que le membre de droite de (4.30) est une forme linéaire continue sur  $H_0^1(\Omega)$ . Par conséquent, il existe un élément  $L_i \in H^{-1}(\Omega)$  tel que

$$L_i(\phi) = - \int_{\Omega} v \frac{\partial \phi}{\partial x_i} dx.$$

On vérifie aussi facilement que l'application  $v \rightarrow L_i$  est linéaire continue de  $L^2(\Omega)$  dans  $H^{-1}(\Omega)$ , et qu'elle prolonge la dérivation usuelle pour les fonctions  $v$  régulières (ou la dérivation faible pour  $v \in H^1(\Omega)$ ). Par conséquent, on peut étendre par continuité la dérivation à toute fonction de  $L^2(\Omega)$  et noter  $L_i = \frac{\partial v}{\partial x_i}$ .  $\square$

**Remarque 4.4.13** Grâce au théorème de représentation de Riesz on sait qu'on peut identifier le dual d'un espace de Hilbert avec lui-même. Cependant, en pratique on n'identifie jamais  $H^{-1}(\Omega)$  et  $H_0^1(\Omega)$ . En effet, on a défini  $H_0^1(\Omega)$  comme un sous-espace strict (mais dense) de  $L^2(\Omega)$ . Or on a déjà décidé d'identifier  $L^2(\Omega)$  (muni de son produit scalaire usuel) et son dual (c'est juste une convention mais elle est universelle), donc on ne peut pas en plus identifier  $H^{-1}(\Omega)$  et  $H_0^1(\Omega)$  (avec un autre produit scalaire). La situation correcte et usuelle est donc

$$H_0^1(\Omega) \subset L^2(\Omega) \equiv \left( L^2(\Omega) \right)' \subset H^{-1}(\Omega),$$

où les inclusions sont strictes.  $\bullet$

## 4.5 Lien avec les distributions

Toutes les notions introduites dans ce chapitre ont un lien avec la théorie des distributions (voir le cours [7]). Stricto sensu nous n'avons pas besoin de la théorie des distributions pour définir les espaces de Sobolev et résoudre des problèmes aux limites (historiquement, les distributions, en tant que théorie mathématique, sont apparues après les espaces de Sobolev et l'approche variationnelle pour résoudre les équations aux dérivées partielles). Cependant, la théorie des distributions est un cadre unificateur pour tous ces espaces, et ceux des lecteurs qui la connaissent ne manqueront pas de se demander quels sont liens entre cette théorie et ce que nous venons d'exposer. Quant aux lecteurs non familiers avec cette théorie, ils peuvent légitimement s'interroger sur les idées essentielles à la base des distributions. C'est pour satisfaire cette double curiosité que nous avons écrit cette section qui se veut un (très) bref et caricatural résumé de la théorie des distributions.

Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$ . On note  $C_c^\infty(\Omega)$  (ou  $\mathcal{D}(\Omega)$ ) l'espace des fonctions de classe  $C^\infty$  à support compact dans  $\Omega$ . On munit  $C_c^\infty(\Omega)$  d'une "pseudo-topologie", c'est-à-dire qu'on définit une notion de convergence dans  $C_c^\infty(\Omega)$ . On dit qu'une suite  $(\phi_n)_{n \geq 1}$  de  $C_c^\infty(\Omega)$  converge vers  $\phi \in C_c^\infty(\Omega)$  si

1. le support de  $\phi_n$  reste dans un compact  $K$  de  $\Omega$ ,
2. pour tout multi-indice  $\alpha$ ,  $\partial^\alpha \phi_n$  converge uniformément dans  $K$  vers  $\partial^\alpha \phi$ .

L'espace des distributions  $\mathcal{D}'(\Omega)$  est le "dual" de  $\mathcal{D}(\Omega)$ , c'est-à-dire l'espace des formes linéaires "continues" sur  $\mathcal{D}(\Omega)$ . Les guillemets sont d'usage car il n'y a pas de norme définie sur  $\mathcal{D}(\Omega)$ , mais simplement une notion de convergence. Néanmoins, on peut définir précisément  $\mathcal{D}'(\Omega)$  dont les éléments sont appelées **distributions**.

**Définition 4.5.1** Une distribution  $T \in \mathcal{D}'(\Omega)$  est une forme linéaire sur  $\mathcal{D}(\Omega)$  qui vérifie

$$\lim_{n \rightarrow +\infty} T(\phi_n) = T(\phi)$$

pour toute suite  $(\phi_n)_{n \geq 1}$  de  $C_c^\infty(\Omega)$  qui converge vers  $\phi \in C_c^\infty(\Omega)$  au sens défini ci-dessus.

On note  $\langle T, \phi \rangle = T(\phi)$  le produit de dualité entre une distribution  $T \in \mathcal{D}'(\Omega)$  et une fonction  $\phi \in \mathcal{D}(\Omega)$  : ce produit de dualité "généralise" l'intégrale usuelle  $\int_\Omega T \phi dx$ . En effet, on vérifie que si  $f$  est une fonction (localement intégrable dans  $\Omega$ ), alors on peut définir une distribution  $T_f$  par

$$\langle T_f, \phi \rangle = \int_\Omega f \phi dx.$$

Par conséquent, on identifie la fonction et la distribution associée  $T_f \equiv f$ .

On peut aussi munir  $\mathcal{D}'(\Omega)$  d'une notion de convergence : on dit qu'une suite  $T_n \in \mathcal{D}'(\Omega)$  **converge au sens des distributions** vers  $T \in \mathcal{D}'(\Omega)$  si, pour tout  $\phi \in \mathcal{D}(\Omega)$ ,

$$\lim_{n \rightarrow +\infty} \langle T_n, \phi \rangle = \langle T, \phi \rangle.$$

Cette convergence au sens des distributions est une convergence extrêmement faible (ou peu exigeante) car elle correspond à une convergence intégrale ou "en moyenne".

Définissons maintenant la **dérivation au sens des distributions** : si  $T \in \mathcal{D}'(\Omega)$ , on définit  $\frac{\partial T}{\partial x_i} \in \mathcal{D}'(\Omega)$  par

$$\left\langle \frac{\partial T}{\partial x_i}, \phi \right\rangle = -\left\langle T, \frac{\partial \phi}{\partial x_i} \right\rangle \quad \forall \phi \in \mathcal{D}(\Omega).$$

On vérifie qu'effectivement la dérivée  $\frac{\partial T}{\partial x_i}$  est une distribution, c'est-à-dire que les distributions sont infiniment dérivables ! C'est là l'un des plus grands intérêts des distributions. On vérifie aussi que si  $f$  est une fonction dérivable au sens classique, alors sa dérivée au sens des distributions coïncide avec sa dérivée usuelle.

Bien sûr, on reconnaît dans la dérivation faible au sens de la Définition 4.2.3 un cas particulier de dérivation au sens des distributions. De plus, tous les espaces  $L^p(\Omega)$  ou les espaces de Sobolev  $H^m(\Omega)$  sont des sous-espaces de l'espace des distributions  $\mathcal{D}'(\Omega)$ . En particulier, on vérifie que toutes les convergences dans ces espaces impliquent la convergence au sens des distributions (mais la réciproque est fausse). Enfin, les égalités dans les formulations variationnelles (que nous avons interprétées comme des égalités presque partout) impliquent encore plus simplement et plus généralement des égalités au sens des distributions.

Lemme 4.2.4 (dérivation faible)	$u \in L^2(\Omega)$ est dérivable au sens faible si, $\forall i$ , $\left  \int_{\Omega} u \frac{\partial \phi}{\partial x_i} dx \right  \leq C \ \phi\ _{L^2(\Omega)} \quad \forall \phi \in C_c^\infty(\Omega)$
Proposition 4.3.2	$H^1(\Omega)$ est un espace de Hilbert pour le produit scalaire $\langle u, v \rangle = \int_{\Omega} (\nabla u \cdot \nabla v + uv) dx$
Théorème 4.3.5 (théorème de densité)	$C_c^\infty(\overline{\Omega})$ est dense dans $H^1(\Omega)$
Proposition 4.3.10 (inégalité de Poincaré)	$\forall u \in H_0^1(\Omega)$ ( $\Omega$ borné) $\ u\ _{L^2(\Omega)} \leq C \ \nabla u\ _{L^2(\Omega)}$
Théorème 4.3.13 (théorème de trace)	$u \rightarrow u _{\partial\Omega}$ application continue de $H^1(\Omega)$ dans $L^2(\partial\Omega)$
Théorème 4.3.15 (formule de Green)	$\forall u, v \in H^1(\Omega)$ $\int_{\Omega} u \frac{\partial v}{\partial x_i} dx = - \int_{\Omega} v \frac{\partial u}{\partial x_i} dx + \int_{\partial\Omega} uv n_i ds$
Corollaire 4.3.16 (caractérisation de $H_0^1(\Omega)$ )	$H_0^1(\Omega)$ est le sous-espace des fonctions de $H^1(\Omega)$ qui s'annulent sur $\partial\Omega$
Théorème 4.3.21 (théorème de Rellich)	l'injection de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte ( $\Omega$ borné régulier)
Théorème 4.3.30 (formule de Green)	$\forall u \in H^2(\Omega), v \in H^1(\Omega)$ $\int_{\Omega} v \Delta u dx = - \int_{\Omega} \nabla u \cdot \nabla v dx + \int_{\partial\Omega} \frac{\partial u}{\partial n} v ds$

TABLE 4.1 – Principaux résultats sur les espaces de Sobolev qu'il faut absolument connaître.

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

## Chapitre 5

# ÉTUDE MATHÉMATIQUE DES PROBLÈMES ELLIPTIQUES

### 5.1 Introduction

Dans ce chapitre nous terminons l'analyse mathématique des équations aux dérivées partielles de type elliptique commencée au Chapitre 3. Pour montrer que les problèmes aux limites sont bien posés pour ces e.d.p. elliptiques, c'est-à-dire qu'elles admettent une solution, unique, et dépendant continûment des données, nous suivons **l'approche variationnelle** présentée au Chapitre 3 et nous utilisons les **espaces de Sobolev** introduits au Chapitre 4.

Le plan de ce chapitre est le suivant. Dans la Section 5.2 nous expliquons en détail le fonctionnement de l'approche variationnelle pour le Laplacien avec divers types de conditions aux limites. Nous démontrons des **résultats d'existence et d'unicité des solutions**. Nous montrons aussi que ces solutions **minimisent une énergie** et qu'elles vérifient un certain nombre de **propriétés qualitatives** très naturelles et importantes du point de vue des applications (principe du maximum, régularité). La Section 5.3 reprend le même programme mais pour d'autres modèles plus compliqués comme celui de **l'élasticité linéarisée** ou celui des **équations de Stokes**. Si la théorie d'existence et d'unicité est très semblable au cas précédent, il n'en est pas de même de toutes les propriétés qualitatives.

## 5.2 Étude du Laplacien

### 5.2.1 Conditions aux limites de Dirichlet

Nous considérons le problème aux limites suivant

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (5.1)$$

où  $\Omega$  est un ouvert borné de l'espace  $\mathbb{R}^N$ , et  $f$  est un second membre qui appartient à l'espace  $L^2(\Omega)$ . L'approche variationnelle pour étudier (5.1) est constituée de trois étapes que nous détaillons.

**Étape 1 :** Établissement d'une formulation variationnelle.

Dans une première étape il faut proposer une formulation variationnelle du problème aux limites (5.1), c'est-à-dire qu'il faut trouver une forme bilinéaire  $a(\cdot, \cdot)$ , une forme linéaire  $L(\cdot)$ , et un espace de Hilbert  $V$  tels que (5.1) soit équivalent à :

$$\text{Trouver } u \in V \text{ tel que } a(u, v) = L(v) \text{ pour tout } v \in V. \quad (5.2)$$

Le but de cette première étape est seulement de trouver la formulation variationnelle (5.2) ; on vérifiera l'équivalence précise avec (5.1) plus tard au cours de la troisième étape.

Pour trouver la formulation variationnelle on multiplie l'équation (5.1) par une fonction test régulière  $v$  et on intègre par parties. Ce calcul est principalement formel au sens où l'on suppose l'existence et la régularité de la solution  $u$  afin que tous les calculs effectués soient licites. À l'aide de la formule de Green (4.22) (voir aussi (3.7)) on trouve

$$\int_{\Omega} f v \, dx = - \int_{\Omega} \Delta u v \, dx = \int_{\Omega} \nabla u \cdot \nabla v \, dx - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \, ds. \quad (5.3)$$

Comme  $u$  doit satisfaire une condition aux limites de Dirichlet,  $u = 0$  sur  $\partial\Omega$ , on choisit un espace de Hilbert  $V$  tel que toute fonction  $v \in V$  vérifie aussi  $v = 0$  sur  $\partial\Omega$ . Dans ce cas, l'égalité (5.3) devient

$$\int_{\Omega} \nabla u(x) \cdot \nabla v(x) \, dx = \int_{\Omega} f(x) v(x) \, dx. \quad (5.4)$$

Pour que le terme de gauche de (5.4) ait un sens il suffit que  $\nabla u$  et  $\nabla v$  appartiennent à  $L^2(\Omega)$  (composante par composante), et pour que le terme de droite de (5.4) ait aussi un sens il suffit que  $v$  appartienne à  $L^2(\Omega)$  (on a supposé que  $f \in L^2(\Omega)$ ). Par conséquent, un choix raisonnable pour l'espace de Hilbert est  $V = H_0^1(\Omega)$ , le sous-espace de  $H^1(\Omega)$  dont les éléments s'annulent sur le bord  $\partial\Omega$ .

En conclusion, la formulation variationnelle proposée pour (5.1) est :

$$\text{trouver } u \in H_0^1(\Omega) \text{ tel que } \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\Omega} f v \, dx \quad \forall v \in H_0^1(\Omega). \quad (5.5)$$

Évidemment, nous avons fait un certain nombre de choix pour arriver à (5.5) ; d'autres choix nous auraient conduit à d'autres formulations variationnelles possibles. La justification de (5.5) s'effectuera donc a posteriori : tout d'abord, la deuxième étape consiste à vérifier que (5.5) admet bien une unique solution, puis la troisième étape que la solution de (5.5) est aussi une solution du problème aux limites (5.1) (dans un sens à préciser).

**Étape 2 :** Résolution de la formulation variationnelle.

Dans cette deuxième étape nous vérifions que la formulation variationnelle (5.5) admet une solution unique. Pour cela nous utilisons le Théorème de Lax-Milgram 3.3.1 dont nous vérifions les hypothèses avec les notations

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

On voit facilement en utilisant l'inégalité de Cauchy-Schwarz que  $a$  est une forme bilinéaire continue sur  $H_0^1(\Omega)$  et que  $L$  est une forme linéaire continue sur  $H_0^1(\Omega)$ . De plus, en vertu de l'inégalité de Poincaré (voir le Corollaire 4.3.12 ; on utilise ici le caractère borné de l'ouvert  $\Omega$ ), la forme bilinéaire  $a$  est coercive, c'est-à-dire qu'il existe  $\nu > 0$  tel que

$$a(v, v) = \int_{\Omega} |\nabla v(x)|^2 dx \geq \nu \|v\|_{H_0^1(\Omega)}^2 \quad \forall v \in H_0^1(\Omega).$$

Comme  $H_0^1(\Omega)$  est un espace de Hilbert (voir la Proposition 4.3.9), toutes les hypothèses du Théorème de Lax-Milgram 3.3.1 sont satisfaites et on peut donc conclure qu'il existe une unique solution  $u \in H_0^1(\Omega)$  de la formulation variationnelle (5.5).

**Remarque 5.2.1** Nous verrons plus loin au Chapitre 9 que, dans le cas présent, comme la forme bilinéaire  $a$  est symétrique, il existe un autre argument que le Théorème de Lax-Milgram pour conclure. En effet, la solution de la formulation variationnelle est dans ce cas l'unique point de minimum de l'énergie définie par

$$J(v) = \frac{1}{2}a(v, v) - L(v) \quad \forall v \in H_0^1(\Omega)$$

(voir la Proposition 5.2.7). Par conséquent, si on démontre que  $J$  a un unique point de minimum, on a ainsi obtenu la solution de la formulation variationnelle. •

**Étape 3 :** Équivalence avec l'équation.

La troisième étape (la dernière et la plus délicate) consiste à vérifier qu'en résolvant la formulation variationnelle (5.5) on a bien résolu le problème aux limites (5.1), et à préciser dans quel sens la solution de (5.5) est aussi une solution de (5.1). En d'autres termes, il s'agit d'interpréter la formulation variationnelle et de retourner à l'équation. Pour cela on procède aux mêmes intégrations par parties qui ont conduit à la formulation variationnelle, mais en sens inverse, et en les justifiant soigneusement.

Cette justification est très facile si l'on suppose que la solution  $u$  de la formulation variationnelle (5.5) est régulière (précisément si  $u \in H^2(\Omega)$ ) et que l'ouvert  $\Omega$  est aussi régulier, ce que nous faisons dans un premier temps. En effet, il suffit d'invoquer la formule de Green (4.22) qui nous donne, pour  $v \in H_0^1(\Omega)$ ,

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = - \int_{\Omega} v \Delta u \, dx$$

puisque  $v = 0$  sur le bord  $\partial\Omega$ . On en déduit alors

$$\int_{\Omega} (\Delta u + f) v \, dx = 0 \quad \forall v \in C_c^\infty(\Omega),$$

ce qui implique, en vertu du Corollaire 4.2.2, que  $-\Delta u = f$  dans  $L^2(\Omega)$ , et on a l'égalité

$$-\Delta u = f \text{ presque partout dans } \Omega. \quad (5.6)$$

De plus, si  $\Omega$  est un ouvert borné régulier de classe  $\mathcal{C}^1$ , alors le Théorème de trace 4.3.13 (ou plus précisément son Corollaire 4.3.16) affirme que toute fonction de  $H_0^1(\Omega)$  a une trace sur  $\partial\Omega$  nulle dans  $L^2(\Omega)$ . On en déduit, en particulier, que

$$u = 0 \text{ presque partout sur } \partial\Omega. \quad (5.7)$$

On a donc bien retrouvé l'équation et la condition aux limites de (5.1).

Si l'on ne suppose plus que la solution  $u$  de (5.5) et l'ouvert  $\Omega$  sont réguliers, il faut travailler davantage (on ne peut plus utiliser la formule de Green (4.22) qui nécessite que  $u \in H^2(\Omega)$ ). On note  $\sigma = \nabla u$  qui est une fonction à valeurs vectorielles dans  $L^2(\Omega)^N$ . Par l'inégalité de Cauchy-Schwarz, on déduit de la formulation variationnelle (5.5) que, pour tout  $v \in H_0^1(\Omega)$ ,

$$\left| \int_{\Omega} \sigma \cdot \nabla v \, dx \right| = \left| \int_{\Omega} f v \, dx \right| \leq C \|v\|_{L^2(\Omega)}. \quad (5.8)$$

Comme  $C_c^\infty(\Omega) \subset H_0^1(\Omega)$ , (5.8) n'est rien d'autre que le critère d'existence d'une divergence faible de  $\sigma$  dans  $L^2(\Omega)$  (voir la Définition 4.2.6 et le Lemme 4.2.7) qui vérifie, pour tout  $v \in C_c^\infty(\Omega)$ ,

$$\int_{\Omega} \sigma \cdot \nabla v \, dx = - \int_{\Omega} \operatorname{div} \sigma v \, dx.$$

On en déduit donc que

$$\int_{\Omega} (\operatorname{div} \sigma + f) v \, dx = 0 \quad \forall v \in C_c^\infty(\Omega),$$

ce qui implique, en vertu du Corollaire 4.2.2, que  $-\operatorname{div} \sigma = f$  dans  $L^2(\Omega)$ . Par conséquent  $\operatorname{div} \sigma = \Delta u$  appartient à  $L^2(\Omega)$  (rappelons que  $\operatorname{div} \nabla = \Delta$ ), et on retrouve



l'équation (5.6). On retrouve la condition aux limites (5.7) comme précédemment si l'ouvert  $\Omega$  est régulier de classe  $\mathcal{C}^1$ . Si  $\Omega$  n'est pas régulier, alors on ne peut pas invoquer le Théorème de trace 4.3.13 pour obtenir (5.7). Néanmoins, le simple fait d'appartenir à  $H_0^1(\Omega)$  est une généralisation de la condition aux limites de Dirichlet pour un ouvert non régulier, et on continuera à écrire **formellement** que  $u = 0$  sur  $\partial\Omega$ .

En conclusion nous avons démontré le résultat suivant.

**Théorème 5.2.2** *Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)$ . Il existe une unique solution  $u \in H_0^1(\Omega)$  de la formulation variationnelle (5.5). De plus,  $u$  vérifie*

$$-\Delta u = f \text{ presque partout dans } \Omega, \quad \text{et } u \in H_0^1(\Omega). \quad (5.9)$$

*Si on suppose en plus que  $\Omega$  est régulier de classe  $\mathcal{C}^1$ , alors  $u$  est solution du problème aux limites (5.1) au sens où*

$$-\Delta u = f \text{ presque partout dans } \Omega, \quad u = 0 \text{ presque partout sur } \partial\Omega.$$

On appelle la solution  $u \in H_0^1(\Omega)$  de la formulation variationnelle (5.5) **solution variationnelle** du problème aux limites (5.1). Par un raccourci de langage bien commode, **on dira que l'unique solution  $u \in H_0^1(\Omega)$  de la formulation variationnelle (5.5) est l'unique solution du problème aux limites (5.1)**. Cette appellation est bien sûr justifiée par le Théorème 5.2.2.

La solution de (5.1), que nous venons d'obtenir, ne vérifie *a priori* l'équation et la condition aux limites que dans un sens "faible", c'est-à-dire presque partout (ou même pire pour la condition aux limites si l'ouvert n'est pas régulier). On parle alors de **solution faible** par opposition aux solutions fortes qu'on aurait pu espérer obtenir dans la formulation classique de (5.1) (voir la Sous-section 3.1.2). De même, on appelle parfois la formulation variationnelle **formulation faible** de l'équation.

**Remarque 5.2.3** En fait, la solution faible peut être une solution forte si le second membre  $f$  est plus régulier. Autrement dit, l'équation et la condition aux limites de (5.1) peuvent être vérifiées en un sens classique, c'est-à-dire pour tout  $x \in \Omega$ , et tout  $x \in \partial\Omega$ , respectivement. C'est ce qu'on appelle un résultat de régularité pour la solution (voir plus loin le Corollaire 5.2.27). •

**Remarque 5.2.4** Il faut bien comprendre le sens précis de l'expression  $\Delta u$  dans l'égalité (5.9) du Théorème 5.2.2. Pour une fonction quelconque  $v$  de  $H_0^1(\Omega)$  nous n'avons pas donné de sens (même faible) à son Laplacien  $\Delta v$ . Par contre, pour la solution  $u \in H_0^1(\Omega)$  de la formulation variationnelle (5.5), nous avons montré que  $\Delta u$  appartient à  $L^2(\Omega)$ . •

Pour que le problème aux limites (5.1) soit bien posé (au sens de Hadamard ; voir la Définition 1.5.3), il faut, en plus de l'existence et de l'unicité de sa solution, montrer que la solution dépend continûment des données. C'est une conséquence immédiate du Théorème de Lax-Milgram 3.3.1 mais nous en donnons un nouvel énoncé et une nouvelle démonstration.

**Proposition 5.2.5** Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ , et soit  $f \in L^2(\Omega)$ . L'application qui à  $f \in L^2(\Omega)$  fait correspondre la solution unique  $u \in H_0^1(\Omega)$  de la formulation variationnelle de (5.1) est linéaire et continue de  $L^2(\Omega)$  dans  $H^1(\Omega)$ . En particulier, il existe une constante  $C > 0$  telle que, pour tout  $f \in L^2(\Omega)$ , on a

$$\|u\|_{H^1(\Omega)} \leq C \|f\|_{L^2(\Omega)}. \quad (5.10)$$

**Remarque 5.2.6** L'inégalité (5.10) est ce qu'on appelle une **estimation d'énergie**. Elle garantit que l'énergie de la solution est contrôlée par celle de la donnée. Les estimations d'énergie sont très naturelles d'un point de vue physique et très utiles d'un point de vue mathématique. •

**Démonstration.** La linéarité de  $f \rightarrow u$  est évidente. Pour obtenir la continuité on prend  $v = u$  dans la formulation variationnelle (5.5)

$$\int_{\Omega} |\nabla u|^2 dx = \int_{\Omega} f u dx.$$

On majore le terme de droite à l'aide de l'inégalité de Cauchy-Schwarz, et on minore celui de gauche par la coercivité de la forme bilinéaire

$$\nu \|u\|_{H^1(\Omega)}^2 \leq \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \|u\|_{H^1(\Omega)},$$

d'où l'on déduit le résultat.  $\square$

Nous avons déjà dit que la formulation variationnelle possède souvent une interprétation physique (c'est, par exemple, le principe des travaux virtuels en mécanique). En fait, la solution de la formulation variationnelle (5.5) réalise le minimum d'une énergie (très naturelle en physique ou en mécanique). Le résultat suivant est une application immédiate de la Proposition 3.3.4.

**Proposition 5.2.7** Soit  $J(v)$  l'énergie définie pour  $v \in H_0^1(\Omega)$  par

$$J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx. \quad (5.11)$$

Soit  $u \in H_0^1(\Omega)$  la solution unique de la formulation variationnelle (5.5). Alors  $u$  est aussi l'unique point de minimum de l'énergie, c'est-à-dire que

$$J(u) = \min_{v \in H_0^1(\Omega)} J(v).$$

Réciproquement, si  $u \in H_0^1(\Omega)$  est un point de minimum de l'énergie  $J(v)$ , alors  $u$  est la solution unique de la formulation variationnelle (5.5).

**Remarque 5.2.8** La Proposition 5.2.7 repose de manière cruciale sur le fait que la forme bilinéaire de la formulation variationnelle est symétrique. Si cela n'est pas le cas, la solution de la formulation variationnelle ne minimise pas l'énergie (voir le contre-exemple de l'Exercice 5.2.3).

Souvent l'origine physique du Laplacien est en fait la recherche des minima de l'énergie  $J(v)$ . Il est remarquable que ce problème de minimisation nécessite de la part de la solution  $u$  moins de régularité que l'équation aux dérivées partielles (une seule dérivée permet de définir  $J(u)$  tandis qu'il en faut deux pour  $\Delta u$ ). Cette constatation confirme le caractère "naturel" de la formulation variationnelle pour analyser une équation aux dérivées partielles. •

**Exercice 5.2.1** A l'aide de l'approche variationnelle démontrer l'existence et l'unicité de la solution de

$$\begin{cases} -\Delta u + u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (5.12)$$

où  $\Omega$  est un ouvert quelconque de l'espace  $\mathbb{R}^N$ , et  $f \in L^2(\Omega)$ . Montrer en particulier que l'ajout d'un terme d'ordre zéro au Laplacien permet de ne pas avoir besoin de l'hypothèse que  $\Omega$  est borné.

**Exercice 5.2.2** Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ . A l'aide de l'approche variationnelle démontrer l'existence et l'unicité de la solution du problème suivant de convection-diffusion

$$\begin{cases} V \cdot \nabla u - \Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (5.13)$$

où  $f \in L^2(\Omega)$  et  $V$  est une fonction régulière à valeurs vectorielles telle que  $\operatorname{div} V = 0$  dans  $\Omega$ .

**Exercice 5.2.3** On reprend les notations et hypothèses de l'Exercice 5.2.2. Montrer que tout  $v \in H_0^1(\Omega)$  vérifie

$$\int_{\Omega} v V \cdot \nabla v \, dx = 0.$$

Montrer que la solution de la formulation variationnelle du problème de convection-diffusion ne minimise pas dans  $H_0^1(\Omega)$  l'énergie

$$J(v) = \frac{1}{2} \int_{\Omega} (|\nabla v|^2 + v V \cdot \nabla v) \, dx - \int_{\Omega} f v \, dx.$$

La partie "unicité" du Théorème 5.2.2 est utile pour démontrer des propriétés de symétrie comme l'indique l'exercice suivant.

**Exercice 5.2.4** On considère à nouveau le problème aux limites (5.1). On suppose que l'ouvert  $\Omega$  est symétrique par rapport à l'hyperplan  $x_N = 0$  de même que la donnée  $f$  (i.e.  $f(x', x_N) = f(x', -x_N)$ ). Montrer que la solution de (5.1) a la même symétrie. Montrer que (5.1) est équivalent à un problème aux limites posé sur  $\Omega^+ = \Omega \cap \{x_N > 0\}$  avec une condition aux limites de Neumann sur  $\Omega \cap \{x_N = 0\}$ .

**Remarque 5.2.9** Jusqu'ici nous avons supposé que le second membre  $f$  de (5.1) appartenait à  $L^2(\Omega)$ , mais une grande partie des résultats reste valable si on suppose seulement que  $f \in H^{-1}(\Omega)$  (espace de "fonctions" moins régulières). Les deux premières étapes restent identiques à condition de remplacer l'intégrale usuelle  $\int_{\Omega} f v \, dx$  par le produit de dualité  $\langle f, v \rangle_{H^{-1}, H_0^1(\Omega)}$  (en particulier  $L(v)$  reste bien une forme linéaire continue sur  $H_0^1(\Omega)$ ). Dans la troisième étape, l'égalité  $-\Delta u = f$  n'a plus lieu qu'au sens de l'égalité entre éléments de  $H^{-1}(\Omega)$  (et non plus presque partout dans  $\Omega$ ).

Ce raffinement mathématique peut correspondre à une modélisation physique pertinente. Prenons comme exemple le cas d'un second membre concentré sur une hypersurface plutôt que distribué dans tout  $\Omega$ . Soit  $\Gamma$  une hypersurface (une variété de dimension  $N - 1$ ) régulière incluse dans  $\Omega$ . Pour modéliser un terme source concentré sur  $\Gamma$ , on prend  $\tilde{f} \in L^2(\Gamma)$  et on définit  $f \in H^{-1}(\Omega)$  par

$$\langle f, v \rangle_{H^{-1}, H_0^1(\Omega)} = \int_{\Gamma} \tilde{f} v \, ds,$$

qui est bien une forme linéaire continue sur  $H_0^1(\Omega)$  grâce au Théorème de trace 4.3.13.

•

**Remarque 5.2.10** Dans (5.1) 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

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = u_0 & \text{sur } \partial\Omega, \end{cases} \quad (5.14)$$

où  $u_0$  est la trace sur  $\partial\Omega$  d'une fonction de  $H^1(\Omega)$ , toujours notée  $u_0$ . Pour analyser (5.14) on pose  $u = u_0 + \tilde{u}$ , et on cherche la solution de

$$\begin{cases} -\Delta \tilde{u} = \tilde{f} = f + \Delta u_0 & \text{dans } \Omega \\ \tilde{u} = 0 & \text{sur } \partial\Omega. \end{cases} \quad (5.15)$$

Suivant la Remarque 5.2.9 on peut résoudre (5.15) par l'approche variationnelle car  $\tilde{f}$  appartient à  $H^{-1}(\Omega)$ . En effet,

$$\langle \tilde{f}, v \rangle_{H^{-1}, H_0^1(\Omega)} = \int_{\Omega} \tilde{f} v \, dx - \int_{\Omega} \nabla u_0 \cdot \nabla v \, dx$$

est bien une forme linéaire continue sur  $H_0^1(\Omega)$ .

## 5.2.2 Conditions aux limites de Neumann

Nous considérons le problème aux limites suivant

$$\begin{cases} -\Delta u + u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = g & \text{sur } \partial\Omega \end{cases} \quad (5.16)$$

où  $\Omega$  est un ouvert (non nécessairement borné) de l'espace  $\mathbb{R}^N$ ,  $f \in L^2(\Omega)$  et  $g \in L^2(\partial\Omega)$ . L'équation de (5.16) est une variante du Laplacien où nous avons ajouté un terme d'ordre zéro afin d'éviter (en premier abord) une difficulté que nous réglerons plus loin au Théorème 5.2.18. L'approche variationnelle pour étudier (5.16) est sensiblement différente de celle présentée à la sous-section précédente dans le traitement des conditions aux limites. C'est pourquoi nous détaillons à nouveau les trois étapes de l'approche.

**Étape 1 :** Établissement d'une formulation variationnelle.

Pour trouver la formulation variationnelle on multiplie l'équation (5.16) par une fonction test régulière  $v$  et on intègre par parties en admettant que la solution  $u$  est suffisamment régulière afin que tous les calculs effectués soient licites. La formule de Green (4.22) (voir aussi (3.7)) donne

$$\begin{aligned} \int_{\Omega} f(x)v(x) dx &= \int_{\Omega} (-\Delta u(x) + u(x)) v(x) dx \\ &= \int_{\Omega} (\nabla u(x) \cdot \nabla v(x) + u(x)v(x)) dx - \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x) ds \quad (5.17) \\ &= \int_{\Omega} (\nabla u(x) \cdot \nabla v(x) + u(x)v(x)) dx - \int_{\partial\Omega} g(x)v(x) ds. \end{aligned}$$

Nous avons utilisé la condition aux limites de Neumann dans (5.17) et il n'est pas nécessaire de l'inscrire dans le choix de l'espace de Hilbert  $V$ . Pour que le premier et les deux derniers termes de (5.17) aient un sens il suffit de prendre  $V = H^1(\Omega)$  (on utilise le Théorème de trace 4.3.13 pour justifier l'intégrale de bord).

En conclusion, la formulation variationnelle proposée pour (5.16) est : trouver  $u \in H^1(\Omega)$  tel que

$$\int_{\Omega} (\nabla u \cdot \nabla v + uv) dx = \int_{\partial\Omega} gv ds + \int_{\Omega} fv dx \quad \forall v \in H^1(\Omega). \quad (5.18)$$

Les étapes suivantes justifieront le choix de (5.18).

**Remarque 5.2.11** La principale différence entre la formulation variationnelle (5.18) pour une condition aux limites de Neumann et celle (5.5) pour une condition aux limites de Dirichlet vient de ce que la condition de Dirichlet est inscrite dans le choix de l'espace alors que la condition de Neumann apparaît dans la forme linéaire mais pas dans l'espace. La condition de Dirichlet est dite **essentielle** (ou explicite) car elle est forcée par l'appartenance à un espace, tandis que la condition de Neumann est dite **naturelle** (ou implicite) car elle découle de l'intégration par partie qui conduit à la formulation variationnelle. •

**Étape 2 :** Résolution de la formulation variationnelle.

Dans cette deuxième étape nous vérifions que la formulation variationnelle (5.18) admet une solution unique. Pour cela nous utilisons le Théorème de Lax-Milgram 3.3.1 dont nous vérifions les hypothèses avec les notations

$$a(u, v) = \int_{\Omega} (\nabla u \cdot \nabla v + uv) dx \text{ et } L(v) = \int_{\partial\Omega} gv ds + \int_{\Omega} fv dx.$$

En utilisant l'inégalité de Cauchy-Schwarz et à l'aide du Théorème de trace 4.3.13, on voit clairement que  $a$  est une forme bilinéaire continue sur  $H^1(\Omega)$  et que  $L$  est une forme linéaire continue sur  $H^1(\Omega)$ . Par ailleurs, la forme bilinéaire  $a$  est manifestement coercive (c'est pour cela qu'on a ajouté un terme d'ordre zéro au Laplacien) car

$$a(v, v) = \|v\|_{H^1(\Omega)}^2 \quad \forall v \in H^1(\Omega).$$

Comme  $H^1(\Omega)$  est un espace de Hilbert (voir la Proposition 4.3.2), toutes les hypothèses du Théorème de Lax-Milgram 3.3.1 sont satisfaites et on peut donc conclure qu'il existe une unique solution  $u \in H^1(\Omega)$  de la formulation variationnelle (5.18).

**Remarque 5.2.12** Pour analyser le problème aux limites (5.16) dans le cas où  $g = 0$ , on aurait pu être tenté d'inscrire la condition aux limites de Neumann dans l'espace de Hilbert  $V$ . Il n'est pas possible de choisir  $V = \{v \in H^1(\Omega), \frac{\partial v}{\partial n} = 0 \text{ sur } \partial\Omega\}$  car, pour une fonction  $v \in H^1(\Omega)$   $\frac{\partial v}{\partial n}$  n'a pas de sens sur  $\partial\Omega$ . En effet,  $\nabla v$  n'est qu'une fonction de  $L^2(\Omega)$  (composante par composante) et on sait qu'il n'y a pas de notion de trace sur  $\partial\Omega$  pour les fonctions de  $L^2(\Omega)$ . On pourrait choisir  $V = \{v \in H^2(\Omega), \frac{\partial v}{\partial n} = 0 \text{ sur } \partial\Omega\}$  qui est bien un sous-espace fermé de  $H^2(\Omega)$  en vertu du Théorème de trace 4.3.28. Mais avec ce dernier choix une nouvelle difficulté surgit : la forme bilinéaire  $a$  n'est pas coercive sur  $V$  et on ne peut pas appliquer le Théorème de Lax-Milgram. Il n'y a donc pas moyen de prendre en compte la condition aux limites de Neumann dans le choix de l'espace de Hilbert. •

### Étape 3 : Équivalence avec l'équation.

Nous interprétons maintenant la formulation variationnelle (5.18) pour vérifier qu'on a bien résolu le problème aux limites (5.16), dans un sens à préciser. Nous allons supposer que les données sont régulières (voir la Remarque 5.2.15 lorsque l'on ne fait pas cette hypothèse). Plus précisément nous allons supposer que nous sommes dans les conditions d'application du lemme suivant de régularité que nous admettrons (voir la sous-section 5.2.4 pour des résultats similaires).

**Lemme 5.2.13** *Soit  $\Omega$  un ouvert régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)$  et  $g$  la trace sur  $\partial\Omega$  d'une fonction de  $H^1(\Omega)$ . Alors la solution  $u$  de la formulation variationnelle (5.18) appartient à  $H^2(\Omega)$ .*

Grâce au Lemme 5.2.13 on peut utiliser la formule de Green du Théorème 4.3.30

$$\int_{\Omega} \Delta u(x) v(x) dx = - \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) ds. \quad (5.19)$$

qui est valable pour  $u \in H^2(\Omega)$  et  $v \in H^1(\Omega)$ . Rappelons que l'intégrale de bord dans (5.19) a bien un sens à cause du Théorème de trace 4.3.28 qui affirme que pour  $u \in H^2(\Omega)$  la dérivée normale  $\frac{\partial u}{\partial n}$  a un sens dans  $L^2(\partial\Omega)$ . On déduit alors de (5.18) et (5.19) que, pour tout  $v \in H^1(\Omega)$ ,

$$\int_{\Omega} (\Delta u - u + f)v \, dx = \int_{\partial\Omega} \left( \frac{\partial u}{\partial n} - g \right) v \, ds. \quad (5.20)$$

Si l'on prend  $v \in C_c^\infty(\Omega) \subset H^1(\Omega)$  dans (5.20), le terme de bord disparaît et l'on déduit, en vertu du Corollaire 4.2.2, que  $\Delta u - u + f = 0$  dans  $L^2(\Omega)$ , donc presque partout dans  $\Omega$ . Par conséquent, le membre de gauche de (5.20) est nul, donc

$$\int_{\partial\Omega} \left( g - \frac{\partial u}{\partial n} \right) v \, ds = 0 \quad \forall v \in H^1(\Omega).$$

Or l'image de  $H^1(\Omega)$  par l'application trace est dense dans  $L^2(\partial\Omega)$  (voir la remarque 4.3.17), ce qui entraîne que  $g - \frac{\partial u}{\partial n} = 0$  dans  $L^2(\partial\Omega)$ , et donc presque partout sur  $\partial\Omega$ . En conclusion nous avons démontré le résultat suivant.

**Théorème 5.2.14** *Soit  $\Omega$  un ouvert régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)$  et  $g$  la trace sur  $\partial\Omega$  d'une fonction de  $H^1(\Omega)$ . Il existe une unique solution  $u \in H^1(\Omega)$  de la formulation variationnelle (5.18). De plus,  $u$  appartient à  $H^2(\Omega)$  et est solution de (5.16) au sens où*

$$-\Delta u + u = f \text{ presque partout dans } \Omega, \quad \frac{\partial u}{\partial n} = g \text{ presque partout sur } \partial\Omega.$$

**Exercice 5.2.5** Démontrer que l'unique solution  $u \in H^1(\Omega)$  de la formulation variationnelle (5.18) vérifie l'estimation d'énergie suivante

$$\|u\|_{H^1(\Omega)} \leq C (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega)}),$$

où  $C > 0$  est une constante qui ne dépend pas de  $u, f$  et  $g$ .

Comme dans la sous-section précédente, par habitude (et lorsque le contexte ne prête pas à confusion) on dit que l'unique solution  $u \in H^1(\Omega)$  de la formulation variationnelle (5.18) est l'unique **solution faible** du problème aux limites (5.16).

**Remarque 5.2.15 (délicate)** Lorsque l'ouvert  $\Omega$  n'est pas régulier et que  $g$  est seulement une fonction de  $L^2(\partial\Omega)$ , on a aussi démontré l'existence d'une unique solution  $u \in H^1(\Omega)$  de la formulation variationnelle (5.18). Pour montrer que cette solution vérifie l'équation  $-\Delta u + u = f$  presque partout dans  $\Omega$ , il faut utiliser un argument un peu plus compliqué. En notant  $\sigma = \nabla u \in L^2(\Omega)^N$ , on déduit de (5.18) que, pour tout  $v \in C_c^\infty(\Omega)$ ,

$$\left| \int_{\Omega} \sigma \cdot \nabla v \, dx \right| \leq \left| \int_{\Omega} uv \, dx \right| + \left| \int_{\Omega} fv \, dx \right| \leq C \|v\|_{L^2(\Omega)}$$

qui n'est rien d'autre que le critère d'existence d'une divergence faible de  $\sigma$  dans  $L^2(\Omega)$  (voir la Définition 4.2.6 et le Lemme 4.2.7). On a donc  $\operatorname{div} \sigma \in L^2(\Omega)$  et

$$\int_{\Omega} (\operatorname{div} \sigma - u + f) v \, dx = 0 \quad \forall v \in C_c^\infty(\Omega),$$

ce qui implique, en vertu du Corollaire 4.2.2, que  $-\operatorname{div} \sigma = -\Delta u = f - u$  dans  $L^2(\Omega)$ .

On peut aussi retrouver la condition aux limites de Neumann, dans un sens très faible, si l'ouvert est régulier (mais pas  $g$ ). Pour ce faire on utilise l'espace  $H(\operatorname{div})$  introduit à la Sous-section 4.4.2 et défini par  $H(\operatorname{div}) = \{\sigma \in L^2(\Omega)^N \text{ tel que } \operatorname{div} \sigma \in L^2(\Omega)\}$ . Le Théorème 4.4.7 affirme que, si  $\Omega$  est un ouvert régulier de classe  $\mathcal{C}^1$ , on a la formule d'intégration par partie suivante

$$\int_{\Omega} \operatorname{div} \sigma v \, dx + \int_{\Omega} \sigma \cdot \nabla v \, dx = \langle \sigma \cdot n, v \rangle_{H^{-1/2}, H^{1/2}(\partial\Omega)}, \quad (5.21)$$

pour  $v \in H^1(\Omega)$  et  $\sigma \in H(\operatorname{div})$ . Le terme de droite de (5.21) désigne le produit de dualité entre  $H^{1/2}(\partial\Omega)$  et son dual, noté  $H^{-1/2}(\partial\Omega)$ . Si  $\sigma$  et  $v$  sont des fonctions régulières, ce “vilain” terme n'est autre que l'intégrale de bord usuelle  $\int_{\partial\Omega} v \sigma \cdot n \, ds$ . Sinon, la formule (5.21) donne un sens à  $\sigma \cdot n$  sur  $\partial\Omega$  (comme élément du dual  $H^{-1/2}(\partial\Omega)$ ) pour  $\sigma \in H(\operatorname{div})$ .

Si on applique ce résultat à la solution de (5.18) (avec  $\sigma = \nabla u$ ), on en déduit que la condition aux limites de Neumann est vérifiée comme égalité entre éléments du dual  $H^{-1/2}(\partial\Omega)$  (car  $g \in L^2(\partial\Omega) \subset H^{-1/2}(\partial\Omega)$ ). Cet argument est passablement compliqué, et en pratique on se contentera de dire que la formulation variationnelle contient une généralisation de la condition aux limites de Neumann, et par habitude on continuera à écrire **formellement** que  $\frac{\partial u}{\partial n} = g$  sur  $\partial\Omega$ . •

Comme dans la sous-section précédente, on peut montrer que la solution de (5.16) minimise une énergie. Remarquons que, si  $g = 0$ , alors l'énergie (5.22) est la même que celle (5.11) définie pour le Laplacien avec conditions aux limites de Dirichlet (modulo le terme d'ordre zéro). Néanmoins, leurs minima ne sont en général pas les mêmes car on minimise sur deux espaces différents, à savoir  $H_0^1(\Omega)$  et  $H^1(\Omega)$ . La Proposition suivante est une application immédiate de la Proposition 3.3.4.

**Proposition 5.2.16** *Soit  $J(v)$  l'énergie définie pour  $v \in H^1(\Omega)$  par*

$$J(v) = \frac{1}{2} \int_{\Omega} (|\nabla v|^2 + |v|^2) \, dx - \int_{\Omega} f v \, dx - \int_{\partial\Omega} g v \, ds. \quad (5.22)$$

*Soit  $u \in H^1(\Omega)$  la solution unique de la formulation variationnelle (5.18). Alors  $u$  est aussi l'unique point de minimum de l'énergie, c'est-à-dire que*

$$J(u) = \min_{v \in H^1(\Omega)} J(v).$$

*Réciproquement, si  $u \in H^1(\Omega)$  est un point de minimum de l'énergie  $J(v)$ , alors  $u$  est la solution unique de la formulation variationnelle (5.18).*



**Exercice 5.2.6** On suppose que  $\Omega$  est un ouvert borné régulier de classe  $\mathcal{C}^1$ . A l'aide de l'approche variationnelle démontrer l'existence et l'unicité de la solution du Laplacien avec une condition aux limites de Fourier

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} + u = g & \text{sur } \partial\Omega \end{cases} \quad (5.23)$$

où  $f \in L^2(\Omega)$  et  $g$  est la trace sur  $\partial\Omega$  d'une fonction de  $H^1(\Omega)$ . On démontrera l'inégalité suivante (qui généralise celle de Poincaré)

$$\|v\|_{L^2(\Omega)} \leq C (\|v\|_{L^2(\partial\Omega)} + \|\nabla v\|_{L^2(\Omega)}) \quad \forall v \in H^1(\Omega).$$

**Exercice 5.2.7** On suppose que  $\Omega$  est un ouvert borné connexe. A l'aide de l'approche variationnelle démontrer l'existence et l'unicité de la solution du Laplacien avec des conditions aux limites mêlées

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega_N \\ u = 0 & \text{sur } \partial\Omega_D \end{cases} \quad (5.24)$$

où  $f \in L^2(\Omega)$ , et  $(\partial\Omega_N, \partial\Omega_D)$  est une partition de  $\partial\Omega$  telle que les mesures superficielles de  $\partial\Omega_N$  et  $\partial\Omega_D$  sont non nulles (voir la Figure 4.1). (Utiliser la Remarque 4.3.18.)

Nous revenons maintenant au véritable opérateur Laplacien (sans ajout d'un terme d'ordre zéro comme dans (5.16)) et nous considérons le problème aux limites

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = g & \text{sur } \partial\Omega \end{cases} \quad (5.25)$$

où  $\Omega$  est un ouvert borné connexe de l'espace  $\mathbb{R}^N$ ,  $f \in L^2(\Omega)$  et  $g \in L^2(\partial\Omega)$ . La difficulté nouvelle dans (5.25) par rapport à (5.16) est qu'il n'existe une solution que si les données  $f$  et  $g$  vérifient une **condition de compatibilité**. En effet, il est facile de voir que s'il existe une solution  $u \in H^2(\Omega)$ , alors intégrant l'équation sur  $\Omega$  (ou bien utilisant la formule de Green (4.22)) on a nécessairement

$$\int_{\Omega} f(x) dx + \int_{\partial\Omega} g(x) ds = 0. \quad (5.26)$$

Remarquons aussi que si  $u$  est solution alors  $u + C$ , avec  $C \in \mathbb{R}$ , est aussi solution. En fait, (5.26) est une condition nécessaire et suffisante d'existence d'une solution dans  $H^1(\Omega)$ , unique à l'addition d'une constante près. Remarquons que, si l'ouvert  $\Omega$  n'est pas connexe, alors il faut écrire (5.26) pour chaque composante connexe de  $\Omega$  et l'unicité de la solution vaudra à l'addition près d'une constante par composante connexe (avec ces modifications tous les résultats qui suivent restent valables).

**Remarque 5.2.17** Physiquement, la condition de compatibilité (5.26) s'interprète comme une **condition d'équilibre** :  $f$  correspond à une source volumique, et  $g$  à un

flux entrant au bord. Pour qu'il existe un état stationnaire ou d'équilibre (c'est-à-dire une solution de (5.25)), il faut que ces deux termes se balancent parfaitement. De même, l'unicité "à une constante additive près" correspond à l'absence d'origine de référence sur l'échelle qui mesure les valeurs de  $u$  (comme pour la température, par exemple). •

**Théorème 5.2.18** *Soit  $\Omega$  un ouvert borné connexe régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)$  et  $g \in L^2(\partial\Omega)$  qui vérifient la condition de compatibilité (5.26). Il existe une solution faible  $u \in H^1(\Omega)$  de (5.25), unique à l'addition d'une constante près.*

**Démonstration.** Pour trouver la formulation variationnelle on procède comme pour l'équation (5.16). Un calcul similaire conduit à

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\partial\Omega} g v \, ds + \int_{\Omega} f v \, dx$$

pour toute fonction test régulière  $v$ . Pour donner un sens à tous les termes de cette égalité, on pourrait choisir  $H^1(\Omega)$  comme espace de Hilbert  $V$ , mais nous ne pourrions montrer la coercivité de la forme bilinéaire. Cette difficulté est intimement liée au fait que, si  $u$  est solution, alors  $u + C$  est aussi solution. Pour éviter cet inconvénient, on lève l'indétermination de cette constante additive en ne travaillant qu'avec des fonctions de moyenne nulle. Autrement dit, on pose

$$V = \left\{ v \in H^1(\Omega), \int_{\Omega} v(x) \, dx = 0 \right\}$$

et la formulation variationnelle de (5.25) est :

$$\text{trouver } u \in V \text{ tel que } \int_{\Omega} \nabla u \cdot \nabla v \, dx = \int_{\partial\Omega} g v \, ds + \int_{\Omega} f v \, dx \quad \forall v \in V. \quad (5.27)$$

On peut également choisir  $V = H^1(\Omega)/\mathbb{R}$ , c'est-à-dire quotienter  $H^1(\Omega)$  par les constantes (les éléments de  $H^1(\Omega)/\mathbb{R}$  sont les classes de fonctions de  $H^1(\Omega)$  égales à une constante près).

Pour pouvoir appliquer le Théorème de Lax-Milgram à la formulation variationnelle (5.27), la seule hypothèse délicate à vérifier est la coercivité de la forme bilinéaire. Celle-ci s'obtient grâce à une généralisation de l'inégalité de Poincaré, connue sous le nom d'inégalité de Poincaré-Wirtinger : si  $\Omega$  est borné et connexe, il existe une constante  $C > 0$  telle que, pour tout  $v \in H^1(\Omega)$ ,

$$\|v - m(v)\|_{L^2(\Omega)} \leq C \|\nabla v\|_{L^2(\Omega)} \text{ avec } m(v) = \frac{\int_{\Omega} v \, dx}{\int_{\Omega} 1 \, dx}. \quad (5.28)$$

L'inégalité (5.28) se démontre par contradiction comme dans la deuxième démonstration de la Proposition 4.3.10 (nous laissons cela au lecteur en exercice). Comme  $m(v) = 0$  pour tout  $v \in V$ , (5.28) prouve que  $\|\nabla v\|_{L^2(\Omega)}$  est une norme dans  $V$ ,

équivalente à la norme usuelle  $\|v\|_{H^1(\Omega)}$ , et donc que la forme bilinéaire est coercive sur  $V$ .

Finalement, pour montrer que la solution unique de (5.27) est bien une solution du problème aux limites (5.25), on procède comme précédemment lors de la démonstration du Théorème 5.2.14. On obtient ainsi pour tout  $v \in V$ ,

$$\int_{\Omega} (\Delta u + f)v \, dx = \int_{\partial\Omega} \left( \frac{\partial u}{\partial n} - g \right) v \, ds. \quad (5.29)$$

Or, quelque soit  $w \in H^1(\Omega)$ , la fonction  $v = w - m(w)$  appartient à  $V$ . En choisissant une telle fonction dans (5.29), en regroupant les termes en facteur de la constante  $m(w)$  et en utilisant la condition de compatibilité (5.26) ainsi que l'égalité  $\int_{\Omega} \Delta u \, dx = \int_{\partial\Omega} \frac{\partial u}{\partial n} \, ds$ , on déduit donc de (5.29)

$$\int_{\Omega} (\Delta u + f)w \, dx = \int_{\partial\Omega} \left( \frac{\partial u}{\partial n} - g \right) w \, ds \quad \forall w \in H^1(\Omega).$$

On peut donc conclure comme d'habitude que  $u$  vérifie bien le problème aux limites (5.25).  $\square$

**Exercice 5.2.8** Démontrer l'inégalité de Poincaré-Wirtinger (5.28).

**Exercice 5.2.9** On suppose que  $\Omega$  est un ouvert borné connexe régulier. Soit  $f \in L^2(\Omega)$ . On considère la formulation variationnelle suivante : trouver  $u \in H^1(\Omega)$  tel que

$$\int_{\Omega} \nabla u \cdot \nabla v \, dx + \left( \int_{\Omega} u \, dx \right) \left( \int_{\Omega} v \, dx \right) = \int_{\Omega} f v \, dx \quad \forall v \in H^1(\Omega).$$

Démontrer l'existence et l'unicité de la solution de cette formulation variationnelle. Quel problème aux limites a-t-on ainsi résolu ? En particulier, si on suppose que  $\int_{\Omega} f \, dx = 0$ , quel problème déjà étudié retrouve-t-on ?

### 5.2.3 Coefficients variables

Dans les deux sous-sections précédentes nous avons considéré des problèmes aux limites pour l'opérateur Laplacien. On peut facilement généraliser les résultats obtenus à des opérateurs plus généraux, dits elliptiques du deuxième ordre **à coefficients variables**. Ce type de problème est issu de la modélisation des milieux hétérogènes. Si l'on reprend l'exemple de la conduction de la chaleur (détaillé dans le Chapitre 1), dans un milieu hétérogène la conductivité  $k(x)$  est une fonction variable d'un point à l'autre du domaine. Dans ce cas, on considère le problème aux limites

$$\begin{cases} -\operatorname{div}(k \nabla u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (5.30)$$

où  $\Omega$  est un ouvert borné de l'espace  $\mathbb{R}^N$ , et  $f \in L^2(\Omega)$ . Bien sûr, si  $k(x) \equiv 1$ , on retrouve le Laplacien. Il est facile de généraliser le Théorème 5.2.2.

**Proposition 5.2.19** Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)$ . On suppose que le coefficient  $k(x)$  est une fonction mesurable et qu'il existe deux constantes strictement positives  $0 < k^- \leq k^+$  telles que

$$0 < k^- \leq k(x) \leq k^+ \text{ presque partout } x \in \Omega. \quad (5.31)$$

Alors, il existe une unique solution (faible)  $u \in H_0^1(\Omega)$  de (5.30).

**Démonstration.** Pour trouver la formulation variationnelle on multiplie l'équation (5.39) par une fonction test  $v$  et on intègre par partie en utilisant la formule de Green de l'Exercice 3.2.1

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

avec  $\sigma = k \nabla u$ . Pour tenir compte de la condition aux limites de Dirichlet on choisit  $H_0^1(\Omega)$  comme espace de Hilbert, et on trouve la formulation variationnelle de (5.30) :

$$\text{trouver } u \in H_0^1(\Omega) \text{ tel que } \int_{\Omega} k \nabla u \cdot \nabla v dx = \int_{\Omega} f v dx \quad \forall v \in H_0^1(\Omega). \quad (5.32)$$

Grâce à l'hypothèse (5.31) on vérifie que la forme bilinéaire de (5.32) est continue

$$\left| \int_{\Omega} k \nabla u \cdot \nabla v dx \right| \leq k^+ \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)},$$

et qu'elle est coercive

$$\int_{\Omega} k \nabla u \cdot \nabla u dx \geq k^- \int_{\Omega} |\nabla u|^2 dx \geq \nu \|u\|_{H_0^1(\Omega)},$$

avec  $\nu > 0$  grâce à l'inégalité de Poincaré. On peut donc appliquer le Théorème de Lax-Milgram, ce qui démontre l'existence et l'unicité de la solution de la formulation variationnelle (5.32). Pour montrer que cette solution variationnelle est bien une solution du problème aux limites (5.30), on procède comme précédemment lors de la démonstration du Théorème 5.2.2.  $\square$

**Remarque 5.2.20** Il est très important de laisser l'équation (5.30) sous forme de divergence : on pourrait *a priori* l'écrire aussi

$$-\operatorname{div}(k \nabla u) = -k \Delta u - \nabla k \cdot \nabla u,$$

mais cette dernière forme n'a de sens que si le coefficient  $k$  est dérivable. Au contraire, l'équation (5.30) a un sens même si  $k$  est discontinu.  $\bullet$

En fait, lorsque l'équation (5.30) est sous forme de divergence, son interprétation "au sens faible" contient plus d'informations que son écriture classique. Plus précisément, la notion de divergence faible contient implicitement ce qu'on appelle les

**conditions aux limites de transmission** entre deux sous-domaines occupés par deux matériaux de conductivité différente. Considérons un exemple où  $(\Omega_1, \Omega_2)$  est une partition de  $\Omega$  sur laquelle  $k(x)$  est constant par morceaux

$$k(x) = k_i > 0 \text{ pour } x \in \Omega_i, i = 1, 2. \quad (5.33)$$

On note  $\Gamma = \partial\Omega_1 \cap \partial\Omega_2$  l'interface (supposée régulière et incluse dans  $\Omega$ ) entre  $\Omega_1$  et  $\Omega_2$  (voir la Figure 5.1), et  $u_i = u|_{\Omega_i}$  la restriction de la solution  $u$  à  $\Omega_i$ .

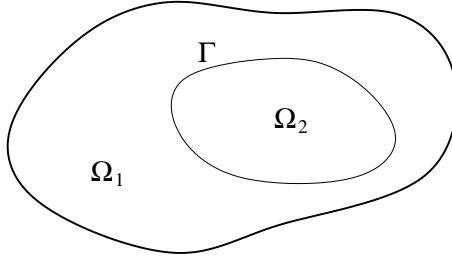


FIGURE 5.1 – Interface entre deux sous-domaines et condition de transmission.

**Lemme 5.2.21** *Sous l'hypothèse (5.33) le problème (5.30) est équivalent à*

$$\begin{cases} -k_i \Delta u_i = f & \text{dans } \Omega_i, \ i = 1, 2, \\ u_1 = 0 & \text{sur } \partial\Omega, \\ u_1 = u_2 & \text{sur } \Gamma, \\ k_1 \nabla u_1 \cdot n = k_2 \nabla u_2 \cdot n & \text{sur } \Gamma. \end{cases} \quad (5.34)$$

Les deux dernières lignes de (5.34) sont appelées **conditions aux limites de transmission** sur l'interface  $\Gamma$ .

**Démonstration.** Si  $u \in H_0^1(\Omega)$  est solution de (5.30), par application du Théorème de trace 4.3.13, on doit avoir  $u_1 = u_2$  sur  $\Gamma$ . Si on pose  $\sigma = k \nabla u$  et  $\sigma_i = \sigma|_{\Omega_i} = k_i \nabla u_i$  sa restriction à  $\Omega_i$ , on sait que  $\sigma$ , ainsi que sa divergence, appartiennent à  $L^2(\Omega)$ . Alors, en vertu du Théorème 4.4.7, la composante normale  $\sigma \cdot n$  a un sens sur  $\Gamma$  et on doit avoir  $\sigma_1 \cdot n = \sigma_2 \cdot n$  sur  $\Gamma$ .

Réciproquement, on construit une formulation variationnelle de (5.34) pour montrer qu'elle admet une unique solution qui coïncide avec la solution  $u$  de (5.30). On cherche  $u_i \in H^1(\Omega_i)$  et on multiplie chaque équation dans  $\Omega_i$  par la même fonction test  $v \in H_0^1(\Omega)$ . En intégrant par partie et en sommant on obtient

$$\int_{\Omega_1} \nabla u_1 \cdot \nabla v \, dx + \int_{\Omega_2} \nabla u_2 \cdot \nabla v \, dx + \int_{\Gamma} \left( k_1 \frac{\partial u_1}{\partial n_1} + k_2 \frac{\partial u_2}{\partial n_2} \right) v \, ds = \int_{\Omega_1} f v \, dx + \int_{\Omega_2} f v \, dx. \quad (5.35)$$

Comme  $n_1 = -n_2$  l'intégrale sur l'interface  $\Gamma$  disparaît à cause de la condition aux limites de transmission. D'autre part, si  $u$  est défini comme  $u_1$  dans  $\Omega_1$  et  $u_2$  dans

$\Omega_2$ , la condition de transmission  $u_1 = u_2$  sur  $\Gamma$  implique que  $u \in H^1(\Omega)$  en vertu du Lemme 4.3.19. Par conséquent, (5.35) n'est rien d'autre que la formulation variationnelle (5.32).  $\square$

**Exercice 5.2.10** Soit  $\Omega$  un ouvert borné et  $K$  un compact connexe de  $\mathbb{R}^N$  inclus dans  $\Omega$  (on suppose que  $\Omega \setminus K$  est régulier). Soit  $f \in L^2(\Omega)$ . On considère un problème de conduction dans  $\Omega$  où  $K$  est une inclusion parfaitement conductrice, c'est-à-dire que l'inconnue  $u$  (la température ou le potentiel électrique, par exemple) est constante dans  $K$  (cette constante est aussi inconnue). On suppose qu'il n'y a pas de terme source dans  $K$ . Ce problème se modélise par

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \setminus K \\ u = C & \text{sur } \partial K \\ \int_{\partial K} \frac{\partial u}{\partial n} ds = 0 & \text{sur } \partial K \\ u = 0 & \text{sur } \partial\Omega, \end{cases}$$

où  $C$  est une constante inconnue à déterminer. Trouver une formulation variationnelle de ce problème aux limites et démontrer l'existence et l'unicité d'une solution  $(u, C)$ .

On peut encore généraliser ce qui précède à des opérateurs plus généraux avec des coefficients tensoriels  $A(x) = (a_{ij}(x))_{1 \leq i, j \leq N}$ . On suppose que la matrice  $A$  est uniformément définie positive sur  $\Omega$  (ou coercive, ou elliptique), c'est-à-dire qu'il existe une constante  $\alpha > 0$  telle que, presque partout dans  $\Omega$ ,

$$A(x)\xi \cdot \xi = \sum_{i,j=1}^N a_{ij}(x)\xi_i\xi_j \geq \alpha|\xi|^2 \text{ pour tout } \xi \in \mathbb{R}^N, \quad (5.36)$$

et qu'elle est uniformément bornée, c'est-à-dire qu'il existe une constante  $\beta > 0$  telle que, presque partout dans  $\Omega$ ,

$$|A(x)\xi| \leq \beta|\xi| \text{ pour tout } \xi \in \mathbb{R}^N. \quad (5.37)$$

On définit alors l'opérateur

$$-\operatorname{div}(A\nabla \cdot) = - \sum_{i,j=1}^N \frac{\partial}{\partial x_i} \left( a_{ij}(x) \frac{\partial}{\partial x_j} \cdot \right), \quad (5.38)$$

et on considère le problème aux limites

$$\begin{cases} -\operatorname{div}(A\nabla u) = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (5.39)$$

Plusieurs motivations physiques conduisent à des modèles du type de (5.39). Dans l'exemple de la conduction de la chaleur, un milieu anisotrope (pour lequel la conductivité n'est pas la même dans toutes les directions) est caractérisé par une matrice symétrique de conductivité  $A(x)$  (non proportionnelle à l'identité). Le cas des matrices  $A(x)$  non symétriques correspond, par exemple, à la prise en compte d'un effet de convection. En effet, si on décompose  $A =$

$A^s + A^a$  en sa partie symétrique  $A^s = (A + A^t)/2$  et sa partie antisymétrique  $A^a = (A - A^t)/2$ , un calcul simple montre que

$$-\operatorname{div}(A\nabla u) = -\operatorname{div}(A^s\nabla u) + V \cdot \nabla u \quad \text{avec} \quad V_j(x) = \sum_{i=1}^N \frac{1}{2} \frac{\partial(a_{ji} - a_{ij})}{\partial x_i}(x),$$

où  $V$  s'interprète comme une vitesse de convection.

**Exercice 5.2.11** Démontrer sous les hypothèses (5.36) et (5.37) que (5.39) admet une unique solution (faible)  $u \in H_0^1(\Omega)$  si  $f \in L^2(\Omega)$ .

On peut remplacer la condition aux limites de Dirichlet dans (5.39) par une condition aux limites de Neumann qui, pour l'opérateur (5.38), s'écrit

$$\frac{\partial u}{\partial n_A} = \left( A(x)\nabla u \right) \cdot n = \sum_{i,j=1}^N a_{ij}(x) \frac{\partial u}{\partial x_j} n_i = 0 \text{ sur } \partial\Omega,$$

où  $\frac{\partial u}{\partial n_A}$  est la dérivée, dite conormale, de  $u$  associée à l'opérateur (5.38). Cette condition de Neumann est très naturelle d'un point de vue physique puisque, si on introduit le flux  $\sigma = A\nabla u$ , elle exprime que la composante normale du flux est nulle sur le bord  $\sigma \cdot n = 0$  sur  $\partial\Omega$ .

**Exercice 5.2.12** Pour  $f \in L^2(\Omega)$ ,  $g \in L^2(\partial\Omega)$ , montrer l'existence et l'unicité de la solution de

$$\begin{cases} -\operatorname{div}(A\nabla u) + u = f & \text{dans } \Omega, \\ \frac{\partial u}{\partial n_A} = g & \text{sur } \partial\Omega. \end{cases}$$

## 5.2.4 Propriétés qualitatives

Dans cette sous-section nous étudions quelques propriétés qualitatives des solutions du Laplacien avec conditions aux limites de Dirichlet. Dans toute cette sous-section,  $\Omega$  est un ouvert borné de  $\mathbb{R}^N$  et  $f \in L^2(\Omega)$ . Le Théorème 5.2.2 fournit une unique solution  $u \in H_0^1(\Omega)$  de

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (5.40)$$

### Principe du maximum

Nous commençons par retrouver cette propriété découverte au Chapitre 1 (grâce à des formules explicites en dimension  $N = 1$ ; voir la Remarque 1.2.10) et exploitée numériquement au Chapitre 2.

**Théorème 5.2.22 (Principe du maximum)** Si  $f \geq 0$  presque partout dans  $\Omega$ , alors  $u \geq 0$  presque partout dans  $\Omega$ .

**Remarque 5.2.23** Le principe du maximum ne fait que traduire une propriété parfaitement naturelle du point de vue physique : par exemple dans le contexte de l'équation stationnaire de la chaleur, si on chauffe ( $f \geq 0$ ), la température intérieure est toujours plus grande que la température au bord ( $u \geq 0$ ). Le principe du maximum reste valable si on remplace le Laplacien par l'opérateur plus général (5.38) à coefficients variables dans  $L^\infty(\Omega)$  ou bien si l'on considère le problème (5.16) avec condition aux limites de Neumann. La validité du principe du maximum est fondamentalement liée au caractère "scalaire" de l'équation (c'est-à-dire que l'inconnue  $u$  est à valeurs dans  $\mathbb{R}$ ). Ce principe du maximum tombe généralement en défaut si l'inconnue  $u$  est à valeurs vectorielles (par exemple pour le système (5.56) de l'élasticité). •

**Démonstration.** On utilise la formulation variationnelle (5.5) de (5.40) avec  $v = u^- = \min(u, 0)$  qui appartient bien à  $H_0^1(\Omega)$  en vertu du Lemme 5.2.24 (car  $u = u^+ + u^-$ ). On a

$$\int_{\Omega} f u^- dx = \int_{\Omega} \nabla u \cdot \nabla u^- dx = \int_{\Omega} 1_{u < 0} \nabla u \cdot \nabla u dx = \int_{\Omega} |\nabla u^-|^2 dx \geq 0. \quad (5.41)$$

Mais  $u^- \leq 0$  et  $f \geq 0$  presque partout dans  $\Omega$ . Par conséquent, tous les termes de (5.41) sont nuls, et comme  $u^- \in H_0^1(\Omega)$  on en déduit que  $u^- = 0$ , c'est-à-dire que  $u \geq 0$  presque partout dans  $\Omega$ .  $\square$

**Lemme 5.2.24** Si  $v \in H_0^1(\Omega)$ , alors  $v^+ = \max(v, 0)$  appartient à  $H_0^1(\Omega)$  et

$$\nabla v^+ = 1_{v > 0} \nabla v \text{ presque partout dans } \Omega,$$

où  $1_{v > 0}(x)$  est la fonction qui vaut 1 là où  $v(x) > 0$  et 0 ailleurs.

**Remarque 5.2.25** La démonstration du Lemme 5.2.24 est assez longue et technique (elle peut être omise en première lecture). Expliquons néanmoins pourquoi cette démonstration n'est pas simple, et en particulier pourquoi on ne peut pas utiliser le Lemme 4.3.19 qui affirme qu'une fonction définie "par morceaux", appartenant à  $H^1$  de chaque sous-domaine et continue aux interfaces entre les sous-domaines appartient en fait à  $H^1$  du domaine entier. En effet, on appliquerait le Lemme 4.3.19 à  $v$  sur le sous-domaine  $v > 0$  et à 0 sur le sous-domaine  $v < 0$ , et le tour serait joué. Le problème est qu'en général la frontière de ces deux sous-domaines (définie par  $v = 0$ ) n'est pas régulière (au sens de la Définition 3.2.5) même si  $v$  est une fonction régulière. •

**Démonstration.** Montrons tout d'abord que, si  $v \in H_0^1(\Omega)$  et si  $G(t)$  est une fonction de  $\mathbb{R}$  dans  $\mathbb{R}$ , de classe  $C^1$  telle que  $G(0) = 0$  et  $G'(t)$  est borné sur  $\mathbb{R}$ , alors  $G(v) \in H_0^1(\Omega)$  et  $\nabla(G(v)) = G'(v) \nabla v$ . Par définition de  $H_0^1(\Omega)$ , il existe une suite de fonctions  $v_n \in C_c^\infty(\Omega)$  qui converge vers  $v$  dans la norme de  $H^1(\Omega)$  (en particulier, pour une sous-suite,  $v_n$  et  $\nabla v_n$  converge presque partout dans  $\Omega$  vers  $v$  et  $\nabla v$  respectivement ; voir le corollaire 3.3.3 de [7]). On a

$$|G(v_n) - G(v)| \leq \left( \sup_{t \in \mathbb{R}} |G'(t)| \right) |v_n - v|, \quad (5.42)$$



donc  $G(v_n)$  converge vers  $G(v)$  dans  $L^2(\Omega)$ . D'autre part, pour  $1 \leq i \leq N$ ,

$$\left| \frac{\partial G(v_n)}{\partial x_i} - G'(v) \frac{\partial v}{\partial x_i} \right| \leq |G'(v_n) - G'(v)| \left| \frac{\partial v}{\partial x_i} \right| + \left( \sup_{t \in \mathbb{R}} |G'(t)| \right) \left| \frac{\partial v_n}{\partial x_i} - \frac{\partial v}{\partial x_i} \right|. \quad (5.43)$$

Comme  $|G'(v_n) - G'(v)| \left| \frac{\partial v}{\partial x_i} \right|$  converge presque partout vers 0 (pour une sous-suite) et est majoré par  $2 \left( \sup |G'(t)| \right) \left| \frac{\partial v}{\partial x_i} \right|$  qui appartient à  $L^2(\Omega)$ , par application du théorème de convergence dominée de Lebesgue (voir le théorème 2.3.7 de [7]) cette suite de fonctions converge vers 0 dans  $L^2(\Omega)$ . Le dernier terme majorant dans (5.43) converge aussi vers 0 dans  $L^2(\Omega)$ , donc  $G(v_n)$  est une suite de Cauchy dans  $H_0^1(\Omega)$  qui est un espace de Hilbert : elle converge vers une limite  $w \in H_0^1(\Omega)$ . Par identification des limites, on trouve que  $w = G(v)$  dans  $L^2(\Omega)$  et que  $\frac{\partial w}{\partial x_i} = G'(v) \frac{\partial v}{\partial x_i}$  dans  $L^2(\Omega)$ , d'où le résultat.

Nous allons maintenant approcher la fonction  $t \rightarrow \max(t, 0)$  par une suite de fonctions  $G_n(t)$  du type précédent pour démontrer que  $v^+$  appartient à  $H_0^1(\Omega)$ . Soit  $G(t)$  une fonction de  $C^1(\mathbb{R})$  telle que

$$G(t) = 0 \text{ si } t \leq \frac{1}{2}, \quad 0 \leq G'(t) \leq 1 \text{ si } \frac{1}{2} \leq t \leq 1, \quad G'(t) = 1 \text{ si } 1 \leq t.$$

On définit  $G_n(t) = G(nt)/n$  pour  $n \geq 1$ , et on sait par l'argument précédent que  $G_n(v) \in H_0^1(\Omega)$  et  $\frac{\partial G_n(v)}{\partial x_i} = G'_n(v) \frac{\partial v}{\partial x_i}$ . D'autre part, on vérifie que

$$|G_n(v) - v^+| \leq \sup_{t \in \mathbb{R}} |G_n(t) - t^+| \leq \frac{1}{n},$$

donc  $G_n(v)$  converge vers  $v^+$  dans  $L^2(\Omega)$ . On a aussi, pour tout  $1 \leq i \leq N$ ,

$$\left| \frac{\partial G_n(v)}{\partial x_i} - 1_{v>0} \frac{\partial v}{\partial x_i} \right| = |G'_n(v) - 1_{v>0}| \left| \frac{\partial v}{\partial x_i} \right| \leq 1_{0 < v < 1/n} \left| \frac{\partial v}{\partial x_i} \right|,$$

et comme  $1_{0 < v < 1/n}$  converge vers 0 presque partout, le théorème de convergence dominée de Lebesgue prouve que  $\frac{\partial G_n(v)}{\partial x_i}$  converge vers  $1_{v>0} \frac{\partial v}{\partial x_i}$  dans  $L^2(\Omega)$ . On en déduit, comme précédemment, que  $G_n(v)$  converge vers  $v^+$  dans  $H_0^1(\Omega)$  et que  $\nabla v^+ = 1_{v>0} \nabla v$ .  $\square$

**Exercice 5.2.13** Montrer que l'application (non-linéaire)  $v \rightarrow v^+$  est continue de  $L^2(\Omega)$  dans lui-même, ainsi que de  $H^1(\Omega)$  dans lui-même (utiliser le fait que  $\nabla u = 0$  presque partout sur l'ensemble  $u^{-1}(0)$ ).

## Régularité

Nous montrons maintenant que la solution d'un problème aux limites elliptique est plus régulière que prévue si les données sont plus régulières que nécessaire.

**Théorème 5.2.26 (de régularité)** Soit un entier  $m \geq 0$ . Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$  de classe  $C^{m+2}$ . Soit  $f \in H^m(\Omega)$ . Alors, l'unique solution  $u \in H_0^1(\Omega)$  de (5.40) appartient à  $H^{m+2}(\Omega)$ . De plus, l'application  $f \rightarrow u$  est linéaire continue de  $H^m(\Omega)$  dans  $H^{m+2}(\Omega)$ , c'est-à-dire qu'il existe une constante  $C > 0$  telle que

$$\|u\|_{H^{m+2}(\Omega)} \leq C \|f\|_{H^m(\Omega)}.$$

Par application immédiate du Théorème de régularité 5.2.26 et du Théorème 4.3.25 (sur la continuité des fonctions de  $H^m(\Omega)$ ), on obtient comme corollaire le résultat annoncé auparavant dans la Remarque 5.2.3, à savoir que les solutions faibles d'équations aux dérivées partielles elliptiques sont en fait des solutions fortes (ou classiques) si les données sont régulières.

**Corollaire 5.2.27** *Si  $\Omega$  est un ouvert borné de  $\mathbb{R}^N$  de classe  $C^{m+2}$ , si  $f \in H^m(\Omega)$ , et si  $m > N/2$ , alors la solution variationnelle  $u \in H_0^1(\Omega)$  de (5.40) est une solution forte car elle appartient à  $C^2(\overline{\Omega})$ .*

*En particulier, si  $\Omega$  est un ouvert borné de  $\mathbb{R}^N$  de classe  $C^\infty$ , et si  $f \in C^\infty(\overline{\Omega})$ , alors la solution  $u \in H_0^1(\Omega)$  de (5.40) est aussi dans  $C^\infty(\overline{\Omega})$ .*

**Remarque 5.2.28** Il est important de bien comprendre la portée de ces résultats de régularité. En supposant que  $\Delta u$ , qui est une combinaison particulière de certaines dérivées secondes de  $u$ , appartient à un certain espace fonctionnel, on en déduit que **toutes** les dérivées secondes de  $u$  appartiennent à ce même espace! Bien sûr, tous ces résultats de régularité sont évidents en dimension  $N = 1$  puisque le Laplacien coïncide avec la dérivée seconde et donc l'équation fournit directement la régularité de cette dérivée seconde. •

**Remarque 5.2.29** Le Théorème de régularité 5.2.26 et son Corollaire 5.2.27 restent valables pour des conditions aux limites de Neumann. Ils se généralisent aussi au cas des opérateurs elliptiques à coefficients variables (comme dans la Sous-section 5.2.3). Dans ce dernier cas, il faut ajouter à l'hypothèse habituelle de coercivité des coefficients, l'hypothèse que les coefficients sont de classe  $C^{m+1}$  dans  $\Omega$  (tandis que  $\Omega$  est borné régulier de classe  $C^{m+2}$  et que  $f \in H^m(\Omega)$ ). •

Nous ne démontrons pas le Théorème de régularité 5.2.26 dans toute sa généralité mais seulement dans un cas particulier plus simple (nous expliquerons dans la Remarque 5.2.32 comment on passe du cas particulier au cas général). On se place dans  $\Omega = \mathbb{R}^N$ , et pour  $f \in L^2(\mathbb{R}^N)$  on considère le problème

$$-\Delta u + u = f \quad \text{dans } \mathbb{R}^N. \quad (5.44)$$

Il n'y a pas de condition aux limites **explicite** dans (5.44) puisqu'il n'y a pas de bord. Néanmoins, le comportement à l'infini de la solution est une sorte de condition aux limites. En prenant  $f$  dans  $L^2(\mathbb{R}^N)$  nous avons choisi une condition aux limites **implicite** qui est de chercher  $u$  dans  $H^1(\mathbb{R}^N)$ , c'est-à-dire que, dans un certain sens,  $u(x)$  "tend vers zéro" à l'infini afin que l'intégrale  $\int_{\mathbb{R}^N} |u|^2 dx$  converge. Une formulation variationnelle de (5.44) est : trouver  $u \in H^1(\mathbb{R}^N)$  tel que

$$\int_{\mathbb{R}^N} (\nabla u \cdot \nabla v + uv) dx = \int_{\mathbb{R}^N} f v dx \quad \forall v \in H^1(\mathbb{R}^N). \quad (5.45)$$

Une application directe du Théorème de Lax-Milgram 3.3.1 prouve qu'il existe une unique solution  $u \in H^1(\mathbb{R}^N)$  de (5.45). Enfin, un raisonnement identique à tous ceux que nous avons déjà fait dans ce chapitre montre que cette solution de la formulation variationnelle est aussi

solution de l'équation aux dérivées partielles (5.44). Nous pouvons maintenant énoncer le résultat de régularité.

**Proposition 5.2.30** *Si  $f \in L^2(\mathbb{R}^N)$ , alors la solution  $u \in H^1(\mathbb{R}^N)$  de (5.44) appartient en fait à  $H^2(\mathbb{R}^N)$ . De même, si  $f \in H^m(\mathbb{R}^N)$  (avec  $m \geq 0$ ), alors  $u$  appartient à  $H^{m+2}(\mathbb{R}^N)$ .*

**Démonstration.** L'ingrédient essentiel de la démonstration est la “méthode des translations”. Pour  $h \in \mathbb{R}^N$ ,  $h \neq 0$ , on définit un quotient différentiel

$$D_h v(x) = \frac{v(x+h) - v(x)}{|h|}$$

qui appartient bien à  $H^1(\mathbb{R}^N)$  si  $v \in H^1(\mathbb{R}^N)$ . On vérifie aisément que  $\nabla(D_h v) = D_h(\nabla v)$  et que, pour  $v, \phi \in L^2(\mathbb{R}^N)$ , on a une “formule d'intégration par partie discrète”

$$\int_{\mathbb{R}^N} (D_h v) \phi \, dx = \int_{\mathbb{R}^N} v (D_{-h} \phi) \, dx.$$

D'autres propriétés du quotient  $D_h v$  sont données dans le Lemme 5.2.31. Dans la formulation variationnelle (5.45) on prend  $v = D_{-h}(D_h u)$ , et appliquant les règles ci-dessus on obtient

$$\int_{\mathbb{R}^N} (|\nabla(D_h u)|^2 + |D_h u|^2) \, dx = \int_{\mathbb{R}^N} f D_{-h}(D_h u) \, dx.$$

On en déduit la majoration

$$\|D_h u\|_{H^1(\mathbb{R}^N)}^2 \leq \|f\|_{L^2(\mathbb{R}^N)} \|D_{-h}(D_h u)\|_{L^2(\mathbb{R}^N)}.$$

Or, par application de (5.48), on a aussi

$$\|D_{-h}(D_h u)\|_{L^2(\mathbb{R}^N)} \leq \|\nabla(D_h u)\|_{L^2(\mathbb{R}^N)} \leq \|D_h u\|_{H^1(\mathbb{R}^N)}.$$

Donc on a  $\|D_h u\|_{H^1(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)}$ , et en particulier, pour  $1 \leq i \leq N$ ,

$$\left\| D_h \frac{\partial u}{\partial x_i} \right\|_{L^2(\mathbb{R}^N)} \leq \|f\|_{L^2(\mathbb{R}^N)},$$

qui implique, d'après le Lemme 5.2.31 ci-dessous, que  $\frac{\partial u}{\partial x_i}$  appartient à  $H^1(\mathbb{R}^N)$ , c'est-à-dire que  $u \in H^2(\mathbb{R}^N)$ .

Supposons maintenant que  $f \in H^1(\mathbb{R}^N)$ . Montrons que  $\frac{\partial u}{\partial x_i}$  est l'unique solution dans  $H^1(\mathbb{R}^N)$  de

$$-\Delta u_i + u_i = \frac{\partial f}{\partial x_i} \text{ dans } \mathbb{R}^N. \quad (5.46)$$

Si c'est vrai, par application de la partie précédente de la démonstration, on en déduira que  $\frac{\partial u}{\partial x_i}$  appartient à  $H^2(\mathbb{R}^N)$ , c'est-à-dire que  $u \in H^3(\mathbb{R}^N)$  comme prévu. Écrivons la formulation variationnelle (5.45) avec la fonction test  $\frac{\partial \phi}{\partial x_i}$  pour  $\phi \in C_c^\infty(\mathbb{R}^N)$

$$\int_{\mathbb{R}^N} \left( \nabla u \cdot \nabla \frac{\partial \phi}{\partial x_i} + u \frac{\partial \phi}{\partial x_i} \right) dx = \int_{\mathbb{R}^N} f \frac{\partial \phi}{\partial x_i} dx.$$

Comme  $u \in H^2(\mathbb{R}^N)$  et  $f \in H^1(\mathbb{R}^N)$ , on peut intégrer par partie pour obtenir

$$\int_{\mathbb{R}^N} \left( \nabla \frac{\partial u}{\partial x_i} \cdot \nabla \phi + \frac{\partial u}{\partial x_i} \phi \right) dx = \int_{\mathbb{R}^N} \frac{\partial f}{\partial x_i} \phi dx \quad (5.47)$$

qui, par densité, reste valable pour tout  $\phi \in H^1(\mathbb{R}^N)$ . On constate que (5.47) est la formulation variationnelle de (5.46). Par conséquent,  $\frac{\partial u}{\partial x_i} = u_i$  est bien l'unique solution dans  $H^1(\mathbb{R}^N)$  de (5.46).

Le cas  $f \in H^m(\mathbb{R}^N) \Rightarrow u \in H^{m+2}(\mathbb{R}^N)$  se démontre par récurrence sur  $m$  comme on vient de le faire pour  $m = 1$ .  $\square$

**Lemme 5.2.31** *Pour  $v \in L^2(\mathbb{R}^N)$ ,  $h \in \mathbb{R}^N$ ,  $h \neq 0$ , on définit un quotient différentiel*

$$D_h v(x) = \frac{v(x+h) - v(x)}{|h|} \in L^2(\mathbb{R}^N).$$

*Si  $v \in H^1(\mathbb{R}^N)$ , on a l'estimation*

$$\|D_h v\|_{L^2(\mathbb{R}^N)} \leq \|\nabla v\|_{L^2(\mathbb{R}^N)}. \quad (5.48)$$

*Réciproquement, soit  $v \in L^2(\mathbb{R}^N)$  : s'il existe une constante  $C$ , telle que, pour tout  $h \neq 0$ , on a*

$$\|D_h v\|_{L^2(\mathbb{R}^N)} \leq C, \quad (5.49)$$

*alors  $v \in H^1(\mathbb{R}^N)$  et  $\|e \cdot \nabla v\|_{L^2(\mathbb{R}^N)} \leq C$  pour tout vecteur unité  $e \in \mathbb{R}^N$ .*

**Démonstration.** Démontrons d'abord (5.48). Pour  $v \in C_c^\infty(\mathbb{R}^N)$ , on écrit

$$D_h v(x) = \int_0^1 \frac{h}{|h|} \cdot \nabla v(x+th) dt,$$

qu'on majore par

$$|D_h v(x)|^2 \leq \int_0^1 |\nabla v(x+th)|^2 dt.$$

Intégrant en  $x$  il vient

$$\|D_h v\|_{L^2(\mathbb{R}^N)}^2 \leq \int_0^1 \int_{\mathbb{R}^N} |\nabla v(x+th)|^2 dx dt \leq \int_0^1 \|\nabla v\|_{L^2(\mathbb{R}^N)}^2 dt = \|\nabla v\|_{L^2(\mathbb{R}^N)}^2.$$

Par densité, l'estimation (5.48) est vraie pour toute fonction de  $H^1(\mathbb{R}^N)$ .

Soit maintenant  $v \in L^2(\mathbb{R}^N)$  qui vérifie l'estimation (5.49). On en déduit que, pour tout  $\phi \in C_c^\infty(\mathbb{R}^N)$ ,

$$\left| \int_{\mathbb{R}^N} D_h v \phi dx \right| \leq C \|\phi\|_{L^2(\mathbb{R}^N)}.$$

Or, par intégration par partie discrète, on a

$$\int_{\mathbb{R}^N} (D_h v) \phi dx = \int_{\mathbb{R}^N} v (D_{-h} \phi) dx,$$

et, comme  $\phi$  est régulière, si on pose  $h = te$  avec  $e \in \mathbb{R}^N$ ,  $e \neq 0$ , on a

$$\lim_{t \rightarrow 0} D_{-h} \phi(x) = -e \cdot \nabla \phi(x).$$

Par conséquent, on en déduit que, pour  $1 \leq i \leq N$  et tout  $\phi \in C_c^\infty(\mathbb{R}^N)$ , on a

$$\left| \int_{\mathbb{R}^N} v \frac{\partial \phi}{\partial x_i} dx \right| \leq C \|\phi\|_{L^2(\mathbb{R}^N)},$$

ce qui n'est rien d'autre que la définition de l'appartenance de  $v$  à  $H^1(\mathbb{R}^N)$  (voir la Définition 4.3.1).  $\square$

**Remarque 5.2.32 (délicate)** Expliquons comment on passe du cas traité dans la Proposition 5.2.30 (avec  $\Omega = \mathbb{R}^N$ ) au cas général du Théorème 5.2.26. On utilise un argument de cartes locales et une partition de l'unité comme dans la démonstration de la Proposition 4.4.2. Suivant les notations de la Définition 3.2.5 d'un ouvert régulier (voir aussi la Figure 4.4), il existe un recouvrement fini de  $\Omega$  par des ouverts  $(\omega_i)_{0 \leq i \leq I}$  et une partition de l'unité  $(\theta_i)_{0 \leq i \leq I}$  telle que

$$\theta_i \in C_c^\infty(\omega_i), \quad 0 \leq \theta_i(x) \leq 1, \quad \sum_{i=0}^I \theta_i(x) = 1 \text{ dans } \overline{\Omega}.$$

L'ouvert  $\omega_0$  est à l'intérieur de  $\Omega$  (en fait  $\overline{\omega_0} \subset \Omega$ ) tandis que les autres ouverts  $\omega_i$ , pour  $i \geq 1$ , recouvrent le bord  $\partial\Omega$ . Soit  $u \in H_0^1(\Omega)$  la solution unique de (5.40). Pour montrer la régularité de  $u = \sum_{i=0}^I \theta_i u$ , on va montrer la régularité de chacun des termes  $\theta_i u$ . Pour le terme  $\theta_0 u$  on parle de la régularité intérieure de  $u$ . Il est immédiat que  $\theta_0 u$  est la solution unique dans  $H^1(\mathbb{R}^N)$  de

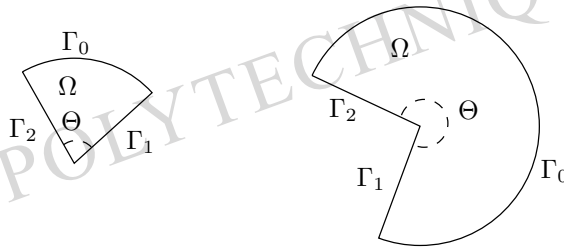
$$-\Delta(\theta_0 u) + \theta_0 u = f_0 \quad \text{dans } \mathbb{R}^N,$$

avec  $f_0 = \theta_0(f - u) - 2\nabla\theta_0 \cdot \nabla u - u\Delta\theta_0$  qui appartient à  $L^2(\mathbb{R}^N)$ . Par application de la Proposition 5.2.30 on en déduit donc que  $\theta_0 u \in H^2(\mathbb{R}^N)$ . Ceci permet d'améliorer la régularité de  $f_0$ , et par application successive de la Proposition 5.2.30 on conclut que  $f \in H^m(\Omega)$  implique que  $\theta_0 u \in H^{m+2}(\Omega)$ . Pour montrer la régularité des autres termes  $\theta_i u$  pour  $i \geq 1$ , il faut d'abord "redresser" le bord pour se ramener au cas  $\Omega = \mathbb{R}_+^N$ . Il faut donc démontrer un résultat du même type que la Proposition 5.2.30 mais pour  $\Omega = \mathbb{R}_+^N$ . C'est un peu plus délicat car en redressant le bord par cartes locales on a changé les coefficients de l'opérateur elliptique (le Laplacien devient un opérateur à coefficients variables comme dans la Sous-section 5.2.3) : nous renvoyons à [8] pour les détails. En résumé, il faut retenir que la régularité dans tout l'espace et dans un demi-espace suffit pour prouver la régularité dans un ouvert borné régulier.  $\bullet$

### Exemple de singularité

Voyons maintenant un exemple de solutions **singulières**, c'est-à-dire non régulières. Il s'agit d'un problème posé dans un ouvert non régulier pour lequel le Théorème de régularité 5.2.26 et son Corollaire 5.2.27 sont faux. En particulier, bien qu'il existe des solutions faibles (c'est-à-dire appartenant à l'espace de Sobolev  $H^1(\Omega)$ ) il n'y a pas de solutions fortes (c'est-à-dire deux fois dérivables) pour ce problème. Nous nous plaçons en dimension  $N = 2$  d'espace, et nous considérons un secteur angulaire  $\Omega$  défini en coordonnées radiales par (voir la Figure 5.2)

$$\Omega = \{(r, \theta) \text{ tel que } 0 \leq r < R \text{ et } 0 < \theta < \Theta\}$$

FIGURE 5.2 – Secteur angulaire  $\Omega$  d'angle  $\Theta$  (plus petit ou plus grand que  $\pi$ ).

avec  $0 < R < +\infty$  et  $0 < \Theta \leq 2\pi$  (rappelons que  $x_1 = r \cos \theta$  et  $x_2 = r \sin \theta$ ). On note  $\Gamma_0$  la partie du bord de  $\Omega$  où  $r = R$ ,  $\Gamma_1$  celle où  $\theta = 0$  et  $\Gamma_2$  celle où  $\theta = \Theta$ . Cet ouvert  $\Omega$  présente trois “coins”, mais seule l'origine (le coin entre les bords  $\Gamma_1$  et  $\Gamma_2$ ) peut poser problème pour la régularité dans les exemples ci-dessous. Physiquement, le cas d'un angle  $\Theta < \pi$  est représentatif d'un effet de pointe, tandis que le cas  $\Theta > \pi$  correspond à une encoche (ou même à une **fissure** si  $\Theta = 2\pi$ ).

Pour un entier  $k \geq 1$ , on étudie les deux problèmes aux limites suivants

$$\begin{cases} -\Delta u = 0 & \text{dans } \Omega \\ u = \cos\left(\frac{k\pi\theta}{\Theta}\right) & \text{sur } \Gamma_0 \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \Gamma_1 \cup \Gamma_2 \end{cases} \quad (5.50)$$

et

$$\begin{cases} -\Delta u = 0 & \text{dans } \Omega \\ u = \sin\left(\frac{k\pi\theta}{\Theta}\right) & \text{sur } \Gamma_0 \\ u = 0 & \text{sur } \Gamma_1 \cup \Gamma_2 \end{cases} \quad (5.51)$$

Nous pourrions étudier la régularité des solutions de (5.50) et (5.51) en termes d'appartenance ou non aux espaces de Sobolev  $H^m(\Omega)$ , mais, par souci de simplicité et pour garder un sens physique évident à nos résultats, nous allons simplement nous intéresser au comportement du gradient  $\nabla u$  au voisinage de l'origine. D'un point de vue physique ou mécanique, ce gradient correspond à un flux de chaleur, au champ électrique, ou à un champ de contraintes : il importe de savoir si cette quantité est bornée continue ou non à l'origine.

**Lemme 5.2.33** *Il existe une unique solution faible de (5.50) dans  $H^1(\Omega)$ , donnée par la formule*

$$u(r, \theta) = \left(\frac{r}{R}\right)^{\frac{k\pi}{\Theta}} \cos\left(\frac{k\pi\theta}{\Theta}\right). \quad (5.52)$$

*De même, il existe une unique solution faible de (5.51) dans  $H^1(\Omega)$ , donnée par la formule*

$$u(r, \theta) = \left(\frac{r}{R}\right)^{\frac{k\pi}{\Theta}} \sin\left(\frac{k\pi\theta}{\Theta}\right). \quad (5.53)$$

Dans les deux cas, si  $k = 1$  et  $\pi < \Theta$ , alors le gradient  $\nabla u$  n'est pas borné à l'origine, tandis que si  $k \geq 2$  ou  $\pi \geq \Theta$ , alors le gradient  $\nabla u$  est continu à l'origine.

**Remarque 5.2.34** Lorsque l'on impose une donnée de Dirichlet plus générale dans les problèmes (5.50) et (5.51), on peut la décomposer en une série de Fourier en  $\theta$  sur  $\Gamma_0$  et appliquer à chaque terme de la série le Lemme 5.2.33 (il faut néanmoins que cette donnée de Dirichlet sur  $\Gamma_0$  soit compatible avec les conditions aux limites sur  $\Gamma_1$  et  $\Gamma_2$ ). On en déduit que, si  $\Theta \leq \pi$ , alors les solutions de (5.50) et (5.51) sont toujours régulières quelle que soit la donnée de Dirichlet sur  $\Gamma_0$ . Au contraire, si  $\Theta > \pi$ , les solutions de (5.50) et (5.51) peuvent être singulières.

Physiquement, on interprète ce résultat de régularité en disant qu'une encoche ( $\Theta > \pi$ ) engendre une singularité, au contraire d'une pointe ( $\Theta \leq \pi$ ). Dans ce cas, il faut parfois revoir la modélisation car un flux de chaleur, un champ électrique, ou un champ de contraintes infini à l'origine n'a pas de sens physique. Deux approximations du modèle peuvent être à l'origine de cette singularité non-physique : d'une part, un angle n'est jamais "parfait" mais souvent un peu "arrondi", d'autre part, lorsque  $\nabla u$  est très grand, on quitte le domaine de validité des équations linéaires que nous étudions (typiquement, une loi constitutive comme la loi de Fourier (1.3) devient non-linéaire car la conductivité thermique est elle-même une fonction de  $\nabla u$ ). En tout état de cause, un résultat de régularité (ou non) donne de précieuses indications sur les limites de pertinence du modèle utilisé.

Dans le cas où  $\Theta = 2\pi$ , l'ouvert  $\Omega$  est un domaine **fissuré** et le Lemme 5.2.33 a une interprétation mécanique très importante si les problèmes (5.50) et (5.51) modélisent le cisaillement anti-plan d'un cylindre de base  $\Omega$  (voir l'Exercice 5.3.7). Dans ce cas, le coefficient du terme d'ordre  $k = 1$  dans la série de Fourier (qui conduit au seul comportement singulier à l'origine) est appelé le **facteur d'intensité des contraintes** qui est souvent utilisé dans des modèles de propagation de fissures ou de rupture (voir par exemple [29]). •

**Remarque 5.2.35** Les solutions singulières fournies par le Lemme 5.2.33 ne sont pas que des contre-exemples théoriques : elles peuvent être mises en évidence numériquement (voir la Figure 6.18). Ajoutons que ces solutions singulières sont une source de difficultés pour les méthodes numériques (voir la Remarque 6.3.15) ce qui confirme l'intérêt de leur étude. •

**Remarque 5.2.36** On peut aussi considérer le cas d'une condition aux limites de Dirichlet sur  $\Gamma_1$  et de Neumann sur  $\Gamma_2$  (dans ce cas, il faut prendre  $u = \sin(\frac{k\pi\theta}{2\Theta})$  sur  $\Gamma_0$ ). La seule différence porte sur le cas  $k\pi = \Theta$  pour lequel le gradient  $\nabla u$  est borné mais pas continu à l'origine. En particulier, on trouvera que pour  $k = 1$  et  $\pi = \Theta$ , **bien qu'il n'y ait pas de coin**, la solution est singulière. Cela est dû au changement de condition aux limites sur une portion régulière de la frontière. •

**Démonstration.** Nous nous bornons à traiter le problème (5.50) avec conditions aux limites de Neumann dans le coin (l'autre problème (5.51) se traite exactement de la

même manière). On commence par “relever” la condition aux limites non-homogène en définissant une fonction  $u_0 \in H^1(\Omega)$  dont la trace sur le bord coïncide avec cette condition aux limites. On vérifie aisément que

$$u_0(r, \theta) = \frac{r^2}{R^2} \cos\left(\frac{k\pi\theta}{\Theta}\right)$$

répond à la question, c’est-à-dire que  $u = u_0 + v$  où  $v$  est la solution d’un problème homogène

$$\begin{cases} -\Delta v = \Delta u_0 & \text{dans } \Omega \\ v = 0 & \text{sur } \Gamma_0 \\ \frac{\partial v}{\partial n} = 0 & \text{sur } \Gamma_1 \cup \Gamma_2. \end{cases} \quad (5.54)$$

Remarquons que ce relèvement est possible car la donnée sur  $\Gamma_0$  est compatible avec les conditions aux limites sur  $\Gamma_1$  et  $\Gamma_2$ , c’est-à-dire dans le cas présent que sa dérivée en  $\theta$  (i.e. sa dérivée normale) s’annule en  $\theta = 0$ , ou  $\Theta$ . Comme  $\Delta u_0$  appartient à  $L^2(\Omega)$ , il existe une unique solution de (5.54) dans  $H^1(\Omega)$ , et par conséquent (5.50) admet aussi une unique solution dans  $H^1(\Omega)$ . Vérifions que (5.52) est précisément cette unique solution. Rappelons que

$$\Delta\phi(r, \theta) = \frac{\partial^2\phi}{\partial r^2} + \frac{1}{r} \frac{\partial\phi}{\partial r} + \frac{1}{r^2} \frac{\partial^2\phi}{\partial\theta^2}.$$

Un calcul simple montre que

$$\left(\frac{\partial^2}{\partial r^2} + \frac{1}{r} \frac{\partial}{\partial r}\right) \left(\frac{r}{R}\right)^{\frac{k\pi}{\Theta}} = \left(\frac{k\pi}{\Theta}\right)^2 \frac{1}{r^2} \left(\frac{r}{R}\right)^{\frac{k\pi}{\Theta}},$$

donc (5.52) est bien solution de (5.50). On vérifie aussi facilement que (5.52) appartient à  $H^1(\Omega)$ . Finalement, la formule du gradient en coordonnées radiales dans la base  $(e_r, e_\theta)$

$$\nabla\phi(r, \theta) = \frac{\partial\phi}{\partial r} e_r + \frac{1}{r} \frac{\partial\phi}{\partial\theta} e_\theta$$

nous donne

$$\nabla u = \left(\frac{r}{R}\right)^{\frac{k\pi}{\Theta}-1} \frac{k\pi}{\Theta} \left( \cos\left(\frac{k\pi\theta}{\Theta}\right) e_r - \sin\left(\frac{k\pi\theta}{\Theta}\right) e_\theta \right)$$

qui, clairement, n’est pas borné à l’origine si  $0 < k\pi < \Theta$  et est continu à l’origine si  $k\pi > \Theta$ . Le cas limite  $k\pi = \Theta$  correspond aussi à un gradient continu car en fait  $u = x_1/R!$   $\square$

## 5.3 Résolution d’autres modèles

### 5.3.1 Système de l’élasticité linéarisée

Nous appliquons l’approche variationnelle à la résolution du système d’équations de l’élasticité linéarisée. Commençons par décrire ce modèle mécanique dont nous



avons vu un cas particulier au Chapitre 1. Ces équations modélisent les déformations d'un solide sous l'hypothèse de petites déformations et de petits déplacements (hypothèse qui permet d'obtenir des équations linéaires; d'où le nom d'élasticité **linéarisée**, voir par exemple [37]). On considère les équations stationnaires de l'élasticité, c'est-à-dire indépendantes du temps. Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ . Soit une force  $f(x)$ , une fonction de  $\Omega$  dans  $\mathbb{R}^N$ . L'inconnue  $u$  (le déplacement) est aussi une fonction de  $\Omega$  dans  $\mathbb{R}^N$ . La modélisation mécanique fait intervenir le tenseur des déformations, noté  $e(u)$ , qui est une fonction à valeurs dans l'ensemble des matrices symétriques

$$e(u) = \frac{1}{2} (\nabla u + (\nabla u)^t) = \frac{1}{2} \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right)_{1 \leq i, j \leq N},$$

ainsi que le tenseur des contraintes  $\sigma$  (une autre fonction à valeurs dans l'ensemble des matrices symétriques) qui est relié à  $e(u)$  par la loi de Hooke

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

où  $\lambda$  et  $\mu$  sont les coefficients de Lamé du matériau homogène isotrope qui occupe  $\Omega$ . Pour des raisons de thermodynamique les coefficients de Lamé vérifient

$$\mu > 0 \quad \text{et} \quad 2\mu + N\lambda > 0.$$

On ajoute à cette loi constitutive le bilan des forces dans le solide

$$-\operatorname{div} \sigma = f \text{ dans } \Omega$$

où, par définition, la divergence de  $\sigma$  est le vecteur de composantes

$$\operatorname{div} \sigma = \left( \sum_{j=1}^N \frac{\partial \sigma_{ij}}{\partial x_j} \right)_{1 \leq i \leq N}.$$

Utilisant le fait que  $\operatorname{tr}(e(u)) = \operatorname{div} u$ , on en déduit les équations pour  $1 \leq i \leq N$

$$-\sum_{j=1}^N \frac{\partial}{\partial x_j} \left( \mu \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) + \lambda (\operatorname{div} u) \delta_{ij} \right) = f_i \text{ dans } \Omega \quad (5.55)$$

avec  $f_i$  et  $u_i$ , pour  $1 \leq i \leq N$ , les composantes de  $f$  et  $u$  dans la base canonique de  $\mathbb{R}^N$ . En ajoutant une condition aux limites de Dirichlet, et en utilisant des notations vectorielles, le problème aux limites considéré est

$$\begin{cases} -\operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (5.56)$$

Nous pouvons énoncer et démontrer un premier résultat d'existence et d'unicité du système de l'élasticité linéarisée.

**Théorème 5.3.1** Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)^N$ . Il existe une unique solution (faible)  $u \in H_0^1(\Omega)^N$  de (5.56).

**Démonstration.** Pour trouver la formulation variationnelle on multiplie chaque équation (5.55) par une fonction test  $v_i$  (qui s'annule sur le bord  $\partial\Omega$  pour prendre en compte la condition aux limites de Dirichlet) et on intègre par partie pour obtenir

$$\int_{\Omega} \mu \sum_{j=1}^N \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial v_i}{\partial x_j} dx + \int_{\Omega} \lambda \operatorname{div} u \frac{\partial v_i}{\partial x_i} dx = \int_{\Omega} f_i v_i dx.$$

On somme alors ces équations, pour  $i$  allant de 1 à  $N$ , afin de faire apparaître la divergence de la fonction  $v = (v_1, \dots, v_N)$  et de simplifier la première intégrale car

$$\sum_{i,j=1}^N \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \frac{\partial v_i}{\partial x_j} = \frac{1}{2} \sum_{i,j=1}^N \left( \frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} \right) \left( \frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) = 2e(u) \cdot e(v).$$

Choissant  $H_0^1(\Omega)^N$  comme espace de Hilbert, on obtient alors la formulation variationnelle : trouver  $u \in H_0^1(\Omega)^N$  tel que

$$\int_{\Omega} 2\mu e(u) \cdot e(v) dx + \int_{\Omega} \lambda \operatorname{div} u \operatorname{div} v dx = \int_{\Omega} f \cdot v dx \quad \forall v \in H_0^1(\Omega)^N. \quad (5.57)$$

On vérifie aisément que chaque terme de (5.57) a bien un sens.

Pour pouvoir appliquer le Théorème de Lax-Milgram 3.3.1 à la formulation variationnelle (5.57), la seule hypothèse délicate à vérifier est la coercivité de la forme bilinéaire. On procède en trois étapes. Premièrement, montrons que

$$\int_{\Omega} 2\mu |e(v)|^2 dx + \int_{\Omega} \lambda |\operatorname{div} v|^2 dx \geq \nu \int_{\Omega} |e(v)|^2 dx,$$

avec  $\nu = \min(2\mu, (2\mu + N\lambda)) > 0$ . Pour cela, on utilise une inégalité algébrique : si on note  $A \cdot B = \sum_{i,j=1}^N a_{ij} b_{ij}$  le produit scalaire usuel des matrices symétriques, on peut décomposer toute matrice réelle symétrique  $A$  sous la forme

$$A = A^d + A^h \text{ avec } A^d = A - \frac{1}{N} \operatorname{tr} A \operatorname{Id} \text{ et } A^h = \frac{1}{N} \operatorname{tr} A \operatorname{Id},$$

de telle manière que  $A^d \cdot A^h = 0$  et  $|A|^2 = |A^d|^2 + |A^h|^2$ . On a alors

$$2\mu |A|^2 + \lambda (\operatorname{tr} A)^2 = 2\mu |A^d|^2 + (2\mu + N\lambda) |A^h|^2 \geq \nu |A|^2$$

avec  $\nu = \min(2\mu, (2\mu + N\lambda))$ , ce qui donne le résultat pour  $A = e(u)$ . Le fait que  $\nu > 0$  n'est pas un hasard : les arguments mécaniques et thermodynamiques qui conduisent aux inégalités  $\mu > 0$  et  $(2\mu + N\lambda) > 0$  sont précisément les mêmes. Deuxièmement, on

utilise l'inégalité de Korn (ou plutôt un cas particulièrement simple de cette inégalité, voir le Lemme 5.3.2 ci-dessous) qui donne une constante  $C > 0$  telle que

$$\int_{\Omega} |e(v)|^2 dx \geq C \int_{\Omega} |\nabla v|^2 dx$$

pour tout  $v \in H_0^1(\Omega)^N$ . Troisièmement, on utilise l'inégalité de Poincaré (composante par composante, voir la Proposition 4.3.10) qui donne une constante  $C > 0$  telle que, pour tout  $v \in H_0^1(\Omega)^N$ ,

$$\int_{\Omega} |v|^2 dx \leq C \int_{\Omega} |\nabla v|^2 dx.$$

Au total, ces trois inégalités conduisent à la coercivité

$$\int_{\Omega} 2\mu |e(v)|^2 dx + \int_{\Omega} \lambda |\operatorname{div} v|^2 dx \geq C \|v\|_{H^1(\Omega)}^2.$$

Le Théorème de Lax-Milgram 3.3.1 donne donc l'existence et l'unicité de la solution de la formulation variationnelle (5.57). Finalement, pour montrer que la solution unique de (5.57) est bien une solution du problème aux limites (5.56), on procède comme lors de la démonstration du Théorème 5.2.2 pour le Laplacien.  $\square$

**Lemme 5.3.2** Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$ . Pour toute fonction  $v \in H_0^1(\Omega)^N$ , on a

$$\|\nabla v\|_{L^2(\Omega)} \leq \sqrt{2} \|e(v)\|_{L^2(\Omega)}. \quad (5.58)$$

**Démonstration.** Soit  $v \in C_c^\infty(\Omega)^N$ . Par intégration par partie on obtient

$$2 \int_{\Omega} |e(v)|^2 dx = \int_{\Omega} |\nabla v|^2 dx + \int_{\Omega} \nabla v \cdot (\nabla v)^t dx = \int_{\Omega} |\nabla v|^2 dx + \int_{\Omega} |\operatorname{div} v|^2 dx.$$

Par densité de  $C_c^\infty(\Omega)$  dans  $H_0^1(\Omega)$ , on en déduit (5.58).  $\square$

**Exercice 5.3.1** Montrer que l'application de  $L^2(\Omega)^N$  dans  $H_0^1(\Omega)^N$  qui à  $f$  fait correspondre  $u$ , l'unique solution faible de (5.56), est linéaire continue.

L'analyse du problème aux limites (5.56) avec une condition aux limites de Dirichlet **sur tout le bord**  $\partial\Omega$  est un peu trompeuse par sa simplicité. En effet, dès que l'on introduit une autre condition aux limites (par exemple, de Neumann) sur une partie du bord, la démonstration de la coercivité de la formulation variationnelle devient beaucoup plus difficile car on doit remplacer l'inégalité élémentaire du Lemme 5.3.2 par sa généralisation, nettement plus technique, dite inégalité de Korn (voir le Lemme 5.3.3 ci-dessous). Rappelons qu'on ne peut pas, en général, se contenter d'une condition aux limites de Dirichlet sur l'intégralité du bord  $\partial\Omega$  car elle s'interprète comme le fait que le solide est fixé et immobile sur son bord. En pratique, tout le bord n'est pas bloqué et souvent une partie du bord est libre de bouger, ou bien des

forces surfaciques sont appliquées sur une autre partie. Ces deux cas de figures sont modélisés par des conditions aux limites de Neumann qui s'écrivent ici

$$\sigma n = g \text{ sur } \partial\Omega, \quad (5.59)$$

où  $g$  est une fonction à valeurs vectorielles. La condition de Neumann (5.59) s'interprète en disant que  $g$  est une force appliquée sur le bord (plus précisément,  $g$  est une densité de forces surfacique, homogène à une pression, tandis que  $f$  est une densité de forces volumique). Si  $g = 0$ , aucune force ne s'applique et le bord peut bouger sans restriction : on dit que le bord est libre.

Nous allons maintenant considérer le système de l'élasticité avec des conditions aux limites mêlées (un mélange de Dirichlet et de Neumann), c'est-à-dire

$$\begin{cases} -\operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega_D \\ \sigma n = g & \text{sur } \partial\Omega_N, \end{cases} \quad (5.60)$$

où  $(\partial\Omega_N, \partial\Omega_D)$  est une partition de  $\partial\Omega$  telle que les mesures superficielles de  $\partial\Omega_N$  et  $\partial\Omega_D$  sont non nulles (voir la Figure 4.1). L'analyse de ce nouveau problème aux limites est plus compliquée que dans le cas de la condition aux limites de Dirichlet : il faudra utiliser l'inégalité de Korn ci-dessous.

**Lemme 5.3.3 (Inégalité de Korn)** *Soit  $\Omega$  un ouvert borné régulier de classe  $\mathcal{C}^1$  de  $\mathbb{R}^N$ . Il existe une constante  $C > 0$  telle que, pour toute fonction  $v \in H^1(\Omega)^N$ , on a*

$$\|v\|_{H^1(\Omega)} \leq C \left( \|v\|_{L^2(\Omega)}^2 + \|e(v)\|_{L^2(\Omega)}^2 \right)^{1/2}. \quad (5.61)$$

L'inégalité (5.61) n'est pas une banalité : en effet, son membre de gauche contient toutes les dérivées partielles de  $v$  alors que son membre de droite fait intervenir seulement certaines combinaisons linéaires des dérivées partielles. Comme l'inégalité inverse de (5.61) est évidente, on en déduit que les deux membres de (5.61) sont des normes équivalentes. La démonstration du Lemme 5.3.3 est compliquée et dépasse le cadre de ce cours (voir, par exemple, [20]). Nous l'admettons donc en remarquant que nous avons bien démontré l'inégalité de Korn (5.61) lorsque  $v$  appartient à  $H_0^1(\Omega)^N$ . En effet, dans ce cas une combinaison du Lemme 5.3.2 et de l'inégalité de Poincaré (pour un ouvert borné) donne bien l'inégalité (5.61).

L'interprétation mécanique de l'inégalité de Korn est la suivante. L'énergie élastique, proportionnelle à la norme du tenseur des déformations  $e(u)$  dans  $L^2(\Omega)$ , contrôle la norme du déplacement  $u$  dans  $H^1(\Omega)^N$ , à l'addition près de la norme de  $u$  dans  $L^2(\Omega)$ . Comme nous allons le voir dans l'Exercice 5.3.2, ce dernier ajout est destiné à prendre en compte les **mouvements de corps rigides** c'est-à-dire les déplacements  $u$  non nuls mais d'énergie élastique nulle.

**Exercice 5.3.2** Soit  $\Omega$  un ouvert connexe de  $\mathbb{R}^N$ . Soit l'ensemble  $\mathcal{R}$  des “mouvements rigides” de  $\Omega$  défini par

$$\mathcal{R} = \{v(x) = b + Mx \text{ avec } b \in \mathbb{R}^N, M = -M^t \text{ matrice antisymétrique}\}. \quad (5.62)$$

Montrer que  $v \in H^1(\Omega)^N$  vérifie  $e(v) = 0$  dans  $\Omega$  si et seulement si  $v \in \mathcal{R}$ .

Nous pouvons maintenant énoncer un deuxième résultat d'existence et d'unicité pour le système de l'élasticité linéarisé avec conditions aux limites mêlées.

**Théorème 5.3.4** Soit  $\Omega$  un ouvert borné connexe régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)^N$  et  $g \in L^2(\partial\Omega_N)^N$ . On définit l'espace

$$V = \{v \in H^1(\Omega)^N \text{ tel que } v = 0 \text{ sur } \partial\Omega_D\}. \quad (5.63)$$

Il existe une unique solution (faible)  $u \in V$  de (5.60) qui dépend linéairement et continûment des données  $f$  et  $g$ .

**Démonstration.** La formulation variationnelle de (5.60) s'obtient comme dans la démonstration du Théorème 5.3.1. L'espace  $V$ , défini par (5.63), contient la condition aux limites de Dirichlet sur  $\partial\Omega_D$  et est bien un espace de Hilbert comme sous-espace fermé de  $H^1(\Omega)^N$  (par application du Théorème de trace 4.3.13). On obtient alors la formulation variationnelle : trouver  $u \in V$  tel que

$$\int_{\Omega} 2\mu e(u) \cdot e(v) dx + \int_{\Omega} \lambda \operatorname{div} u \operatorname{div} v dx = \int_{\Omega} f \cdot v dx + \int_{\partial\Omega_N} g \cdot v ds \quad \forall v \in V. \quad (5.64)$$

Pour pouvoir appliquer le Théorème de Lax-Milgram 3.3.1 à la formulation variationnelle (5.64), la seule hypothèse délicate à vérifier est encore une fois la coercivité de la forme bilinéaire. Autrement dit, il faut montrer qu'il existe une constante  $C > 0$  telle que, pour toute fonction  $v \in V$ , on a

$$\|v\|_{H^1(\Omega)} \leq C \|e(v)\|_{L^2(\Omega)}. \quad (5.65)$$

Tout d'abord, on note que  $\|e(v)\|_{L^2(\Omega)}$  est une norme sur  $V$ . Le seul point à vérifier qui mérite que l'on s'y attarde est que  $\|e(v)\|_{L^2(\Omega)} = 0$  implique  $v = 0$ . Supposons donc que  $\|e(v)\|_{L^2(\Omega)} = 0$  : alors l'Exercice 5.3.2 montre que  $v$  est un déplacement rigide, c'est-à-dire que  $v(x) = b + Mx$  avec  $M = -M^t$ . On vérifie facilement que, si  $M \neq 0$ , alors les points  $x$ , solutions de  $b + Mx = 0$ , forment une droite dans  $\mathbb{R}^3$  et un point dans  $\mathbb{R}^2$ . Or  $v(x) = 0$  sur  $\partial\Omega_D$ , qui est de mesure surfacique non nulle, donc nécessairement  $M = 0$  et  $b = 0$ . Démontrons maintenant (5.65) par un argument de contradiction (voir la Remarque 4.3.18). Si (5.65) est faux, il existe une suite  $v_n \in V$  telle que

$$\|v_n\|_{H^1(\Omega)} = 1 > n \|e(v_n)\|_{L^2(\Omega)}.$$

En particulier, la suite  $e(v_n)$  tend vers 0 dans  $L^2(\Omega)^{N^2}$ . D'autre part, comme  $v_n$  est bornée dans  $H^1(\Omega)^N$ , par application du Théorème de Rellich 4.3.21, il existe

une sous-suite  $v_{n'}$  qui converge dans  $L^2(\Omega)^N$ . L'inégalité de Korn du Lemme 5.3.3 implique que

$$\|v_{n'} - v_{p'}\|_{H^1(\Omega)}^2 \leq C \|v_{n'} - v_{p'}\|_{L^2(\Omega)}^2 + \|e(v_{n'}) - e(v_{p'})\|_{L^2(\Omega)}^2,$$

d'où l'on déduit que la suite  $v_{n'}$  est de Cauchy dans  $H^1(\Omega)^N$ , donc convergente vers une limite  $v_\infty$  qui vérifie  $\|e(v_\infty)\|_{L^2(\Omega)} = 0$ . Comme il s'agit d'une norme on en déduit que la limite est nulle  $v_\infty = 0$ , ce qui est une contradiction avec le fait que  $\|v_{n'}\|_{H^1(\Omega)} = 1$ .

L'interprétation de la formulation variationnelle (5.64) pour retrouver l'équation (5.60) est semblable à celle que nous avons fait lors de la démonstration du Théorème 5.2.14 sur le Laplacien avec condition aux limites de Neumann. Finalement, l'application  $(f, g) \rightarrow u$  est linéaire. Pour montrer qu'elle est continue de  $L^2(\Omega)^N \times L^2(\partial\Omega)^N$  dans  $H^1(\Omega)^N$ , nous prenons  $v = u$  dans la formulation variationnelle (5.64). En utilisant la coercivité de la forme bilinéaire et en majorant la forme linéaire, on obtient **l'estimation d'énergie**

$$C \|u\|_{H^1(\Omega)}^2 \leq \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_N)} \|u\|_{L^2(\partial\Omega)}. \quad (5.66)$$

Grâce à l'inégalité de Poincaré et au théorème de trace, on peut majorer le terme de droite de (5.66) par  $C (\|f\|_{L^2(\Omega)} + \|g\|_{L^2(\partial\Omega_N)}) \|u\|_{H^1(\Omega)}$ , ce qui prouve la continuité.  $\square$

**Remarque 5.3.5** Pour le système de l'élasticité avec des conditions aux limites de Dirichlet (ou de Neumann) on a les mêmes résultats de régularité qu'avec le Laplacien (voir le Théorème 5.2.26). Par contre, à la différence du Laplacien, il n'y a pas de principe du maximum pour le système de l'élasticité (comme pour la plupart des systèmes de plusieurs équations). Nous donnons un contre-exemple numérique à la Figure 5.3 : en l'absence de forces,  $f = 0$ , les conditions aux limites sont de Neumann sur les faces supérieure et inférieure du domaine, de Dirichlet  $u = 0$  à droite et  $u = e_1$  à gauche. Cela revient à étirer le domaine qui, du coup, s'amincit, conduisant ainsi à un déplacement vertical qui change de signe.  $\bullet$

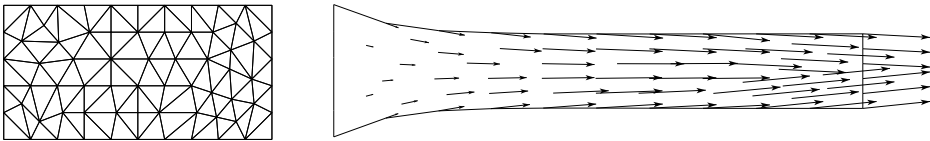


FIGURE 5.3 – Contre-exemple numérique pour le principe du maximum en élasticité. A gauche le domaine maillé au repos, et à droite le domaine déformé où les flèches représentent le déplacement (sa composante verticale change de signe).

Nous avons déjà dit que la formulation variationnelle n'est rien d'autre que le **principe des travaux virtuels** en mécanique. Poursuivant cette analogie, l'espace

$V$  est l'espace des **déplacements  $v$  cinématiquement admissibles**, et l'espace des tenseurs symétriques  $\sigma \in L^2(\Omega)^{N^2}$ , tels que  $-\operatorname{div} \sigma = f$  dans  $\Omega$  et  $\sigma n = g$  sur  $\partial\Omega_N$ , est celui des tenseurs de **contraintes statiquement admissibles**. Comme pour le Laplacien, la solution de la formulation variationnelle (5.64) réalise le minimum d'une énergie mécanique définie pour  $v \in V$  par

$$J(v) = \frac{1}{2} \int_{\Omega} (2\mu|e(v)|^2 + \lambda|\operatorname{div} v|^2) dx - \int_{\Omega} f \cdot v dx - \int_{\partial\Omega_N} g \cdot v ds. \quad (5.67)$$

En termes mécaniques  $J(v)$  est la somme de **l'énergie de déformation**

$$\frac{1}{2} \int_{\Omega} (2\mu|e(v)|^2 + \lambda|\operatorname{div} v|^2) dx$$

et de **l'énergie potentielle des forces extérieures** (ou travail des forces extérieures au signe près)

$$- \int_{\Omega} f \cdot v dx - \int_{\partial\Omega_N} g \cdot v ds.$$

**Exercice 5.3.3** Montrer que  $u \in V$  est l'unique solution de la formulation variationnelle (5.64) si et seulement si  $u$  réalise le minimum sur  $V$  de l'énergie  $J(v)$  définie par (5.67). (Indication : on pourra s'inspirer de la Proposition 3.3.4).

**Exercice 5.3.4** Soit  $\Omega$  un ouvert borné connexe de  $\mathbb{R}^N$ . On considère le système de l'élasticité avec la condition de Neumann (5.59) sur tout le bord  $\partial\Omega$ . Montrer que la condition d'équilibre

$$\int_{\Omega} f \cdot (Mx + b) dx + \int_{\partial\Omega} g \cdot (Mx + b) ds = 0 \quad \forall b \in \mathbb{R}^N, \forall M = -M^t \in \mathbb{R}^{N \times N}$$

est une condition nécessaire et suffisante d'existence et d'unicité d'une solution dans  $H^1(\Omega)^N$  (l'unicité étant obtenue à l'addition d'un "mouvement de corps rigide" près, voir (5.62)).

**Remarque 5.3.6** Lorsque les coefficients de Lamé sont constants et que les conditions aux limites sont du type de Dirichlet homogène, les équations de l'élasticité peuvent se réarranger pour donner le système de Lamé (présenté au Chapitre 1)

$$\begin{cases} -\mu\Delta u - (\mu + \lambda)\nabla(\operatorname{div} u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (5.68)$$

L'avantage de (5.68) par rapport à (5.56) est que le tenseur  $e(u)$  a disparu et qu'on peut donc se passer de l'inégalité de Korn. Il faut cependant faire attention que (5.68) n'est plus équivalent à (5.56) si les coefficients de Lamé dépendent de  $x$  ou si on place des conditions aux limites de Neumann sur une partie du bord. D'un point de vue mécanique, le seul modèle correct dans ce cas est (5.56). •

**Exercice 5.3.5** On suppose que  $\Omega$  est un ouvert borné de  $\mathbb{R}^N$  et que  $f \in L^2(\Omega)^N$ . Montrer l'existence et l'unicité de la solution de (5.68) dans  $H_0^1(\Omega)^N$  sans utiliser l'inégalité de Korn. Vérifier qu'on peut affaiblir les hypothèses de positivité sur les coefficients de Lamé en supposant seulement que  $\mu > 0$  et  $2\mu + \lambda > 0$ .

**Exercice 5.3.6** Vérifier l'équivalence de (5.68) et (5.56) si  $\lambda$  et  $\mu$  sont constants. Montrer que (5.68) et (5.56) ne sont plus équivalents si  $\lambda$  et  $\mu$  sont des fonctions (régulières), même si on remplace l'équation vectorielle de (5.68) par

$$-\operatorname{div}(\mu \nabla u) - \nabla((\mu + \lambda) \operatorname{div} u) = f \text{ dans } \Omega.$$

Dans un cas très particulier, appelé **problème du cisaillement anti-plan**, le système de l'élasticité linéarisée se simplifie considérablement puisqu'il se ramène à la résolution d'un problème aux limites pour le Laplacien. Cet exemple permet donc de faire un lien direct entre les équations de l'élasticité et le Laplacien ce qui explique d'une certaine manière pourquoi les résultats pour les deux modèles sont très semblables dans l'ensemble. Ce cas particulier du cisaillement anti-plan est étudié dans l'exercice suivant.

**Exercice 5.3.7** Le but de cet exercice est de trouver une solution particulière du système de l'élasticité linéarisée dans le cas d'une force de cisaillement anti-plan. On considère un domaine cylindrique homogène  $\Omega$  de longueur  $L > 0$  et de section  $\omega$ , où  $\omega$  est un ouvert borné connexe régulier de  $\mathbb{R}^{N-1}$  (les coefficients de Lamé  $\lambda$  et  $\mu$  sont constants). Autrement dit,  $\Omega = \omega \times (0, L)$ , et pour  $x \in \Omega$ , on note  $x = (x', x_N)$  avec  $0 < x_N < L$  et  $x' \in \omega$ . On considère le problème aux limites suivant

$$\begin{cases} -\operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = 0 & \text{dans } \Omega \\ \sigma n = g & \text{sur } \partial\omega \times (0, L) \\ u' = 0 & \text{sur } \omega \times \{0, L\} \\ (\sigma n) \cdot n = 0 & \text{sur } \omega \times \{0, L\} \end{cases} \quad (5.69)$$

où on a utilisé la notation, pour un vecteur  $v = (v_1, \dots, v_N)$ ,  $v = (v', v_N)$  avec  $v' \in \mathbb{R}^{N-1}$  et  $v_N \in \mathbb{R}$ . On suppose que la force surfacique  $g$  est du type "cisaillement anti-plan", c'est-à-dire que  $g' = (g_1, \dots, g_{N-1}) = 0$  et  $g_N$  ne dépend que de  $x'$ .

Montrer que la solution unique de (5.69) est donnée par  $u = (0, \dots, 0, u_N)$  où  $u_N(x')$  est la solution du Laplacien suivant

$$\begin{cases} -\Delta' u_N = 0 & \text{dans } \omega \\ \mu \frac{\partial u_N}{\partial n} = g_N & \text{sur } \partial\omega \end{cases}$$

où  $\Delta'$  est le Laplacien dans la variable  $x' \in \mathbb{R}^{N-1}$ .

**Exercice 5.3.8** Généraliser l'Exercice 5.3.7 au cas d'une condition aux limites latérale du type

$$u' = 0 \text{ et } (\sigma n) \cdot e_N = g_N \text{ sur } \partial\omega \times (0, L).$$



**Exercice 5.3.9** A l'aide de l'approche variationnelle démontrer l'existence et l'unicité de la solution de l'équation des plaques

$$\begin{cases} \Delta(\Delta u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega \end{cases} \quad (5.70)$$

où  $f \in L^2(\Omega)$ . On pourra remarquer que, si  $u \in H_0^2(\Omega)$ , alors  $\frac{\partial u}{\partial x_i} \in H_0^1(\Omega)$  et

$$\int_{\Omega} |\Delta u|^2 dx = \sum_{i,j=1}^N \int_{\Omega} \left| \frac{\partial^2 u}{\partial x_i \partial x_j} \right|^2 dx.$$

On admettra le résultat de régularité suivant : si  $w \in L^2(\Omega)$  et  $f \in L^2(\Omega)$  vérifient pour tout  $v \in C_c^\infty(\Omega)$

$$-\int_{\Omega} w \Delta v dx = \int_{\Omega} f v dx,$$

alors  $(\theta w) \in H^2(\Omega)$  quelle que soit la fonction  $\theta \in C_c^\infty(\Omega)$ .

### 5.3.2 Équations de Stokes

Nous appliquons l'approche variationnelle à la résolution du système d'équations de Stokes. Soit  $\Omega$  un ouvert borné connexe de  $\mathbb{R}^N$ . Soit une force  $f(x)$ , une fonction de  $\Omega$  dans  $\mathbb{R}^N$ . Il y a deux inconnues : la vitesse  $u$  qui est une fonction vectorielle, et la pression  $p$  qui est une fonction scalaire. En notation vectorielle, le problème aux limites considéré est

$$\begin{cases} \nabla p - \mu \Delta u = f & \text{dans } \Omega \\ \operatorname{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (5.71)$$

où  $\mu > 0$  est la viscosité du fluide. La deuxième équation de (5.71) est la contrainte d'incompressibilité du fluide, tandis que la première est le bilan des forces. La condition aux limites de Dirichlet modélise l'adhérence du fluide sur les parois.

**Remarque 5.3.7** Le problème de Stokes (5.71) est un modèle **simplifié** d'écoulement d'un fluide visqueux incompressible (en régime stationnaire). En effet, les "vraies" équations du mouvement d'un tel fluide sont les équations de Navier-Stokes stationnaires (voir par exemple [39])

$$\begin{cases} (u \cdot \nabla)u + \nabla p - \mu \Delta u = f & \text{dans } \Omega \\ \operatorname{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (5.72)$$

Lorsque la vitesse du fluide  $u$  est faible, le terme non-linéaire  $(u \cdot \nabla)u$  étant quadratique en  $u$  devient négligeable. On obtient alors les équations de Stokes. Il faut donc avoir à l'esprit que le domaine de validité de ce modèle est limité par cette hypothèse de petite vitesse. •

**Théorème 5.3.8** Soit  $\Omega$  un ouvert borné connexe régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Soit  $f \in L^2(\Omega)^N$ . Il existe une unique solution (faible)  $u \in H_0^1(\Omega)^N$  et  $p \in L^2(\Omega)/\mathbb{R}$  de (5.71) (la pression est unique à une constante additive près dans  $\Omega$ ).

**Démonstration.** Pour trouver la formulation variationnelle on multiplie chaque équation du système (5.71) par une fonction test  $v_i$  (qui s'annule sur le bord  $\partial\Omega$  pour prendre en compte la condition aux limites de Dirichlet), on intègre par partie et on somme pour  $i$  allant de 1 à  $N$  (pour faire apparaître la divergence de la fonction  $v = (v_1, \dots, v_N)$ )

$$\int_{\Omega} \mu \nabla u \cdot \nabla v \, dx - \int_{\Omega} p \operatorname{div} v \, dx = \int_{\Omega} f \cdot v \, dx,$$

avec la notation  $\nabla u \cdot \nabla v = \sum_{i=1}^N \nabla u_i \cdot \nabla v_i$ . Afin de tenir compte de la condition d'incompressibilité  $\operatorname{div} u = 0$ , on choisit comme espace de Hilbert le sous-espace suivant de  $H_0^1(\Omega)^N$

$$V = \{v \in H_0^1(\Omega)^N \text{ tel que } \operatorname{div} v = 0 \text{ p.p. dans } \Omega\}, \quad (5.73)$$

qui est bien un espace de Hilbert comme sous-espace fermé de  $H_0^1(\Omega)^N$ . On trouve alors la formulation variationnelle :

$$\text{trouver } u \in V \text{ tel que } \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx = \int_{\Omega} f \cdot v \, dx \quad \forall v \in V, \quad (5.74)$$

dans laquelle la pression a disparu ! On vérifie aisément que chaque terme de (5.74) a bien un sens.

L'application du Théorème de Lax-Milgram à la formulation variationnelle (5.74) ne pose pas de problème. En particulier, la coercivité de la forme bilinéaire est évidente dans  $H_0^1(\Omega)^N$  (grâce à l'inégalité de Poincaré) donc dans  $V$  qui est un sous-espace de  $H_0^1(\Omega)^N$ .

Le point le plus délicat est ici de montrer que la solution unique de (5.74) est bien une solution du problème aux limites (5.71). Expliquons la difficulté en supposant même momentanément que  $u$  est régulière, c'est-à-dire appartient à  $H^2(\Omega)^N$ . Par intégration par parties, (5.74) implique que

$$\int_{\Omega} (\mu \Delta u + f) \cdot v \, dx = 0 \quad \forall v \in V,$$

mais on ne peut pas en déduire que  $(\mu \Delta u + f) = 0$  car l'orthogonal de  $V$  dans  $L^2(\Omega)^N$  n'est pas réduit au seul vecteur nul ! En effet, on voit facilement que si  $\phi$  est une fonction régulière, alors, pour tout  $v \in V$ , on a

$$\int_{\Omega} \nabla \phi \cdot v \, dx = - \int_{\Omega} \phi \operatorname{div} v \, dx = 0,$$

c'est-à-dire que l'orthogonal de  $V$  contient au moins tous les gradients. En fait le Théorème de de Rham 5.3.9 nous dit que l'orthogonal de  $V$  coïncide exactement avec l'espace des gradients.

Grâce au Théorème de de Rham 5.3.9 nous pouvons conclure comme suit. On pose

$$L(v) = \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx - \int_{\Omega} f \cdot v \, dx,$$

qui est bien une forme linéaire continue sur  $H_0^1(\Omega)^N$  et nulle sur  $V$ . Par conséquent, il existe  $p \in L^2(\Omega)$ , unique à une constante additive près, tel que

$$L(v) = \int_{\Omega} p \operatorname{div} v \, dx \quad \forall v \in H_0^1(\Omega)^N.$$

Si on pose  $\sigma = \mu \nabla u - p \operatorname{Id}$  qui appartient à  $L^2(\Omega)^{N^2}$ , on a donc

$$\left| \int_{\Omega} \sigma \cdot \nabla v \, dx \right| = \left| \int_{\Omega} f \cdot v \, dx \right| \leq C \|v\|_{L^2(\Omega)},$$

ce qui prouve que  $\sigma$  admet une divergence faible dans  $L^2(\Omega)^N$ , et on a  $-\operatorname{div} \sigma = f$ . Par conséquent, on en déduit que

$$\nabla p - \mu \Delta u = f \text{ presque partout dans } \Omega. \quad (5.75)$$

D'autre part, comme  $u \in V$ ,  $\operatorname{div} u$  est nul dans  $L^2(\Omega)$ , ce qui implique

$$\operatorname{div} u = 0 \text{ presque partout dans } \Omega.$$

De même, la condition aux limites s'interprète par le théorème de trace et on obtient que  $u = 0$  presque partout sur  $\partial\Omega$ .  $\square$

Nous énonçons maintenant le résultat très profond et difficile qui nous a permis de retrouver la pression à partir de la formulation variationnelle (5.74) du système de Stokes où elle avait disparu ! Sa démonstration dépasse très largement le cadre de ce cours (signalons qu'il n'est pas facile de trouver dans la littérature une démonstration "élémentaire" et auto-contenue; voir néanmoins [25]).

**Théorème 5.3.9 (de de Rham)** *Soit  $\Omega$  un ouvert borné connexe régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Soit  $L$  une forme linéaire continue sur  $H_0^1(\Omega)^N$ . Alors  $L$  s'annule sur  $V$  si et seulement si il existe une fonction  $p \in L^2(\Omega)$  telle que*

$$L(v) = \int_{\Omega} p \operatorname{div} v \, dx \quad \forall v \in H_0^1(\Omega)^N. \quad (5.76)$$

*De plus  $p$  est unique à une constante additive près.*

**Remarque 5.3.10** Pour Stokes comme pour le Laplacien ou pour l'élasticité on peut aussi définir des conditions aux limites de Neumann qui s'écrivent

$$\mu \frac{\partial u}{\partial n} - pn = g \text{ sur } \partial\Omega, \quad (5.77)$$

où  $g$  est une fonction (à valeurs vectorielles) de  $L^2(\partial\Omega)^N$ . Comme pour l'élasticité, la condition de Neumann (5.77) s'interprète en disant que  $g$  est une force appliquée sur le bord.  $\bullet$

**Remarque 5.3.11** Pour le système de Stokes (5.71) on a les mêmes résultats de régularité que pour le Laplacien ou le système de l'élasticité linéarisée (voir le Théorème 5.2.26). Par contre, les équations de Stokes étant un système de plusieurs équations, il n'y a pas de principe du maximum (c'est la même situation que pour le système de l'élasticité). Physiquement, à cause de la condition d'incompressibilité du fluide cela se comprend très bien. En effet, si l'on étudie un écoulement de Stokes dans une tuyère présentant un resserrement (ou col) prononcé, le débit étant constant dans la section, la vitesse du fluide est nécessairement plus élevée au niveau de ce col que sur l'entrée ou la sortie de la tuyère (ce qui viole le principe du maximum en l'absence de force extérieure). •

Comme pour le système de l'élasticité il existe un principe de minimisation de l'énergie (dite de **dissipation visqueuse**) pour les équations de Stokes.

**Exercice 5.3.10** Soit  $V$  l'espace des champs de vitesse à divergence nulle défini par (5.73). Soit  $J(v)$  l'énergie définie pour  $v \in V$  par

$$J(v) = \frac{1}{2} \int_{\Omega} \mu |\nabla v|^2 dx - \int_{\Omega} f \cdot v dx. \quad (5.78)$$

Soit  $u \in V$  la solution unique de la formulation variationnelle (5.74). Montrer que  $u$  est aussi l'unique point de minimum de l'énergie, c'est-à-dire que  $J(u) = \min_{v \in V} J(v)$ . Réciproquement, montrer que, si  $u \in V$  est un point de minimum de l'énergie  $J(v)$ , alors  $u$  est la solution unique de la formulation variationnelle (5.74).

Dans l'exercice suivant nous allons voir que dans certains cas très particuliers les équations de Stokes se réduisent au Laplacien.

**Exercice 5.3.11** Le but de cet exercice est de trouver une solution particulière des équations de Stokes dans un canal rectiligne de section uniforme, appelée profil de Poiseuille. Soit  $\Omega = \omega \times (0, L)$  où  $L > 0$  est la longueur du canal et  $\omega$  sa section, un ouvert borné connexe régulier de  $\mathbb{R}^{N-1}$ . Pour  $x \in \Omega$ , on note  $x = (x', x_N)$  avec  $0 < x_N < L$  et  $x' \in \omega$ . On considère le problème aux limites suivant

$$\begin{cases} \nabla p - \mu \Delta u = 0 & \text{dans } \Omega \\ \operatorname{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\omega \times (0, L) \\ pn - \mu \frac{\partial u}{\partial n} = p_0 n & \text{sur } \omega \times \{0\} \\ pn - \mu \frac{\partial u}{\partial n} = p_L n & \text{sur } \omega \times \{L\} \end{cases} \quad (5.79)$$

où  $p_0$  et  $p_L$  sont deux pressions constantes. Montrer que la solution unique de (5.79) est  $p(x) = p_0 + \frac{x_N}{L}(p_L - p_0)$ , et  $u = (0, \dots, 0, u_N)$  où  $u_N$  est la solution du Laplacien suivant

$$\begin{cases} -\mu \Delta' u_N = -\frac{(p_L - p_0)}{L} & \text{dans } \omega \\ u_N = 0 & \text{sur } \partial\omega \end{cases}$$

où  $\Delta'$  est le Laplacien dans la variable  $x' \in \mathbb{R}^{N-1}$ .

**Exercice 5.3.12** Généraliser l'Exercice 5.3.11 au cas des équations de Navier-Stokes (5.72) avec les conditions aux limites proposées dans (5.79).

## Chapitre 6

# MÉTHODE DES ÉLÉMENTS FINIS

### 6.1 Approximation variationnelle

#### 6.1.1 Introduction

Dans ce chapitre nous présentons la méthode des **éléments finis** qui est la méthode numérique de référence pour le calcul des solutions de problèmes aux limites elliptiques, mais aussi paraboliques ou hyperboliques comme nous le verrons par la suite. Le principe de cette méthode est directement issu de l'**approche variationnelle** que nous avons étudiée en détail dans les chapitres précédents.

L'idée de base de la méthode des éléments finis est de remplacer l'espace de Hilbert  $V$  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  $V$ .

Historiquement, les premières prémices de la méthode des éléments finis ont été proposées par le mathématicien Richard Courant (sans utiliser cette dénomination) dans les années 1940, mais ce sont les mécaniciens qui ont développé, popularisé, et démontré l'efficacité de cette méthode dans les années 1950-1960 (en plus de lui donner son nom actuel). Après ces premiers succès pratiques, les mathématiciens ont alors considérablement développé les fondations théoriques de la méthode et proposé des améliorations significatives. C'est en tout cas un bel exemple de coopération interdisciplinaire où les efforts conjugués des mécaniciens et des mathématiciens appliqués ont fait faire des progrès immenses à la simulation numérique (sans négliger non plus les avancées encore plus spectaculaires de la puissance des ordinateurs).

Le plan de ce chapitre est le suivant. Dans la suite de cette section nous détaillons le processus **d'approximation variationnelle interne**. La Section 6.2 présente les éléments finis en une dimension d'espace où, sans trahir les idées générales valables en dimensions supérieures, les aspects techniques sont nettement plus simples. 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). La Section 6.3 est dédiée aux éléments finis en dimension supérieure ( $N \geq 2$ ). On introduit les notions de **maillage** (triangulaire ou quadrangulaire) et de **degrés de liberté** qui permettent de construire plusieurs familles de méthodes d'éléments finis. On reprend alors les aspects pratiques et théoriques déjà ébauchés en dimension  $N = 1$ .

Terminons cette introduction en disant qu'en plus des méthodes de différences finies et d'éléments finis, qui sont les seules exposées dans ce cours, il existe bien d'autres méthodes numériques de résolution d'équations aux dérivées partielles comme les méthodes de volumes finis, d'éléments finis de frontière (ou méthode intégrale), spectrale, de Fourier, etc. (voir les encyclopédies [12], [18]). Pour plus de détails sur la méthode des éléments finis nous renvoyons à [4], [11], [22], [25], [35], [36] (voir aussi [16], [17], [32] pour des aspects pratiques de programmation informatique).

### 6.1.2 Approximation interne générale

Nous considérons à nouveau le cadre général du formalisme variationnel introduit au Chapitre 3. Étant donné un espace de Hilbert  $V$ , une forme bilinéaire continue et coercive  $a(u, v)$ , et une forme linéaire continue  $L(v)$ , on considère la formulation variationnelle :

$$\text{trouver } u \in V \text{ tel que } a(u, v) = L(v) \quad \forall v \in V, \quad (6.1)$$

dont on sait qu'elle admet une unique solution par le Théorème 3.3.1 de Lax-Milgram.

**L'approximation interne** de (6.1) consiste à remplacer l'espace de Hilbert  $V$  par un sous-espace de dimension finie  $V_h$ , c'est-à-dire à chercher la solution de :

$$\text{trouver } u_h \in V_h \text{ tel que } a(u_h, v_h) = L(v_h) \quad \forall v_h \in V_h. \quad (6.2)$$

La résolution de l'approximation interne (6.2) est facile comme le montre le lemme suivant.

**Lemme 6.1.1** *Soit  $V$  un espace de Hilbert réel, et  $V_h$  un sous-espace de dimension finie. Soit  $a(u, v)$  une forme bilinéaire continue et coercive sur  $V$ , et  $L(v)$  une forme linéaire continue sur  $V$ . Alors l'approximation interne (6.2) 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.** L'existence et l'unicité de  $u_h \in V_h$ , solution de (6.2), découle du Théorème 3.3.1 de Lax-Milgram appliqué à  $V_h$ . Pour mettre le problème sous une

forme plus simple, on introduit une base  $(\phi_j)_{1 \leq j \leq N_h}$  de  $V_h$ . Si  $u_h = \sum_{j=1}^{N_h} u_j \phi_j$ , on pose  $U_h = (u_1, \dots, u_{N_h})$  le vecteur dans  $\mathbb{R}^{N_h}$  des coordonnées de  $u_h$ . Le problème (6.2) est équivalent à :

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

c'est qui s'écrit sous la forme d'un système linéaire

$$\mathcal{K}_h U_h = b_h, \quad (6.3)$$

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  $U_h \in \mathbb{R}^{N_h}$ , on a

$$\mathcal{K}_h U_h \cdot U_h \geq \nu \left\| \sum_{j=1}^{N_h} u_j \phi_j \right\|^2 \geq C |U_h|^2 \text{ avec } C > 0,$$

car toutes les normes sont équivalentes en dimension finie ( $\|\cdot\|$  désigne la norme euclidienne dans  $\mathbb{R}^{N_h}$ ). 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**.  $\square$

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 (6.1) et  $u_h$  celle dans  $V_h$  de (6.2). Précisons auparavant quelques notations : on note  $\nu > 0$  la constante de coercivité et  $M > 0$  la constante de continuité de la forme bilinéaire  $a(u, v)$  qui vérifient

$$\begin{aligned} a(u, u) &\geq \nu \|u\|^2 \quad \forall u \in V, \\ |a(u, v)| &\leq M \|u\| \|v\| \quad \forall u, v \in V \end{aligned}$$

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 6.1.2 (de Céa)** *On se place sous les hypothèses du Lemme 6.1.1. Soit  $u$  la solution de (6.1) et  $u_h$  celle de (6.2). On a*

$$\|u - u_h\| \leq \frac{M}{\nu} \inf_{v_h \in V_h} \|u - v_h\|. \quad (6.4)$$

**Démonstration.** Puisque  $V_h \subset V$ , on déduit, par soustraction des formulations variationnelles (6.1) et (6.2), 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 (6.4).  $\square$

**Exercice 6.1.1** Dans le cadre du Lemme de Céa 6.1.2, démontrer que, si la forme bilinéaire  $a(u, v)$  est symétrique, alors on améliore (6.4) 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 (6.2) réalise aussi le minimum d'une énergie.

Finalement, pour démontrer la convergence de cette approximation variationnelle, nous donnons un dernier lemme général. Rappelons que dans la notation  $V_h$  le paramètre  $h > 0$  n'a pas encore de signification pratique. Néanmoins, nous supposons que c'est dans la limite  $h \rightarrow 0$  que l'approximation interne (6.2) "converge" vers la formulation variationnelle (6.1).

**Lemme 6.1.3** *On se place sous les hypothèses du Lemme 6.1.1. On suppose qu'il existe un sous-espace  $\mathcal{V} \subset V$  dense dans  $V$  et une application  $r_h$  de  $\mathcal{V}$  dans  $V_h$  (appelée **opérateur d'interpolation**) tels que*

$$\lim_{h \rightarrow 0} \|v - r_h(v)\| = 0 \quad \forall v \in \mathcal{V}. \quad (6.5)$$

*Alors la méthode d'approximation variationnelle interne converge, c'est-à-dire que*

$$\lim_{h \rightarrow 0} \|u - u_h\| = 0. \quad (6.6)$$

**Démonstration.** Soit  $\epsilon > 0$ . Par densité de  $\mathcal{V}$ , il existe  $v \in \mathcal{V}$  tel que  $\|u - v\| \leq \epsilon$ . Par ailleurs, il existe un  $h_0 > 0$  (dépendant de  $\epsilon$ ) tel que, pour cet élément  $v \in \mathcal{V}$ , on a

$$\|v - r_h(v)\| \leq \epsilon \quad \forall h \leq h_0.$$

En vertu du Lemme 6.1.2, on a

$$\|u - u_h\| \leq C \|u - r_h(v)\| \leq C (\|u - v\| + \|v - r_h(v)\|) \leq 2C\epsilon,$$

d'où l'on déduit le résultat.  $\square$



La stratégie indiquée par les Lemmes 6.1.1, 6.1.2 et 6.1.3 ci-dessus est maintenant claire. **Pour obtenir une approximation numérique de la solution exacte du problème variationnel (6.1), il faut introduire un espace  $V_h$  de dimension finie puis résoudre un simple système linéaire associé à l'approximation variationnelle interne (6.2).** Néanmoins, le choix de  $V_h$  n'est pas évident. Il faut qu'il respecte deux critères :

1. on doit pouvoir construire un opérateur d'interpolation  $r_h$  de  $\mathcal{V}$  dans  $V_h$  satisfaisant (6.5) (où typiquement  $\mathcal{V}$  est un espace de fonctions régulières),
2. il faut que la résolution du système linéaire  $\mathcal{K}_h U_h = b_h$  soit économique (en pratique ces systèmes linéaires sont de très grandes tailles).

La méthode des éléments finis consiste précisément à fournir de tels “bons” espaces  $V_h$ . Avant d'entrer dans les détails, disons quelques mots de la méthode de Galerkin qui rentre aussi dans ce cadre.

### 6.1.3 Méthode de Galerkin

La méthode de Galerkin a été un précurseur de la méthode des éléments finis. Bien qu'elle n'a pas d'intérêt numérique en général, elle est très utile d'un point de vue théorique (notamment pour l'étude des problèmes non-linéaires). Elle rentre dans le cadre de l'approximation variationnelle interne décrit ci-dessus.

On suppose que l'espace de Hilbert  $V$  est séparable de dimension infinie, ce qui entraîne, par la Proposition 12.1.15, qu'il existe une base hilbertienne  $(e_i)_{i \geq 1}$  de  $V$ . On choisit alors  $\mathcal{V}$  comme le sous-espace engendré par cette base hilbertienne (engendré par combinaison linéaires finies) qui est bien sûr dense dans  $V$ . En posant  $h = 1/n$ , on définit  $V_h$  comme le sous-espace de dimension finie engendré par  $(e_1, \dots, e_n)$ . Finalement, l'opérateur d'interpolation  $r_h$  est simplement la projection orthogonale sur  $V_h$  (qui est ici définie dans tout  $V$  et pas seulement dans  $\mathcal{V}$ ).

Toutes les hypothèses des Lemmes 6.1.1, 6.1.2 et 6.1.3 sont donc satisfaites et on en déduit que la solution approchée  $u_h$  converge vers la solution exacte  $u$ . Rappelons que  $u_h$  est calculé en résolvant le système linéaire  $\mathcal{K}_h U_h = b_h$  où  $U_h$  est le vecteur dans  $\mathbb{R}^n$  des coordonnées de  $u_h$  dans la base  $(e_1, \dots, e_n)$ .

Malgré son adéquation au cadre théorique développé ci-dessus, la méthode de Galerkin est peu commode d'un point de vue numérique. En effet, la matrice  $\mathcal{K}_h$  que l'on obtient ainsi est généralement “pleine”, c'est-à-dire que tous ces coefficients sont non nuls en général, et “mal-conditionnée”, c'est-à-dire que la résolution numérique du système linéaire sera instable car très sensible aux erreurs d'arrondi du calcul sur ordinateur. De ce point de vue, la méthode des éléments finis est bien plus performante et de loin préférable à la méthode de Galerkin.

### 6.1.4 Méthode des éléments finis (principes généraux)

Le principe de la méthode des éléments finis est de construire des espaces d'approximation interne  $V_h$  des espaces fonctionnels usuels  $H^1(\Omega)$ ,  $H_0^1(\Omega)$ ,  $H^2(\Omega)$ , ... dont

la définition est basée sur la notion géométrique de **maillage** du domaine  $\Omega$ . Un maillage est un pavage de l'espace en volumes élémentaires très simples : triangles, tétraèdres, parallélépipèdes (voir, par exemple, la Figure 6.7). Nous donnerons plus loin une définition précise d'un maillage dans le cadre de la méthode des éléments finis.

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  $h \rightarrow 0$ , 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 (6.3) sera **creuse**, c'est-à-dire que la plupart de ses coefficients seront nuls (ce qui limitera le coût de la résolution numérique).

La méthode des éléments finis est une des méthodes les plus efficace et les plus populaire pour résoudre numériquement des problèmes aux limites. Elle est à la base d'innombrables logiciels de calculs industriels.

## 6.2 Éléments finis en dimension $N = 1$

Pour simplifier l'exposition nous commençons par présenter la méthode des éléments finis en une dimension d'espace. 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

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

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

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

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

$$\begin{cases} -u'' = f & \text{dans } ]0, 1[ \\ u(0) = u(1) = 0, \end{cases} \quad (6.7)$$

dont nous savons qu'il admet une solution unique dans  $H_0^1(\Omega)$  si  $f \in L^2(\Omega)$  (voir le Chapitre 5). 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  $k$ .

### 6.2.1 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

$$V_h = \{v \in C([0, 1]) \text{ tel que } v|_{[x_j, x_{j+1}]} \in \mathbb{P}_1 \text{ pour tout } 0 \leq j \leq n\}, \quad (6.8)$$

et sur son sous-espace

$$V_{0h} = \{v \in V_h \text{ tel que } v(0) = v(1) = 0\}. \quad (6.9)$$

La méthode des éléments finis  $\mathbb{P}_1$  est alors simplement la méthode d'approximation variationnelle interne de la Sous-section 6.1.2 appliquée aux espaces  $V_h$  ou  $V_{0h}$  définis par (6.8) ou (6.9).

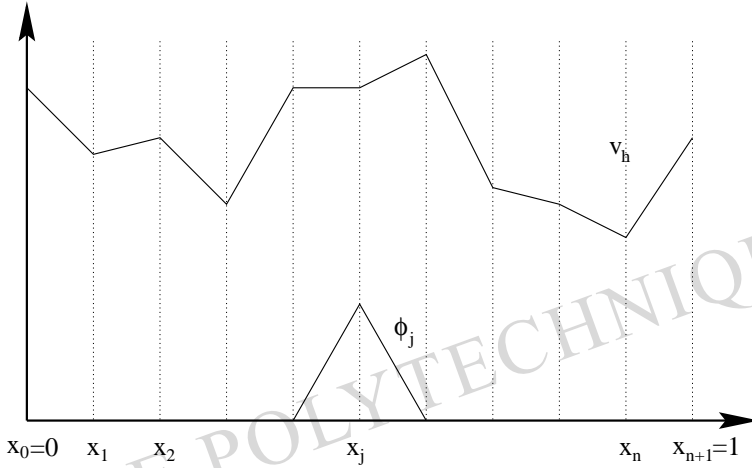


FIGURE 6.1 – Maillage de  $\Omega = ]0, 1[$  et fonction de base en éléments finis  $\mathbb{P}_1$ .

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”  $\phi$  définie par

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

Si le maillage est uniforme, pour  $0 \leq j \leq n+1$  on définit les fonctions de base (voir la Figure 6.1)

$$\phi_j(x) = \phi\left(\frac{x - x_j}{h}\right). \quad (6.10)$$

**Lemme 6.2.1** *L'espace  $V_h$ , défini par (6.8), est un sous-espace de  $H^1(0, 1)$  de dimension  $n+2$ , et toute fonction  $v_h \in V_h$  est définie de manière unique par ses valeurs aux sommets  $(x_j)_{0 \leq j \leq n+1}$*

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

De même,  $V_{0h}$ , défini par (6.9), est un sous-espace de  $H_0^1(0,1)$  de dimension  $n$ , et toute fonction  $v_h \in V_{0h}$  est définie de manière unique par ses valeurs aux sommets  $(x_j)_{1 \leq j \leq n}$

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

**Démonstration.** Rappelons qu'en vertu du Lemme 4.3.19, les fonctions continues et de classe  $C^1$  par morceaux appartiennent à  $H^1(\Omega)$ . Donc  $V_h$  et  $V_{0h}$  sont bien des sous-espaces de  $H^1(0,1)$ . Le reste de la preuve est immédiat en remarquant que  $\phi_j(x_i) = \delta_{ij}$ , où  $\delta_{ij}$  est le symbole de Kronecker qui vaut 1 si  $i = j$  et 0 sinon (voir la Figure 6.1).  $\square$

**Remarque 6.2.2** La base  $(\phi_j)$ , définie par (6.10), permet de caractériser une fonction de  $V_h$  par ses valeurs aux nœuds du maillage. Dans ce cas on parle **d'éléments finis de Lagrange**. Nous verrons plus loin à la Sous-section 6.2.5 qu'on peut introduire d'autres espaces  $V_h$  pour lesquels une fonction sera caractérisée, non seulement par ses valeurs, mais aussi par les valeurs de sa dérivée. On parle alors **d'éléments finis de Hermite**. Ici, comme les fonctions sont localement  $\mathbb{P}_1$ , on dit que l'espace  $V_h$ , défini par (6.8), 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 (6.7) (en fait la dérivée seconde d'une fonction de  $V_h$  est une somme de masses de Dirac aux nœuds du maillage!). Au contraire, il est parfaitement légitime d'utiliser des fonctions de  $V_h$  dans la formulation variationnelle (6.2) qui ne requiert qu'une seule dérivée.  $\bullet$

Décrivons la **résolution pratique** du problème de Dirichlet (6.7) par la méthode des éléments finis  $\mathbb{P}_1$ . La formulation variationnelle (6.2) de l'approximation interne devient ici :

$$\text{trouver } u_h \in V_{0h} \text{ tel que } \int_0^1 u_h'(x) v_h'(x) dx = \int_0^1 f(x) v_h(x) dx \quad \forall v_h \in V_{0h}. \quad (6.11)$$

On décompose  $u_h$  sur la base des  $(\phi_j)_{1 \leq j \leq n}$  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  $U_h = (u_h(x_j))_{1 \leq j \leq n}$ ,  $b_h = \left( \int_0^1 f(x) \phi_i(x) dx \right)_{1 \leq i \leq n}$ , et en introduisant la **matrice de rigidité**

$$K_h = \left( \int_0^1 \phi_j'(x) \phi_i'(x) dx \right)_{1 \leq i, j \leq n},$$

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  $\phi_j$  ont un “petit” support, l’intersection des supports de  $\phi_j$  et  $\phi_i$  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

$$\mathcal{K}_h = h^{-1} \begin{pmatrix} 2 & -1 & & 0 \\ -1 & 2 & -1 & \\ & \ddots & \ddots & \ddots \\ 0 & & -1 & 2 & -1 \\ & 0 & & -1 & 2 \end{pmatrix}. \quad (6.12)$$

Pour obtenir le second membre  $b_h$  il faut calculer les intégrales

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

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”

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

ou bien encore la formule de Simpson

$$\frac{1}{x_{i+1} - x_i} \int_{x_i}^{x_{i+1}} \psi(x) dx \approx \left( \frac{1}{6} \psi(x_{i+1}) + \frac{1}{6} \psi(x_i) + \frac{2}{3} \psi\left(\frac{x_{i+1} + x_i}{2}\right) \right).$$

Les deux premières formules sont exactes pour les fonctions  $\psi$  affines, et la troisième est exacte pour les polynômes du troisième degré (ce qui donne un calcul exact de  $b_h$  si

$f \in V_h$ ). Si la fonction  $\psi$  est régulière quelconque, alors ces formules sont simplement approchées avec un reste de l'ordre de  $\mathcal{O}(h^2)$ , de  $\mathcal{O}(h^2)$ , et de  $\mathcal{O}(h^4)$  respectivement.

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 en annexe dans la Section 13.1 des méthodes performantes de résolution. Rappelons que la matrice  $\mathcal{K}_h$  est nécessairement inversible par application du Lemme 6.1.1.

**Remarque 6.2.3** 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 (2.14) (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 6.2.3 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

$$\begin{cases} -u'' + au = f \text{ dans } ]0, 1[ \\ u'(0) = \alpha, u'(1) = \beta. \end{cases} \quad (6.13)$$

Rappelons que (6.13) admet une solution unique dans  $H^1(\Omega)$  si  $f \in L^2(\Omega)$ ,  $\alpha, \beta \in \mathbb{R}$ , et  $a \in L^\infty(\Omega)$  tel que  $a(x) \geq a_0 > 0$  p.p. dans  $\Omega$  (voir le Chapitre 5). La formulation variationnelle (6.2) 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

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

et

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

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 6.2.1** Appliquer la méthode des éléments finis  $\mathbb{P}_1$  au problème

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

Vérifier que les conditions aux limites de Dirichlet non-homogènes apparaissent dans le second membre du système linéaire qui en découle.

**Exercice 6.2.2** On reprend le problème de Neumann (6.13) en supposant que la fonction  $a(x) = 0$  dans  $\Omega$ . 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

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

Comparer ce résultat avec le Théorème 5.2.18.

**Exercice 6.2.3** Appliquer la méthode des différences finies (voir le Chapitre 2) au problème de Dirichlet (6.7). 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 (6.13).

**Exercice 6.2.4** On considère  $(n+2)$  masses ponctuelles (alignées) situées aux points  $x_j = j/(n+1)$  pour  $0 \leq j \leq n+1$  et reliées entre voisines par des ressorts de même raideur  $k > 0$ . On applique à chaque masse ponctuelle une force longitudinale  $f_j$ . Dans l'hypothèse de petits déplacements (longitudinaux) écrire l'énergie totale du système qu'il faut minimiser (on discutera le cas des extrémités libres ou fixées). Interpréter la recherche de la position d'équilibre du système en termes d'éléments finis.

## 6.2.2 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 dans la Sous-section 6.1.2. Nous définissons tout d'abord un **opérateur d'interpolation**  $r_h$  (comme dans le Lemme 6.1.3).

**Définition 6.2.4** On appelle opérateur d'interpolation  $\mathbb{P}_1$  l'application linéaire  $r_h$  de  $H^1(0, 1)$  dans  $V_h$  définie, pour tout  $v \in H^1(0, 1)$ , par

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

Cette définition a bien un sens car, en vertu du Lemme 4.3.3, les fonctions de  $H^1(0,1)$  sont continues et leurs valeurs ponctuelles sont donc bien définies. 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 6.2). Remarquons qu'en une dimension d'espace l'interpolée est définie pour toute fonction de  $H^1(0,1)$ , et non pas seulement pour les fonctions régulières de  $H^1(0,1)$  (ce qui sera le cas en dimension supérieure).

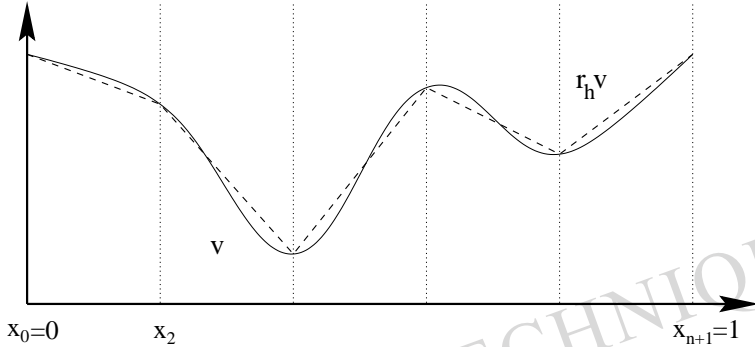


FIGURE 6.2 – Interpolation  $\mathbb{P}_1$  d'une fonction de  $H^1(0,1)$ .

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

**Lemme 6.2.5 (d'interpolation)** *Soit  $r_h$  l'opérateur d'interpolation  $\mathbb{P}_1$ . Pour tout  $v \in H^1(0,1)$ , il vérifie*

$$\lim_{h \rightarrow 0} \|v - r_h v\|_{H^1(0,1)} = 0.$$

*De plus, si  $v \in H^2(0,1)$ , alors il existe une constante  $C$  indépendante de  $h$  telle que*

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

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.

**Théorème 6.2.6** *Soit  $u \in H_0^1(0,1)$  et  $u_h \in V_{0h}$  les solutions de (6.7) et (6.11), respectivement. Alors, la méthode des éléments finis  $\mathbb{P}_1$  converge, c'est-à-dire que*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(0,1)} = 0. \quad (6.14)$$

*De plus, si  $u \in H^2(0,1)$  (ce qui est vrai si  $f \in L^2(0,1)$ ), alors il existe une constante  $C$  indépendante de  $h$  telle que*

$$\|u - u_h\|_{H^1(0,1)} \leq Ch \|u''\|_{L^2(0,1)} = Ch \|f\|_{L^2(0,1)}. \quad (6.15)$$



**Remarque 6.2.7** La première conclusion (6.14) du Théorème 6.2.6 est vrai aussi si le second membre  $f$  appartient seulement à  $H^{-1}(0, 1)$ . L'estimation (6.15) indique la vitesse de convergence de la méthode des éléments finis  $\mathbb{P}_1$ . Comme cette majoration est proportionnelle à  $h$ , on dit que la méthode des éléments finis  $\mathbb{P}_1$  converge linéairement.

Remarquons que le Théorème 6.2.6 de convergence est valable lorsque la matrice de rigidité  $\mathcal{K}_h$  et le second membre  $b_h$  sont évalués exactement. Cependant, la méthode des éléments finis  $\mathbb{P}_1$  converge aussi lorsqu'on utilise des formules de quadrature adéquates pour calculer  $\mathcal{K}_h$  et  $b_h$  (voir [36]). •

**Remarque 6.2.8** Nous démontrons le Théorème 6.2.6 dans le cas d'un maillage uniforme, c'est-à-dire de points  $x_j$  équidistants dans le segment  $[0, 1]$  (autrement dit  $x_{j+1} - x_j = h$ ). Le résultat peut néanmoins se généraliser à des maillages non uniforme mais réguliers (au sens de la Définition 6.3.11), et dans ce cas  $h$  est la distance maximum entre deux points :  $h = \max_{0 \leq j \leq n} (x_{j+1} - x_j)$ . •

**Remarque 6.2.9** 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 2.2.20, 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 6.2.5, 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.** Le Lemme 6.2.5 permet d'appliquer le résultat de convergence du Lemme 6.1.3 qui entraîne immédiatement (6.14). Pour obtenir (6.15), on majore l'estimation du Lemme 6.1.2 de Céa

$$\|u - u_h\|_{H^1(0,1)} \leq C \inf_{v_h \in V_h} \|u - v_h\|_{H^1(0,1)} \leq C \|u - r_h u\|_{H^1(0,1)},$$

ce qui permet de conclure grâce au Lemme 6.2.5.  $\square$

Nous donnons maintenant la démonstration du Lemme 6.2.5 sous la forme de deux autres lemmes techniques.

**Lemme 6.2.10** *Il existe une constante  $C$  indépendante de  $h$  telle que, pour tout  $v \in H^2(0, 1)$ ,*

$$\|v - r_h v\|_{L^2(0,1)} \leq C h^2 \|v''\|_{L^2(0,1)}, \quad (6.16)$$

et

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq C h \|v''\|_{L^2(0,1)}. \quad (6.17)$$

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

$$\begin{aligned}
 v(x) - r_h v(x) &= v(x) - \left( v(x_j) + \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j} (x - x_j) \right) \\
 &= \int_{x_j}^x v'(t) dt - \frac{x - x_j}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} v'(t) dt \\
 &= (x - x_j) v'(x_j + \theta_x) - (x - x_j) v'(x_j + \theta_j) \\
 &= (x - x_j) \int_{x_j + \theta_j}^{x_j + \theta_x} v''(t) dt,
 \end{aligned} \tag{6.18}$$

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

$$|v(x) - r_h v(x)|^2 \leq h^2 \left( \int_{x_j}^{x_{j+1}} |v''(t)| dt \right)^2 \leq h^3 \int_{x_j}^{x_{j+1}} |v''(t)|^2 dt. \tag{6.19}$$

En intégrant (6.19) 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 (6.16). Par densité ce résultat est encore vrai pour tout  $v \in H^2(0, 1)$ . La démonstration de (6.17) est tout à fait similaire : pour  $v \in C^\infty([0, 1])$  et  $x \in ]x_j, x_{j+1}[$  on écrit

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

Élevant au carré cette inégalité, appliquant Cauchy-Schwarz deux fois et sommant en  $j$  on obtient (6.17), qui est aussi valide pour tout  $v \in H^2(0, 1)$  par densité.  $\square$

**Lemme 6.2.11** *Il existe une constante  $C$  indépendante de  $h$  telle que, pour tout  $v \in H^1(0, 1)$ ,*

$$\|r_h v\|_{H^1(0,1)} \leq C \|v\|_{H^1(0,1)}, \tag{6.20}$$

et

$$\|v - r_h v\|_{L^2(0,1)} \leq Ch \|v'\|_{L^2(0,1)}. \tag{6.21}$$

De plus, pour tout  $v \in H^1(0, 1)$ , on a

$$\lim_{h \rightarrow 0} \|v' - (r_h v)'\|_{L^2(0,1)} = 0. \tag{6.22}$$

**Démonstration.** Les preuves de (6.20) et (6.21) sont dans le même esprit que celles du lemme précédent. Soit  $v \in H^1(0, 1)$ . Tout d'abord on a

$$\|r_h v\|_{L^2(0,1)} \leq \max_{x \in [0,1]} |r_h v(x)| \leq \max_{x \in [0,1]} |v(x)| \leq C \|v\|_{H^1(0,1)},$$

en vertu du Lemme 4.3.3. D'autre part, comme  $r_h v$  est affine, et grâce à la propriété (4.8) du Lemme 4.3.3 qui affirme que  $v$  est bien la primitive de  $v'$ , on a

$$\begin{aligned} \int_{x_j}^{x_{j+1}} |(r_h v)'(x)|^2 dx &= \frac{(v(x_{j+1}) - v(x_j))^2}{h} \\ &= \frac{1}{h} \left( \int_{x_j}^{x_{j+1}} v'(x) dx \right)^2 \leq \int_{x_j}^{x_{j+1}} |v'(x)|^2 dx, \end{aligned}$$

par Cauchy-Schwarz, ce qui, par sommation en  $j$ , conduit à (6.20). Pour obtenir (6.21) on reprend la deuxième égalité de (6.18) d'où l'on déduit

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

En élevant au carré, en utilisant Cauchy-Schwarz, en intégrant par rapport à  $x$ , puis en sommant en  $j$ , on obtient bien (6.21).

Passons à la démonstration de (6.22). Soit  $\epsilon > 0$ . Comme  $C^\infty([0, 1])$  est dense dans  $H^1(0, 1)$ , pour tout  $v \in H^1(0, 1)$  il existe  $\phi \in C^\infty([0, 1])$  tel que

$$\|v' - \phi'\|_{L^2(0,1)} \leq \epsilon.$$

Or  $r_h$  est une application linéaire qui vérifie (6.20), donc on en déduit

$$\|(r_h v)' - (r_h \phi)'\|_{L^2(0,1)} \leq C \|v' - \phi'\|_{L^2(0,1)} \leq C\epsilon.$$

Le choix de  $\phi$  et de  $\epsilon$  étant fixé, on déduit de (6.17) appliqué à  $\phi$  que, pour  $h$  suffisamment petit,

$$\|\phi' - (r_h \phi)'\|_{L^2(0,1)} \leq \epsilon.$$

Par conséquent, en sommant ces trois dernières inégalités on obtient

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leq \|v' - \phi'\|_{L^2} + \|\phi' - (r_h \phi)'\|_{L^2} + \|(r_h v)' - (r_h \phi)'\|_{L^2} \leq C\epsilon,$$

ce qui implique (6.22).  $\square$

**Exercice 6.2.5** Démontrer l'équivalent du Théorème 6.2.6 de convergence pour le problème de Neumann (6.13).

### 6.2.3 Éléments finis $\mathbb{P}_2$

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

$$V_h = \{v \in C([0, 1]) \text{ tel que } v|_{[x_j, x_{j+1}]} \in \mathbb{P}_2 \text{ pour tout } 0 \leq j \leq n\}, \quad (6.23)$$

et sur son sous-espace

$$V_{0h} = \{v \in V_h \text{ tel que } v(0) = v(1) = 0\}. \quad (6.24)$$

La méthode des éléments finis  $\mathbb{P}_2$  est la méthode d'approximation variationnelle interne de la Sous-section 6.1.2 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.

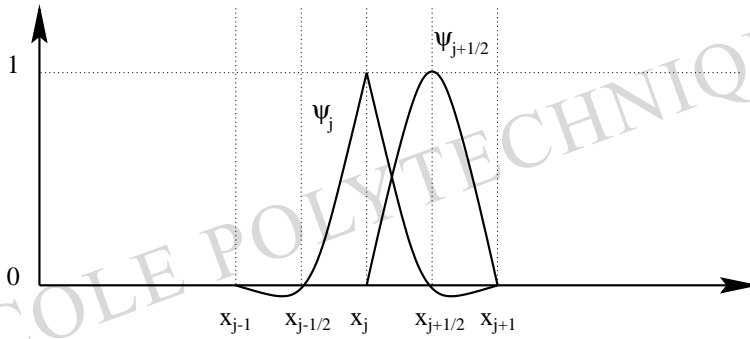


FIGURE 6.3 – Les fonctions de base des éléments finis  $\mathbb{P}_2$ .

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  $0 \leq j \leq n$ . On définit aussi deux fonctions "mères"

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

et

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

Si le maillage est uniforme, pour  $0 \leq j \leq n+1$  on définit les fonctions de base (voir la Figure 6.3)

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

**Lemme 6.2.12** *L'espace  $V_h$ , défini par (6.23), est un sous-espace de  $H^1(0,1)$  de dimension  $2n+3$ , et toute fonction  $v_h \in V_h$  est définie de manière unique par ses valeurs aux sommets  $(x_j)_{0 \leq j \leq n+1}$  et aux milieux  $(x_{j+1/2})_{0 \leq j \leq n}$*

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

*De même,  $V_{0h}$ , défini par (6.24), est un sous-espace de  $H_0^1(0,1)$  de dimension  $2n+1$ , et toute fonction  $v_h \in V_{0h}$  est définie de manière unique par ses valeurs aux sommets  $(x_j)_{1 \leq j \leq n}$  et aux milieux  $(x_{j+1/2})_{0 \leq j \leq n}$*

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

**Remarque 6.2.13** Ici encore,  $V_h$  est un espace d'éléments finis de Lagrange (cf. la Remarque 6.2.2). Comme les fonctions sont localement  $\mathbb{P}_2$ , on dit que l'espace  $V_h$ , défini par (6.23), est l'espace des éléments finis de Lagrange d'ordre 2. •

**Démonstration.** Par application du Lemme 4.3.19  $V_h$  et  $V_{0h}$  sont bien des sous-espaces de  $H^1(0,1)$ . Leur dimension et les bases proposées 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 6.3). □

Décrivons la **résolution pratique** du problème de Dirichlet (6.7) par la méthode des éléments finis  $\mathbb{P}_2$ . La formulation variationnelle (6.2) de l'approximation interne revient à résoudre dans  $\mathbb{R}^{2n+1}$  le système linéaire

$$\mathcal{K}_h U_h = b_h. \quad (6.25)$$

Pour expliciter ce système linéaire il est commode de changer d'indice en notant désormais les points  $(x_{1/2}, x_1, x_{3/2}, x_2, \dots, x_{n+1/2})$  sous la forme  $(x_{k/2})_{1 \leq k \leq 2n+1}$ , et la base  $(\psi_{1/2}, \psi_1, \psi_{3/2}, \psi_2, \dots, \psi_{n+1/2})$  de  $V_{0h}$  sous la forme  $(\psi_{k/2})_{1 \leq k \leq 2n+1}$ . Dans cette base  $U_h \in \mathbb{R}^{2n+1}$  est le vecteur des coordonnées de la solution approchée  $u_h$  qui vérifie

$$u_h(x) = \sum_{k=1}^{2n+1} (U_h)_{k/2} \psi_{k/2}(x) \text{ avec } (U_h)_{k/2} = u_h(x_{k/2}), \quad (6.26)$$

et on a

$$\mathcal{K}_h = \left( \int_0^1 \psi'_{k/2}(x) \psi'_{l/2}(x) dx \right)_{1 \leq k, l \leq 2n+1}, \quad b_h = \left( \int_0^1 f(x) \psi_{k/2}(x) dx \right)_{1 \leq k \leq 2n+1}.$$

Les fonctions de base  $\psi_{k/2}$  ont un "petit" support, et la plupart des coefficients de  $\mathcal{K}_h$  sont donc nuls. Un calcul simple montre que la matrice de rigidité  $\mathcal{K}_h$  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 mêmes formules de quadrature (ou formules d’intégration numérique) que celles présentées dans la méthode  $\mathbb{P}_1$ .

**Théorème 6.2.14** *Soit  $u \in H_0^1(0,1)$  et  $u_h \in V_{0h}$  les solutions de (6.7) et (6.25)-(6.26), respectivement. Alors, la méthode des éléments finis  $\mathbb{P}_2$  converge, c’est-à-dire que*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(0,1)} = 0.$$

*De plus, si  $u \in H^3(0,1)$  (ce qui est vrai si  $f \in H^1(0,1)$ ), alors il existe une constante  $C$  indépendante de  $h$  telle que*

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

**Exercice 6.2.6** En généralisant les arguments précédents, démontrer le Théorème 6.2.14.

Le Théorème 6.2.14 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 n’est pas régulière ( $u \in H^3(0,1)$ ) il n’y a aucun avantage théorique (mais aussi pratique) à utiliser des éléments finis  $\mathbb{P}_2$  plutôt que  $\mathbb{P}_1$ .

## 6.2.4 Propriétés qualitatives

Nous savons que la solution d’un problème de Dirichlet vérifie le principe du maximum (voir le Théorème 5.2.22). Il est important de savoir si cette propriété est conservée par l’approximation variationnelle interne.

**Proposition 6.2.15 (principe du maximum discret)** *On suppose que  $f \geq 0$  presque partout dans  $]0, 1[$ . Alors, la solution  $u_h$  de l'approximation variationnelle (6.11) par la méthode des éléments finis  $\mathbb{P}_1$  vérifie  $u_h \geq 0$  dans  $[0, 1]$ .*

**Démonstration.** Soit  $u_h$  la solution de (6.11). En vertu du Lemme 6.2.1, on a

$$u_h(x) = \sum_{j=1}^n u_h(x_j) \phi_j(x),$$

où les fonctions  $\phi_j$  sont les fonctions de base des éléments finis  $\mathbb{P}_1$  dans  $V_{0h}$  et  $U_h = (u_h(x_j))_{1 \leq j \leq n}$  est solution du système linéaire

$$\mathcal{K}_h U_h = b_h. \quad (6.27)$$

Les fonctions  $\phi_j$  sont des fonctions “chapeaux” (voir la Figure 6.1) qui sont positives : il suffit donc de montrer que toutes les composantes du vecteur  $U_h = (U_h^j)_{1 \leq j \leq n}$  sont positives pour prouver que la fonction  $u_h$  est positive sur  $[0, 1]$ . Rappelons que, en posant  $U_h^0 = U_h^{n+1} = 0$ , le système linéaire (6.27) est équivalent à

$$-U_h^{j-1} + 2U_h^j - U_h^{j+1} = h b_h^j \text{ pour tout } 1 \leq j \leq n. \quad (6.28)$$

Soit  $U_h^{j_0} = \min_j U_h^j$  la plus petite composante de  $U_h$  : s'il y a plusieurs plus petites composantes, on choisit celle de plus petit indice  $j_0$ . Si  $j_0 = 0$ , alors  $U_h^j \geq U_h^0 = 0$  pour tout  $j$ , ce qui est le résultat recherché. Si  $j_0 \geq 1$ , alors  $U_h^{j_0} < U_h^0 = 0$ , et comme  $U_h^{n+1} = 0$  on en déduit que  $j_0 \leq n$ . Comme  $b_h^j = \int_0^1 f \psi_j dx \geq 0$  par hypothèse sur  $f$ , on peut alors déduire de la relation (6.28) pour  $j_0$  que

$$\left( U_h^{j_0} - U_h^{j_0-1} \right) + \left( U_h^{j_0} - U_h^{j_0+1} \right) \geq 0,$$

ce qui est une contradiction avec le caractère minimal (strict) de  $U_h^{j_0}$ . Par conséquent la méthode des éléments finis  $\mathbb{P}_1$  vérifie le principe du maximum discret.  $\square$

Nous avons démontré dans les sous-sections précédentes des résultats théoriques de convergence. Nous pouvons **vérifier numériquement les vitesses de convergence** prédites en résolvant le problème de Dirichlet (6.7) par la méthode des éléments finis avec des maillages de tailles distinctes. Considérons l'exemple suivant

$$\begin{cases} -((1+x)u')' + (1 + \cos(\pi x))u = f & \text{pour } 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases} \quad (6.29)$$

avec  $f(x) = -\pi \cos(\pi x) + \sin(\pi x)(1 + \cos(\pi x) + \pi^2(1+x))$ , dont la solution exacte est  $u(x) = \sin(\pi x)$ . L'idéal serait de calculer l'erreur exacte  $\|u - u_h\|_{H^1(0,1)}$ , mais cela nécessite de faire des calculs précis d'intégrales, ce qui n'est pas commode si la solution  $u$  est compliquée. En pratique (et c'est ce que nous faisons ici) on se contente de calculer l'erreur projetée dans  $V_h$ , c'est-à-dire qu'on calcule  $\|r_h(u - u_h)\|_{H^1(0,1)}$  (on

dit aussi qu'il s'agit de la **norme discrète** dans  $V_h$ ). L'intérêt de cette approche est que l'on peut calculer exactement les intégrales puisque  $r_h(u - u_h) = r_h u - u_h \in V_h$  (cela revient à ne pas tenir compte des erreurs d'interpolation entre  $H^1(0, 1)$  et  $V_h$ ). On trace cette erreur discrète  $\|r_h(u - u_h)\|_{H^1(0,1)}$  en fonction du pas du maillage  $h$ . Lorsque la solution est régulière, le Théorème 6.2.6 prévoit une convergence linéaire (en  $h$ ) de l'erreur par la méthode des éléments finis  $\mathbb{P}_1$ , tandis que le Théorème 6.2.14 prévoit une convergence quadratique (en  $h^2$ ) de l'erreur par la méthode des éléments finis  $\mathbb{P}_2$ . Dans le cas de l'exemple (6.29) on trace cette erreur pour différentes valeurs de  $h$  sur la Figure 6.4 (en échelle logarithmique). Les croix ou les ronds correspondent à des résultats de calcul, les lignes sont des droites de référence correspondant aux fonctions  $h^2$  et  $h^3$  respectivement. Remarquons que l'utilisation d'une échelle logarithmique permet de bien visualiser les vitesses de convergence comme la pente du logarithme de l'erreur en fonction du logarithme de  $h$ . On observe donc un phénomène de **super-convergence**, c'est-à-dire que les éléments finis convergent plus rapidement que ce qui est prévu par la théorie : l'erreur est en  $h^2$  pour la méthode  $\mathbb{P}_1$  et  $h^3$  pour celle  $\mathbb{P}_2$ . Ce gain est dû à l'uniformité du maillage et au choix de la norme discrète dans  $V_h$ .

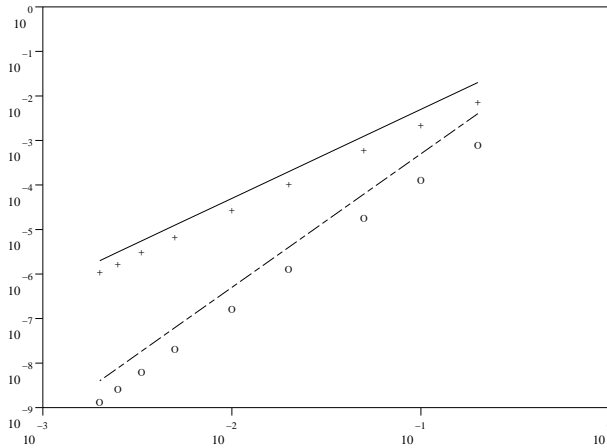


FIGURE 6.4 – Cas d'une solution régulière : exemple (6.29). Norme discrète  $H^1$  de l'erreur en fonction du pas  $h$  du maillage (les croix correspondent aux éléments finis  $\mathbb{P}_1$ , les ronds aux éléments finis  $\mathbb{P}_2$ , les droites sont les tracés de  $h \rightarrow h^2$  et  $h \rightarrow h^3$ ).

Si la solution n'est pas régulière, il y a toujours convergence mais avec une vitesse plus faible que ce qui est prédit dans le cas régulier par les Théorèmes 6.2.6 et 6.2.14. Pour obtenir une solution non régulière, on prend un second membre dans  $H^{-1}(0, 1)$  qui n'appartient pas à  $L^2(0, 1)$ . En une dimension d'espace on peut ainsi prendre une



masse de Dirac. Considérons donc l'exemple

$$\begin{cases} -u'' = 6x - 2 + \delta_{1/2} & \text{pour } 0 < x < 1 \\ u(0) = u(1) = 0 \end{cases} \quad (6.30)$$

avec  $\delta_{1/2}$  la masse de Dirac au point  $x = 1/2$ , dont la solution exacte est  $u(x) = 1/2 - |x - 1/2| + x^2(1 - x)$ .

On trace l'erreur  $\|r_h(u - u_h)\|_{H^1(0,1)}$  en fonction de  $h$  sur la Figure 6.5 (en échelle logarithmique). Les croix ou les ronds correspondent à des résultats de calcul ( $\mathbb{P}_1$  ou  $\mathbb{P}_2$  respectivement), les lignes sont des droites de référence correspondant aux fonctions  $\sqrt{h}$  et  $h$  respectivement. On voit que les éléments finis  $\mathbb{P}_1$  et  $\mathbb{P}_2$  convergent à la même vitesse proportionnelle à  $\sqrt{h}$ , ce qui est bien inférieur à la vitesse de  $h$  (ou même  $h^2$ ) prédite dans le cas régulier. En particulier, il n'y a aucun intérêt à utiliser les éléments finis  $\mathbb{P}_2$ , plutôt que  $\mathbb{P}_1$ , dans un tel cas.

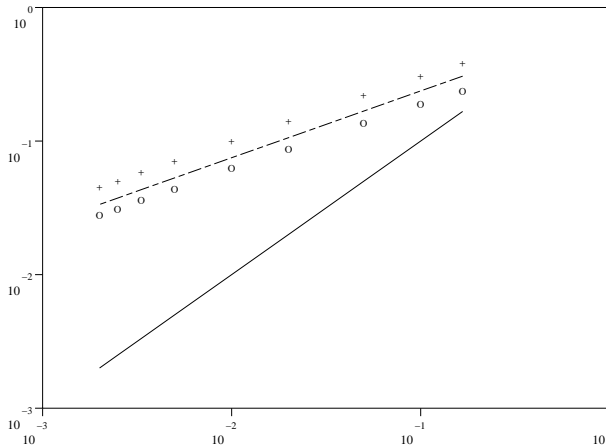


FIGURE 6.5 – Cas d’une solution non régulière : exemple (6.30). Norme discrète  $H^1$  de l’erreur en fonction du pas  $h$  du maillage (les croix correspondent aux éléments finis  $\mathbb{P}_1$ , les ronds aux éléments finis  $\mathbb{P}_2$ , les droites sont les tracés de  $h \rightarrow \sqrt{h}$  et  $h \rightarrow h$ ).

Pour calculer l’erreur dans les Figures 6.4 et 6.5 nous avons utilisé une solution exacte. Cependant, si celle-ci n’est pas connue, on peut la remplacer par la solution approchée obtenue avec le maillage le plus fin (supposée être la plus convergée). Cette procédure de **convergence numérique** peut aussi être mise en oeuvre pour d’autres méthodes numériques, y compris (et surtout) lorsqu’on ne dispose d’aucun théorème de convergence. C’est souvent le seul moyen “heuristique” de vérifier si un algorithme converge et à quelle vitesse par rapport au raffinement du maillage.

### 6.2.5 Éléments finis d'Hermite

Après avoir défini des éléments finis  $\mathbb{P}_1$  et  $\mathbb{P}_2$ , le lecteur imagine facilement comment généraliser et définir des éléments finis  $\mathbb{P}_k$  avec  $k \in \mathbb{N}^*$ . Ces éléments finis, dits de Lagrange, utilisent des fonctions de base qui sont seulement continues mais pas continûment dérivables. Toutefois, il est clair que des polynômes de  $\mathbb{P}_3$  peuvent se raccorder de manière continûment dérivables. Dans ce cas les valeurs des dérivées seront aussi utilisées pour caractériser les fonctions (voir la Remarque 6.2.2). On introduit donc une méthode des **éléments finis de Hermite**  $\mathbb{P}_3$  qui repose sur l'espace discret

$$V_h = \{v \in C^1([0, 1]) \text{ tel que } v|_{[x_j, x_{j+1}]} \in \mathbb{P}_3 \text{ pour tout } 0 \leq j \leq n\}. \quad (6.31)$$

Il faut bien faire attention que dans la définition (6.31) de  $V_h$  on demande aux fonctions d'appartenir à  $C^1([0, 1])$ , et non plus seulement à  $C([0, 1])$ . C'est ce qui fait la différence entre les éléments finis de Hermite et de Lagrange, respectivement.

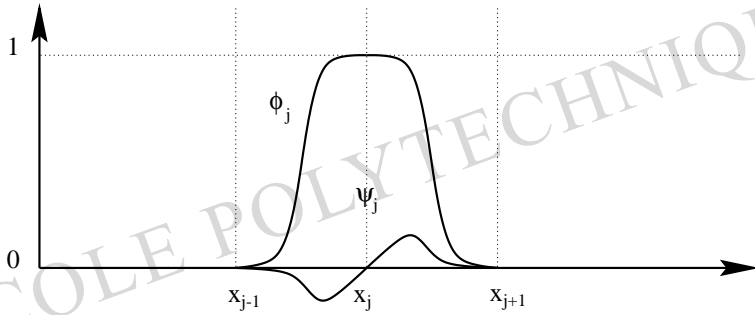


FIGURE 6.6 – Les fonctions de base des éléments finis d'Hermite  $\mathbb{P}_3$ .

On peut représenter les fonctions de  $V_h$  à l'aide de fonctions de base très simples. On définit deux fonctions “mères”

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

et

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

Si le maillage est uniforme, pour  $0 \leq j \leq n+1$  on définit les fonctions de base (voir la Figure 6.6)

$$\phi_j(x) = \phi\left(\frac{x-x_j}{h}\right) \text{ pour } 0 \leq j \leq n+1, \quad \psi_j(x) = h\psi\left(\frac{x-x_j}{h}\right) \text{ pour } 0 \leq j \leq n+1.$$

**Lemme 6.2.16** *L'espace  $V_h$ , défini par (6.31), est un sous-espace de  $H^1(0,1)$  de dimension  $2(n+2)$ . Toute fonction  $v_h$  de  $V_h$  est définie de manière unique par ses valeurs et celles de sa dérivée aux sommets  $(x_j)_{0 \leq j \leq n+1}$ , et on a pour tout  $x \in [0,1]$*

$$v_h(x) = \sum_{j=0}^{n+1} v_h(x_j) \phi_j(x) + \sum_{j=0}^{n+1} (v_h)'(x_j) \psi_j(x). \quad (6.32)$$

**Démonstration.** Les fonctions de  $V_h$  étant de classe  $C^1$ , il s'agit bien d'un sous-espace de  $H^1(0,1)$ . On vérifie facilement que les  $(\phi_j, \psi_j)$  forment une base de  $V_h$  en remarquant que  $\phi_j(x_i) = \delta_{ij}$ ,  $\psi_j(x_i) = 0$ ,  $\phi_j'(x_i) = 0$ ,  $\psi_j'(x_i) = \delta_{ij}$  (voir la Figure 6.6).  $\square$

On peut bien sûr utiliser l'espace  $V_h$  (ou du moins son sous-espace des fonctions qui s'annulent en 0 et 1) pour résoudre le problème de Dirichlet (6.7), mais ce n'est pas l'utilisation la plus courante de  $V_h$ . En pratique on utilise  $V_h$  pour **résoudre l'équation des plaques** (voir (5.70) et le Chapitre 1), ou plutôt des poutres en dimension  $N = 1$ ,

$$\begin{cases} u'''' = f & \text{dans } ]0,1[ \\ u(0) = u(1) = u'(0) = u'(1) = 0, \end{cases} \quad (6.33)$$

(qui admet une unique solution  $u \in H_0^2(0,1)$  si  $f \in L^2(0,1)$ ). En effet,  $V_h$  n'est pas seulement un sous-espace de  $H^1(0,1)$ , mais est aussi un sous-espace de  $H^2(0,1)$  (ce qui n'est pas le cas pour les éléments finis de Lagrange). Pour résoudre (6.33) nous aurons besoin du sous-espace

$$V_{0h} = \{v \in V_h \text{ tel que } v(0) = v(1) = v'(0) = v'(1) = 0\}. \quad (6.34)$$

**Lemme 6.2.17** *L'espace  $V_h$ , et son sous-espace  $V_{0h}$  défini par (6.34), sont des sous-espaces de  $H^2(0,1)$ , et de  $H_0^2(0,1)$  respectivement, de dimensions  $2(n+2)$ , et  $2n$  respectivement. Toute fonction  $v_h$  de  $V_{0h}$  est définie de manière unique par ses valeurs et celles de sa dérivée aux sommets  $(x_j)_{1 \leq j \leq n}$ , et on a pour tout  $x \in [0,1]$*

$$v_h(x) = \sum_{j=1}^n v_h(x_j) \phi_j(x) + \sum_{j=1}^n (v_h)'(x_j) \psi_j(x).$$

**Démonstration.** Soit  $v_h \in V_h$  : elle est de classe  $C^1$  sur  $[0,1]$  et  $C^2$  par morceaux. Donc sa dérivée  $v_h'$ , étant continue et  $C^1$  par morceaux, appartient bien à  $H^1(0,1)$  (en vertu du Lemme 4.3.19). Par conséquent,  $v_h$  est un élément de  $H^2(0,1)$ . Le reste du lemme est semblable au Lemme 6.2.16  $\square$

Décrivons brièvement la résolution pratique de l'équation des plaques (6.33) par la méthode des éléments finis d'Hermite  $\mathbb{P}_3$ . La formulation variationnelle de l'approximation interne est

$$\text{trouver } u_h \in V_{0h} \text{ tel que } \int_0^1 u_h''(x) v_h''(x) dx = \int_0^1 f(x) v_h(x) dx \quad \forall v_h \in V_{0h}. \quad (6.35)$$

On décompose  $u_h$  sur la base des  $(\phi_j, \psi_j)_{1 \leq j \leq n}$  et on note  $U_h = (u_h(x_j), u'_h(x_j))_{1 \leq j \leq n}$  le vecteur de ses coordonnées dans cette base. La formulation variationnelle (6.35) revient à résoudre dans  $\mathbb{R}^{2n}$  un système linéaire

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

**Exercice 6.2.7** Calculer explicitement la matrice de rigidité  $\mathcal{K}_h$  pour (6.35).

### 6.3 Éléments finis en dimension $N \geq 2$

Nous nous plaçons maintenant en dimension d'espace  $N \geq 2$  (en pratique  $N = 2, 3$ ). Pour simplifier l'exposé, certains résultats ne seront démontrés qu'en dimension  $N = 2$ , mais ils s'étendent à la dimension  $N = 3$  (au prix, parfois, de complications techniques et pratiques importantes).

Nous considérons le problème modèle de Dirichlet

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (6.36)$$

dont nous savons qu'il admet une solution unique dans  $H_0^1(\Omega)$ , si  $f \in L^2(\Omega)$  (voir le Chapitre 5).

Dans tout ce qui suit nous supposons que le domaine  $\Omega$  est **polyédrique** (polygonal si  $N = 2$ ), c'est-à-dire que  $\bar{\Omega}$  est une réunion finie de polyèdres de  $\mathbb{R}^N$ . Rappelons qu'un polyèdre est une intersection finie de demi-espaces de  $\mathbb{R}^N$  et que les parties de son bord qui appartiennent à un seul hyperplan sont appelées ses faces. La raison de cette hypothèse est qu'il n'est possible de mailler exactement que de tels ouverts. Nous dirons plus loin ce qui se passe pour des domaines généraux à bords "courbes" (voir la Remarque 6.3.18).

#### 6.3.1 Éléments finis triangulaires

Tout commence par la définition d'un maillage du domaine  $\Omega$  par des triangles en dimension  $N = 2$  et des tétraèdres en dimension  $N = 3$ . On regroupe les triangles et les tétraèdres dans la famille plus générale des  $N$ -simplexes. On appelle  $N$ -simplexe  $K$  de  $\mathbb{R}^N$  l'enveloppe convexe de  $(N+1)$  points  $(a_j)_{1 \leq j \leq N+1}$  de  $\mathbb{R}^N$ , appelés sommets de  $K$ . Bien sûr un 2-simplexe est simplement un triangle et un 3-simplexe un tétraèdre (voir la Figure 6.9). On dit que le  $N$ -simplexe  $K$  est non dégénéré si les points  $(a_j)_{1 \leq j \leq N+1}$  n'appartiennent pas à un même hyperplan de  $\mathbb{R}^N$  (le triangle ou le tétraèdre est non "plat"). Si on note  $(a_{i,j})_{1 \leq i \leq N}$  les coordonnées du vecteur  $a_j$ , la condition de non

dégénérescence de  $K$  est que la matrice

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,N+1} \\ a_{2,1} & a_{2,2} & \dots & a_{2,N+1} \\ \vdots & \vdots & & \vdots \\ a_{N,1} & a_{N,2} & \dots & a_{N,N+1} \\ 1 & 1 & \dots & 1 \end{pmatrix} \quad (6.37)$$

soit inversible (ce que l'on supposera toujours par la suite). Un  $N$ -simplexe a autant de faces que de sommets, qui sont elles-mêmes des  $(N-1)$ -simplexes.

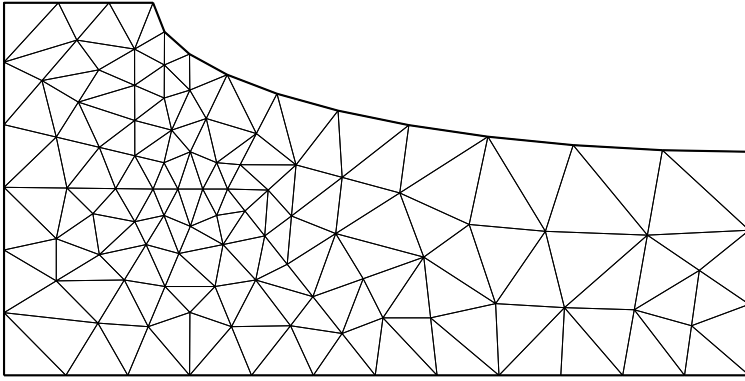


FIGURE 6.7 – Exemple de maillage triangulaire en dimension  $N = 2$ .

**Définition 6.3.1** Soit  $\Omega$  un ouvert connexe polyédrique de  $\mathbb{R}^N$ . Un **maillage triangulaire** ou une **triangulation** de  $\overline{\Omega}$  est un ensemble  $\mathcal{T}_h$  de  $N$ -simplexes (non dégénérés)  $(K_i)_{1 \leq i \leq n}$  qui vérifient

1.  $K_i \subset \overline{\Omega}$  et  $\overline{\Omega} = \cup_{i=1}^n K_i$ ,
2. l'intersection  $K_i \cap K_j$  de deux  $N$ -simplexes distincts est un  $m$ -simplexe, avec  $0 \leq m \leq N-1$ , dont tous les sommets sont aussi des sommets de  $K_i$  et  $K_j$ . (En dimension  $N = 2$ , l'intersection de deux triangles est soit vide, soit réduite à un sommet commun, soit une arête commune **entière** ; en dimension  $N = 3$ , l'intersection de deux tétraèdres est soit vide, soit un sommet commun, soit une arête commune entière, soit une face commune entière.)

Les **sommets** ou **nœuds** du maillage  $\mathcal{T}_h$  sont les sommets des  $N$ -simplexes  $K_i$  qui le composent. Par convention, le paramètre  $h$  désigne le maximum des diamètres des  $N$ -simplexes  $K_i$ .

Il est clair que la Définition 6.3.1 ne peut s'appliquer qu'à un ouvert polyédrique et pas à un ouvert quelconque. La Définition 6.3.1 contient un certain nombre de

restrictions sur le maillage : dans ce cas on parle souvent de **maillage conforme**. Un exemple de maillage conforme est donné à la Figure 6.7, tandis que la Figure 6.8 présente des situations interdites par la Définition 6.3.1.

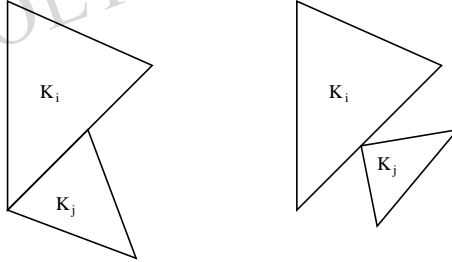


FIGURE 6.8 – Exemples de situations interdites pour un maillage triangulaire.

**Remarque 6.3.2** Nous ne disons rien ici des algorithmes qui permettent de construire un maillage triangulaire. Contentons nous de dire que, s'il est relativement facile de mailler des domaines plans (il existe de nombreux logiciels libres qui permettent de le faire), il est encore assez compliqué de mailler des domaines tridimensionnels. Nous renvoyons à l'ouvrage [24] le lecteur intéressé par ce sujet. •

**Exercice 6.3.1** Soit  $\mathcal{T}_h$  un maillage de  $\overline{\Omega}$  pour  $\Omega$  ouvert simplement connexe polygonal de  $\mathbb{R}^2$ . On note  $n_t$  le nombre de triangles de  $\mathcal{T}_h$ ,  $n_c$  le nombre de faces ou cotés des triangles (un coté commun à deux triangles n'est compté qu'une seule fois),  $n_s$  le nombre de sommets du maillage, et  $n_{0s}$  le nombre de sommets intérieurs du maillage (qui ne sont pas sur  $\partial\Omega$ ). Démontrer les relations, dites d'Euler,  $n_t + n_s = n_c + 1$  et  $3n_t + n_s = 2n_c + n_{0s}$ .

Dans un  $N$ -simplexe  $K$  il est commode d'utiliser des coordonnées barycentriques au lieu des coordonnées cartésiennes usuelles. Rappelons que, si  $K$  est un  $N$ -simplexe non dégénéré de sommets  $(a_j)_{1 \leq j \leq N+1}$ , les **coordonnées barycentriques**  $(\lambda_j)_{1 \leq j \leq N+1}$  de  $x \in \mathbb{R}^N$  sont définies par

$$\sum_{j=1}^{N+1} \lambda_j = 1, \quad \sum_{j=1}^{N+1} a_{i,j} \lambda_j = x_i \quad \text{pour } 1 \leq i \leq N, \quad (6.38)$$

qui admet bien une unique solution car la matrice  $A$ , définie par (6.37), est inversible. Remarquons que les  $\lambda_j$  sont des fonctions affines de  $x$ . On vérifie alors que

$$K = \{x \in \mathbb{R}^N \text{ tel que } \lambda_j(x) \geq 0 \text{ pour } 1 \leq j \leq N+1\},$$

et que les  $(N+1)$  faces de  $K$  sont les intersections de  $K$  et des hyperplans  $\lambda_j(x) = 0$ ,  $1 \leq j \leq N+1$ . On peut alors définir un ensemble de points de  $K$  qui vont jouer un

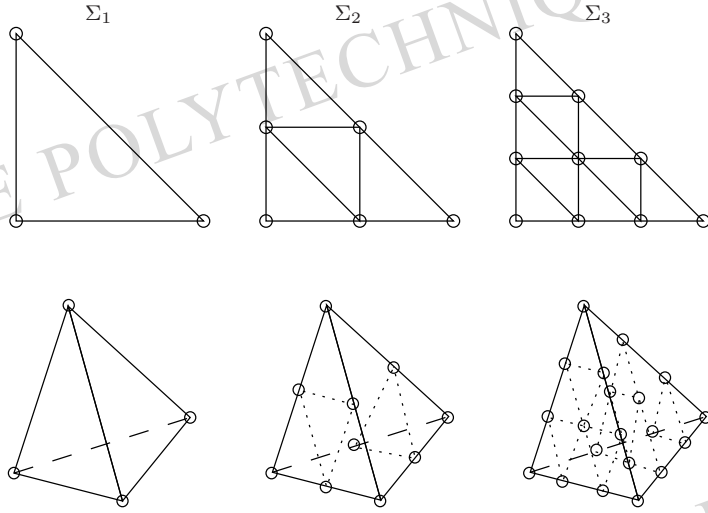


FIGURE 6.9 – Treillis d'ordre 1, 2, et 3 pour un triangle (en haut) et un tétraèdre (en bas). Les ronds représentent les points du treillis.

rôle particulier pour la suite : pour tout entier  $k \geq 1$  on appelle **treillis d'ordre  $k$**  l'ensemble

$$\Sigma_k = \left\{ x \in K \text{ tel que } \lambda_j(x) \in \left\{ 0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1 \right\} \text{ pour } 1 \leq j \leq N \right\}. \quad (6.39)$$

Pour  $k = 1$  il s'agit de l'ensemble des sommets de  $K$ , et pour  $k = 2$  des sommets et des points milieux des arêtes reliant deux sommets (voir la Figure 6.9). Dans le cas général,  $\Sigma_k$  est un ensemble fini de points  $(\sigma_j)_{1 \leq j \leq n_k}$ .

Nous définissons maintenant l'ensemble  $\mathbb{P}_k$  des polynômes à coefficients réels de  $\mathbb{R}^N$  dans  $\mathbb{R}$  de degré inférieur ou égal à  $k$ , c'est-à-dire que tout  $p \in \mathbb{P}_k$  s'écrit sous la forme

$$p(x) = \sum_{\substack{i_1, \dots, i_N \geq 0 \\ i_1 + \dots + i_N \leq k}} \alpha_{i_1, \dots, i_N} x_1^{i_1} \cdots x_N^{i_N} \text{ avec } x = (x_1, \dots, x_N).$$

L'intérêt de la notion de treillis  $\Sigma_k$  d'un  $N$ -simplexe  $K$  est qu'il permet de caractériser tous les polynômes de  $\mathbb{P}_k$  (on dit que  $\Sigma_k$  est **unisolvant** pour  $\mathbb{P}_k$ ).

**Lemme 6.3.3** *Soit  $K$  un  $N$ -simplexe. Pour un entier  $k \geq 1$ , soit  $\Sigma_k$  le treillis d'ordre  $k$ , défini par (6.39), dont les points sont notés  $(\sigma_j)_{1 \leq j \leq n_k}$ . Alors, tout polynôme de  $\mathbb{P}_k$  est déterminé de manière unique par ses valeurs aux points  $(\sigma_j)_{1 \leq j \leq n_k}$ . Autrement dit, il existe une base  $(\psi_j)_{1 \leq j \leq n_k}$  de  $\mathbb{P}_k$  telle que*

$$\psi_j(\sigma_i) = \delta_{ij} \quad 1 \leq i, j \leq n_k.$$

**Démonstration.** Le cardinal de  $\Sigma_k$  et la dimension de  $\mathbb{P}_k$  coïncident

$$\text{card}(\Sigma_k) = \dim(\mathbb{P}_k) = \frac{(N+k)!}{N!k!}$$

(le vérifier en guise d'exercice, au moins pour  $k = 1, 2$ ). Comme l'application qui, à tout polynôme de  $\mathbb{P}_k$ , fait correspondre ses valeurs sur le treillis  $\Sigma_k$  est linéaire, il suffit de montrer qu'elle est injective pour montrer qu'elle est bijective. Soit donc un polynôme  $p \in \mathbb{P}_k$  qui s'annule sur  $\Sigma_k$ . Montrons, par récurrence sur la dimension  $N$ , que  $p$  est identiquement nul sur  $\mathbb{R}^N$ . Pour  $N = 1$ , il est clair qu'un polynôme de degré  $k$  qui s'annule en  $(k+1)$  points distincts est nul. Supposons le résultat vrai à l'ordre  $N-1$ . Comme  $x$  dépend linéairement des coordonnées barycentriques  $(\lambda_j(x))_{1 \leq j \leq N+1}$ , on peut définir un polynôme  $q(\lambda) = p(x)$  de degré au plus  $k$  en la variable  $\lambda \in \mathbb{R}^{N+1}$ . Si l'on fixe une coordonnée  $\lambda_j$  dans l'ensemble  $\{0, 1/k, \dots, (k-1)/k, 1\}$  et que l'on pose  $\lambda = (\lambda', \lambda_j)$ , on obtient un polynôme  $q_j(\lambda') = q(\lambda)$  qui dépend de  $N-1$  variables indépendantes (car on a la relation  $\sum_{j=1}^{N+1} \lambda_j = 1$ ) et qui est nul sur la section du treillis  $\Sigma_k$  correspondant à la valeur fixée de  $\lambda_j$ . Comme cette section est aussi le treillis d'ordre  $k$  d'un  $(N-1)$ -simplexe dans l'hyperplan  $\lambda_j$  fixé, on peut appliquer l'hypothèse de récurrence et en déduire que  $q_j = 0$ . Autrement dit, le facteur  $\lambda_j(\lambda_j - 1/k) \cdots (\lambda_j - (k-1)/k)(\lambda_j - 1)$  divise  $q$ , ce qui est une contradiction avec le fait que le degré de  $q(\lambda)$  est inférieur ou égal à  $k$ , sauf si  $q = 0$ , ce qui est le résultat désiré.  $\square$

**Lemme 6.3.4** Soit  $K$  et  $K'$  deux  $N$ -simplexes ayant une face commune  $\Gamma = \partial K \cap \partial K'$ . Soit un entier  $k \geq 1$ . Alors, leurs treillis d'ordre  $k$ ,  $\Sigma_k$  et  $\Sigma'_k$  coïncident sur cette face  $\Gamma$ . De plus, étant donné  $p_K$  et  $p_{K'}$  deux polynômes de  $\mathbb{P}_k$ , la fonction  $v$  définie par

$$v(x) = \begin{cases} p_K(x) & \text{si } x \in K \\ p_{K'}(x) & \text{si } x \in K' \end{cases}$$

est continue sur  $K \cup K'$ , si et seulement si  $p_K$  et  $p_{K'}$  ont des valeurs qui coïncident aux points du treillis sur la face commune  $\Gamma$ .

**Démonstration.** Il est clair que la restriction à une face de  $K$  de son treillis d'ordre  $\Sigma_k$  est aussi un treillis d'ordre  $k$  dans l'hyperplan contenant cette face, qui ne dépend que des sommets de cette face. Par conséquent, les treillis  $\Sigma_k$  et  $\Sigma'_k$  coïncident sur leur face commune  $\Gamma$ . Si les polynômes  $p_K$  et  $p_{K'}$  coïncident aux points de  $\Sigma_k \cap \Gamma$ , alors par application du Lemme 6.3.3 ils sont égaux sur  $\Gamma$ , ce qui prouve la continuité de  $v$ .  $\square$

En pratique, on utilise surtout des polynômes de degré 1 ou 2. Dans ce cas on a les caractérisations suivantes de  $\mathbb{P}_1$  et  $\mathbb{P}_2$  dans un  $N$ -simplexe  $K$ .



**Exercice 6.3.2** Soit  $K$  un  $N$ -simplexe de sommets  $(a_j)_{1 \leq j \leq N+1}$ . Montrer que tout polynôme  $p \in \mathbb{P}_1$  se met sous la forme

$$p(x) = \sum_{j=1}^{N+1} p(a_j) \lambda_j(x),$$

où les  $(\lambda_j(x))_{1 \leq j \leq N+1}$  sont les coordonnées barycentriques de  $x \in \mathbb{R}^N$ .

**Exercice 6.3.3** Soit  $K$  un  $N$ -simplexe de sommets  $(a_j)_{1 \leq j \leq N+1}$ . On définit les points milieux  $(a_{jj'})_{1 \leq j < j' \leq N+1}$  des arêtes de  $K$  par leur coordonnées barycentriques

$$\lambda_j(a_{jj'}) = \lambda_{j'}(a_{jj'}) = \frac{1}{2}, \quad \lambda_l(a_{jj'}) = 0 \text{ pour } l \neq j, j'.$$

Vérifier que  $\Sigma_2$  est précisément constitué des sommets et des points milieux des arêtes et que tout polynôme  $p \in \mathbb{P}_2$  se met sous la forme

$$p(x) = \sum_{j=1}^{N+1} p(a_j) \lambda_j(x) (2\lambda_j(x) - 1) + \sum_{1 \leq j < j' \leq N+1} 4p(a_{jj'}) \lambda_j(x) \lambda_{j'}(x),$$

où les  $(\lambda_j(x))_{1 \leq j \leq N+1}$  sont les coordonnées barycentriques de  $x \in \mathbb{R}^N$ .

Nous avons maintenant tous les outils pour définir la méthode des éléments finis  $\mathbb{P}_k$ .

**Définition 6.3.5** Étant donné un maillage  $\mathcal{T}_h$  d'un ouvert connexe polyédrique  $\Omega$ , la méthode des éléments finis  $\mathbb{P}_k$ , ou **éléments finis triangulaires de Lagrange d'ordre  $k$** , associée à ce maillage, est définie par l'espace discret

$$V_h = \{v \in C(\overline{\Omega}) \text{ tel que } v|_{K_i} \in \mathbb{P}_k \text{ pour tout } K_i \in \mathcal{T}_h\}. \quad (6.40)$$

On appelle **nœuds des degrés de liberté** l'ensemble des points  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$  des treillis d'ordre  $k$  de chacun des  $N$ -simplexes  $K_i \in \mathcal{T}_h$ . On ne compte qu'une seule fois les points qui coïncident et  $n_{dl}$  est le nombre de degrés de liberté de la méthode des éléments finis  $\mathbb{P}_k$ . On appelle **degrés de liberté** d'une fonction  $v \in V_h$  l'ensemble des valeurs de  $v$  en ces nœuds  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$ . On définit aussi le sous-espace  $V_{0h}$  par

$$V_{0h} = \{v \in V_h \text{ tel que } v = 0 \text{ sur } \partial\Omega\}. \quad (6.41)$$

Lorsque  $k = 1$  les nœuds des degrés de liberté coïncident avec les sommets du maillage. Lorsque  $k = 2$  ces nœuds sont constitués d'une part des sommets du maillage et d'autre part des points milieux des arêtes reliant deux sommets.

**Remarque 6.3.6** L'appellation "éléments finis de Lagrange" correspond aux éléments finis dont les degrés de liberté sont des valeurs ponctuelles des fonctions de l'espace  $V_h$ . On peut définir d'autres types d'éléments finis, par exemple les éléments finis de Hermite (voir la Sous-section 6.2.5) pour lesquels les degrés de liberté sont les valeurs ponctuelles de la fonction et de ses dérivées. •

**Proposition 6.3.7** *L'espace  $V_h$ , défini par (6.40), est un sous-espace de  $H^1(\Omega)$  dont la dimension est finie, égale au nombre de degrés de liberté. De plus, il existe une base de  $V_h$   $(\phi_i)_{1 \leq i \leq n_{dl}}$  définie par*

$$\phi_i(\hat{a}_j) = \delta_{ij} \quad 1 \leq i, j \leq n_{dl},$$

telle que

$$v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i) \phi_i(x).$$

**Démonstration.** Les éléments de  $V_h$ , étant réguliers sur chaque maille  $K_i$  et continus sur  $\bar{\Omega}$ , appartiennent à  $H^1(\Omega)$  (voir le Lemme 4.3.19). Grâce au Lemme 6.3.4 les éléments de  $V_h$  sont exactement obtenus en assemblant sur chaque  $K_i \in \mathcal{T}_h$  des polynômes de  $\mathbb{P}_k$  qui coïncident sur les degrés de liberté des faces (ce qui prouve au passage que  $V_h$  n'est pas réduit aux seules fonctions constantes). Enfin, en assemblant les bases  $(\psi_j)_{1 \leq j \leq n_k}$  de  $\mathbb{P}_k$  sur chaque maille  $K_i$  (fournies par le Lemme 6.3.3) on obtient la base annoncée  $(\phi_i)_{1 \leq i \leq n_{dl}}$  de  $V_h$ .  $\square$

**Remarque 6.3.8** On obtient un résultat semblable pour le sous-espace  $V_{0h}$ , défini par (6.41), qui est un sous-espace de  $H_0^1(\Omega)$  de dimension finie égale au nombre de degrés de liberté intérieurs (on ne compte pas les nœuds sur le bord  $\partial\Omega$ ).  $\bullet$

**Exercice 6.3.4** Soit  $\mathcal{T}_h$  un maillage de  $\bar{\Omega}$  pour  $\Omega$  ouvert simplement connexe polygonal de  $\mathbb{R}^2$ . On note  $n_t$  le nombre de triangles de  $\mathcal{T}_h$ ,  $n_c$  le nombre de faces ou cotés des triangles (un coté commun à deux triangles n'est compté qu'une seule fois),  $n_s$  le nombre de sommets du maillage, et  $n_{0s}$  le nombre de sommets intérieurs du maillage. Montrer que les dimensions des espaces  $V_h$  et  $V_{0h}$  sont

$$\dim V_h = \frac{k(k-1)}{2} n_t + k n_s - k + 1, \quad \dim V_{0h} = \frac{k(k-1)}{2} n_t - k n_s + k + 1.$$

Décrivons la **résolution pratique** du problème de Dirichlet (6.36) par la méthode des éléments finis  $\mathbb{P}_k$ . La formulation variationnelle (6.2) de l'approximation interne devient ici :

$$\text{trouver } u_h \in V_{0h} \text{ tel que } \int_{\Omega} \nabla u_h \cdot \nabla v_h \, dx = \int_{\Omega} f v_h \, dx \quad \forall v_h \in V_{0h}. \quad (6.42)$$

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

$$\sum_{j=1}^{n_{dl}} u_h(\hat{a}_j) \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dx = \int_{\Omega} f \phi_i \, dx.$$

En notant  $U_h = (u_h(\hat{a}_j))_{1 \leq j \leq n_{dl}}$ ,  $b_h = (\int_{\Omega} f \phi_i \, dx)_{1 \leq i \leq n_{dl}}$ , et en introduisant la **matrice de rigidité**

$$\mathcal{K}_h = \left( \int_{\Omega} \nabla \phi_j \cdot \nabla \phi_i \, dx \right)_{1 \leq i, j \leq n_{dl}},$$

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

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

Comme les fonctions de base  $\phi_j$  ont un “petit” support autour du nœud  $\hat{a}_i$  (voir la Figure 6.10), l’intersection des supports de  $\phi_j$  et  $\phi_i$  est souvent vide et la plupart des coefficients de  $\mathcal{K}_h$  sont nuls. On dit que la matrice  $\mathcal{K}_h$  est **creuse**.

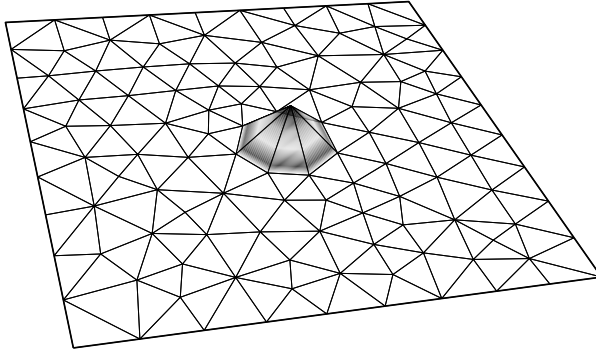


FIGURE 6.10 – Fonction de base  $\mathbb{P}_1$  en dimension  $N = 2$ .

Pour calculer les coefficients de  $\mathcal{K}_h$ , on peut utiliser la formule d’intégration exacte suivante. On note  $(\lambda_i(x))_{1 \leq i \leq N+1}$  les coordonnées barycentriques du point courant  $x$  d’un  $N$ -simplexe  $K$ . Pour tout  $\alpha_1, \dots, \alpha_{N+1} \in \mathbb{N}$ , on a

$$\int_K \lambda_1(x)^{\alpha_1} \dots \lambda_{N+1}(x)^{\alpha_{N+1}} dx = \text{Volume}(K) \frac{\alpha_1! \dots \alpha_{N+1}! N!}{(\alpha_1 + \dots + \alpha_{N+1} + N)!}. \quad (6.43)$$

Pour calculer le second membre  $b_h$  (et même éventuellement la matrice  $\mathcal{K}_h$ ), on utilise des **formules de quadrature** (ou formules d’intégration numérique) qui donnent une approximation des intégrales sur chaque  $N$ -simplexe  $K_i \in \mathcal{T}_h$ . Par exemple, si  $K$  est un  $N$ -simplexe de sommets  $(a_i)_{1 \leq i \leq N+1}$ , les formules suivantes généralisent les formules en dimension 1, dites du “point milieu” et des “trapèzes” :

$$\int_K \psi(x) dx \approx \text{Volume}(K) \psi(a_0), \quad (6.44)$$

avec  $a_0 = (N+1)^{-1} \sum_{i=1}^{N+1} a_i$ , le barycentre de  $K$ , et

$$\int_K \psi(x) dx \approx \frac{\text{Volume}(K)}{N+1} \sum_{i=1}^{N+1} \psi(a_i). \quad (6.45)$$

Comme le montre les Exercices 6.3.6 et 6.3.8, ces formules sont exactes pour des fonctions affines et sont donc approchées à l’ordre 2 en  $h$  pour des fonctions régulières.

La construction de la matrice  $\mathcal{K}_h$  est appelée **assemblage de la matrice**. La mise en oeuvre informatique de cette étape du calcul peut être assez compliquée, mais son coût en terme de temps de calcul est faible. Ce n'est pas le cas de la résolution du système linéaire  $\mathcal{K}_h U_h = b_h$  qui est l'étape la **plus coûteuse** de la méthode en temps de calcul (et en place mémoire). En particulier, les calculs tridimensionnels sont encore très chers de nos jours dès que l'on utilise des maillages fins. L'Exercice 6.3.11 permet de s'en rendre compte. Heureusement, la matrice de rigidité  $\mathcal{K}_h$  est **creuse** (c'est-à-dire que la plupart de ses éléments sont nuls), ce qui permet de minimiser les calculs (pour plus de détails voir les algorithmes de résolution de systèmes linéaires dans la Section 13.1 de l'annexe). Rappelons que la matrice  $\mathcal{K}_h$  est nécessairement inversible par application du Lemme 6.1.1 et qu'elle est symétrique.

**Exercice 6.3.5** Démontrer la formule (6.43) en dimension  $N = 2$ .

**Exercice 6.3.6** Montrer que les formules (6.44) et (6.45) sont exactes pour  $\psi \in \mathbb{P}_1$ .

**Exercice 6.3.7** Soit  $K$  un triangle de  $\mathbb{R}^2$  de sommets  $(a_i)_{1 \leq i \leq 3}$  et de barycentre  $a_0$ . Soit  $(a_{ij})_{1 \leq i < j \leq 3}$  les points milieux des segments d'extrémités  $a_i, a_j$ . Montrer que la formule de quadrature

$$\int_K \psi(x) dx \approx \frac{\text{Aire}(K)}{3} \sum_{1 \leq i < j \leq 3} \psi(a_{ij})$$

est exacte pour  $\psi \in \mathbb{P}_2$ , tandis que la formule

$$\int_K \psi(x) dx \approx \frac{\text{Aire}(K)}{60} \left( 3 \sum_{i=1}^3 \psi(a_i) + 8 \sum_{1 \leq i < j \leq 3} \psi(a_{ij}) + 27 \psi(a_0) \right)$$

est exacte pour  $\psi \in \mathbb{P}_3$ .

**Exercice 6.3.8** Soit  $(b_i)_{1 \leq i \leq I}$  des points d'un  $N$ -simplexe  $K$  et  $(\omega_i)_{1 \leq i \leq I}$  des poids réels. Soit une formule de quadrature

$$\int_K \psi(x) dx \approx \text{Volume}(K) \sum_{i=1}^I \omega_i \psi(b_i)$$

qui soit exacte pour  $\psi \in \mathbb{P}_k$ . Montrer que, pour une fonction régulière  $\psi$ , on a

$$\frac{1}{\text{Volume}(K)} \int_K \psi(x) dx = \sum_{i=1}^I \omega_i \psi(b_i) + \mathcal{O}(h^{k+1}),$$

où  $h$  est le diamètre de  $K$ .

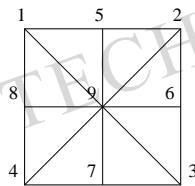


FIGURE 6.11 – Exemple de maillage et de numérotation des nœuds.

**Exercice 6.3.9** On considère le carré  $\Omega = ]-1, +1[^2$  maillé suivant la Figure 6.11. Calculer la matrice de rigidité  $\mathcal{K}_h$  des éléments finis  $\mathbb{P}_1$  appliqués au Laplacien avec condition aux limites de Neumann (on utilisera les symétries du maillage).

**Exercice 6.3.10** Appliquer la méthode des éléments finis  $\mathbb{P}_1$  au problème de Dirichlet (6.36) dans le carré  $\Omega = ]0, 1[^2$  avec le maillage triangulaire uniforme de la Figure 6.12. Montrer que la matrice de rigidité  $\mathcal{K}_h$  est la même matrice que celle que l'on obtiendrait par application de la méthode des différences finies (à un facteur multiplicatif  $h^2$  près), mais que le second membre  $b_h$  est différent.

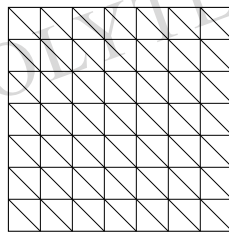


FIGURE 6.12 – Maillage triangulaire uniforme d'un carré.

**Exercice 6.3.11** On reprend les notations de l'Exercice 6.3.10. On note  $n$  le nombre de points du maillage sur un côté du carré (supposé être le même pour chaque côté). On numérote "ligne par ligne" les nœuds du maillage (ou les degrés de liberté). Montrer que la matrice de rigidité  $\mathcal{K}_h$  des éléments finis  $\mathbb{P}_1$  est de taille de l'ordre de  $n^2$  et de largeur de bande de l'ordre de  $2n$  (pour  $n$  grand).

Montrer que la même méthode et le même type de maillage pour le cube  $\Omega = ]0, 1[^3$  conduisent à une matrice de taille de l'ordre de  $n^3$  et de largeur de bande de l'ordre de  $2n^2$  (où  $n$  est le nombre de nœuds le long d'une arête du cube  $\Omega$ ).

**Remarque 6.3.9** Comme le montre l'exemple de l'Exercice 6.3.10 la manière de numérotter les nœuds des degrés de liberté (ou de façon équivalente les fonctions de base) a une influence sur la structure creuse de la matrice  $\mathcal{K}_h$ , c'est-à-dire sur l'emplacement des ses éléments non nuls. Comme expliqué dans la Section 13.1 (voir,

par exemple, le Lemme 13.1.4), cette structure creuse de la matrice a une grande influence sur la performance de la résolution du système linéaire  $\mathcal{K}_h U_h = b_h$ . Par exemple, si l'on résout ce système linéaire par une méthode du type "élimination de Gauss", il est avantageux de choisir une numérotation qui regroupe les éléments non nuls près de la diagonale. •

**Remarque 6.3.10** Pour simplifier l'analyse (aussi bien que la mise en oeuvre) on peut utiliser une transformation affine pour ramener tout  $N$ -simplexe  $K$  du maillage  $\mathcal{T}_h$  à un  $N$ -simplexe de "référence"  $K_0$ . Par ce simple changement de variable, tous les calculs se ramènent à des calculs sur  $K_0$ . En pratique, on choisit souvent

$$K_0 = \left\{ x \in \mathbb{R}^N \text{ tel que } \sum_{i=1}^N x_i \leq 1, x_i \geq 0 \ 1 \leq i \leq N \right\}, \quad (6.46)$$

et on vérifie sans peine que tout  $N$ -simplexe  $K$  est l'image par une transformation affine de  $K_0$ . En effet, les coordonnées barycentriques, définies par (6.38) sont les mêmes pour  $K$  et  $K_0$  et, en notant  $\lambda = (\lambda_j)_{1 \leq j \leq N+1}$ ,  $\tilde{x} = (x, 1)$  le point courant dans  $K$ ,  $\tilde{x}_0 = (x_0, 1)$  le point courant dans  $K_0$ , on a  $A\lambda = \tilde{x}$ , et  $A_0\lambda = \tilde{x}_0$ , où les matrices  $A$  et  $A_0$  sont définies par (6.37) et inversibles. On en déduit donc que  $\tilde{x} = AA_0^{-1}\tilde{x}_0$ , c'est-à-dire qu'il existe une matrice  $B$ , inversible d'ordre  $N$ , et un vecteur  $b \in \mathbb{R}^N$  tels que  $x = Bx_0 + b$ . •

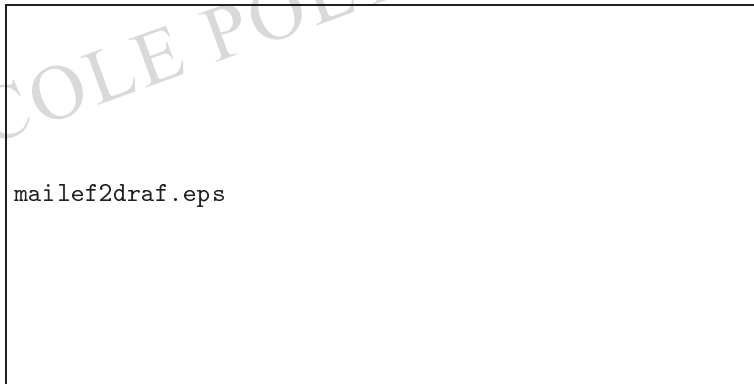


FIGURE 6.13 – Maillage triangulaire plus fin que celui de la Figure 6.7.

L'exercice suivant montre que la méthode des éléments finis  $\mathbb{P}_1$  vérifie le principe du maximum.

**Exercice 6.3.12** On dit qu'une matrice carrée réelle  $B = (b_{ij})_{1 \leq i, j \leq n}$  est une M-matrice si, pour tout  $i$ ,

$$b_{ii} > 0, \quad \sum_{k=1}^n b_{ik} > 0, \quad b_{ij} \leq 0 \quad \forall j \neq i.$$

Montrer que toute M-matrice est inversible et que tous les coefficients de son inverse sont positifs ou nuls.

**Exercice 6.3.13** On se place en dimension  $N = 2$ . Soit  $u_h$  la solution approchée du problème de Dirichlet (6.36) obtenue par la méthode des éléments finis  $\mathbb{P}_1$ . On suppose que tous les angles des triangles  $K_i \in \mathcal{T}_h$  sont inférieurs ou égaux à  $\pi/2$ . Montrer que  $u_h(x) \geq 0$  dans  $\Omega$  si  $f(x) \geq 0$  dans  $\Omega$ . Indication : on montrera que, pour tout  $\epsilon > 0$ ,  $\mathcal{K}_h + \epsilon \text{Id}$  est une M-matrice, où  $\mathcal{K}_h$  est la matrice de rigidité.

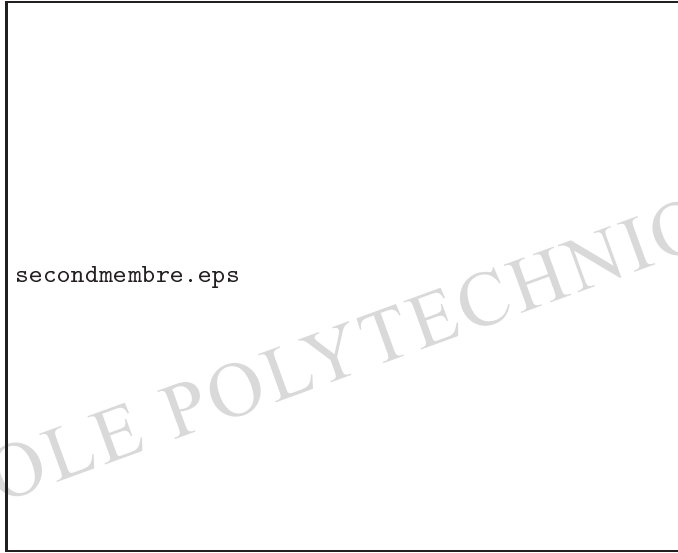


FIGURE 6.14 – Terme source  $f$  dans l'équation (6.36).

Il n'y a évidemment aucune difficulté à étendre la méthode des éléments finis  $\mathbb{P}_k$  à d'autres problèmes que (6.36).

**Exercice 6.3.14** Appliquer la méthode des éléments finis  $\mathbb{P}_k$  au système de l'élasticité (5.56). Montrer en particulier que la matrice de rigidité  $\mathcal{K}_h$  est dans ce cas d'ordre  $Nn_{dl}$  où  $N$  est la dimension d'espace et  $n_{dl}$  est le nombre de nœuds de degrés de liberté.

**Exercice 6.3.15** Expliciter la matrice de rigidité  $\mathcal{K}_h$  obtenue par application de la méthode des éléments finis  $\mathbb{P}_k$  au problème de Neumann

$$\begin{cases} -\Delta u + au = f & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = g & \text{sur } \partial\Omega, \end{cases} \quad (6.47)$$

avec  $f \in L^2(\Omega)$ ,  $g \in L^2(\partial\Omega)$ , et  $a \in L^\infty(\Omega)$  tel que  $a(x) \geq a_0 > 0$  p.p. dans  $\Omega$ .

**Exercice 6.3.16** Montrer que la matrice de rigidité  $\mathcal{K}_h$  obtenue par application de la méthode des éléments finis  $\mathbb{P}_k$  au problème de convection-diffusion de l'Exercice 5.2.2 est inversible mais pas symétrique.

**Exercice 6.3.17** On se propose de résoudre numériquement l'équation des plaques (5.70) par une méthode d'éléments finis (de type Hermite) en dimension  $N = 2$ . Pour un maillage triangulaire  $\mathcal{T}_h$  on introduit l'espace discret

$$V_h = \{v \in C^1(\overline{\Omega}) \text{ tel que } v|_{K_i} \in P_5 \text{ pour tout } K_i \in \mathcal{T}_h\}.$$

Montrer que tout polynôme  $p \in P_5$  est caractérisé de manière unique sur un triangle  $K$  par les 21 valeurs réelles suivantes

$$v(a_j), \nabla v(a_j), \nabla \nabla v(a_j), \frac{\partial p(b_j)}{\partial n} \quad j = 1, 2, 3, \quad (6.48)$$

où  $(a_1, a_2, a_3)$  sont les sommets de  $K$ ,  $(b_1, b_2, b_3)$  les milieux des cotés de  $K$ , et  $\partial p(b_j)/\partial n$  désigne la dérivée normale au côté de  $b_j$ . Montrer que  $V_h$  est un sous-espace de  $H^2(\Omega)$  dont les éléments  $v$  sont caractérisés de manière unique par les valeurs (6.48) pour chaque sommet et milieu d'arête du maillage. En déduire une méthode d'éléments finis (dite d'Argyris) pour résoudre (5.70).

Nous terminons cette sous-section en l'illustrant par un résultat numérique obtenu par la méthode des éléments finis  $\mathbb{P}_1$  appliquée au problème de Dirichlet (6.36). Le second membre est donné par la Figure 6.14. On peut interpréter ce problème comme la modélisation de la diffusion dans l'atmosphère d'un polluant émis par une source localisée. Le domaine de calcul représente une région autour de la source (la direction verticale est "moyennée" et absente du calcul) et on suppose que la concentration est nulle sur son bord. Deux maillages ont été utilisés : le maillage "grossier" de la Figure 6.7, et le maillage "fin" de la Figure 6.13. Les résultats correspondant sont montrés sur la Figure 6.15, respectivement. On remarque que la valeur maximale de la solution numérique  $u_h$  est plus élevée pour le maillage fin que pour le maillage grossier (les échelles ne sont pas les mêmes). C'est la manifestation du fait que le pas  $h$  du maillage grossier n'est pas encore assez petit pour que la solution approchée  $u_h$  ait convergé vers la solution exacte. Si l'on rajoute en plus de la diffusion un effet de convection (modélisant un vent constant dans la direction horizontale, voir (5.13)), on peut voir l'effet de panache ainsi produit sur la concentration dans la Figure 6.16 (obtenue avec le maillage fin). La valeur maximale de la solution est plus petite en présence d'un terme de convection, ce qui correspond bien à l'intuition physique que le vent "dilue" les concentrations élevées de polluant.

### 6.3.2 Convergence et estimation d'erreur

Nous démontrons la convergence des méthodes d'éléments finis  $\mathbb{P}_k$  pour le problème de Dirichlet (6.36). Insistons sur le fait qu'il s'agit seulement d'un problème modèle,



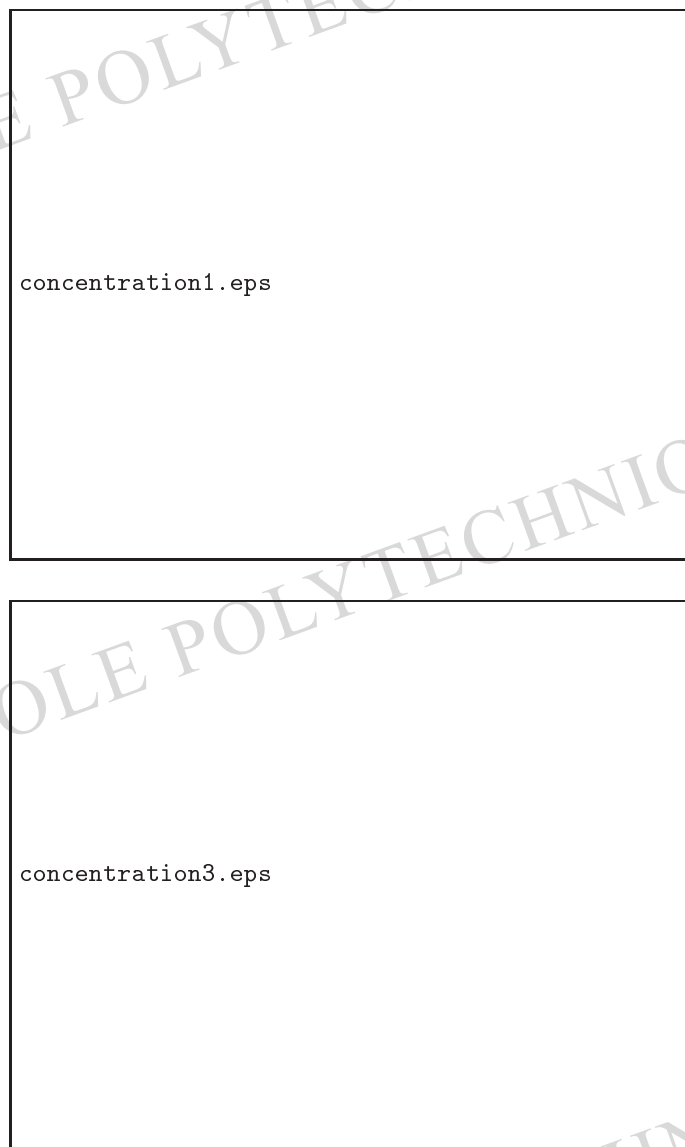
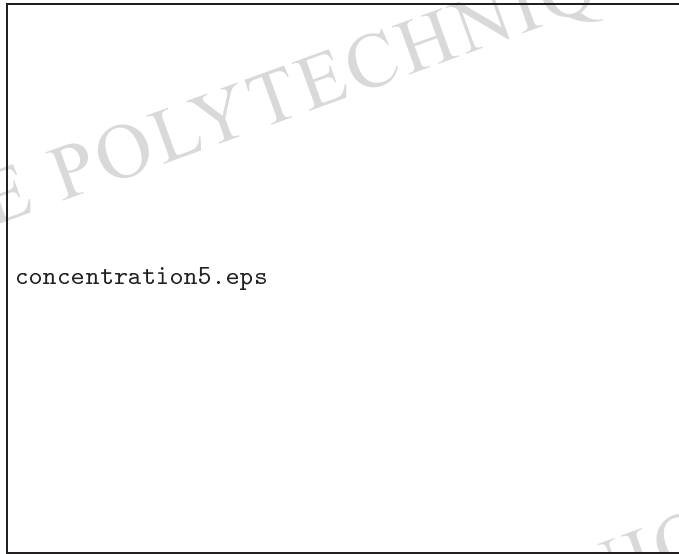


FIGURE 6.15 – Solution approchée  $u_h$  de l'équation de diffusion (6.36) pour le maillage grossier de la Figure 6.7 (haut) et pour le maillage fin de la Figure 6.13 (bas).

FIGURE 6.16 – Solution approchée  $u_h$  de l'équation de convection-diffusion (5.13)).

et que ces méthodes convergent pour d'autres problèmes, comme celui de Neumann (6.47). Nous allons avoir besoin d'hypothèses géométriques sur la qualité du maillage. Pour tout  $N$ -simplexe  $K$  on introduit deux paramètres géométriques : le **diamètre**  $\text{diam}(K)$  et la **rondeur**  $\rho(K)$ , définie comme le diamètre de la plus grande boule contenue dans  $K$ ,

$$\text{diam}(K) = \max_{x,y \in K} \|x - y\|, \quad \rho(K) = \max_{B_r \subset K} (2r).$$

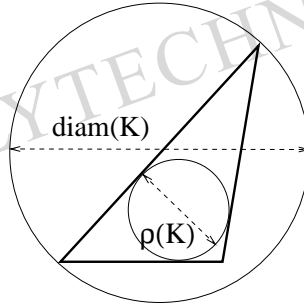
Bien sûr, on a toujours  $\text{diam}(K)/\rho(K) > 1$ . Ce rapport est d'autant plus grand que  $K$  est "aplatis" : il mesure en quelque sorte la tendance à la dégénérescence de  $K$ . En pratique, comme en théorie, il faut éviter d'utiliser des  $N$ -simplexes  $K$  trop aplatis.

**Définition 6.3.11** Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages de  $\Omega$ . On dit qu'il s'agit d'une suite de **maillages réguliers** si

1. la suite  $h = \max_{K_i \in \mathcal{T}_h} \text{diam}(K_i)$  tend vers 0,
2. il existe une constante  $C$  telle que, pour tout  $h > 0$  et tout  $K \in \mathcal{T}_h$ ,

$$\frac{\text{diam}(K)}{\rho(K)} \leq C. \quad (6.49)$$

**Remarque 6.3.12** En dimension  $N = 2$  la condition (6.49) est équivalente à la condition suivante sur les angles du triangle  $K$  : il existe un angle minimum  $\theta_0 > 0$  qui minore (uniformément en  $h$ ) tous les angles de tout  $K \in \mathcal{T}_h$ . Insistons sur le fait

FIGURE 6.17 – Diamètre  $\text{diam}(K)$  et rondeur  $\rho(K)$  d'un triangle  $K$ .

que la condition (6.49) est tout aussi importante en pratique que pour l'analyse de convergence qui va suivre. •

Nous pouvons maintenant énoncer le résultat principal de cette sous-section qui affirme la convergence de la méthode des éléments finis  $\mathbb{P}_k$  et qui donne une estimation de la vitesse de convergence si la solution est régulière.

**Théorème 6.3.13** *Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages réguliers de  $\Omega$ . Soit  $u \in H_0^1(\Omega)$ , la solution du problème de Dirichlet (6.36), et  $u_h \in V_{0h}$ , celle de son approximation interne (6.42) par la méthode des éléments finis  $\mathbb{P}_k$ . Alors la méthode des éléments finis  $\mathbb{P}_k$  converge, c'est-à-dire que*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0. \quad (6.50)$$

De plus, si  $u \in H^{k+1}(\Omega)$  et si  $k+1 > N/2$ , alors on a l'estimation d'erreur

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^k \|u\|_{H^{k+1}(\Omega)}, \quad (6.51)$$

où  $C$  est une constante indépendante de  $h$  et de  $u$ .

**Remarque 6.3.14** Le Théorème 6.3.13 s'applique en fait à toute méthode d'éléments finis de type Lagrange (par exemple, les éléments finis rectangulaires de la Sous-section 6.3.3). En effet, le seul argument utilisé est la construction d'un opérateur d'interpolation basé sur la caractérisation des fonctions de  $V_h$  par leurs valeurs aux nœuds des degrés de liberté, ce qui est toujours possible pour des éléments finis de type Lagrange (voir la Remarque 6.3.6). Remarquons que, pour les cas physiquement pertinents  $N = 2$  ou  $N = 3$ , la condition  $k+1 > N/2$  est toujours satisfaite dès que  $k \geq 1$ . •

**Remarque 6.3.15** L'estimation d'erreur (6.51) du Théorème 6.3.13 n'est vraie que si la solution exacte  $u$  est régulière, ce qui n'est pas toujours le cas. Si  $u$  n'est pas

régulière, on constate en pratique que la convergence est plus lente (voir la Figure 6.5 en une dimension d'espace). D'autre part, la convergence (6.50), qui a lieu dans l'espace "d'énergie", n'implique pas la convergence ponctuelle de  $u_h$  ou de ses dérivées. La Figure 6.18 illustre ce fait pour le problème de Dirichlet (6.36) avec  $f \equiv 1$  dans la géométrie du "coin rentrant" où la solution est singulière (voir le Lemme 5.2.33). Numériquement, le module du gradient de  $u_h$  croît vers l'infini dans le coin lorsque  $h$  tend vers zéro (le maximum de  $|\nabla u_h|$  vaut 0.92 pour le maillage de gauche à 1187 sommets, 1.18 pour le maillage du milieu à 4606 sommets, et 1.50 pour celui de droite à 18572 sommets) •

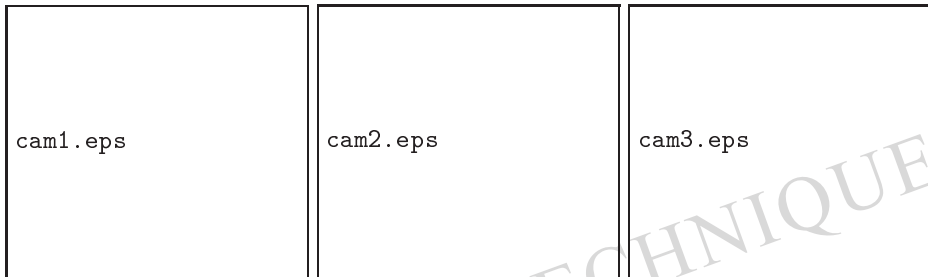


FIGURE 6.18 – Module du gradient de  $u_h$  pour trois maillages (de plus en plus fin de gauche à droite).

La démonstration du Théorème 6.3.13 repose sur la définition suivante d'un **opérateur d'interpolation**  $r_h$  et sur le résultat d'interpolation de la Proposition 6.3.16. Rappelons que nous avons noté  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$  la famille des nœuds des degrés de liberté et  $(\phi_i)_{1 \leq i \leq n_{dl}}$  la base de  $V_{0h}$  de la méthode des éléments finis  $\mathbb{P}_k$  (voir la Proposition 6.3.7). Pour toute fonction continue  $v$ , on définit son interpolée

$$r_h v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i) \phi_i(x). \quad (6.52)$$

La différence principale avec l'étude faite en dimension  $N = 1$  est que, les fonctions de  $H^1(\Omega)$  n'étant pas continues lorsque  $N \geq 2$ , l'opérateur d'interpolation  $r_h$  n'est pas défini sur  $H^1(\Omega)$  (les valeurs ponctuelles d'une fonction de  $H^1(\Omega)$  n'ont a priori pas de sens). Néanmoins, et c'est la raison de l'hypothèse  $k + 1 > N/2$ ,  **$r_h$  est bien défini sur  $H^{k+1}(\Omega)$**  car les fonctions de  $H^{k+1}(\Omega)$  sont continues ( $H^{k+1}(\Omega) \subset C(\bar{\Omega})$ ) d'après le Théorème 4.3.25).

**Proposition 6.3.16** *Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages réguliers de  $\Omega$ . On suppose que  $k + 1 > N/2$ . Alors, pour tout  $v \in H^{k+1}(\Omega)$  l'interpolée  $r_h v$  est bien définie, et il existe une constante  $C$ , indépendante de  $h$  et de  $v$ , telle que*

$$\|v - r_h v\|_{H^1(\Omega)} \leq C h^k \|v\|_{H^{k+1}(\Omega)}. \quad (6.53)$$

Admettant pour l'instant la Proposition 6.3.16 nous pouvons conclure quant à la convergence de la méthode des éléments finis  $\mathbb{P}_k$ .

**Démonstration du Théorème 6.3.13.** On applique le cadre abstrait de la Sous-section 6.1.2. Pour démontrer (6.50) on utilise le Lemme 6.1.3 avec  $\mathcal{V} = C_c^\infty(\Omega)$  qui est bien dense dans  $H_0^1(\Omega)$ . Comme  $C_c^\infty(\Omega) \subset H^{k+1}(\Omega)$ , l'estimation (6.53) de la Proposition 6.3.16 permet de vérifier l'hypothèse (6.5) du Lemme 6.1.3 (pour des fonctions régulières on n'a pas besoin de la condition  $k+1 > N/2$  dans la Proposition 6.3.16).

Pour obtenir l'estimation d'erreur (6.51) on utilise le Lemme de Céa 6.1.2 qui nous dit que

$$\|u - u_h\|_{H^1(\Omega)} \leq C \inf_{v_h \in V_{0h}} \|u - v_h\|_{H^1(\Omega)} \leq C \|u - r_h u\|_{H^1(\Omega)},$$

si  $r_h u$  appartient bien à  $H^1(\Omega)$ . Par application de la Proposition 6.3.16 à  $u$  on obtient (6.51).  $\square$

**Remarque 6.3.17** Le Théorème 6.3.13 est valable lorsque  $u_h$  est la solution **exacte** de l'approximation interne (6.42) dans  $V_{0h}$ . Cela nécessite de calculer exactement toutes les intégrales intervenant dans la matrice  $\mathcal{K}_h$  et le second membre  $b_h$ . En pratique, on ne les évalue pas exactement car on a recourt à une intégration numérique. Néanmoins, si on utilise des formules de quadrature "raisonnables", la méthode des éléments finis  $\mathbb{P}_k$  converge encore (voir [36]). En particulier, si la formule de quadrature utilisée pour calculer des intégrales sur un  $N$ -simplexe  $K$  est exacte pour les polynômes de  $P_{2k-2}$ , alors l'estimation d'erreur (6.51) est toujours valable (où  $u_h$  est la solution discrète calculée avec intégration numérique). Par exemple, pour les éléments finis  $\mathbb{P}_1$  on peut utiliser les formules de quadrature (6.44) ou (6.45) sans perte de précision ou de vitesse de convergence.  $\bullet$

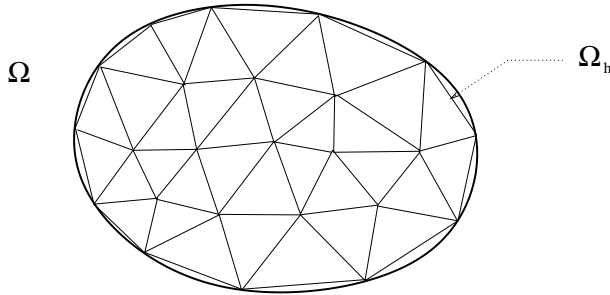


FIGURE 6.19 – Approximation par un domaine polyédrique  $\Omega_h$  d'un ouvert régulier  $\Omega$ .

**Remarque 6.3.18** Indiquons brièvement ce qui se passe lorsque le domaine  $\Omega$  n'est pas polyédrique (mais suffisamment régulier). On commence par approcher  $\Omega$  par un domaine polyédrique  $\Omega_h$  que l'on maille par un maillage  $\mathcal{T}_h$  (voir la Figure 6.19). On peut choisir  $\Omega_h$

et son maillage (avec  $h$  le diamètre maximum des mailles) de telle manière qu'il existe une constante  $C$  (qui ne dépend que de la courbure de  $\Omega$ ) vérifiant

$$\text{dist}(\partial\Omega, \partial\Omega_h) \leq Ch^2.$$

On appelle  $u_h$  la solution de l'approximation variationnelle dans l'espace  $V_h$  associé au maillage  $\mathcal{T}_h$  et à l'élément fini  $\mathbb{P}_k$ . En général, même si on choisit  $\Omega_h \subset \Omega$ ,  $V_h$  **n'est pas un sous-espace** de l'espace de Sobolev  $V$  dans lequel on cherche la solution exacte  $u$  (par exemple, si les conditions aux limites sont de Neumann), ce qui complique sérieusement l'analyse. Néanmoins, en dimension  $N = 2$  on peut montrer (voir [36]) que, pour les éléments finis  $\mathbb{P}_1$ , si  $u \in H^2(\Omega)$ , alors on a toujours

$$\|u - u_h\|_{H^1(\Omega_h)} \leq Ch\|u\|_{H^2(\Omega)}, \quad (6.54)$$

tandis que, pour les éléments finis  $\mathbb{P}_2$ , si  $u \in H^{k+1}(\Omega)$ , alors on a seulement

$$\|u - u_h\|_{H^1(\Omega_h)} \leq Ch^{3/2}\|u\|_{H^{k+1}(\Omega)}. \quad (6.55)$$

Par conséquent, cette méthode est satisfaisante pour des éléments finis  $\mathbb{P}_1$ , puisque la convergence (6.54) est du même ordre que (6.51), mais décevante et non optimale pour des éléments finis  $\mathbb{P}_k$  avec  $k \geq 2$ . On peut remédier à cette situation en introduisant des “éléments finis isoparamétriques” : il s'agit de mailler la partie de  $\Omega$  près du bord par des mailles à bords “courbes” obtenus par déformation de  $N$ -simplexes standards (cette déformation est une généralisation de la transformation affine introduite à la Remarque 6.3.10). Par exemple, en dimension  $N = 2$  on utilise souvent une transformation polynomiale de degré 2 qui déforme un triangle de référence en un “triangle” dont un des cotés est un arc de parabole. Cela permet de mieux approcher la frontière  $\partial\Omega$  par  $\partial\Omega_h$ . On peut alors démontrer une estimation d'erreur optimale du même ordre que (6.51) (voir [11], [36]). •

Nous passons maintenant à la démonstration de la Proposition 6.3.16 que l'on peut admettre en première lecture. Elle passe par la construction d'un opérateur d'interpolation local dans chaque maille du maillage. Soit  $K$  un  $N$ -simplexe de treillis d'ordre  $k$ ,  $\Sigma_k$ . On définit l'opérateur d'interpolation  $r_K$ , pour toute fonction  $v$  continue sur  $K$ ,

$$r_K v = p \in \mathbb{P}_k \text{ tel que } p(x) = v(x) \quad \forall x \in \Sigma_k. \quad (6.56)$$

D'après le Lemme 6.3.3 on sait que tout polynôme de  $\mathbb{P}_k$  est déterminé de manière unique par ses valeurs aux points de  $\Sigma_k$  : par conséquent, (6.56) définit bien  $r_K$  comme une application (linéaire de surcroît).

**Lemme 6.3.19 (de Bramble-Hilbert)** *On suppose que  $k + 1 > N/2$ . L'opérateur d'interpolation  $r_K$  est linéaire continu de  $H^{k+1}(K)$  dans  $H^{k+1}(K)$ , et il existe une constante  $C(K)$  telle que, pour tout  $v \in H^{k+1}(K)$  on a*

$$\|v - r_K v\|_{H^{k+1}(K)} \leq C(K)|v|_{H^{k+1}(K)}, \quad (6.57)$$

où  $|v|_{H^{k+1}(K)}$  est la semi-norme définie par

$$|v|_{H^{k+1}(K)}^2 = \sum_{|\alpha|=k+1} \int_K |\partial^\alpha v|^2 dx = \|v\|_{H^{k+1}(K)}^2 - \|v\|_{H^k(K)}^2.$$

**Démonstration.** Pour  $k+1 > N/2$  le Théorème 4.3.25 indique que  $H^{k+1}(K) \subset C(K)$ , donc les valeurs ponctuelles des fonctions de  $H^{k+1}(K)$  sont bien définies comme formes linéaires continues. Par conséquent,  $r_K v$  est un polynôme dont les coefficients dépendent linéairement et continûment de  $v \in H^{k+1}(K)$ , et ceci dans n'importe quel espace  $H^m(K)$  avec  $m \in \mathbb{N}$ . On en déduit que  $r_K$  est linéaire continu dans  $H^{k+1}(K)$ . Démontrons maintenant l'inégalité

$$\|v\|_{H^{k+1}(K)} \leq C(K) (|v|_{H^{k+1}(K)} + \|r_K v\|_{H^{k+1}(K)}), \quad (6.58)$$

en procédant par contradiction (comme nous l'avons déjà fait pour d'autres inégalités; voir, par exemple, la démonstration (4.15) de l'inégalité de Poincaré). Il existe donc une suite  $v_n \in H^{k+1}(K)$  telle que

$$1 = \|v_n\|_{H^{k+1}(K)} > n (|v_n|_{H^{k+1}(K)} + \|r_K v_n\|_{H^{k+1}(K)}). \quad (6.59)$$

Le terme de gauche de (6.59) implique que la suite  $v_n$  est bornée dans  $H^{k+1}(K)$ . Par application du Théorème de Rellich 4.3.21, il existe une sous-suite  $v_{n'}$  qui converge dans  $H^k(K)$ . Le terme de droite de (6.59) indique que la suite des dérivées  $\partial^\alpha v_{n'}$ , pour tout multi-indice  $|\alpha| = k+1$ , converge vers zéro dans  $L^2(K)$ . Par conséquent,  $v_{n'}$  converge dans  $H^{k+1}(K)$  vers une limite  $v$  qui vérifie (en passant à la limite dans (6.59))

$$|v|_{H^{k+1}(K)} = 0, \quad \|r_K v\|_{H^{k+1}(K)} = 0. \quad (6.60)$$

La première égalité de (6.60) montre que  $v \in \mathbb{P}_k$  car  $K$  est connexe (par application réitérée de la Proposition 4.2.5). Par la définition (6.56) de  $r_K$  on a  $r_K v = v$  pour  $v \in \mathbb{P}_k$ . La deuxième égalité de (6.60) montre donc que  $r_K v = v = 0$ , ce qui est une contradiction avec la limite du terme de gauche de (6.59). Pour obtenir (6.57) on applique (6.58) à  $(v - r_K v)$  en remarquant que  $r_K(v - r_K v) = 0$  et que  $|v - r_K v|_{H^{k+1}(K)} = |v|_{H^{k+1}(K)}$  puisque les dérivées d'ordre  $k+1$  d'un polynôme de  $\mathbb{P}_k$  sont nulles.  $\square$

L'inconvénient du Lemme 6.3.19 de Bramble-Hilbert est que la constante dans l'inégalité (6.57) dépend de  $K$  de manière non explicite. On en précise la dépendance dans le lemme suivant.

**Lemme 6.3.20** *On suppose que  $k+1 > N/2$  et que  $\text{diam}(K) \leq 1$ . Il existe une constante  $C$  indépendante de  $K$  telle que, pour tout  $v \in H^{k+1}(K)$  on a*

$$\|v - r_K v\|_{H^1(K)} \leq C \frac{(\text{diam}(K))^{k+1}}{\rho(K)} |v|_{H^{k+1}(K)}. \quad (6.61)$$

**Démonstration.** On utilise la Remarque 6.3.10 qui affirme que tout  $N$ -simplexe  $K$  est l'image par une transformation affine du  $N$ -simplexe de référence  $K_0$ , défini par (6.46). Autrement dit, il existe une matrice inversible  $B$  et un vecteur  $b$  (dépendant de  $K$ ) tel que, pour tout  $x \in K$ , il existe  $x_0 \in K_0$  vérifiant

$$x = Bx_0 + b. \quad (6.62)$$

Pour obtenir (6.61), on part de l'inégalité (6.57) établie dans  $K_0$  et on applique le changement de variable (6.62). Cela va permettre de trouver la dépendance par rapport à  $K$  de la constante dans cette inégalité. Nous ne détaillons pas ce calcul dont nous indiquons simplement les principales étapes (en cas de besoin le lecteur pourra consulter [36]). Le Jacobien du changement de variable étant  $\det(B)$ , et les dérivées dans  $K$  s'obtenant à partir des dérivées

dans  $K_0$  par composition avec  $B^{-1}$ , il existe une constante  $C$ , indépendante de  $K$ , telle que, pour toute fonction régulière  $v(x)$  avec  $v_0(x_0) = v(Bx_0 + b)$ , on a

$$\begin{aligned} |v_0|_{H^l(K_0)} &\leq C \|B\|^l |\det(B)|^{-1/2} |v|_{H^l(K)} \\ |v|_{H^l(K)} &\leq C \|B^{-1}\|^l |\det(B)|^{1/2} |v_0|_{H^l(K_0)}. \end{aligned}$$

On déduit donc de (6.57)

$$\begin{aligned} |v - r_K v|_{H^1(K)} &\leq C \|B\|^{k+1} \|B^{-1}\| |v|_{H^{k+1}(K)} \\ \|v - r_K v\|_{L^2(K)} &\leq C \|B\|^{k+1} |v|_{H^{k+1}(K)}. \end{aligned}$$

Par ailleurs, on vérifie facilement que

$$\|B\| \leq \frac{\text{diam}(K)}{\rho(K_0)}, \quad \|B^{-1}\| \leq \frac{\text{diam}(K_0)}{\rho(K)}.$$

En combinant ces résultats on obtient (6.61).  $\square$

**Démonstration de la Proposition 6.3.16.** Par construction, si  $v \in H^{k+1}(\Omega)$ , son interpolée  $r_h v$  restreinte au  $N$ -simplexe  $K$  est simplement  $r_K v$ . Par conséquent,

$$\|v - r_h v\|_{H^1(\Omega)}^2 = \sum_{K_i \in \mathcal{T}_h} \|v - r_{K_i} v\|_{H^1(K_i)}^2.$$

On applique la majoration (6.61) à chaque maille  $K_i$  (avec la même constante  $C$  pour toutes), et comme le maillage est régulier l'inégalité (6.49) permet de majorer uniformément le rapport  $\text{diam}(K_i)/\rho(K_i)$ . On en déduit

$$\|v - r_h v\|_{H^1(\Omega)}^2 \leq Ch^{2k} \sum_{K_i \in \mathcal{T}_h} |v|_{H^{k+1}(K_i)}^2 \leq Ch^{2k} \|v\|_{H^{k+1}(\Omega)}^2$$

ce qui est le résultat désiré.  $\square$

**Exercice 6.3.18** Montrer que pour une suite de maillages réguliers, et pour des éléments finis  $\mathbb{P}_1$ , l'opérateur d'interpolation  $r_h$  vérifie en dimension  $N = 2$  ou  $3$

$$\|v - r_h v\|_{L^2(\Omega)} \leq Ch^2 \|v\|_{H^2(\Omega)}.$$

### 6.3.3 Éléments finis rectangulaires

Si le domaine  $\Omega$  est de type rectangulaire (c'est-à-dire que  $\Omega$  est un ouvert polyédrique dont les faces sont perpendiculaires aux axes), on peut le mailler par des rectangles (voir la Figure 6.20) et utiliser une méthode d'éléments finis adaptée. Nous allons définir des éléments finis de type Lagrange (c'est-à-dire dont les degrés de liberté sont des valeurs ponctuelles de fonctions), dits éléments finis  $\mathbb{Q}_k$ . Commençons par définir un  $N$ -rectangle  $K$  de  $\mathbb{R}^N$  comme le pavé (non-dégénéré)  $\prod_{i=1}^N [l_i, L_i]$  avec  $-\infty < l_i < L_i < +\infty$ . On note  $(a_j)_{1 \leq j \leq 2^N}$  les sommets de  $K$ .

**Définition 6.3.21** Soit  $\Omega$  un ouvert connexe polyédrique de  $\mathbb{R}^N$ . Un **maillage rectangulaire** de  $\bar{\Omega}$  est un ensemble  $\mathcal{T}_h$  de  $N$ -rectangles (non dégénérés)  $(K_i)_{1 \leq i \leq n}$  qui vérifient



FIGURE 6.20 – Exemple de maillage rectangulaire en dimension  $N = 2$ .

1.  $K_i \subset \overline{\Omega}$  et  $\overline{\Omega} = \cup_{i=1}^n K_i$ ,
2. l'intersection  $K_i \cap K_j$  de deux  $N$ -rectangles distincts est un  $m$ -rectangle, avec  $0 \leq m \leq N - 1$ , dont tous les sommets sont aussi des sommets de  $K_i$  et  $K_j$ . (En dimension  $N = 2$ , l'intersection de deux rectangles est soit vide, soit un sommet commun, soit une face commune entière.)

Les **sommets** ou **nœuds** du maillage  $\mathcal{T}_h$  sont les sommets des  $N$ -rectangles  $K_i$  qui le composent. Par convention, le paramètre  $h$  désigne le maximum des diamètres des  $N$ -rectangles  $K_i$ .

Nous définissons l'ensemble  $\mathbb{Q}_k$  des polynômes à coefficients réels de  $\mathbb{R}^N$  dans  $\mathbb{R}$  de degré inférieur ou égal à  $k$  **par rapport à chaque variable**, c'est-à-dire que tout  $p \in \mathbb{Q}_k$  s'écrit sous la forme

$$p(x) = \sum_{0 \leq i_1 \leq k, \dots, 0 \leq i_N \leq k} \alpha_{i_1, \dots, i_N} x_1^{i_1} \cdots x_N^{i_N} \text{ avec } x = (x_1, \dots, x_N).$$

Remarquons que le degré total de  $p$  peut être supérieur à  $k$ , ce qui différencie l'espace  $\mathbb{Q}_k$  de  $\mathbb{P}_k$ .

Pour tout entier  $k \geq 1$  on définit le **treillis d'ordre  $k$**  du  $N$ -rectangle  $K$  comme l'ensemble

$$\Sigma_k = \left\{ x \in K \text{ tel que } \frac{x_j - l_j}{L_j - l_j} \in \left\{ 0, \frac{1}{k}, \dots, \frac{k-1}{k}, 1 \right\} \text{ pour } 1 \leq j \leq N \right\}. \quad (6.63)$$

Pour  $k = 1$  il s'agit de l'ensemble des sommets de  $K$ , et pour  $k = 2$  et  $N = 2$  des sommets, des points milieux des arêtes reliant deux sommets, et du barycentre (voir la Figure 6.21).

Le treillis  $\Sigma_k$  d'un  $N$ -rectangle  $K$  est **unisolvant** pour  $\mathbb{Q}_k$ , c'est-à-dire qu'il permet de caractériser tous les polynômes de  $\mathbb{Q}_k$ .

**Lemme 6.3.22** Soit  $K$  un  $N$ -rectangle. Soit un entier  $k \geq 1$ . Alors, tout polynôme de  $\mathbb{Q}_k$  est déterminé de manière unique par ses valeurs aux points du treillis d'ordre  $k$ ,  $\Sigma_k$ , défini par (6.63).

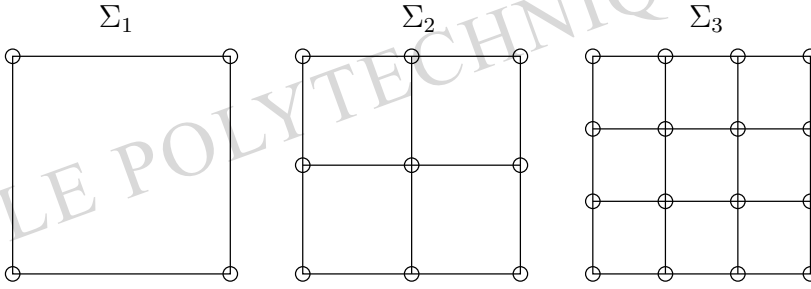


FIGURE 6.21 – Treillis d'ordre 1, 2, et 3 pour un rectangle (les ronds représentent les points du treillis).

**Démonstration.** On vérifie que le cardinal de  $\Sigma_k$  et la dimension de  $\mathbb{Q}_k$  coïncident

$$\text{card}(\Sigma_k) = \dim(\mathbb{Q}_k) = (k+1)^N.$$

Comme l'application qui, à tout polynôme de  $\mathbb{Q}_k$ , fait correspondre ses valeurs sur le treillis  $\Sigma_k$  est linéaire, il suffit d'exhiber une base de  $\mathbb{Q}_k$  dont les éléments valent 1 en un point du treillis et 0 ailleurs pour démontrer le résultat. Soit un point  $x^\mu$  de  $\Sigma_k$  défini par

$$\frac{x_j^\mu - l_j}{L_j - l_j} = \frac{\mu_j}{k} \quad \text{avec } 0 \leq \mu_j \leq k, \quad \forall j \in \{1, \dots, N\}.$$

On définit le polynôme  $p \in \mathbb{Q}_k$  par

$$p(x) = \prod_{j=1}^N \left( \prod_{\substack{i=0 \\ i \neq \mu_j}}^k \frac{k(x_j - l_j) - i(L_j - l_j)}{(\mu_j - i)(L_j - l_j)} \right) \quad \text{avec } x = (x_1, \dots, x_N).$$

On vérifie facilement que  $p(x^\mu) = 1$  tandis que  $p$  s'annule sur tous les autres points de  $\Sigma_k$ , ce qui est le résultat désiré.  $\square$

Comme dans le cas triangulaire nous avons la condition suivante de continuité à travers une face (nous laissons la démonstration, tout à fait similaire à celle du Lemme 6.3.4, au lecteur).

**Lemme 6.3.23** *Soit  $K$  et  $K'$  deux  $N$ -rectangles ayant une face commune  $\Gamma = \partial K \cap \partial K'$ . Soit un entier  $k \geq 1$ . Alors, leur treillis d'ordre  $k$   $\Sigma_k$  et  $\Sigma'_k$  coïncident sur cette face  $\Gamma$ . De plus, étant donné  $p_K$  et  $p_{K'}$  deux polynômes de  $\mathbb{Q}_k$ , la fonction  $v$  définie par*

$$v(x) = \begin{cases} p_K(x) & \text{si } x \in K \\ p_{K'}(x) & \text{si } x \in K' \end{cases}$$

*est continue sur  $K \cup K'$ , si et seulement si  $p_K$  et  $p_{K'}$  ont des valeurs qui coïncident aux points du treillis sur la face commune  $\Gamma$ .*

En pratique, on utilise surtout les espaces  $\mathbb{Q}_1$  et  $\mathbb{Q}_2$ . La Figure 6.22 montre une fonction de base  $\mathbb{Q}_1$  en dimension  $N = 2$  (on peut y vérifier que les fonctions de  $\mathbb{Q}_1$  ne sont pas affines par morceaux comme celles de  $\mathbb{P}_1$ ).

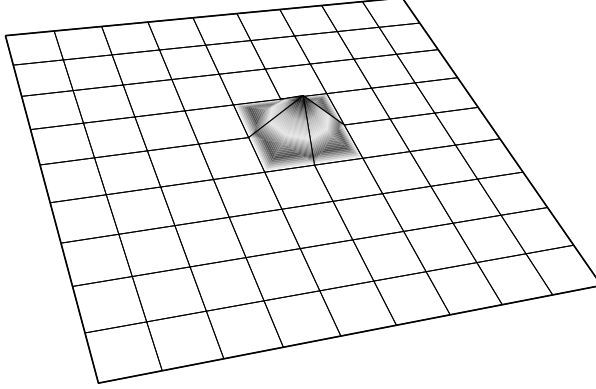


FIGURE 6.22 – Fonction de base  $\mathbb{Q}_1$  en dimension  $N = 2$ .

**Exercice 6.3.19** Soit  $K = [0, 1]^2$  le cube unité en dimension  $N = 2$  de sommets  $a^1 = (0, 0)$ ,  $a^2 = (1, 0)$ ,  $a^3 = (1, 1)$ ,  $a^4 = (0, 1)$ . On définit  $x_3 = 1 - x_1$ ,  $x_4 = 1 - x_2$ , et  $\bar{i}$  comme la valeur de  $i$  modulo 4. Grâce à ses notations, chaque sommet  $a^i$  est défini par  $x_{\bar{i}} = x_{\bar{i}+1} = 0$ . Vérifier que les fonctions de base de  $\mathbb{Q}_1$  sont

$$p_i(x) = x_{\bar{i}+2}x_{\bar{i}+3} \quad \text{pour } 1 \leq i \leq 4,$$

et que celles de  $\mathbb{Q}_2$  sont

$$\begin{aligned} P_i(x) &= x_{\bar{i}+2}(2x_{\bar{i}+2} - 1)x_{\bar{i}+3}(2x_{\bar{i}+3} - 1) & \text{pour } 1 \leq i \leq 4 \\ P_i(x) &= -4x_{\bar{i}+2}(x_{\bar{i}+2} - 1)x_{\bar{i}+3}(2x_{\bar{i}+3} - 1) & \text{pour } 5 \leq i \leq 8 \\ P_9(x) &= 16x_1x_2x_3x_4. \end{aligned}$$

**Remarque 6.3.24** En pratique, on remplace parfois l'élément fini  $\mathbb{Q}_2$  par un autre élément fini plus simple, et tout aussi efficace, noté  $\mathbb{Q}_2^*$ . En dimension  $N = 2$ , l'élément fini  $\mathbb{Q}_2^*$  est défini par les 8 fonctions de base  $(p_i)_{1 \leq i \leq 8}$  de l'Exercice 6.3.19 (on a enlevé la dernière  $p_9$ ). On vérifie que les degrés de liberté de  $\mathbb{Q}_2^*$  sont les sommets et les milieux des arêtes du rectangle (mais pas son barycentre). En dimension  $N = 3$ , l'élément fini  $\mathbb{Q}_2^*$  est défini par ses degrés de liberté qui sont les 8 sommets et les 12 milieux des arêtes du cube (il n'y a pas de degrés de liberté à l'intérieur). •

**Définition 6.3.25** Étant donné un maillage rectangulaire  $\mathcal{T}_h$  d'un ouvert  $\Omega$ , la méthode des éléments finis  $\mathbb{Q}_k$  est définie par l'espace discret

$$V_h = \{v \in C(\bar{\Omega}) \text{ tel que } v|_{K_i} \in \mathbb{Q}_k \text{ pour tout } K_i \in \mathcal{T}_h\}. \quad (6.64)$$

On appelle **nœuds des degrés de liberté** l'ensemble des points  $(\hat{a}_i)_{1 \leq i \leq n_{dl}}$  des treillis d'ordre  $k$  de chacun des  $N$ -rectangles  $K_i \in \mathcal{T}_h$ .

Comme dans le cas triangulaire, la Définition 6.3.25 a un sens grâce à la proposition suivante (dont nous laissons la démonstration au lecteur en guise d'exercice).

**Proposition 6.3.26** *L'espace  $V_h$ , défini par (6.64), est un sous-espace de  $H^1(\Omega)$  dont la dimension est le nombre de degrés de liberté  $n_{dl}$ . De plus, il existe une base de  $V_h$   $(\phi_i)_{1 \leq i \leq n_{dl}}$  définie par*

$$\phi_i(\hat{a}_j) = \delta_{ij} \quad 1 \leq i, j \leq n_{dl},$$

telle que

$$v(x) = \sum_{i=1}^{n_{dl}} v(\hat{a}_i) \phi_i(x).$$

Comme les éléments finis  $\mathbb{Q}_k$  sont des éléments finis de type Lagrange, on peut démontrer les mêmes résultats de convergence que pour la méthode des éléments finis  $\mathbb{P}_k$ . Nous laissons au lecteur le soin de vérifier que la démonstration du Théorème 6.3.13 s'applique "mutatis mutandis" au théorème suivant (la Définition 6.3.11 de maillages réguliers s'étend aisément aux maillages rectangulaires).

**Théorème 6.3.27** *Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages rectangulaires réguliers de  $\Omega$ . Soit  $u \in H_0^1(\Omega)$ , la solution exacte du problème de Dirichlet (6.36), et  $u_h \in V_{0h}$ , la solution approchée par la méthode des éléments finis  $\mathbb{Q}_k$ . Alors la méthode des éléments finis  $\mathbb{Q}_k$  converge, c'est-à-dire que*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{H^1(\Omega)} = 0.$$

De plus, si  $u \in H^{k+1}(\Omega)$  et si  $k+1 > N/2$ , alors on a l'estimation d'erreur

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch^k \|u\|_{H^{k+1}(\Omega)},$$

où  $C$  est une constante indépendante de  $h$  et de  $u$ .

**Remarque 6.3.28** On peut généraliser un peu la notion de maillage rectangulaire et d'éléments finis  $\mathbb{Q}_k$  en utilisant la notion de transformation affine. On appelle  $N$ -parallélotope l'image par une application affine  $F$  du cube unité  $[0, 1]^N$  (un 2-parallélotope est un parallélogramme). On vérifie qu'un  $N$ -parallélotope a  $2N$  faces, parallèles 2 à 2, et que son treillis est bien l'image du treillis du cube unité. On peut alors mailler un domaine  $\Omega$  par des  $N$ -parallélotopes et définir une méthode d'éléments finis basée sur l'image  $F(\mathbb{Q}_k)$ , dans chaque  $N$ -parallélotope, de l'espace  $\mathbb{Q}_k$  pour le cube unité (en général  $F(\mathbb{Q}_k) \neq \mathbb{Q}_k$ ). On peut aussi utiliser des transformations plus compliquées (non affines) : c'est la méthode des éléments finis isoparamétriques (voir la Remarque 6.3.18 à ce sujet). Par exemple, en dimension  $N = 2$ , l'utilisation de transformations  $\mathbb{Q}_1$  permet de mailler un domaine avec des quadrangles quelconques (à faces non parallèles). Pour plus de détails, nous renvoyons à [36].

•

**Remarque 6.3.29** On peut aussi mailler une partie de  $\Omega$  en  $N$ -simplexes, et une autre en  $N$ -rectangles et construire une méthode d'éléments finis mélangeant les deux types  $\mathbb{P}_k$  et  $\mathbb{Q}_k$ . Pour plus de détails, nous renvoyons encore à [36]. •

**Remarque 6.3.30** On peut définir des éléments finis intermédiaires entre  $\mathbb{P}_k$  et  $\mathbb{Q}_k$  en dimension  $N = 3$ , appelés “éléments finis prismatiques d'ordre  $k$ ”. Supposons que  $\Omega = \omega \times ]0, L[$  avec  $\omega$  un ouvert de  $\mathbb{R}^2$ . On maille  $\omega$  par des triangles  $T_i$ , et  $]0, L[$  par des segments  $[z_j, z_{j+1}]$ . On définit alors des prismes de  $\mathbb{R}^3$  comme le produit  $T_i \times [z_j, z_{j+1}]$ , avec lesquels on maille  $\Omega$ . On construit alors des fonctions de base intermédiaires entre celles de  $\mathbb{P}_k$  et  $\mathbb{Q}_k$  sur ces prismes. Pour plus de détails, nous renvoyons à [36]. •

### 6.3.4 Éléments finis pour Stokes

La généralisation de la méthode des éléments finis à des systèmes d'équations aux dérivées partielles (comme le système de l'élasticité linéarisée) ne pose pas de problèmes particuliers. Ce n'est pas le cas pour le système des équations de Stokes (5.71) à cause de la condition d'incompressibilité du fluide (ou condition de divergence nulle pour la vitesse). L'importance pratique considérable des simulations numériques en mécanique des fluides incompressibles justifie que nous parlions brièvement de ce cas particulier (qui permet aussi de montrer que l'analyse numérique n'est pas toujours ce long fleuve tranquille que l'on imagine à la lecture de ce polycopié).

Rappelons que, dans un domaine borné connexe  $\Omega \subset \mathbb{R}^N$ , en présence de forces extérieures  $f(x)$ , et pour des conditions aux limites d'adhérence du fluide aux parois, les équations de Stokes s'écrivent

$$\begin{cases} \nabla p - \mu \Delta u = f & \text{dans } \Omega \\ \operatorname{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \quad (6.65)$$

où  $\mu > 0$  est la viscosité du fluide. Dans la Sous-section 5.3.2 nous avons proposé comme formulation variationnelle de (6.65)

$$\text{Trouver } u \in V \text{ tel que } \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx = \int_{\Omega} f \cdot v \, dx \quad \forall v \in V, \quad (6.66)$$

où  $V$  est l'espace de Hilbert défini par

$$V = \{v \in H_0^1(\Omega)^N \text{ tel que } \operatorname{div} v = 0 \text{ p.p. dans } \Omega\}. \quad (6.67)$$

Comme  $V$  contient la contrainte d'incompressibilité  $\operatorname{div} v = 0$ , il est très difficile en pratique de construire des approximations variationnelles internes de (6.66) comme nous l'avons fait jusqu'ici. Plus précisément, la difficulté est de définir simplement (explicitement) un sous-espace  $V_h$  de  $V$  de dimension finie dont les éléments s'écrivent grâce aux fonctions de base des éléments finis  $\mathbb{P}_k$  ou  $\mathbb{Q}_k$ . Par exemple, si  $\mathcal{T}_h = (K_i)_{1 \leq i \leq n}$  est un maillage triangulaire de l'ouvert connexe polyédrique  $\Omega$ , on peut définir

$$V_h = \{v \in C(\overline{\Omega})^N \text{ tel que } \operatorname{div} v = 0 \text{ dans } \Omega, \, v|_{K_i} \in \mathbb{P}_k^N \text{ pour tout } K_i \in \mathcal{T}_h\},$$

mais il n'est pas clair que  $V_h$  ne soit pas "trop petit" et de quelle manière on peut caractériser ses éléments en termes de degré de liberté. En particulier, la condition  $\operatorname{div} v = 0$  dans la définition de  $V_h$  mélange toutes les composantes de  $v$ , ce qui rend très difficile et compliqué la caractérisation d'une base explicite de  $V_h$ . On n'utilise donc pas la formulation variationnelle (6.67) pour définir une méthode d'éléments finis.

En pratique, on introduit une autre formulation variationnelle des équations de Stokes qui consiste à ne pas forcer l'incompressibilité dans la définition de l'espace et à garder la pression comme inconnue dans la formulation variationnelle. En multipliant la première équation de (6.65) par une fonction test  $v \in H_0^1(\Omega)^N$  et la deuxième équation par une autre fonction test  $q \in L^2(\Omega)$ , on obtient après intégration par parties : trouver  $(u, p) \in H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$  tel que

$$\begin{cases} \int_{\Omega} \mu \nabla u \cdot \nabla v \, dx - \int_{\Omega} p \operatorname{div} v \, dx = \int_{\Omega} f \cdot v \, dx \\ \int_{\Omega} q \operatorname{div} u \, dx = 0, \end{cases} \quad (6.68)$$

pour tout  $(v, q) \in H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$ . Un intérêt supplémentaire de (6.68) est que la pression n'y est pas éliminée comme dans (6.66). Il sera donc possible de calculer cette variable (physiquement importante) avec (6.68). Nous laissons au lecteur le soin de vérifier en guise d'exercice le résultat suivant.

**Lemme 6.3.31** *Soit  $(u, p) \in H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$ . Le couple  $(u, p)$  est solution de (6.68) si et seulement s'il est solution (faible, au sens du Théorème 5.3.8) des équations de Stokes (6.65).*

Il est alors facile de construire une approximation variationnelle interne de (6.68). On introduit les espaces discrets

$$\begin{cases} V_{0h} = \{v \in C(\overline{\Omega})^N \text{ tel que } v|_{K_i} \in \mathbb{P}_k^N \text{ pour tout } K_i \in \mathcal{T}_h \text{ et } v = 0 \text{ sur } \partial\Omega\}, \\ Q_h = \{q \in C(\overline{\Omega})/\mathbb{R} \text{ tel que } q|_{K_i} \in P_{k'} \text{ pour tout } K_i \in \mathcal{T}_h\}, \end{cases}$$

qui vérifient bien que  $V_{0h} \times Q_h$  est un sous-espace de  $H_0^1(\Omega)^N \times L^2(\Omega)/\mathbb{R}$  de dimension finie. L'approximation variationnelle interne de (6.68) est simplement

$$\begin{cases} \int_{\Omega} \mu \nabla u_h \cdot \nabla v_h \, dx - \int_{\Omega} p_h \operatorname{div} v_h \, dx = \int_{\Omega} f \cdot v_h \, dx \\ \int_{\Omega} q_h \operatorname{div} u_h \, dx = 0, \end{cases} \quad (6.69)$$

pour tout  $(v_h, q_h) \in V_{0h} \times Q_h$ . Expliquons comment on résout (6.69) en pratique. En notant  $n_V$  la dimension de  $V_{0h}$  et  $n_Q$  celle de  $Q_h$ , on introduit la base  $(\phi_j)_{1 \leq j \leq n_V}$  de  $V_{0h}$  et la base  $(\psi_j)_{1 \leq j \leq n_Q}$  de  $Q_h$  construite avec les fonctions de base des éléments

finis (voir la Proposition 6.3.7). On décompose  $u_h$  et  $p_h$  sur ces bases

$$u_h(x) = \sum_{j=1}^{n_V} u_h(\hat{a}_j) \phi_j(x), \quad p_h(x) = \sum_{j=1}^{n_Q} p_h(\hat{a}'_j) \psi_j(x),$$

En notant  $U_h = (u_h(\hat{a}_j))_{1 \leq j \leq n_V}$  et  $P_h = (p_h(\hat{a}'_j))_{1 \leq j \leq n_Q}$ , on obtient le système linéaire suivant

$$\begin{pmatrix} A_h & B_h^* \\ B_h & 0 \end{pmatrix} \begin{pmatrix} U_h \\ P_h \end{pmatrix} = \begin{pmatrix} b_h \\ 0 \end{pmatrix}, \quad (6.70)$$

où  $B_h^*$  est la matrice adjointe (ou transposée) de  $B_h$ ,  $b_h = (\int_{\Omega} f \cdot \phi_i dx)_{1 \leq i \leq n_V}$ , et

$$A_h = \left( \mu \int_{\Omega} \nabla \phi_i \cdot \nabla \phi_j dx \right)_{1 \leq i, j \leq n_V}, \quad B_h = \left( - \int_{\Omega} \psi_i \operatorname{div} \phi_j dx \right)_{1 \leq i \leq n_Q, 1 \leq j \leq n_V}.$$

Les choses se compliquent lorsqu'il s'agit de savoir si on peut toujours résoudre le système linéaire (6.70) de manière unique. Remarquons que la matrice  $A_h$  est symétrique définie positive d'ordre  $n_V$ , que la matrice  $B_h$  est rectangulaire de taille  $n_Q \times n_V$ , et que, si la matrice globale de (6.70) est symétrique d'ordre  $n_V + n_Q$ , elle n'est pas définie positive. Néanmoins on a le résultat suivant.

**Lemme 6.3.32** *Le système linéaire (6.70) admet toujours une solution  $(U_h, P_h)$  dans  $\mathbb{R}^{n_V} \times \mathbb{R}^{n_Q}$ . Le vecteur  $U_h$  est unique, tandis que  $P_h$  est unique à l'addition près d'un élément de  $\operatorname{Ker} B_h^*$ .*

**Démonstration.** Comme  $(\operatorname{Ker} B_h)^\perp = \operatorname{Im} B_h^*$ , il est facile de voir que (6.70) est équivalent à

trouver  $U_h \in \operatorname{Ker} B_h$  tel que  $A_h U_h \cdot W_h = b_h \cdot W_h$  pour tout  $W_h \in \operatorname{Ker} B_h$ .

Il suffit alors d'appliquer le Théorème de Lax-Milgram 3.3.1 pour obtenir l'existence et l'unicité de  $U_h$  dans  $\operatorname{Ker} B_h$ . Par conséquent, (6.70) admet au moins une solution  $(U_h, P_h)$  dans  $\mathbb{R}^{n_V} \times \mathbb{R}^{n_Q}$ . Comme  $U_h$  doit appartenir à  $\operatorname{Ker} B_h$ , il est unique dans  $\mathbb{R}^{n_V}$ . Par ailleurs, on vérifie aisément que  $P_h$  est unique à l'addition près d'un élément de  $\operatorname{Ker} B_h^*$ .  $\square$

Là où les choses se compliquent, c'est que le noyau  $\operatorname{Ker} B_h^*$  n'est jamais réduit au vecteur nul et qu'il peut parfois être très "gros". Tout dépend du choix qui est fait des ordres  $k$  et  $k'$  des éléments finis pour la vitesse et pour la pression.

**Lemme 6.3.33** *Le noyau  $\operatorname{Ker} B_h^*$  contient au moins le vecteur  $\mathbb{I}$  de  $\mathbb{R}^{n_Q}$  dont toutes les composantes sont égales à 1. Autrement dit, la pression discrète  $p_h$  est au moins définie à une constante près.*

**Démonstration.** Soit  $r_h \in Q_h$  et  $w_h \in V_{0h}$ . Par définition

$$W_h \cdot B_h^* R_h = B_h W_h \cdot R_h = \int_{\Omega} r_h \operatorname{div} w_h \, dx.$$

Or  $r_h = 1$  appartient toujours à  $Q_h$ , et comme

$$\int_{\Omega} \operatorname{div} w_h \, dx = \int_{\partial\Omega} w_h \cdot n \, ds = 0$$

pour tout  $w_h \in V_{0h}$ , on en déduit que  $R_h = \mathbb{I} = (1, \dots, 1)$  appartient à  $\operatorname{Ker} B_h^*$ .  $\square$

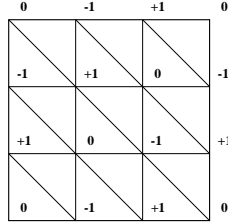


FIGURE 6.23 – Mode instable de la pression sur un maillage triangulaire uniforme (éléments finis  $\mathbb{P}_1$  pour la vitesse et la pression).

**Lemme 6.3.34** *Lorsque  $k = 2$  et  $k' = 1$  (éléments finis  $\mathbb{P}_2$  pour la vitesse et  $\mathbb{P}_1$  pour la pression), le noyau  $\operatorname{Ker} B_h^*$  est de dimension un, engendré par le vecteur  $\mathbb{I}$  (autrement dit, la pression discrète  $p_h$  est unique à une constante près).*

*Lorsque  $k = k' = 1$  (éléments finis  $\mathbb{P}_1$  pour la vitesse et la pression), le noyau  $\operatorname{Ker} B_h^*$  est en général de dimension strictement plus grande que un (autrement dit, la pression discrète  $p_h$  n'est pas unique, même à une constante près).*

**Démonstration.** Soit  $r_h \in Q_h$  et  $w_h \in V_{0h}$ . Par définition

$$W_h \cdot B_h^* R_h = B_h W_h \cdot R_h = \int_{\Omega} r_h \operatorname{div} w_h \, dx = - \int_{\Omega} \nabla r_h \cdot w_h \, dx.$$

Lorsque  $k = 2$  et  $k' = 1$ , le gradient  $\nabla r_h$  est constant dans chaque maille  $K_i$ , donc

$$\int_{\Omega} \nabla r_h \cdot w_h \, dx = \sum_{i=1}^n \nabla r_h(K_i) \cdot \int_{K_i} w_h \, dx.$$

Or la formule de quadrature de l'Exercice 6.3.7 nous dit que, pour  $w_h \in \mathbb{P}_2$ ,

$$\int_K w_h \, dx = \frac{|K|}{N(N+1)/2} \sum_j w_h(a_{ij})$$



où les  $(a_{ij})$  sont les  $N(N+1)/2$  points milieux des arêtes reliant les sommets  $a_i$  et  $a_j$  de  $K$ . En regroupant la somme suivant ces points milieux (qui sont commun à deux mailles), on obtient

$$\int_{\Omega} \nabla r_h \cdot w_h dx = \sum_{a_{ij}} w_h(a_{ij}) \cdot \left( \frac{|K_i|}{N(N+1)/2} \nabla r_h(K_i) + \frac{|K_j|}{N(N+1)/2} \nabla r_h(K_j) \right).$$

En prenant  $w_h$  qui vaut 1 sur le point milieu  $a_{ij}$  et 0 ailleurs, on en déduit que  $W_h \cdot B_h^* R_h = 0$  implique que

$$|K_i| \nabla r_h(K_i) + |K_j| \nabla r_h(K_j) = 0. \quad (6.71)$$

La fonction  $r_h$  a donc un gradient de direction constante mais dont l'orientation change de sens d'une maille à l'autre. Comme  $r_h$  appartient à  $\mathbb{P}_1$  et est continue sur  $\Omega$ , son gradient tangentiel est continu à l'interface entre deux mailles. Par conséquent, il est nul, et la seule possibilité dans (6.71) est que le gradient de  $r_h$  soit nul partout, c'est-à-dire que  $r_h$  soit une fonction constante. En conclusion, on a montré que  $W_h \cdot B_h^* R_h = 0$  pour tout  $W_h$  implique que  $R_h = \text{Cste} \mathbb{I}$ , qui est le résultat recherché.

Donnons un contre-exemple en dimension deux d'espace lorsque  $k = k' = 1$ . On reprend le maillage triangulaire uniforme du carré  $\Omega = ]0, 1]^2$  (voir la Figure 6.12). On définit la fonction  $p_0 \in Q_h$ , avec  $k' = 1$ , par ses valeurs  $-1, 0, +1$  aux trois sommets de chaque triangle  $K_i$  (voir la Figure 6.23). Toujours par définition, on a

$$B_h W_h \cdot R_h = \int_{\Omega} r_h \operatorname{div} w_h dx.$$

Mais comme  $w_h$  est affine par morceaux sur chaque maille  $K_i$ , sa divergence est constante dans chaque  $K_i$  et on a

$$\int_{\Omega} r_h \operatorname{div} w_h dx = \sum_{i=1}^n \operatorname{div} w_h(K_i) \int_{K_i} r_h dx$$

qui vaut zéro pour  $r_h = p_0$  car  $\int_{K_i} p_0 dx = \frac{|K_i|}{3}(0 + 1 - 1) = 0$ . Par conséquent  $p_0$  engendre un nouveau vecteur de  $\operatorname{Ker} B_h^*$ , en plus de  $\mathbb{I}$ .  $\square$

Dans la pratique, **dès que la dimension de  $\operatorname{Ker} B_h^*$  est strictement plus grande que un, la méthode des éléments finis correspondante est inutilisable**. En effet, si  $\dim(\operatorname{Ker} B_h^*) > 1$ , le calcul numérique des solutions du système linéaire (6.70) va conduire à des oscillations numériques sur la pression : l'algorithme hésite entre plusieurs pressions discrètes  $P_h$  dont la différence appartient à  $\operatorname{Ker} B_h^*$ . On dit que la méthode est **instable**. Remarquons justement que l'élément  $p_0$  de la démonstration du Lemme 6.3.34 s'interprète comme une oscillation de la pression à l'échelle du maillage. Si  $\dim(\operatorname{Ker} B_h^*) = 1$ , on élimine facilement l'indétermination sur la pression discrète  $P_h$  en imposant sa valeur en un nœud, ou bien en précisant sa moyenne sur le domaine  $\Omega$ . Dans tous les cas, il n'y a pas d'indétermination sur la

vitesse  $U_h$  qui est définie de manière unique. Pour plus de détails sur les méthodes d'éléments finis en mécanique des fluides, nous renvoyons à [35].

Nous n'avons rien dit pour l'instant de la résolution pratique du système linéaire (6.70). Pour cela on utilise un algorithme, dit d'Uzawa, issu de la théorie de l'optimisation, que nous verrons au Chapitre 10. Il s'agit là d'un très bel exemple d'interaction entre l'analyse numérique et l'optimisation. Expliquons brièvement l'idée principale (nous y reviendrons en détail au Chapitre 10). On sait que les équations de Stokes sont équivalentes à un problème de minimisation d'une énergie (voir l'Exercice 5.3.10). Nous verrons que, de la même manière, la résolution du système linéaire (6.70) est équivalente à la minimisation suivante

$$J(U_h) = \min_{V_h \in \text{Ker} B_h} J(V_h) \text{ avec } J(V_h) = \frac{1}{2} A_h V_h \cdot V_h - b_h \cdot V_h.$$

L'algorithme d'Uzawa permet précisément de résoudre ce problème de minimisation sous contrainte.

Pour des raisons de coût de calcul, on n'utilise que très rarement la méthode des éléments finis  $\mathbb{P}_2$  pour la vitesse et  $\mathbb{P}_1$  pour la pression. On lui préfère une autre méthode, dite  $\mathbb{P}_1$ /bulle pour la vitesse et  $\mathbb{P}_1$  pour la pression. Il s'agit de la méthode des éléments finis  $\mathbb{P}_1$  pour la vitesse et la pression dans laquelle on enrichit l'espace  $V_{0h}$  des vitesses en ajoutant dans sa base pour chaque maille et pour chaque composante dans  $\mathbb{R}^N$  une fonction **bulle** définie comme le produit  $\lambda_1(x) \dots \lambda_{N+1}(x)$ , où les  $\lambda_j(x)$  sont les coordonnées barycentriques de  $x$  dans la maille  $K_i$ . Comme cette fonction bulle est nulle sur le bord de  $K_i$  et positive à l'intérieur, on lui associe comme degré de liberté le barycentre de la maille. Cette méthode est stable comme le montre l'exercice suivant.

**Exercice 6.3.20** Montrer que pour la méthode des éléments finis  $\mathbb{P}_1$ /bulle pour la vitesse et  $\mathbb{P}_1$  pour la pression on a  $\dim(\text{Ker} B_h^*) = 1$ .

Les instabilités de pression ne sont pas l'apanage des méthodes d'éléments finis. Il existe aussi des méthodes de différences finies qui présentent le même type d'inconvénient, comme le montre l'exercice suivant.

**Exercice 6.3.21** On considère les équations de Stokes (6.65) en dimension  $N = 1$  (ce modèle n'a aucun intérêt puisque sa solution explicite est  $u = 0$  et  $p$  une primitive de  $f$ , mais il permet de bien comprendre les problèmes de discrétisation). Pour  $\Omega = (0, 1)$ , on considère le maillage de points  $x_j = jh$  avec  $h = 1/(n+1)$  et  $0 \leq j \leq n+1$ . On définit la méthode de différences finies centrées (d'ordre 2) suivante

$$\begin{cases} \mu \frac{-u_{j+1} + 2u_j - u_{j-1}}{h^2} + \frac{p_{j+1} - p_{j-1}}{2h} = f(x_j) \text{ pour } 1 \leq j \leq n \\ \frac{u_{j+1} - u_{j-1}}{2h} = 0 \text{ pour } 1 \leq j \leq n \\ u_0 = u_{n+1} = 0. \end{cases}$$

Montrer que ce système d'équations algébriques est mal posé, et en particulier que la pression  $(p_j)$  est définie à l'addition d'une constante près ou d'un multiple d'une pression définie par ses composantes  $(1, 0, 1, 0, \dots, 1, 0)$ .

**Remarque 6.3.35** L'idée de la formulation variationnelle (6.68) s'étend sans problème au Laplacien ou à tout opérateur elliptique. Pour résoudre

$$\begin{cases} -\operatorname{div}(A\nabla u) = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases}$$

en posant  $\sigma = A\nabla u$ , on introduit la formulation variationnelle

$$\begin{cases} -\int_{\Omega} \operatorname{div} \sigma v \, dx = \int_{\Omega} f v \, dx \\ \int_{\Omega} A^{-1} \sigma \cdot \tau \, dx + \int_{\Omega} u \operatorname{div} \tau \, dx = 0, \end{cases}$$

pour tout  $(v, \tau) \in L^2(\Omega) \times H(\operatorname{div})$ . La méthode d'éléments finis qui en découle est différente de celles qui ont été exposées dans ce chapitre. Elle est appelée méthode des éléments finis mixtes. •

### 6.3.5 Visualisation des résultats numériques

Dans cette sous-section nous disons rapidement quelques mots sur la **visualisation** des résultats obtenus par la méthode des éléments finis. Les figures ci-dessous ont été tracées à l'aide du logiciel (libre) graphique xd3d.



FIGURE 6.24 – Carte d'isovaleurs dans une coupe plane d'un problème de diffusion 3-D

La visualisation des résultats pour un problème scalaire (où l'inconnue est à valeurs dans  $\mathbb{R}$ ) est assez simple. En dimension  $N = 2$  on peut tracer des lignes d'isovaleurs et/ou des cartes de grisé d'intensité, comme on peut le voir sur les Figures 6.15 et 6.16. Il faut néanmoins faire attention aux échelles de valeurs comme le prouve la Figure 6.15 où les deux tracés (correspondant au même problème sur deux maillages distincts) sont d'allure comparable mais d'échelles de référence très différentes. En dimension  $N = 3$  on trace des isosurfaces (surfaces où l'inconnue est constante) ou bien des coupes planes.

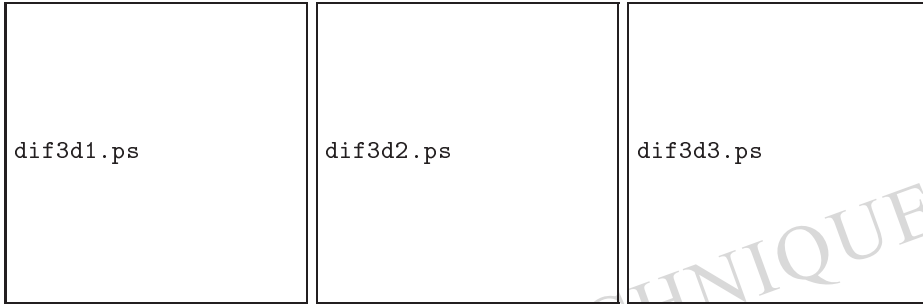


FIGURE 6.25 – Isosurfaces pour un problème de diffusion 3-D (valeurs diminuant de gauche à droite).

En guise d'exemple, nous considérons un problème de diffusion dans l'espace d'une concentration (un polluant par exemple) émis par une source localisée sur le sol (la base du cube). On résout par la méthode des éléments finis  $\mathbb{Q}_1$  le problème de Dirichlet (6.36) avec un second membre nul et une condition aux limites de Dirichlet partout sauf sur la base du cube. Au centre de la base on impose une condition aux limites de Dirichlet "non-homogène"  $u = 1$ , et sur le reste de la base une condition aux limites de Neumann. La Figure 6.24 représente les valeurs de  $u$  dans une coupe, et la Figure 6.25 des isosurfaces de  $u$ .

La visualisation des résultats pour un problème vectoriel (où l'inconnue est à valeurs dans  $\mathbb{R}^N$ ) est différente. Prenons, par exemple, le cas du système de l'élasticité (voir la Sous-section 5.3.1). On peut tracer des flèches représentant le vecteur calculé, mais ce type d'image est difficile à lire et à interpréter. Il est bien plus parlant de tracer la "déformée" du domaine en utilisant l'interprétation physique de la solution (en dimension  $N = 2$  ou 3). Rappelons que le vecteur inconnu  $u(x)$  est le déplacement du point  $x$  sous l'action des forces exercées : par conséquent, le point courant du domaine déformé est  $x + u(x)$ . Nous illustrons ces deux manières de représenter les résultats sur l'exemple d'une poutre encastree sur son bord vertical gauche (condition aux limites de Dirichlet) et libre sur les autres bords (condition aux limites de Neumann) soumise à son poids propre (la force  $f$  est un vecteur constant vertical) ; voir la Figure 6.26. L'avantage de tracer la configuration déformée est qu'on peut y superposer le tracé d'une autre grandeur scalaire comme la norme du tenseur des contraintes.

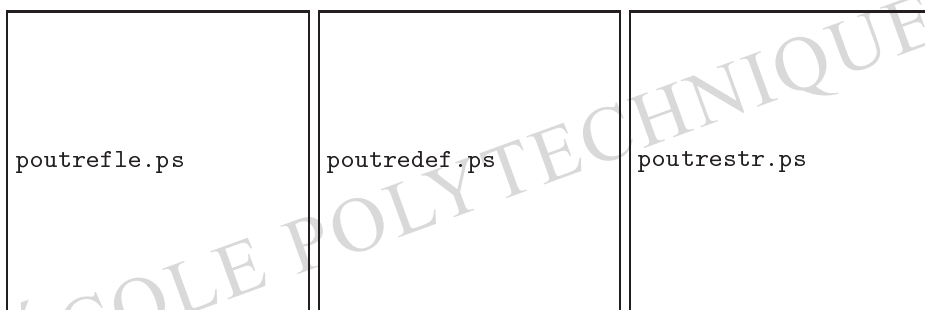


FIGURE 6.26 – De gauche à droite, déplacement, configuration déformée, et norme du tenseur des contraintes (les valeurs les élevées sont les plus foncées) dans une poutre encastrée soumise à son poids propre.

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

## Chapitre 7

# PROBLÈMES AUX VALEURS PROPRES

### 7.1 Motivation et exemples

#### 7.1.1 Introduction

Ce chapitre est consacré 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, nous en déduirons une méthode de résolution générale de ces mêmes problèmes d'évolution que nous mettrons en oeuvre dans le Chapitre 8.

Donnons tout de suite un exemple de **problème aux valeurs propres** pour le Laplacien avec condition aux limites de Dirichlet. Si  $\Omega$  est un ouvert borné de  $\mathbb{R}^N$  on cherche les couples  $(\lambda, u) \in \mathbb{R} \times H_0^1(\Omega)$ , avec  $u \neq 0$ , solutions de

$$\begin{cases} -\Delta u = \lambda u & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (7.1)$$

Le réel  $\lambda$  est appelé **valeur propre**, et la fonction  $u(x)$  **mode propre** ou **fonction propre**. L'ensemble des valeurs propres est appelé le spectre de (7.1). On peut faire l'analogie entre (7.1) et le problème plus simple de détermination des valeurs et vecteurs propres d'une matrice  $A$  d'ordre  $n$ ,

$$Au = \lambda u \quad \text{avec} \quad (\lambda, u) \in \mathbb{R} \times \mathbb{R}^n, \quad (7.2)$$

en affirmant que l'opérateur  $-\Delta$  est une "généralisation" en dimension infinie d'une matrice  $A$  en dimension finie. La résolution de (7.1) 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 (7.5) ou l'équation des ondes (7.7). Néanmoins, les solutions de (7.1) ont aussi une interprétation physique qui leur est propre, par exemple comme modes propres de vibration.

Le plan de ce chapitre est le suivant. Après avoir motivé plus amplement le problème aux valeurs propres (7.1), nous développons dans la Section 7.2 une **théorie spectrale abstraite** dans les espaces de Hilbert. Le but de cette section est de généraliser en dimension infinie le résultat bien connu en dimension finie qui affirme que toute matrice symétrique réelle est diagonalisable dans une base orthonormée. Cette section relève en partie d'un cours de mathématiques "pures", aussi nous insistons à nouveau sur le fait que c'est l'esprit des résultats plus que la lettre des démonstrations qui importe ici. Nous appliquons cette théorie spectrale aux équations aux dérivées partielles elliptiques dans la Section 7.3. En particulier, nous démontrons que le problème spectral (7.1) **admet une infinité dénombrable de solutions**. Enfin, la Section 7.4 est consacrée aux questions d'**approximation numérique** des valeurs propres et fonctions propres d'une équation aux dérivées partielles. En particulier, nous introduisons la notion de **matrice de masse**  $\mathcal{M}$  qui vient compléter celle de matrice de rigidité  $\mathcal{K}$ , et nous montrons que des valeurs propres approchées de (7.1) se calculent comme les valeurs propres du système  $\mathcal{K}u = \lambda \mathcal{M}u$ , ce qui confirme l'analogie entre (7.1) et sa version discrète (7.2).

### 7.1.2 Résolution des problèmes instationnaires

Avant de nous lancer dans les développements abstraits de la prochaine section, montrons en quoi la résolution d'un problème aux valeurs propres permet de résoudre aussi un problème d'évolution. Pour cela nous allons faire une analogie avec la résolution de systèmes différentiels en dimension finie. Dans tout ce qui suit  $A$  désigne une matrice symétrique réelle, définie positive, d'ordre  $n$ . On note  $\lambda_k$  ses valeurs propres et  $r_k$  ses vecteurs propres,  $1 \leq k \leq n$ , tels que  $Ar_k = \lambda_k r_k$ .

On commence par un système différentiel du premier ordre

$$\begin{cases} \frac{\partial u}{\partial t} + Au = 0 & \text{pour } t \geq 0 \\ u(t=0) = u_0, \end{cases} \quad (7.3)$$

où  $u(t)$  est une fonction de classe  $C^1$  de  $\mathbb{R}^+$  dans  $\mathbb{R}^n$ , et  $u_0 \in \mathbb{R}^n$ . Il est bien connu que (7.3) admet une solution unique obtenue en diagonalisant la matrice  $A$ . Plus précisément, la donnée initiale se décompose sous la forme  $u_0 = \sum_{k=1}^n u_k^0 r_k$ , ce qui donne

$$u(t) = \sum_{k=1}^n u_k^0 e^{-\lambda_k t} r_k.$$



Un deuxième exemple est le système différentiel du deuxième ordre

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} + Au = 0 & \text{pour } t \geq 0 \\ u(t=0) = u_0, \\ \frac{\partial u}{\partial t}(t=0) = u_1, \end{cases} \quad (7.4)$$

où  $u(t)$  est une fonction de classe  $C^2$  de  $\mathbb{R}^+$  dans  $\mathbb{R}^n$ , et  $u_0, u_1 \in \mathbb{R}^n$ . En décomposant les données initiales sous la forme  $u_0 = \sum_{k=1}^n u_k^0 r_k$  et  $u_1 = \sum_{k=1}^n u_k^1 r_k$ , (7.4) admet comme solution unique

$$u(t) = \sum_{k=1}^n \left( u_k^0 \cos(\sqrt{\lambda_k} t) + \frac{u_k^1}{\sqrt{\lambda_k}} \sin(\sqrt{\lambda_k} t) \right) r_k.$$

Il est clair sur ces deux exemples que la connaissance du spectre de la matrice  $A$  permet de résoudre les problèmes d'évolution (7.3) et (7.4). Aussi évident soient-ils, ces exemples sont tout à fait représentatifs de la démarche que nous allons suivre dans la suite. Nous allons remplacer la matrice  $A$  par l'opérateur  $-\Delta$ , l'espace  $\mathbb{R}^n$  par l'espace de Hilbert  $L^2(\Omega)$ , et nous allons "diagonaliser" le Laplacien pour résoudre l'équation de la chaleur ou l'équation des ondes.

Afin de se convaincre que (7.1) 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

$$\frac{\partial \mathbf{u}}{\partial t} - \Delta \mathbf{u} = 0, \quad (7.5)$$

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

$$-\Delta u = \lambda u \quad (7.6)$$

muni de conditions aux limites adéquates.

De la même manière, si  $\mathbf{u}$  est solution de l'équation des ondes

$$\frac{\partial^2 \mathbf{u}}{\partial t^2} - \Delta \mathbf{u} = 0, \quad (7.7)$$

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 (7.6). 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.

**Remarque 7.1.1** Nous venons de voir que l'équation des ondes admet des solutions oscillantes périodiquement en temps du type

$$\mathbf{u}(x, t) = e^{-i\omega t} u(x),$$

où  $\omega = \sqrt{\lambda}$  et la fréquence des oscillations et  $u(x)$  est leur amplitude. Il s'agit là d'une caractéristique générale des équations aux dérivées partielles linéaires de type hyperbolique. Ces solutions oscillantes ont un intérêt physique évident et indépendant de la résolution générale d'une équation d'évolution hyperbolique. Elles modélisent typiquement des vibrations (par exemple, élastiques) ou des ondes (par exemple, électromagnétiques), et elles se rencontrent généralement en l'absence de terme source et après un temps d'établissement qui permet "d'oublier" la condition initiale. •

**Exercice 7.1.1** Soit  $\Omega = \mathbb{R}^N$ . Montrer que  $u(x) = \exp(ik \cdot x)$  est une solution de (7.6) si  $|k|^2 = \lambda$ . Une telle solution est appelée onde plane.

Prenons un autre exemple, à savoir l'équation de Schrödinger issue de la mécanique quantique (voir le Chapitre 1).

**Exercice 7.1.2** Soit un potentiel régulier  $V(x)$ . Montrer que, si  $\mathbf{u}(x, t) = e^{-i\omega t} u(x)$  est solution de

$$i \frac{\partial \mathbf{u}}{\partial t} + \Delta \mathbf{u} - V \mathbf{u} = 0 \quad \text{dans } \mathbb{R}^N \times \mathbb{R}_*^+, \quad (7.8)$$

alors  $u(x)$  est solution de

$$-\Delta u + Vu = \omega u \quad \text{dans } \mathbb{R}^N. \quad (7.9)$$

On retrouve le même type de problème spectral que (7.6), à l'addition d'un terme d'ordre zéro près. Pour l'équation de Schrödinger la valeur propre  $\omega$  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 (7.8). 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 7.1.3** 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 (7.9) si  $\omega = \text{tr}(A^{1/2})$ . Une telle solution est appelée état fondamental.

## 7.2 Théorie spectrale

Dans cette section nous introduisons une théorie spectrale abstraite dans les espaces de Hilbert (voir par exemple [27]). Le but ultime des développements qui suivent est de généraliser à la dimension infinie le résultat bien connu en dimension finie qui affirme que toute matrice symétrique réelle est diagonalisable dans une base orthonormée. En première lecture on pourra admettre tous les résultats de cette section.

### 7.2.1 Généralités

Dans tout ce qui suit  $V$  désigne un espace de Hilbert réel muni d'un produit scalaire  $\langle x, y \rangle$ .

**Définition 7.2.1** Soit  $A$  une application linéaire continue de  $V$  dans  $V$ . On appelle valeur propre de  $A$  un réel  $\lambda \in \mathbb{R}$  tel qu'il existe un élément non nul  $x \in V$  qui vérifie  $Ax = \lambda x$ . Un tel vecteur  $x$  est appelé vecteur propre associé à la valeur propre  $\lambda$ .

**Théorème 7.2.2** Soit  $A$  une application linéaire continue de  $V$  dans  $V$ . Il existe une unique application linéaire continue  $A^*$  de  $V$  dans  $V$ , dite adjointe, telle que

$$\langle Ax, y \rangle = \langle x, A^*y \rangle \quad \forall x, y \in V.$$

**Démonstration.** Pour  $y \in V$  fixé, soit  $L \in V'$  la forme linéaire continue définie par  $L(x) = \langle Ax, y \rangle$ . Par application du Théorème 12.1.18 de Riesz, il existe un unique  $z \in V$  tel que  $L(x) = \langle z, x \rangle$ . On définit alors l'application  $A^*$  de  $V$  dans  $V$  qui, à chaque  $y$  associe le  $z$  correspondant. On vérifie facilement que  $A^*$  est linéaire continue, et on a bien  $L(x) = \langle Ax, y \rangle = \langle x, A^*y \rangle$ .  $\square$

**Définition 7.2.3** Soit  $A$  une application linéaire continue de  $V$  dans  $V$ . On dit que  $A$  est auto-adjointe si elle coïncide avec son adjointe, c'est-à-dire que  $A^* = A$ .

**Définition 7.2.4** Soit  $A$  une application linéaire continue de  $V$  dans  $V$ . On dit que  $A$  est définie positive si  $\langle Ax, x \rangle > 0$  pour tout  $x \in V$  non nul.

On sait qu'en dimension finie toutes les applications linéaires auto-adjointes sont diagonalisables dans une base orthonormée. Nous allons voir qu'en dimension infinie ce résultat se généralise aux applications linéaires continues auto-adjointes qui sont en plus **compactes**. Introduisons maintenant les notions qui permettent de définir la compacité d'une application linéaire continue.

**Définition 7.2.5** *Un sous-ensemble  $K \subset V$  est dit compact si, de toute suite  $(u_n)_{n \geq 1}$  d'éléments de  $K$ , on peut extraire une sous-suite  $u_{n'}$  convergente dans  $K$ .*

*Un sous-ensemble  $K \subset V$  est dit relativement compact si, de toute suite  $(u_n)_{n \geq 1}$  d'éléments de  $K$ , on peut extraire une sous-suite  $u_{n'}$  convergente dans  $V$ .*

Il est bien connu que, si  $V$  est de dimension finie, alors les sous-ensembles compacts de  $V$  sont les fermés bornés. Malheureusement, ce résultat n'est plus vrai en dimension infinie. En effet, un sous-ensemble compact est toujours fermé borné mais la réciproque n'est pas vraie comme le montre le lemme suivant.

**Lemme 7.2.6** *Dans un espace de Hilbert  $V$  de dimension infinie, la boule unité fermée n'est jamais compacte.*

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

$$\|e_n - e_p\|^2 = \|e_n\|^2 + \|e_p\|^2 - 2\langle e_n, e_p \rangle = 2,$$

ce qui prouve qu'aucune sous-suite de  $e_n$  n'est une suite de Cauchy.  $\square$

**Définition 7.2.7** *Soit  $V$  et  $W$  deux espaces de Hilbert et  $A$  une application linéaire continue de  $V$  dans  $W$ . On dit que  $A$  est compacte si l'image par  $A$  de la boule unité de  $V$  est relativement compacte dans  $W$ .*

De manière équivalente, une application linéaire continue  $A$  est compacte si, pour toute suite bornée  $x_n$  de  $V$ , on peut extraire une sous-suite telle que  $Ax_{n'}$  converge dans  $W$ . Si  $W$  ou  $V$  est de dimension finie, alors toute application linéaire continue est compacte. Ce n'est plus vrai si  $W$  et  $V$  sont de dimension infinie, comme le montre l'exercice suivant.

**Exercice 7.2.1** Montrer que l'application identité  $\text{Id}$  dans un espace de Hilbert  $V$  de dimension infinie n'est jamais compacte (utiliser le Lemme 7.2.6).

**Exercice 7.2.2** Soit l'espace de Hilbert  $\ell_2$  des suites réelles  $x = (x_i)_{i \geq 1}$  telles que  $\sum_{i \geq 1} |x_i|^2 < +\infty$ , muni du produit scalaire  $\langle x, y \rangle = \sum_{i \geq 1} x_i y_i$ . Soit  $(a_i)_{i \geq 1}$  une suite de réels bornés,  $|a_i| \leq C < +\infty$  pour tout  $i \geq 1$ . On définit l'application linéaire  $A$  par  $Ax = (a_i x_i)_{i \geq 1}$ . Vérifier que  $A$  est continue. Montrer que  $A$  est compacte si et seulement si  $\lim_{i \rightarrow +\infty} a_i = 0$ .

**Exercice 7.2.3** Soit  $U$ ,  $V$  et  $W$  trois espaces de Hilbert de dimension infinie,  $A$  une application linéaire continue de  $V$  dans  $W$ , et  $B$  une application linéaire continue de  $U$  dans  $V$ . Montrer que l'application  $AB$  est compacte dès que  $A$  ou  $B$  est compacte. En déduire qu'une application linéaire continue compacte n'est jamais inversible d'inverse continu en dimension infinie.

## 7.2.2 Décomposition spectrale d'un opérateur compact

Le résultat principal de cette sous-section est le suivant.

**Théorème 7.2.8** Soit  $V$  un espace de Hilbert réel de dimension infinie et  $A$  une application linéaire continue, définie positive, auto-adjointe, compacte de  $V$  dans  $V$ . Alors les valeurs propres de  $A$  forment une suite  $(\lambda_k)_{k \geq 1}$  de réels strictement positifs qui tend vers 0, et il existe une base hilbertienne  $(u_k)_{k \geq 1}$  de  $V$  formée de vecteurs propres de  $A$ , avec

$$Au_k = \lambda_k u_k \text{ pour } k \geq 1.$$

**Remarque 7.2.9** Comme conséquence du Théorème 7.2.8, et avec les mêmes notations, on obtient la **décomposition spectrale** de tout élément  $v \in V$

$$v = \sum_{k=1}^{+\infty} \langle v, u_k \rangle u_k \text{ avec } \|v\|^2 = \sum_{k=1}^{+\infty} |\langle v, u_k \rangle|^2.$$

**Exercice 7.2.4** On reprend les notations et les hypothèses du Théorème 7.2.8. Montrer que, pour  $v \in V$ , l'équation  $Au = v$  admet une unique solution  $u \in V$  si et seulement si  $v$  vérifie

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

Lorsque l'application linéaire  $A$  n'est pas compacte le Théorème 7.2.8 tombe en défaut comme le montre l'exercice suivant.

**Exercice 7.2.5** Soit  $V = L^2(0,1)$  et  $A$  l'application linéaire de  $V$  dans  $V$  définie par  $(Af)(x) = (x^2 + 1)f(x)$ . Vérifier que  $A$  est continue, définie positive, auto-adjointe mais pas compacte. Montrer que  $A$  n'a pas de valeurs propres. On pourra vérifier aussi que  $(A - \lambda \text{Id})$  est inversible d'inverse continu si et seulement si  $\lambda \notin [1, 2]$ .

Pour démontrer le Théorème 7.2.8 nous avons besoin de deux lemmes préliminaires.

**Lemme 7.2.10** Soit  $V$  un espace de Hilbert réel (non réduit au seul vecteur nul) et  $A$  une application linéaire continue auto-adjointe compacte de  $V$  dans  $V$ . On définit

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

Alors,  $\|A\| = \max(|m|, |M|)$ , et soit  $m$ , soit  $M$ , est valeur propre de  $A$ .

**Démonstration.** On voit facilement que  $|\langle Au, u \rangle| \leq \|A\| \|u\|^2$ , donc  $\max(|m|, |M|) \leq \|A\|$ . D'autre part, comme  $A$  est auto-adjoint, on obtient pour tout  $u, v \in V$

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

Or,  $\|A\| = \sup_{\|u\|=\|v\|=1} \langle Au, v \rangle$  puisque  $\|Au\| = \sup_{\|v\|=1} \langle Au, v \rangle$ . On en déduit donc que  $\|A\| \leq \max(|m|, |M|)$ , d'où l'égalité entre ces deux termes.

Par ailleurs, comme  $m \leq M$ , un des deux cas suivants a lieu : soit  $\|A\| = M \geq 0$ , soit  $\|A\| = -m$  avec  $m \leq 0$ . Considérons le cas  $\|A\| = M \geq 0$  (l'autre cas  $\|A\| = -m$  est complètement symétrique en remplaçant  $A$  par  $-A$ ). Soit une suite  $(u_n)_{n \geq 1}$  de vecteurs unitaires de  $V$  qui est maximisante dans la définition de  $M$ , c'est-à-dire que

$$\lim_{n \rightarrow +\infty} \langle Au_n, u_n \rangle = M \text{ et } \|u_n\| = 1.$$

Comme  $A$  est compacte, il existe une sous-suite telle que  $Au_{n'}$  converge dans  $V$  vers une limite  $v$ . D'autre part, on a

$$\langle Au_n, u_n \rangle \leq \|Au_n\| \leq \|A\| = M,$$

d'où l'on déduit que  $\lim_{n \rightarrow +\infty} \|Au_n\| = M$ , c'est-à-dire que  $\|v\| = M$ . Finalement, comme

$$\|Au_n - Mu_n\|^2 = \|Au_n\|^2 + M^2 - 2M\langle Au_n, u_n \rangle,$$

on obtient que  $\lim_{n \rightarrow +\infty} \|Au_n - Mu_n\| = 0$ . Pour la sous-suite  $n'$ , cela implique que  $u_{n'}$  converge vers  $v/M$  (du moins, si  $M \neq 0$ ; le cas  $M = 0$  est trivial puisqu'il implique que  $A = 0$ ). Par continuité de  $A$ , on en déduit donc que  $Au_{n'}$  converge aussi vers  $Av/M$  (en plus de vers  $v$ ). L'unicité de la limite montre que  $Av/M = v$ , c'est-à-dire que  $v$  est un vecteur propre (non nul car  $\|v\| = M \neq 0$ ) associé à la valeur propre  $M$ .  $\square$

**Lemme 7.2.11** *Soit  $V$  un espace de Hilbert et  $A$  une application linéaire continue compacte de  $V$  dans  $V$ . Pour tout réel  $\delta > 0$ , il n'existe au plus qu'un nombre fini de valeurs propres en dehors de l'intervalle  $]-\delta, +\delta[$ , et le sous-espace des vecteurs propres associés à chacune de ces valeurs propres est de dimension finie.*

**Démonstration.** Montrons qu'on ne peut pas avoir une infinité d'éléments de  $V$ , linéairement indépendants, qui soient vecteurs propres de  $A$  pour des valeurs propres  $\lambda$  telles que  $|\lambda| \geq \delta > 0$ . On procède par contradiction. Supposons donc qu'il existe une suite infinie  $(u_k)_{k \geq 1}$  d'éléments de  $V$ , linéairement indépendants, et une suite de valeurs propres  $(\lambda_k)_{k \geq 1}$  tels que

$$Au_k = \lambda_k u_k \text{ et } |\lambda_k| \geq \delta \text{ pour tout } k \geq 1.$$

On note  $E_k$  le sous-espace vectoriel engendré par la famille  $(u_1, u_2, \dots, u_k)$ . Comme  $E_{k-1}$  est inclus strictement dans  $E_k$ , il existe un vecteur unitaire  $v_k \in E_k$  qui est orthogonal à  $E_{k-1}$ . Comme  $|\lambda_k| \geq \delta$ , la suite  $v_k/\lambda_k$  est bornée dans  $V$ , et, puisque  $A$  est compact, on peut extraire une sous-suite telle que  $Av_{k'}/\lambda_{k'}$  converge dans  $V$ . Cependant, pour  $j < k$  on peut écrire

$$\frac{Av_k}{\lambda_k} - \frac{Av_j}{\lambda_j} = v_k + (A - \lambda_k \text{Id}) \frac{v_k}{\lambda_k} - \frac{Av_j}{\lambda_j}. \quad (7.10)$$

Or, on vérifie facilement que  $AE_k \subset E_k$  et que  $(A - \lambda_k \text{Id})E_k \subset E_{k-1}$ , donc les deux derniers termes à droite de (7.10) appartiennent à  $E_{k-1}$ . Comme  $v_k$  est orthogonal à  $E_{k-1}$ , on déduit de (7.10)

$$\left\| \frac{Av_k}{\lambda_k} - \frac{Av_j}{\lambda_j} \right\| \geq \|v_k\| = 1,$$

ce qui est une contradiction avec la convergence de la sous-suite  $Av_{k'}/\lambda_{k'}$ .  $\square$

**Démonstration du Théorème 7.2.8.** Le Lemme 7.2.10 montre que l'ensemble des valeurs propres de  $A$  n'est pas vide, tandis que le Lemme 7.2.11 montre que cet ensemble est soit fini, soit infini dénombrable avec 0 comme seul point d'accumulation. Par ailleurs, comme  $A$  est définie positive, toutes les valeurs propres sont strictement positives. Notons  $(\lambda_k)$  les valeurs propres de  $A$  et  $V_k = \text{Ker}(A - \lambda_k \text{Id})$  les sous-espaces propres associés (le Lemme 7.2.11 nous dit aussi que chaque  $V_k$  est de dimension finie). On remarque que les sous-espaces propres  $V_k$  sont orthogonaux deux à deux : en effet, si  $v_k \in V_k$  et  $v_j \in V_j$  avec  $k \neq j$ , alors, comme  $A$  est auto-adjoint, on a

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

d'où l'on déduit que  $\langle v_k, v_j \rangle = 0$  puisque  $\lambda_k \neq \lambda_j$ . Soit  $W$  l'adhérence dans  $V$  de l'union des  $V_k$

$$W = \overline{\left\{ u \in V, \exists K \geq 1 \text{ tel que } u = \sum_{i=1}^K u_k, u_k \in V_k \right\}}.$$

On construit facilement une base hilbertienne de  $W$  par réunion des bases orthonormales de chaque  $V_k$  (chacun de dimension finie et orthogonaux entre eux). Montrons qu'en fait  $W = V$  (ce qui prouvera aussi que la suite  $(\lambda_k)$  est infinie puisque  $V$  est de dimension infinie). On introduit l'orthogonal de  $W$  défini par

$$W^\perp = \{u \in V \text{ tel que } \langle u, v \rangle = 0 \forall v \in W\}.$$

Comme  $W$  est stable par  $A$  ( $AW \subset W$ ), on vérifie que  $W^\perp$  est aussi stable par  $A$  car  $\langle Au, v \rangle = \langle u, Av \rangle = 0$  si  $u \in W^\perp$  et  $v \in W$ . On peut donc définir la restriction de  $A$  à  $W^\perp$  qui est aussi une application linéaire continue auto-adjointe compacte. Par application du Lemme 7.2.10, si  $W^\perp \neq \{0\}$ , cette restriction admet aussi une valeur propre et un vecteur propre  $u \in W^\perp$  qui sont aussi valeur et vecteur propres de  $A$ . C'est bien sûr une contradiction avec le fait que, par définition,  $W$  contient déjà tous les vecteurs propres de  $A$  et que  $W \cap W^\perp = \{0\}$ . Par conséquent, on a nécessairement  $W^\perp = \{0\}$ , et comme  $W$  est fermé on en déduit que  $W = \{0\}^\perp = V$ .  $\square$

**Remarque 7.2.12** La démonstration du Théorème 7.2.8 est encore valable si  $A$  n'est pas définie positive aux restrictions suivantes près : les valeurs propres ne sont pas nécessairement positives, les valeurs propres non nulles peuvent être en nombre fini, et  $\text{Ker} A$  (le sous-espace propre associé à la valeur propre nulle) peut être de dimension infinie. •

## 7.3 Valeurs propres d'un problème elliptique

### 7.3.1 Problème variationnel

Nous revenons au cadre variationnel introduit au Chapitre 3. L'intérêt de ce cadre assez général est qu'il s'appliquera à de nombreux modèles différents. Dans un espace

de Hilbert  $V$  nous considérons une forme bilinéaire  $a(\cdot, \cdot)$ , **symétrique**, continue et coercive, c'est-à-dire que  $a(w, v) = a(v, w)$ , et il existe  $M > 0$  et  $\nu > 0$  tels que

$$|a(w, v)| \leq M \|w\|_V \|v\|_V \text{ pour tout } w, v \in V$$

et

$$a(v, v) \geq \nu \|v\|_V^2 \text{ pour tout } v \in V.$$

Pour pouvoir appliquer les résultats de la section précédente, nous introduisons un nouvel ingrédient, à savoir un autre espace de Hilbert  $H$ . Nous faisons l'hypothèse fondamentale suivante

$$\begin{cases} V \subset H \text{ avec injection compacte} \\ V \text{ est dense dans } H. \end{cases} \quad (7.11)$$

L'expression "injection compacte" veut dire précisément que l'opérateur d'inclusion  $\mathcal{I}$  qui à  $v \in V$  associe  $\mathcal{I}v = v \in H$  est continu et compact (voir la Définition 7.2.7). Autrement dit, l'hypothèse (7.11) implique que de toute suite bornée de  $V$  on peut extraire une sous-suite convergente dans  $H$ . Les espaces  $H$  et  $V$  ne partagent pas le même produit scalaire, et nous les noterons  $\langle \cdot, \cdot \rangle_H$  et  $\langle \cdot, \cdot \rangle_V$  pour éviter toute confusion.

Nous considérons le problème variationnel de valeurs propres suivant (ou problème spectral) : trouver  $\lambda \in \mathbb{R}$  et  $u \in V \setminus \{0\}$  tels que

$$a(u, v) = \lambda \langle u, v \rangle_H \quad \forall v \in V. \quad (7.12)$$

On dira que  $\lambda$  est une valeur propre du problème variationnel (7.12) (ou de la forme bilinéaire  $a$ ) et que  $u$  est le vecteur propre associé.

**Remarque 7.3.1** Sous l'hypothèse (7.11) les espaces  $H$  et  $V$  ne peuvent jamais avoir le même produit scalaire. Sinon ils seraient égaux puisque  $V$  est dense dans  $H$ . Mais cela est impossible car alors l'injection de  $V$  dans  $H$  serait l'identité qui n'est pas compacte (voir l'Exercice 7.2.1). •

Donnons tout de suite un exemple concret et typique d'une telle situation. Pour un ouvert borné  $\Omega$ , on pose  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ , et la forme bilinéaire symétrique est définie par

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

Comme  $C_c^\infty(\Omega)$  est dense à la fois dans  $H_0^1(\Omega)$  et dans  $L^2(\Omega)$ , et grâce au Théorème 4.3.21 de Rellich, l'hypothèse (7.11) est vérifiée, et on a vu au Chapitre 5 que cette forme bilinéaire  $a$  est bien continue et coercive sur  $V$ . Par une simple intégration par parties, on voit facilement que (7.12) est équivalent à

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

c'est-à-dire que  $\lambda$  et  $u$  sont valeur propre et fonction propre du Laplacien.

Les solutions de (7.12) sont données par le résultat suivant.



**Théorème 7.3.2** Soit  $V$  et  $H$  deux espaces de Hilbert réels de dimension infinie. On suppose que  $V \subset H$  avec injection compacte et que  $V$  est dense dans  $H$ . Soit  $a(\cdot, \cdot)$  une forme bilinéaire symétrique continue et coercive sur  $V$ . Alors les valeurs propres de (7.12) forment une suite croissante  $(\lambda_k)_{k \geq 1}$  de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de  $H$   $(u_k)_{k \geq 1}$  de vecteurs propres associés, c'est-à-dire que

$$u_k \in V, \text{ et } a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall v \in V.$$

De plus,  $(u_k / \sqrt{\lambda_k})_{k \geq 1}$  est une base hilbertienne de  $V$  pour le produit scalaire  $a(\cdot, \cdot)$ .

**Démonstration.** Pour  $f \in H$ , nous résolvons le problème variationnel

$$\text{trouver } u \in V \text{ tel que } a(u, v) = \langle f, v \rangle_H \text{ pour toute fonction } v \in V. \quad (7.13)$$

Il est facile de vérifier les hypothèses du Théorème 3.3.1 de Lax-Milgram pour (7.13) qui admet donc une unique solution  $u \in V$ . On définit une application linéaire  $\mathcal{A}$  de  $H$  dans  $V$  qui à  $f$  associe la solution  $u = \mathcal{A}f$ . Autrement dit, l'application linéaire  $\mathcal{A}$  est définie par

$$\mathcal{A}f \in V \text{ tel que } a(\mathcal{A}f, v) = \langle f, v \rangle_H \text{ pour tout } v \in V. \quad (7.14)$$

En prenant  $v = \mathcal{A}f$  dans (7.14), on obtient

$$\nu \|\mathcal{A}f\|_V^2 \leq a(\mathcal{A}f, \mathcal{A}f) = \langle f, \mathcal{A}f \rangle_H \leq \|f\|_H \|\mathcal{A}f\|_H \leq C \|f\|_H \|\mathcal{A}f\|_V$$

car l'opérateur d'injection  $\mathcal{I}$  de  $V$  dans  $H$  est continu. Par conséquent, l'application linéaire  $\mathcal{A}$  est continue de  $H$  dans  $V$ . On définit maintenant une application linéaire  $A = \mathcal{I}\mathcal{A}$  de  $H$  dans  $H$ , qui est bien continue. Comme  $\mathcal{I}$  est compact, le produit  $A$  est aussi compact (voir l'Exercice 7.2.3). Pour montrer que  $A$  est auto-adjoint, on prend  $v = \mathcal{A}g$  dans (7.14) et on obtient, pour tout  $f, g \in H$ ,

$$\langle f, \mathcal{A}g \rangle_H = \langle f, \mathcal{A}g \rangle_H = a(\mathcal{A}f, \mathcal{A}g) = a(\mathcal{A}g, \mathcal{A}f) = \langle g, \mathcal{A}f \rangle_H = \langle g, \mathcal{A}f \rangle_H,$$

à cause de la symétrie de  $a$ , ce qui prouve que  $A$  est auto-adjoint défini positif dans  $H$ . On peut donc appliquer le Théorème 7.2.8 à l'opérateur  $A$  qui en vérifie toutes les hypothèses. Il existe une suite décroissante  $(\mu_k)_{k \geq 1}$  de réels positifs qui tend vers 0, et il existe une base hilbertienne  $(u_k)_{k \geq 1}$  de  $H$  formée de vecteurs propres de  $A$ , avec

$$Au_k = \mu_k u_k \text{ pour } k \geq 1.$$

Remarquons que, par cette égalité, les vecteurs propres  $u_k$  appartiennent non seulement à  $H$  mais aussi à  $V$ . Revenons maintenant au problème aux valeurs propres (7.12) qui peut s'écrire

$$a(u, v) = \lambda \langle u, v \rangle_H = \lambda a(\mathcal{A}u, v) \quad \forall v \in V,$$

à cause de la définition (7.14), c'est-à-dire  $a(u - \lambda \mathcal{A}u, v) = 0$ , donc

$$u = \lambda \mathcal{A}u = \lambda Au.$$

Par conséquent, les valeurs propres  $(\lambda_k)_{k \geq 1}$  du problème variationnel (7.12) sont exactement les inverses des valeurs propres  $(\mu_k)_{k \geq 1}$  de  $A$ , et leurs vecteurs propres sont les mêmes. On pose

$$\lambda_k = \frac{1}{\mu_k} \quad \text{et} \quad v_k = \frac{u_k}{\sqrt{\lambda_k}}.$$

Par construction, les vecteurs propres  $u_k$  forment une base hilbertienne de  $H$ . On vérifie que

$$a(v_k, v_j) = \frac{a(u_k, u_j)}{\sqrt{\lambda_k \lambda_j}} = \lambda_k \frac{\langle u_k, u_j \rangle_H}{\sqrt{\lambda_k \lambda_j}} = \delta_{kj},$$

et comme l'orthogonal des  $(v_k)_{k \geq 1}$  dans  $V$  est contenu dans l'orthogonal des  $(u_k)_{k \geq 1}$  dans  $H$  (qui est réduit au vecteur nul), on en déduit que les  $(v_k)_{k \geq 1}$  forment une base hilbertienne de  $V$  pour le produit scalaire  $a(u, v)$ .  $\square$

**Remarque 7.3.3** Insistons sur le fait que l'opérateur  $\mathcal{A}$ , défini par (7.14), est l'opérateur de résolution de la formulation variationnelle, c'est-à-dire qu'il est en quelque sorte **l'inverse** de la forme bilinéaire  $a$ . C'est pour cette raison que les valeurs propres  $\lambda_k$  de la formulation variationnelle sont les inverses des valeurs propres  $\mu_k$  de  $\mathcal{A}$ . Par exemple, en dimension finie la forme bilinéaire s'écrit  $a(u, v) = \mathcal{K}u \cdot v$  et on a  $\mathcal{A} = \mathcal{K}^{-1}$ . De même, pour le Laplacien on a  $\mathcal{A} = (-\Delta)^{-1}$  (seul l'inverse du Laplacien est compact, pas le Laplacien lui-même; voir l'Exercice 7.2.3). En fait, c'est le gain de régularité de la solution du Laplacien par rapport au second membre qui est à l'origine de la compacité de l'opérateur  $(-\Delta)^{-1}$ .  $\bullet$

**Exercice 7.3.1** Démontrer une variante du Théorème 7.3.2 où l'on remplace l'hypothèse de coercivité de la forme bilinéaire  $a(\cdot, \cdot)$  par l'hypothèse plus faible qu'il existe deux constantes positives  $\eta > 0$  et  $\nu > 0$  telles que

$$a(v, v) + \eta \|v\|_H^2 \geq \nu \|v\|_V^2 \quad \text{pour tout } v \in V.$$

(Dans ce cas les valeurs propres  $(\lambda_k)_{k \geq 1}$  ne sont pas forcément positives, mais vérifient seulement  $\lambda_k + \eta > 0$ .)

Nous donnons au passage une caractérisation très utile des valeurs propres du problème variationnel (7.12), appelée **principe du min-max ou de Courant-Fisher**. Pour cela on introduit le quotient de Rayleigh défini, pour chaque fonction  $v \in V \setminus \{0\}$ , par

$$R(v) = \frac{a(v, v)}{\|v\|_H^2}.$$

**Proposition 7.3.4 (Courant-Fisher)** Soit  $V$  et  $H$  deux espaces de Hilbert réels de dimension infinie. On suppose que  $V \subset H$  avec injection compacte et que  $V$  est dense dans  $H$ . Soit  $a(\cdot, \cdot)$  une forme bilinéaire symétrique continue et coercive sur  $V$ . Pour  $k \geq 0$  on note  $\mathcal{E}_k$  l'ensemble des sous-espaces vectoriels de dimension  $k$  de  $V$ . On

note  $(\lambda_k)_{k \geq 1}$  la suite **croissante** des valeurs propres du problème variationnel (7.12). Alors, pour tout  $k \geq 1$ , la  $k$ -ème valeur propre est donnée par

$$\lambda_k = \min_{W \in \mathcal{E}_k} \left( \max_{v \in W \setminus \{0\}} R(v) \right) = \max_{W \in \mathcal{E}_{k-1}} \left( \min_{v \in W^\perp \setminus \{0\}} R(v) \right). \quad (7.15)$$

En particulier, la première valeur propre vérifie

$$\lambda_1 = \min_{v \in V \setminus \{0\}} R(v), \quad (7.16)$$

et tout point de minimum dans (7.16) est un vecteur propre associé à  $\lambda_1$ .

**Démonstration.** Soit  $(u_k)_{k \geq 1}$  la base hilbertienne de  $H$  formée des vecteurs propres de (7.12). D'après le Théorème 7.2.8,  $(u_k/\sqrt{\lambda_k})_{k \geq 1}$  est une base hilbertienne de  $V$ . On peut donc caractériser les espaces  $H$  et  $V$  à partir de leur décomposition spectrale (voir la Remarque 7.2.9)

$$H = \left\{ v = \sum_{k=1}^{+\infty} \alpha_k u_k, \ \|v\|_H^2 = \sum_{k=1}^{+\infty} \alpha_k^2 < +\infty \right\},$$

$$V = \left\{ v = \sum_{k=1}^{+\infty} \alpha_k u_k, \ \|v\|_V^2 = \sum_{k=1}^{+\infty} \lambda_k \alpha_k^2 < +\infty \right\}.$$

On remarque au passage que, comme les valeurs propres  $\lambda_k$  sont minorées par  $\lambda_1 > 0$ , cette caractérisation fait bien apparaître  $V$  comme un sous-espace de  $H$ . On peut alors réécrire le quotient de Rayleigh

$$R(v) = \frac{\sum_{k=1}^{+\infty} \lambda_k \alpha_k^2}{\sum_{k=1}^{+\infty} \alpha_k^2},$$

ce qui démontre immédiatement le résultat pour la première valeur propre. Introduisons le sous-espace  $W_k \in \mathcal{E}_k$  engendré par  $(u_1, u_2, \dots, u_k)$ . On a

$$R(v) = \frac{\sum_{j=1}^k \lambda_j \alpha_j^2}{\sum_{j=1}^k \alpha_j^2} \quad \forall v \in W_k \quad \text{et} \quad R(v) = \frac{\sum_{j=k}^{+\infty} \lambda_j \alpha_j^2}{\sum_{j=k}^{+\infty} \alpha_j^2} \quad \forall v \in W_{k-1}^\perp,$$

d'où l'on déduit

$$\lambda_k = \max_{v \in W_k \setminus \{0\}} R(v) = \min_{v \in W_{k-1}^\perp \setminus \{0\}} R(v).$$

Soit  $W$  un sous-espace quelconque dans  $\mathcal{E}_k$ . Comme  $W$  est de dimension  $k$  et  $W_{k-1}$  de dimension  $k-1$ , l'intersection  $W \cap W_{k-1}^\perp$  n'est pas réduite à  $\{0\}$ . Par conséquent,

$$\max_{v \in W \setminus \{0\}} R(v) \geq \max_{v \in W \cap W_{k-1}^\perp \setminus \{0\}} R(v) \geq \min_{v \in W \cap W_{k-1}^\perp \setminus \{0\}} R(v) \geq \min_{v \in W_{k-1}^\perp \setminus \{0\}} R(v) = \lambda_k,$$

ce qui prouve la première égalité dans (7.15). De même, si  $W$  est un sous-espace dans  $\mathcal{E}_{k-1}$ , alors  $W^\perp \cap W_k$  n'est pas réduit à  $\{0\}$ , et

$$\min_{v \in W^\perp \setminus \{0\}} R(v) \leq \min_{v \in W^\perp \cap W_k \setminus \{0\}} R(v) \leq \max_{v \in W^\perp \cap W_k \setminus \{0\}} R(v) \leq \max_{v \in W_k \setminus \{0\}} R(v) = \lambda_k,$$

ce qui prouve la deuxième égalité dans (7.15). Soit maintenant  $u$  un point de minimum dans (7.16). Pour  $v \in V$ , on introduit la fonction  $f(t) = R(u + tv)$  d'une variable réelle  $t \in \mathbb{R}$  qui admet un minimum en  $t = 0$ . Par conséquent sa dérivée s'annule en  $t = 0$ . En tenant compte de ce que  $f(0) = \lambda_1$ , un simple calcul montre que

$$f'(0) = 2 \frac{a(u, v) - \lambda_1 \langle u, v \rangle_H}{\|u\|_H^2}.$$

Comme  $v$  est quelconque dans  $V$ , la condition  $f'(0) = 0$  n'est rien d'autre que la formulation variationnelle (7.12), c'est-à-dire que  $u$  est un vecteur propre associé à la valeur propre  $\lambda_1$ .  $\square$

### 7.3.2 Valeurs propres du Laplacien

On peut immédiatement appliquer le Théorème 7.2.8 à la formulation variationnelle du Laplacien avec conditions aux limites de Dirichlet, ce qui nous donne le résultat suivant.

**Théorème 7.3.5** *Soit  $\Omega$  un ouvert borné régulier de classe  $\mathcal{C}^1$  de  $\mathbb{R}^N$ . Il existe une suite croissante  $(\lambda_k)_{k \geq 1}$  de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de  $L^2(\Omega)$   $(u_k)_{k \geq 1}$ , telle que chaque  $u_k$  appartient à  $H_0^1(\Omega)$  et vérifie*

$$\begin{cases} -\Delta u_k = \lambda_k u_k & p.p. \text{ dans } \Omega \\ u_k = 0 & p.p. \text{ sur } \partial\Omega. \end{cases} \quad (7.17)$$

**Démonstration.** Pour le Laplacien avec conditions aux limites de Dirichlet, on choisit  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ , et la forme bilinéaire symétrique est définie par

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

et le produit scalaire sur  $L^2(\Omega)$  est bien sûr

$$\langle u, v \rangle_H = \int_{\Omega} uv \, dx.$$

On vérifie aisément les hypothèses du Théorème 7.2.8. Grâce au Théorème 4.3.21 de Rellich,  $V$  est bien compactement inclus dans  $H$ . Comme  $C_c^\infty(\Omega)$  est dense à la fois dans  $H$  et dans  $V$ ,  $V$  est bien dense dans  $H$ . Enfin, on a vu au Chapitre 5 que la forme bilinéaire  $a$  est bien continue et coercive sur  $V$ . Par conséquent, il existe une

suite croissante  $(\lambda_k)_{k \geq 1}$  de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de  $L^2(\Omega)$   $(u_k)_{k \geq 1}$ , tels que  $u_k \in H_0^1(\Omega)$  et

$$\int_{\Omega} \nabla u_k \cdot \nabla v \, dx = \lambda_k \int_{\Omega} u_k v \, dx \quad \forall v \in H_0^1(\Omega).$$

Par une simple intégration par parties (du même type que celle pratiquée dans la démonstration du Théorème 5.2.2) on obtient (7.17). Remarquons que nous n'utilisons la régularité de  $\Omega$  que pour pouvoir appliquer le Théorème de trace 4.3.13 et donner un sens "presque partout" à la condition aux limites de Dirichlet.  $\square$

**Remarque 7.3.6** L'hypothèse sur le caractère borné de l'ouvert  $\Omega$  est absolument fondamentale dans le Théorème 7.3.5. Si elle n'est pas satisfaite, le Théorème 4.3.21 de Rellich (sur l'injection compacte de  $H^1(\Omega)$  dans  $L^2(\Omega)$ ) est en général faux, et on peut montrer que le Théorème 7.3.5 n'a pas lieu. En fait, il se peut qu'il existe une infinité (non dénombrable) de valeurs propres "généralisées" au sens où les fonctions propres n'appartiennent pas à  $L^2(\Omega)$ . A la lumière de l'Exercice 7.1.1 on méditera le cas du Laplacien dans  $\Omega = \mathbb{R}^N$ . •

**Exercice 7.3.2** En dimension  $N = 1$ , on considère  $\Omega = ]0, 1[$ . Calculer explicitement toutes les valeurs propres et les fonctions propres du Laplacien avec conditions aux limites de Dirichlet (7.17). A l'aide de la décomposition spectrale de ce problème (voir la Remarque 7.2.9), montrer que la série

$$\sum_{k=1}^{+\infty} a_k \sin(k\pi x)$$

converge dans  $L^2(0, 1)$  si et seulement si  $\sum_{k=1}^{+\infty} a_k^2 < +\infty$ , et dans  $H^1(0, 1)$  si et seulement si  $\sum_{k=1}^{+\infty} k^2 a_k^2 < +\infty$ .

**Exercice 7.3.3** On considère un parallélépipède  $\Omega = ]0, L_1[ \times ]0, L_2[ \times \cdots \times ]0, L_N[$ , où les  $(L_i > 0)_{1 \leq i \leq N}$  sont des constantes positives. Calculer explicitement toutes les valeurs propres et les fonctions propres du Laplacien avec conditions aux limites de Dirichlet (7.17).

Le Théorème 7.3.5 se généralise aisément au cas d'autres conditions aux limites. Nous laissons au lecteur le soin de démontrer le corollaire suivant.

**Corollaire 7.3.7** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$  dont la frontière  $\partial\Omega$  se décompose en deux parties disjointes régulières  $\partial\Omega_N$  et  $\partial\Omega_D$  (voir la Figure 4.1). Il existe une suite croissante  $(\lambda_k)_{k \geq 1}$  de réels positifs ou nuls qui tend vers l'infini, et il existe une base hilbertienne de  $L^2(\Omega)$   $(u_k)_{k \geq 1}$ , telle que chaque  $u_k$  appartient à  $H^1(\Omega)$  et vérifie

$$\begin{cases} -\Delta u_k = \lambda_k u_k & \text{dans } \Omega \\ u_k = 0 & \text{sur } \partial\Omega_D \\ \frac{\partial u_k}{\partial n} = 0 & \text{sur } \partial\Omega_N. \end{cases}$$

**Remarque 7.3.8** Dans le cas d'une condition aux limites purement de Neumann, c'est-à-dire que  $\partial\Omega_D = \emptyset$ , la forme bilinéaire n'est plus coercive sur  $H^1(\Omega)$ . Pour démontrer le Corollaire 7.3.7 il faut alors utiliser l'Exercice 7.3.1. •

**Exercice 7.3.4** On considère à nouveau un ouvert  $\Omega$  parallélépipédique comme dans l'Exercice 7.3.3. Calculer explicitement toutes les valeurs propres et les fonctions propres du Laplacien avec conditions aux limites de Neumann sur tout le bord  $\partial\Omega$ .

La caractérisation des valeurs propres par le principe de Courant-Fisher est souvent fort utile, comme le montre l'exercice suivant.

**Exercice 7.3.5** On reprend les notations et les hypothèses du Théorème 7.3.5. Montrer que la meilleure (i.e. la plus petite) constante  $C$  dans l'inégalité de Poincaré (voir la Proposition 4.3.10) est précisément la première valeur propre  $\lambda_1$  de (7.17).

On peut aussi montrer que les fonctions propres du Laplacien, avec conditions aux limites de Dirichlet ou de Neumann, sont régulières.

**Proposition 7.3.9** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^\infty$ . Alors les fonctions propres solutions de (7.17) appartiennent à  $C^\infty(\overline{\Omega})$ .*

**Démonstration.** Soit  $u_k$  la  $k$ -ème fonction propre solution dans  $H_0^1(\Omega)$  de (7.17). On peut considérer que  $u_k$  est solution du problème aux limites suivant

$$\begin{cases} -\Delta u_k = f_k & \text{dans } \Omega \\ u_k = 0 & \text{sur } \partial\Omega, \end{cases}$$

avec  $f_k = \lambda_k u_k$ . Comme  $f_k$  appartient à  $H^1(\Omega)$ , par application du Théorème 5.2.26 de régularité on en déduit que la solution  $u_k$  appartient à  $H^3(\Omega)$ . Du coup, le second membre  $f_k$  est plus régulier ce qui permet d'augmenter encore la régularité de  $u_k$ . Par une récurrence facile on montre ainsi que  $u_k$  appartient à  $H^m(\Omega)$  pour tout  $m \geq 1$ . En vertu du Théorème 4.3.25 sur la continuité des fonctions de  $H^m(\Omega)$  (voir aussi la Remarque 4.3.26), on en déduit que  $u_k$  appartient donc  $C^\infty(\overline{\Omega})$ . □

Nous démontrons maintenant un résultat qualitatif très important à propos de la première valeur propre.

**Théorème 7.3.10 (de Krein-Rutman)** *On reprend les notations et les hypothèses du Théorème 7.3.5. On suppose que l'ouvert  $\Omega$  est connexe. Alors la première valeur propre  $\lambda_1$  est simple (i.e. le sous-espace propre correspondant est de dimension 1) et le premier vecteur propre peut être choisi positif presque partout dans  $\Omega$ .*

**Remarque 7.3.11** Le Théorème 7.3.10 de Krein-Rutman est spécifique au cas des équations "scalaires" (c'est-à-dire que l'inconnue  $u$  est à valeurs dans  $\mathbb{R}$ ). Ce résultat est faux en général si l'inconnue  $u$  est à valeurs vectorielles (voir plus loin l'exemple du

système de l'élasticité). La raison de cette différence entre le cas scalaire et vectoriel est que ce théorème s'appuie sur le principe du maximum (voir le Théorème 5.2.22) qui n'est valable que dans le cas scalaire. •

**Démonstration.** Soit  $u \in H_0^1(\Omega)$  un vecteur propre non nul associé à la première valeur propre  $\lambda_1$ . D'après le Lemme 5.2.24 on sait que  $u^+ = \max(u, 0)$  appartient à  $H_0^1(\Omega)$  et  $\nabla u^+ = 1_{u>0} \nabla u$  (de même pour  $u^- = \min(u, 0)$ ). Par conséquent, la fonction  $|u| = u^+ - u^-$  appartient à  $H_0^1(\Omega)$  et on a  $\nabla |u| = \text{sign}(u) \nabla u$ . En vertu de la Proposition 7.3.4 de Courant-Fisher on a

$$\lambda_1 = \min_{v \in H_0^1(\Omega) \setminus \{0\}} \left\{ R(v) \equiv \frac{\int_{\Omega} |\nabla v|^2 dx}{\int_{\Omega} v^2 dx} \right\},$$

et tout point de minimum est vecteur propre. Or  $\lambda_1 = R(u) = R(|u|)$ , donc  $|u|$  est aussi vecteur propre associé à  $\lambda_1$ . Comme  $u^+$  et  $u^-$  sont combinaisons linéaires de  $u$  et  $|u|$ , ce sont aussi des fonctions propres associées à  $\lambda_1$ .

En fait, on peut montrer que  $u$  ne s'annule pas dans  $\Omega$  grâce au principe du maximum "fort" qui affirme que, si  $w \in C^2(\overline{\Omega})$  vérifie

$$-\Delta w \geq 0 \text{ dans } \Omega, \text{ et } w = 0 \text{ sur } \partial\Omega,$$

alors soit  $w \equiv 0$  dans  $\Omega$ , soit  $w > 0$  dans  $\Omega$ . On applique ce résultat à  $u^+$  et  $u^-$  (qui sont régulières à cause de la Proposition 7.3.9) qui ne peuvent pas être toutes les deux non nulles, donc l'une des deux est nulle. Supposons alors que le sous-espace propre associé à  $\lambda_1$  soit de dimension strictement plus grande que 1. On peut y trouver deux fonctions propres  $u_1$  et  $u_2$  orthogonales, c'est-à-dire que

$$\int_{\Omega} u_1 u_2 dx = 0,$$

ce qui est impossible puisqu'elles sont de signe constant, partout non nulles. □

**Exercice 7.3.6** Soit  $\Omega$  un ouvert borné régulier et connexe. Montrer que la première valeur propre du Laplacien dans  $\Omega$  avec condition aux limites de Neumann est nulle et qu'elle est simple.

**Remarque 7.3.12** L'ensemble des résultats de cette sous-section se généralise sans difficulté aux opérateurs elliptiques généraux du second ordre, c'est-à-dire au problème aux valeurs propres suivant

$$\begin{cases} -\text{div}(A \nabla u) = \lambda u & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases}$$

où  $A(x)$  est une matrice symétrique coercive (voir la Sous-section 5.2.3). •

### 7.3.3 Autres modèles

L'extension des résultats de la sous-section précédente à des équations aux dérivées partielles elliptiques plus compliquées que le Laplacien ne pose pas de problèmes conceptuels nouveaux. Nous décrivons brièvement cette généralisation pour deux exemples significatifs : le système de l'élasticité linéarisée et les équations de Stokes.

Les équations (5.56) de l'élasticité linéarisée décrivent en fait le régime stationnaire des équations dynamiques suivantes (très semblables à l'équation des ondes)

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} - \operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+, \end{cases} \quad (7.18)$$

où  $\rho > 0$  est la densité volumique du matériau et  $e(u) = (\nabla u + (\nabla u)^t)/2$ . Rappelons que les coefficients de Lamé du matériau vérifient  $\mu > 0$  et  $2\mu + N\lambda > 0$ . En l'absence de forces extérieures  $f$  (et en ne tenant pas compte d'éventuelles conditions initiales) on peut aussi chercher des solutions oscillantes en temps de (7.18) comme nous l'avons décrit pour l'équation des ondes dans Sous-section 7.1.2. Cela conduit à chercher des solutions  $(\ell, u)$  du problème aux valeurs propres suivant

$$\begin{cases} -\operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = \ell u & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (7.19)$$

où  $\ell = \omega^2$  est le carré de la fréquence de vibration (nous avons changé la notation de la valeur propre pour éviter une confusion avec le coefficient de Lamé  $\lambda$ ). En mécanique, la fonction propre  $u$  est aussi appelée **mode propre de vibration**.

En suivant la méthode appliquée ci-dessus au Laplacien on peut démontrer le résultat suivant (nous laissons les détails au lecteur en guise d'exercice).

**Proposition 7.3.13** *Soit  $\Omega$  un ouvert borné régulier de classe  $C^1$  de  $\mathbb{R}^N$ . Il existe une suite croissante  $(\ell_k)_{k \geq 1}$  de réels positifs qui tend vers l'infini, et il existe une base hilbertienne de  $L^2(\Omega)^N$   $(u_k)_{k \geq 1}$ , telle que chaque  $u_k$  appartient à  $H_0^1(\Omega)^N$  et vérifie*

$$\begin{cases} -\operatorname{div} (2\mu e(u_k) + \lambda \operatorname{tr}(e(u_k)) \operatorname{Id}) = \ell_k u_k & \text{p.p. dans } \Omega \\ u_k = 0 & \text{p.p. sur } \partial\Omega. \end{cases}$$

Le résultat de régularité sur les fonctions propres  $u_k$  de la Proposition 7.3.9 s'étend aussi facilement au cas de l'élasticité et du problème (7.19). Par contre le Théorème 7.3.10 sur la simplicité de la première valeur propre et la positivité de la première fonction propre est faux en général (comme est faux le principe du maximum). Comme exemple nous calculons par la méthode des éléments finis  $Q_1$  les 4 premiers modes propres d'une "tour" dont la base est fixée (condition aux limites de Dirichlet) et dont les autres parois sont libres (condition aux limites de Neumann). Les 2 premiers modes, dits de battement, correspondent à la même valeur propre (ils sont indépendants mais symétriques par rotation de  $90^\circ$  suivant l'axe  $z$ ) (voir la Figure 7.1 et le Tableau 7.3.3). La première valeur propre est donc "double".



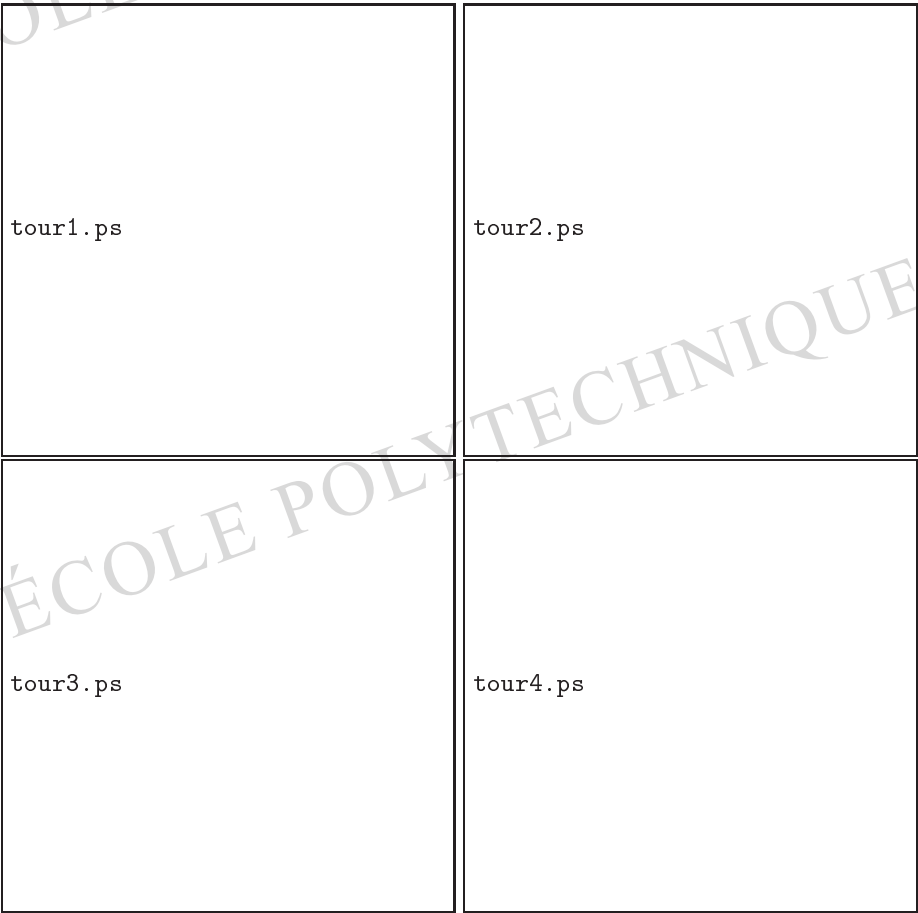


FIGURE 7.1 – Les 4 premiers modes propres d’une “tour” en élasticité.

Rang du mode propre	1	2	3	4
Valeur propre	102.54	102.54	1885.2	2961.2

TABLE 7.1 – Valeurs propres correspondant aux modes propres de la Figure 7.1.

Nous passons maintenant aux équations de Stokes (5.71) qui sont une version stationnaire d'un problème d'évolution de type parabolique (voir plus loin (8.2)). Pour résoudre ce problème d'évolution il sera intéressant d'utiliser les valeurs et fonctions propres  $(\lambda, u, p)$  du problème suivant

$$\begin{cases} \nabla p - \mu \Delta u = \lambda u & \text{dans } \Omega \\ \operatorname{div} u = 0 & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (7.20)$$

où  $\mu > 0$  est la viscosité,  $u$  la vitesse et  $p$  la pression du fluide. En suivant la méthode appliquée ci-dessus au Laplacien, le lecteur pourra résoudre l'exercice suivant.

**Exercice 7.3.7** Soit  $\Omega$  un ouvert borné régulier connexe de classe  $C^1$  de  $\mathbb{R}^N$ . Montrer qu'il existe une suite croissante  $(\lambda_k)_{k \geq 1}$  de réels positifs qui tend vers l'infini, et une base hilbertienne  $(u_k)_{k \geq 1}$  du sous-espace de  $L^2(\Omega)^N$  des fonctions à divergence nulle, telle que chaque  $u_k$  appartient à  $H_0^1(\Omega)^N$ , et il existe une famille de pressions  $p_k \in L^2(\Omega)$  qui vérifient

$$\begin{cases} \nabla p_k - \mu \Delta u_k = \lambda_k u_k & \text{p.p. dans } \Omega \\ \operatorname{div} u_k = 0 & \text{p.p. dans } \Omega \\ u_k = 0 & \text{p.p. sur } \partial\Omega. \end{cases}$$

Le résultat de régularité sur les fonctions propres de la Proposition 7.3.9 s'étend aussi facilement au cas des équations de Stokes (7.20). Par contre le Théorème 7.3.10 sur la simplicité de la première valeur propre et la positivité de la première fonction propre est faux en général (comme est faux le principe du maximum).

**Exercice 7.3.8** On considère le problème aux valeurs propres pour l'équation de Schrödinger avec un potentiel quadratique  $V(x) = Ax \cdot x$  où  $A$  est une matrice symétrique définie positive (modèle de l'oscillateur harmonique)

$$-\Delta u + Vu = \lambda u \quad \text{dans } \mathbb{R}^N. \quad (7.21)$$

On définit les espaces  $H = L^2(\mathbb{R}^N)$  et

$$V = \{v \in H^1(\mathbb{R}^N) \text{ tel que } |x|v(x) \in L^2(\mathbb{R}^N)\}.$$

Montrer que  $V$  est un espace de Hilbert pour le produit scalaire

$$\langle u, v \rangle_V = \int_{\mathbb{R}^N} \nabla u(x) \cdot \nabla v(x) dx + \int_{\mathbb{R}^N} |x|^2 u(x)v(x) dx,$$

et que l'injection de  $V$  dans  $H$  est compacte. En déduire qu'il existe une suite croissante  $(\lambda_k)_{k \geq 1}$  de réels positifs qui tend vers l'infini et une base hilbertienne de  $L^2(\mathbb{R}^N)$   $(u_k)_{k \geq 1}$  qui sont les valeurs propres et les fonctions propres de (7.21). Calculer explicitement ses valeurs et fonctions propres (on cherchera  $u_k$  sous la forme  $p_k(x) \exp(-Ax \cdot x/2)$  où  $p_k$  est un polynôme de degré  $k-1$ ). Interpréter physiquement les résultats.

**Exercice 7.3.9** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . On considère le problème de vibrations pour l'équation des plaques avec condition aux limites d'encastrement

$$\begin{cases} \Delta(\Delta u) = \lambda u & \text{dans } \Omega \\ \frac{\partial u}{\partial n} = u = 0 & \text{sur } \partial\Omega. \end{cases}$$

Montrer qu'il existe une suite croissante  $(\lambda_k)_{k \geq 1}$  de valeurs propres positives qui tend vers l'infini et une base hilbertienne dans  $L^2(\Omega)$  de fonctions propres  $(u_k)_{k \geq 1}$  qui appartiennent à  $H_0^2(\Omega)$ .

## 7.4 Méthodes numériques

### 7.4.1 Discrétisation par éléments finis

On va considérer une approximation interne de la formulation variationnelle introduite à la Sous-section 7.3.1. Étant donné un sous-espace  $V_h$  de l'espace de Hilbert  $V$ , de dimension finie, on cherche les solutions  $(\lambda_h, u_h) \in \mathbb{R} \times V_h$  de

$$a(u_h, v_h) = \lambda_h \langle u_h, v_h \rangle_H \quad \forall v_h \in V_h. \quad (7.22)$$

Typiquement,  $V_h$  est un espace d'éléments finis comme ceux introduits par les Définitions 6.3.5 et 6.3.25, et  $H$  est l'espace  $L^2(\Omega)$ . La résolution de l'approximation interne (7.22) est facile comme le montre le lemme suivant.

**Lemme 7.4.1** *On se place sous les hypothèses du Théorème 7.3.2. Alors les valeurs propres de (7.22) forment une suite croissante finie*

$$0 < \lambda_1 \leq \dots \leq \lambda_{n_{dl}} \quad \text{avec } n_{dl} = \dim V_h,$$

*et il existe une base de  $V_h$ , orthonormale dans  $H$ ,  $(u_{k,h})_{1 \leq k \leq n_{dl}}$  de vecteurs propres associés, c'est-à-dire que*

$$u_{k,h} \in V_h, \quad \text{et } a(u_{k,h}, v_h) = \lambda_k \langle u_{k,h}, v_h \rangle_H \quad \forall v_h \in V_h.$$

**Démonstration.** Ce lemme peut être considéré comme une variante évidente du Théorème 7.3.2 (à la différence près qu'en dimension finie il existe un nombre fini de valeurs propres). Néanmoins nous en donnons une démonstration différente, purement algébrique, qui correspond plus à la démarche suivie en pratique. Soit  $(\phi_i)_{1 \leq i \leq n_{dl}}$  une

base de  $V_h$  (par exemple, les fonctions de base d'une méthode d'éléments finis, voir la Proposition 6.3.7). On cherche  $u_h$  solution de (7.22) 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 (7.22) est équivalent à trouver  $(\lambda_h, U_h) \in \mathbb{R} \times \mathbb{R}^{n_{dl}}$  solution de

$$\mathcal{K}_h U_h = \lambda_h \mathcal{M}_h U_h. \quad (7.23)$$

Les appellations “matrices de masse et de rigidité” proviennent des applications en mécanique des solides. Remarquons que, dans le cas où  $V_h$  est un espace d'éléments finis, la matrice de rigidité  $\mathcal{K}_h$  est exactement la même matrice que celle rencontrée au Chapitre 6 dans l'application de la méthode des éléments finis aux problèmes elliptiques. On vérifie immédiatement que les matrices  $\mathcal{M}_h$  et  $\mathcal{K}_h$  sont symétriques et définies positives. Le système (7.23) 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

$$\mathcal{M}_h = P_h P_h^*, \text{ et } \mathcal{K}_h = P_h \text{diag}(\lambda_k) P_h^*.$$

Par conséquent, les solutions de (7.23) sont les valeurs propres  $(\lambda_k)$  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  $\mathcal{M}_h$  (nous indiquerons brièvement à la Remarque 7.4.3 comment calculer cette base). 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  $H$ .  $\square$

**Remarque 7.4.2** Dans le Lemme 7.4.1 on a repris les hypothèses du Théorème 7.3.2 : en particulier, la forme bilinéaire  $a(u, v)$  est supposée **symétrique**. On voit bien l'importance de cette hypothèse dans la démonstration. En effet, si elle n'était pas symétrique, on ne saurait pas si le système (7.23) est diagonalisable, c'est-à-dire s'il existe des solutions du problème aux valeurs propres (7.22).  $\bullet$

L'application du Lemme 7.4.1 à l'approximation variationnelle par éléments finis du problème de Dirichlet (7.17) est immédiate. On prend  $V = H_0^1(\Omega)$ ,  $H = L^2(\Omega)$ , et l'espace discret  $V_{0h}$  de la Définition 6.3.5 (rappelons que  $V_{0h}$  contient la condition aux limites de Dirichlet).

**Exercice 7.4.1** On considère le problème aux valeurs propres en dimension  $N = 1$

$$\begin{cases} -u_k'' = \lambda_k u_k & \text{pour } 0 < x < 1 \\ u_k(0) = u_k(1) = 0. \end{cases}$$

On se propose de calculer la matrice de masse pour la méthode des éléments finis  $P_1$ . On reprend les notations de la Section 6.2. Montrer que la matrice de masse  $\mathcal{M}_h$  est donnée par

$$\mathcal{M}_h = h \begin{pmatrix} 2/3 & 1/6 & & & 0 \\ 1/6 & 2/3 & 1/6 & & \\ & \ddots & \ddots & \ddots & \\ & & 1/6 & 2/3 & 1/6 \\ 0 & & & 1/6 & 2/3 \end{pmatrix},$$

et que ses valeurs propres sont

$$\lambda_k(\mathcal{M}_h) = \frac{h}{3} (2 + \cos(k\pi h)) \quad \text{pour } 1 \leq k \leq n.$$

Montrer que, si on utilise la formule de quadrature (6.45), alors on trouve que  $\mathcal{M}_h = h \text{Id}$ . Dans ce dernier cas, calculer les valeurs propres du problème spectral discret.

**Remarque 7.4.3** Pour calculer les valeurs et vecteurs propres du problème spectral matriciel (7.23) il faut, en général, commencer par calculer la factorisation de Cholesky de la matrice de masse  $\mathcal{M}_h = \mathcal{L}_h \mathcal{L}_h^*$ , pour se ramener au cas classique

$$\tilde{\mathcal{K}}_h \tilde{U}_h = \lambda_h \tilde{U}_h \quad \text{avec } \tilde{\mathcal{K}}_h = \mathcal{L}_h^{-1} \mathcal{K}_h (\mathcal{L}_h^*)^{-1} \text{ et } \tilde{U}_h = \mathcal{L}_h^* U_h,$$

pour lequel on dispose d'algorithmes de calcul de valeurs et vecteurs propres. Nous renvoyons à la Section 13.2 pour plus de détails sur ces algorithmes : disons seulement que c'est l'étape la plus coûteuse en temps de calcul.

On peut éviter de construire la matrice  $\tilde{\mathcal{K}}_h$  et faire l'économie de la factorisation de Cholesky de  $\mathcal{M}_h$  si on utilise une formule de quadrature pour évaluer les coefficients de la matrice  $\mathcal{M}_h$  qui la rende **diagonale**. Ce procédé d'intégration numérique est appelé **condensation de masse** (ou "mass lumping" en anglais) et est fréquemment utilisé. Par exemple, si on utilise la formule de quadrature (6.45) (qui utilise uniquement les valeurs aux noeuds d'une fonction pour calculer une intégrale), on voit facilement que la matrice de masse  $\mathcal{M}_h$  ainsi obtenue est diagonale (voir l'Exercice 7.4.1). •

Nous verrons dans la sous-section suivante que **seules les premières valeurs propres** discrètes  $\lambda_{k,h}$  (les plus petites) sont des approximations correctes des valeurs propres exactes  $\lambda_k$  (même chose pour les vecteurs propres). Il faut donc faire attention au fait que les dernières valeurs propres (les plus grandes) du problème discret (7.23) n'ont aucune signification physique! Par conséquent, si on est intéressé par la millième valeur propre du problème de Dirichlet (7.17), il faut prendre un maillage suffisamment

fin du domaine  $\Omega$  afin que la dimension de l'espace d'éléments finis  $V_h$  soit bien plus grande que mille.

Nous illustrons cette sous-section par le calcul des 6 premiers modes propres de vibration d'un tambour (modélisé comme une membrane circulaire fixée sur son bord). On résout donc le problème de Dirichlet (7.17) dans un disque de rayon 1. On utilise une méthode d'éléments finis  $P_1$ . Les résultats sont présentés dans la Figure 7.2 et le Tableau 7.4.1). On remarque que la première valeur propre est simple tandis que la deuxième et la troisième sont "doubles".

Rang du mode propre	1	2	3	4	5	6
Valeur propre	5.78	14.69	14.69	26.42	26.42	30.53

TABLE 7.2 – Valeurs propres correspondant aux modes propres de la Figure 7.2.

## 7.4.2 Convergence et estimations d'erreur

Dans cette sous-section, nous nous contentons d'énoncer un résultat de convergence de la méthode des éléments finis triangulaires  $P_k$  pour le calcul des valeurs et vecteurs propres du problème de Dirichlet (7.17). Il est bien clair que ce résultat se généralise aisément à d'autres problèmes et à d'autres types d'éléments finis.

**Théorème 7.4.4** *Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages triangulaires réguliers de  $\Omega$ . Soit  $V_{0h}$  le sous-espace de  $H_0^1(\Omega)$ , défini par la méthode des éléments finis  $P_k$ , de dimension  $n_{dl}$ . Soit  $(\lambda_i, u_i) \in \mathbb{R} \times H_0^1(\Omega)$ , pour  $i \geq 1$ , les valeurs et vecteurs propres (orthonormés dans  $L^2(\Omega)$ ) du problème de Dirichlet (6.36), rangés par ordre croissant*

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_i \leq \lambda_{i+1} \dots$$

*Soit les valeurs propres*

$$0 < \lambda_{1,h} \leq \lambda_{2,h} \leq \dots \leq \lambda_{n_{dl},h},$$

*de l'approximation variationnelle (7.22) dans  $V_{0h}$ . Pour tout  $i \geq 1$  fixé, on a*

$$\lim_{h \rightarrow 0} |\lambda_i - \lambda_{i,h}| = 0. \quad (7.24)$$

*Il existe une famille de vecteurs propres  $(u_{i,h})_{1 \leq i \leq n_{dl}}$  de (7.22) dans  $V_{0h}$  telle que, si  $\lambda_i$  est une valeur propre simple, on a*

$$\lim_{h \rightarrow 0} \|u_i - u_{i,h}\|_{H^1(\Omega)} = 0. \quad (7.25)$$

*De plus, si le sous-espace engendré par  $(u_1, \dots, u_i)$  est inclus dans  $H^{k+1}(\Omega)$  et si  $k+1 > N/2$ , alors on a l'estimation d'erreur*

$$|\lambda_i - \lambda_{i,h}| \leq C_i h^{2k}, \quad (7.26)$$

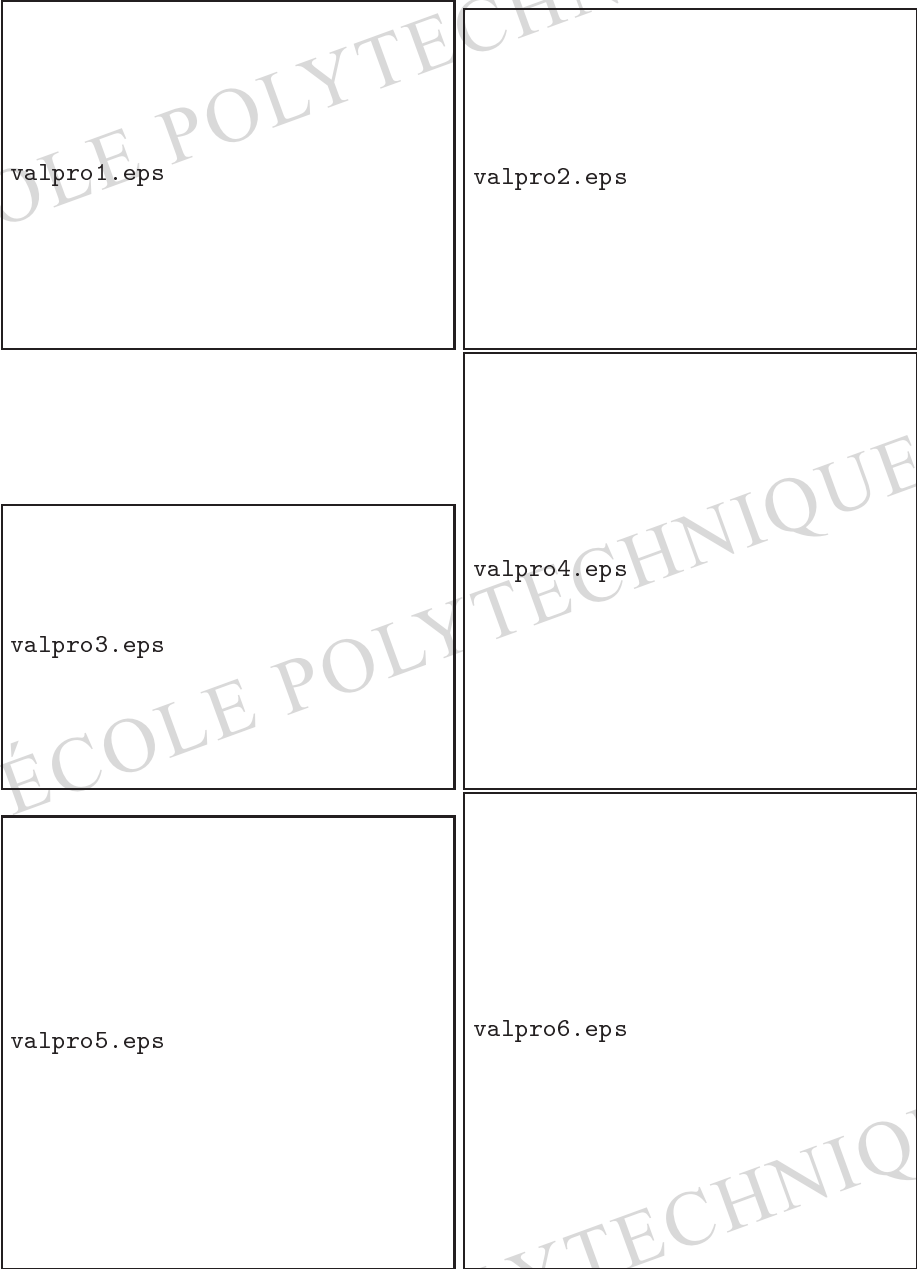


FIGURE 7.2 – Les 6 premiers modes propres d’un tambour.

où  $C_i$  ne dépend pas de  $h$ , et si  $\lambda_i$  est une valeur propre simple, on a

$$\|u_i - u_{i,h}\|_{H^1(\Omega)} \leq C_i h^k. \quad (7.27)$$

**Remarque 7.4.5** La constante  $C_i$  dans (7.26) ou (7.27) tends vers  $+\infty$  lorsque  $i \rightarrow +\infty$ , ce qui fait qu'il n'y a aucune garantie que les plus grandes valeurs propres discrètes (par exemple,  $\lambda_{n_{dl},h}$ ) soient proches des valeurs propres exactes  $\lambda_i$ .

L'ordre de convergence des valeurs propres est le double de celui de la convergence des vecteurs propres. Il s'agit là d'un phénomène général d'approximation spectrale d'opérateurs auto-adjoints qu'on retrouvera aussi dans les algorithmes numériques de la Section 13.2 (voir la Proposition 13.2.1).

La convergence des vecteurs propres ne peut s'obtenir que si la valeur propre correspondante est simple. En effet, si la valeur propre  $\lambda_i$  est multiple, la suite des vecteurs propres approchés  $u_{i,h}$  peut ne pas converger et admettre plusieurs points d'accumulation qui sont différentes combinaisons linéaires des vecteurs propres du sous-espace propre associé à  $\lambda_i$ . •



## Chapitre 8

# PROBLÈMES D'ÉVOLUTION

### 8.1 Motivation et exemples

#### 8.1.1 Introduction

Ce chapitre est consacré à l'analyse mathématique et numérique des problèmes d'évolution en temps (dans les chapitres précédents nous avons étudié des problèmes stationnaires sans variable de temps). Nous allons plus particulièrement analyser deux types différents d'équations aux dérivées partielles : celles de type parabolique, et celles de type hyperbolique. L'exemple typique d'équation parabolique est l'équation de la chaleur que nous étudierons en détail (mais notre analyse s'étend à des modèles plus compliqués comme les équations de Stokes instationnaires). Le prototype d'équation hyperbolique est l'équation des ondes sur laquelle nous nous concentrerons (mais encore une fois notre analyse s'étend à des modèles plus compliqués comme les équations de l'élastodynamique ou celles de l'électromagnétisme). Plus généralement, l'approche développée ici s'étend à beaucoup d'autres problèmes d'évolution en temps, pas nécessairement de type parabolique ou hyperbolique, comme, par exemple, l'équation de Schrödinger de la mécanique quantique.

Le plan de ce chapitre est le suivant. Le reste de cette section est consacré à quelques questions liées à la modélisation. Dans les Sections 8.2 et 8.3 nous démontrons l'existence et l'unicité de la solution de l'équation de la chaleur ou des ondes en utilisant à nouveau le concept de **formulation variationnelle**. Comme nous l'avons laissé entendre dans la Sous-section 7.1.2, nous utilisons pour cela des **bases hilbertiennes de fonctions propres** construites au Chapitre 7. Nous insistons aussi sur la notion d'**estimations d'énergie** qui exprime un bilan d'énergie physique et qui justifie en partie les espaces utilisés par la théorie. Dans les Sections 8.4 et 8.5 nous étudierons certaines **propriétés qualitatives** des solutions. Indiquons tout de suite que, si les résultats d'existence et d'unicité sont très semblables pour les équations de la chaleur et des ondes, **leurs propriétés qualitatives sont par contre très différentes**. Nous verrons aussi que, si certaines de ces propriétés qualitatives sont

conformes à l'intuition physique (comme le principe du maximum pour la chaleur, et la conservation de l'énergie pour les ondes), d'autres sont plus surprenantes, et à ce titre particulièrement intéressantes (comme la vitesse de propagation "infinie" de la chaleur, et la réversibilité en temps des ondes). Tous ces résultats seront démontrés pour une équation posée dans un domaine borné. Néanmoins, nous dirons un mot de la situation lorsque l'équation est posée dans l'espace tout entier  $\mathbb{R}^N$ ; dans ce cas on peut obtenir une formule explicite de la solution en utilisant une fonction de Green.

Les Sections 8.6 et 8.7 sont consacrées à la **résolution numérique** des équations de la chaleur et des ondes. Nous avons déjà exploré largement la méthode des différences finies pour ces problèmes (voir les Chapitres 1 et 2). Aussi nous nous concentrerons sur l'utilisation de la méthode des **éléments finis** dans ce contexte. Plus précisément, conformément à l'usage, nous utiliserons des éléments finis pour la discrétisation spatiale, mais des différences finies pour la discrétisation temporelle.

### 8.1.2 Modélisation et exemples d'équations paraboliques

Présentons rapidement les principaux problèmes paraboliques que nous étudierons dans ce chapitre, en disant quelques mots de leur origine physique ou mécanique. L'archétype de ces modèles est **l'équation de la chaleur** dont l'origine physique a déjà été discutée au Chapitre 1. Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$  de frontière  $\partial\Omega$ . Pour des conditions aux limites de Dirichlet ce modèle s'écrit

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_+^* \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_+^* \\ u(x, 0) = u_0(x) & \text{pour } x \in \Omega. \end{cases} \quad (8.1)$$

Le problème aux limites (8.1) modélise l'évolution de la température  $u(x, t)$  dans un corps thermiquement conducteur qui occupe le domaine  $\Omega$ . 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 (8.1) puisqu'il s'agit d'une équation aux dérivées partielles du premier ordre en  $t$  et du deuxième ordre en  $x$  (le Laplacien ne porte que sur la variable spatiale).

Indiquons qu'il existe d'autres origines physiques du système (8.1). Par exemple, (8.1) modélise aussi la diffusion d'une concentration  $u$  dans le domaine  $\Omega$ , 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  $\Omega$ .

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  $\Omega$  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 (voir la Sous-section 5.2.3). 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 que nous avons rapidement évoqué au chapitre précédent. En notant  $u$  la vitesse et  $p$  la pression d'un fluide visqueux soumis à des forces  $f$ , ce système s'écrit

$$\begin{cases} \frac{\partial u}{\partial t} + \nabla p - \mu \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ \operatorname{div} u = 0 & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(x, t = 0) = u_0(x) & \text{dans } \Omega \end{cases} \quad (8.2)$$

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  $\Omega$  (voir le Chapitre 1), et que le système de Stokes n'est valable que pour des vitesses faibles (voir la Remarque 5.3.7). La plupart des résultats que nous verrons dans ce chapitre se généralise à de tels modèles.

### 8.1.3 Modélisation et exemples d'équations hyperboliques

Présentons rapidement les deux principaux modèles hyperboliques que nous étudierons dans ce chapitre. Le premier modèle est **l'équation des ondes** dont l'origine physique a déjà été discutée au Chapitre 1. Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$  de frontière  $\partial\Omega$ . Pour des conditions aux limites de Dirichlet ce modèle s'écrit

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0(x) & \text{dans } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1(x) & \text{dans } \Omega. \end{cases} \quad (8.3)$$

Le problème aux limites (8.3) 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.

Le deuxième modèle est **l'élastodynamique** qui est la version d'évolution en temps des équations de l'élasticité linéarisée (voir les Chapitres 1 et 5). 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 (8.3). Néanmoins, une différence importante avec (8.3) 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

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} - \operatorname{div} (2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(t = 0) = u_0(x) & \text{dans } \Omega \\ \frac{\partial u}{\partial t}(t = 0) = u_1(x) & \text{dans } \Omega, \end{cases} \quad (8.4)$$

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  $\Omega$ , sa densité est constante  $\rho > 0$ , de même que ses modules de Lamé qui vérifient  $\mu > 0$  et  $2\mu + N\lambda > 0$ .

**Remarque 8.1.1** On peut ajouter aux équations (8.3) et (8.4) un terme du premier ordre en temps, ce qui donne

$$\frac{\partial^2 u}{\partial t^2} + \eta \frac{\partial u}{\partial t} - \Delta u = f \text{ dans } \Omega \times \mathbb{R}_*^+.$$

Lorsque le coefficient  $\eta$  est positif, ce terme du premier ordre correspond à une force de freinage proportionnelle à la vitesse. On dit aussi qu'il s'agit d'un terme d'amortissement. On parle alors **d'équation des ondes amorties**. •

**Remarque 8.1.2** Il existe d'autres modèles physiques donnant lieu à des équations aux dérivées partielles hyperboliques. Cependant, tout modèle hyperbolique n'est pas nécessairement un problème d'évolution d'ordre deux. C'est le cas notamment des équations d'Euler linéarisées en acoustique, ou des équations de Maxwell en électromagnétisme, qui sont des systèmes d'équations hyperboliques d'ordre un seulement en temps. Les idées contenues dans ce chapitre s'étendent à ces problèmes, mais la différence d'ordre en temps en change la présentation. •

## 8.2 Existence et unicité dans le cas parabolique

Nous allons suivre une démarche similaire dans l'esprit à celle qui nous a guidée au Chapitre 5 pour établir l'existence et l'unicité de la solution d'un problème elliptique. Cette démarche se décompose en trois étapes : premièrement, établir une formulation variationnelle (Sous-section 8.2.1), deuxièmement, démontrer l'existence et l'unicité de la solution de cette formulation variationnelle en utilisant une base hilbertienne de fonctions propres (Sous-section 8.2.2), troisièmement, montrer que cette solution vérifie bien le problème aux limites étudié (Sous-section 8.2.3).

### 8.2.1 Formulation variationnelle

L'idée est d'écrire une formulation variationnelle qui ressemble à une **équation différentielle ordinaire** du premier ordre, similaire à (7.3). Pour cela nous multiplions l'équation de la chaleur (8.1) par une fonction test  $v(x)$  qui ne dépend pas du temps  $t$ . A cause de la condition aux limites nous allons demander à ce que  $v$  s'annule sur le bord de l'ouvert  $\Omega$ , ce qui va nous permettre d'effectuer une intégration par partie simple (sans terme de bord). Pour l'instant ce calcul est principalement formel. Nous obtenons donc

$$\int_{\Omega} \frac{\partial u}{\partial t}(x, t) v(x) dx + \int_{\Omega} \nabla u(x, t) \cdot \nabla v(x) dx = \int_{\Omega} f(x, t) v(x) dx. \quad (8.5)$$

Comme ni  $\Omega$  ni  $v(x)$  ne varient avec le temps  $t$ , on peut réécrire cette équation sous la forme

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

Exploitant le fait que les variables  $x$  et  $t$  jouent des rôles très différents, nous séparons ces variables en considérant désormais la solution  $u(t, x)$  comme une fonction du temps  $t$  à valeurs dans un espace de fonctions définies sur  $\Omega$  (même chose pour  $f(t, x)$ ). Plus précisément, si l'on se donne un temps final  $T > 0$  (éventuellement égal à  $+\infty$ ), on considère que  $u$  est définie par

$$\begin{aligned} u : ]0, T[ &\rightarrow H_0^1(\Omega) \\ t &\rightarrow u(t), \end{aligned}$$

et nous continuerons à noter  $u(x, t)$  la valeur  $u(t)(x)$ . Le choix de l'espace  $H_0^1(\Omega)$  est évidemment dicté par la nature du problème et peut varier d'un modèle à un autre. En général il s'agit de l'espace qui convient pour la formulation variationnelle du problème stationnaire associé. De même, le terme source  $f$  est désormais considéré comme une fonction de  $t$  à valeurs dans  $L^2(\Omega)$ .

On introduit alors le produit scalaire de  $L^2(\Omega)$  et la forme bilinéaire  $a(w, v)$  définis par

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

En choisissant la fonction test dans l'espace  $H_0^1(\Omega)$ , on peut alors mettre (8.5) sous la forme d'une sorte **d'équation différentielle ordinaire** en  $t$ . On obtient ainsi la formulation variationnelle suivante : trouver  $u(t)$  fonction de  $]0, T[$  à valeurs dans  $H_0^1(\Omega)$  telle que

$$\begin{cases} \frac{d}{dt} \langle u(t), v \rangle_{L^2(\Omega)} + a(u(t), v) = \langle f(t), v \rangle_{L^2(\Omega)} & \forall v \in H_0^1(\Omega), \quad 0 < t < T, \\ u(t=0) = u_0. \end{cases} \quad (8.6)$$

Il reste plusieurs points à préciser dans la formulation variationnelle (8.6) pour lui donner un sens mathématique précis : quelle est la régularité en temps de  $f$  et de  $u$ , et quel sens donner à la dérivée en temps ? En particulier, il faudra absolument que  $u(t)$  soit continue en  $t = 0$  pour donner un sens correct à la donnée initiale  $u_0$ .

Pour cela nous avons besoin d'introduire une famille d'espaces fonctionnels de fonctions de  $t$  à valeurs dans des espaces de fonctions de  $x$ .

**Définition 8.2.1** Soit  $X$  un espace de Hilbert, ou plus généralement, un espace de Banach défini sur  $\Omega$  (typiquement,  $X = L^2(\Omega)$ ,  $H_0^1(\Omega)$ , ou  $C(\overline{\Omega})$ ). Soit un temps final  $0 < T \leq +\infty$ . Pour un entier  $k \geq 0$ , on note  $C^k([0, T]; X)$  l'espace des fonctions  $k$  fois continûment dérivables de  $[0, T]$  dans  $X$ . Si on note  $\|v\|_X$  la norme dans  $X$ , il

est classique (voir [31]) que  $C^k([0, T]; X)$  est un espace de Banach pour la norme

$$\|v\|_{C^k([0, T]; X)} = \sum_{m=0}^k \left( \sup_{0 \leq t \leq T} \left\| \frac{d^m v}{dt^m}(t) \right\|_X \right).$$

On note  $L^2(]0, T[; X)$  l'espace des fonctions de  $]0, T[$  dans  $X$  telles que la fonction  $t \rightarrow \|v(t)\|_X$  soit mesurable et de carré intégrable, c'est-à-dire que

$$\|v\|_{L^2(]0, T[; X)} = \sqrt{\int_0^T \|v(t)\|_X^2 dt} < +\infty.$$

Muni de cette norme  $L^2(]0, T[; X)$  est aussi un espace de Banach. De plus, si  $X$  est un espace de Hilbert, alors  $L^2(]0, T[; X)$  est un espace de Hilbert pour le produit scalaire

$$\langle u, v \rangle_{L^2(]0, T[; X)} = \int_0^T \langle u(t), v(t) \rangle_X dt.$$

**Remarque 8.2.2** Si  $X$  est l'espace  $L^2(\Omega)$ , alors  $L^2(]0, T[; L^2(\Omega))$  s'identifie à l'espace  $L^2(]0, T[ \times \Omega)$  puisque, par le théorème de Fubini, on a

$$\|v\|_{L^2(]0, T[; L^2(\Omega))}^2 = \int_0^T \left( \int_{\Omega} |v(t)|^2(x) dx \right) dt = \int_0^T \int_{\Omega} |v(x, t)|^2 dx dt = \|v\|_{L^2(]0, T[ \times \Omega)}^2.$$

Pour  $1 \leq p < +\infty$ , on peut généraliser la Définition 8.2.1 en introduisant l'espace de Banach  $L^p(]0, T[; X)$  des fonctions de  $]0, T[$  dans  $X$  telles que la fonction  $t \rightarrow \|v(t)\|_X$  soit mesurable et de puissance  $p$ -ème intégrable. •

Dans la suite on prendra le terme source  $f$  dans l'espace  $L^2(]0, T[; L^2(\Omega))$ , et on cherchera la solution  $u$  dans l'espace d'énergie  $L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ . Ce choix peut paraître arbitraire, mais il sera justifié, non seulement par la démonstration de l'existence d'une solution pour la formulation variationnelle (8.6), mais aussi par son lien avec des estimations d'énergie (voir l'Exercice 8.2.1). Notons déjà que, les fonctions de cet espace étant continues en temps à valeurs dans  $L^2(\Omega)$ , la condition initiale a bien un sens.

Finalement, la dérivée en temps dans la formulation variationnelle (8.6) doit être prise au sens faible puisqu'a priori la fonction  $t \rightarrow \langle u(t), v \rangle_{L^2(\Omega)}$  n'appartient qu'à  $L^2(0, T)$  (voir le Lemme 4.4.12 pour une définition précise de cette notion de dérivée). Fort heureusement, s'il existe une solution de (8.6), alors l'égalité dans (8.6) nous dit que cette dérivée en temps est tout ce qu'il y a de plus classique puisqu'elle appartient à  $L^2(]0, T[)$ .

### 8.2.2 Un résultat général

Pour démontrer l'existence et l'unicité de la solution de la formulation variationnelle (8.6), nous revenons au cadre général introduit dans la Section 7.3. Nous allons

pouvoir ainsi “diagonaliser” l’opérateur Laplacien et nous ramener à la résolution d’une famille de simples équations différentielles ordinaires du premier ordre. On introduit donc deux espaces de Hilbert  $V$  et  $H$  tels que  $V \subset H$  avec injection dense et compacte (voir (7.11) et l’explication qui s’en suit). Typiquement on aura  $V = H_0^1(\Omega)$  et  $H = L^2(\Omega)$ .

**Théorème 8.2.3** *Soient  $V$  et  $H$  deux espaces de Hilbert tels que  $V \subset H$  avec injection compacte et  $V$  est dense dans  $H$ . Soit  $a(u, v)$  une forme bilinéaire symétrique continue et coercive dans  $V$ . Soit un temps final  $T > 0$ , une donnée initiale  $u_0 \in H$ , et un terme source  $f \in L^2([0, T]; H)$ . Alors le problème*

$$\begin{cases} \frac{d}{dt} \langle u(t), v \rangle_H + a(u(t), v) = \langle f(t), v \rangle_H & \forall v \in V, \quad 0 < t < T, \\ u(t=0) = u_0, \end{cases} \quad (8.7)$$

(où l’équation de (8.7) a lieu au sens faible dans  $]0, T[$ ) a une unique solution  $u \in L^2([0, T]; V) \cap C([0, T]; H)$ . De plus, il existe une constante  $C > 0$  (qui ne dépend que de  $\Omega$ ) telle que

$$\|u\|_{L^2([0, T]; V)} + \|u\|_{C([0, T]; H)} \leq C (\|u_0\|_H + \|f\|_{L^2([0, T]; H)}). \quad (8.8)$$

**Remarque 8.2.4 L’estimation d’énergie** (8.8) prouve que la solution de (8.7) dépend continûment des données, et donc que le problème parabolique (8.7) est bien posé au sens de Hadamard. •

**Remarque 8.2.5** Dans le Théorème 8.2.3 on peut affaiblir l’hypothèse de coercivité de la forme bilinéaire symétrique  $a(u, v)$  (comme on l’a déjà proposé dans l’Exercice 7.3.1). On obtient les mêmes conclusions en supposant seulement qu’il existe deux constantes positives  $\nu > 0$  et  $\eta > 0$  telles que

$$a(v, v) + \eta \|v\|_H^2 \geq \nu \|v\|_V^2 \text{ pour tout } v \in V.$$

En effet, si l’on effectue le changement de fonction inconnue  $u(t) = e^{\eta t} w(t)$ , on voit que (8.7) est équivalent à

$$\begin{cases} \frac{d}{dt} \langle w(t), v \rangle_H + a(w(t), v) + \eta \langle w(t), v \rangle_H = \langle f(t), v \rangle_H & \forall v \in V, \quad 0 < t < T, \\ w(t=0) = u_0, \end{cases}$$

où la forme bilinéaire  $a(w, v) + \eta \langle w, v \rangle_H$  est bien coercive sur  $V$ . Cette hypothèse affaiblie est utile, par exemple, dans la résolution de l’Exercice 8.2.4. •

**Démonstration.** La démonstration est divisée en deux étapes. Dans une première étape, en supposant l’existence d’une solution  $u$ , nous obtenons une formule explicite pour  $u$  sous la forme d’une série obtenue par décomposition spectrale des espaces  $H$

et  $V$ . En particulier, cette formule prouve l'unicité de la solution. Dans une deuxième étape, nous démontrons que cette série converge dans les espaces  $L^2(]0, T[; V)$  et  $C([0, T]; H)$ , et que la somme est bien une solution de (8.7).

**Étape 1.** Supposons que  $u \in L^2(]0, T[; V) \cap C([0, T]; H)$  est solution de (8.7). Les hypothèses permettent d'appliquer le Théorème 7.3.2 sur la résolution du problème aux valeurs propres associé à la forme bilinéaire symétrique  $a(u, v)$ . Par conséquent, il existe une base hilbertienne  $(u_k)_{k \geq 1}$  de  $H$  composée de vecteurs propres de (7.12)

$$u_k \in V, \quad \text{et} \quad a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall v \in V.$$

On définit

$$\alpha_k(t) = \langle u(t), u_k \rangle_H, \quad \alpha_k^0 = \langle u_0, u_k \rangle_H, \quad \beta_k(t) = \langle f(t), u_k \rangle_H.$$

Puisque  $u \in L^2(]0, T[; V) \cap C([0, T]; H)$  et  $f \in L^2(]0, T[; H)$ , on en déduit que  $\alpha_k(t) \in C([0, T])$  et  $\beta_k(t) \in L^2(]0, T[)$ . Comme  $(u_k)_{k \geq 1}$  est une base hilbertienne de  $H$ , on a

$$u(t) = \sum_{k=1}^{+\infty} \alpha_k(t) u_k,$$

et choisissant  $v = u_k$  dans (8.7) on obtient

$$\begin{cases} \frac{d\alpha_k}{dt} + \lambda_k \alpha_k = \beta_k & \text{dans } ]0, T[ \\ \alpha_k(t=0) = \alpha_k^0. \end{cases} \quad (8.9)$$

On vérifie immédiatement que l'unique solution de (8.9) est

$$\alpha_k(t) = \alpha_k^0 e^{-\lambda_k t} + \int_0^t \beta_k(s) e^{-\lambda_k(t-s)} ds \quad \text{pour } t > 0,$$

ce qui donne une formule explicite pour la solution  $u$  (qui est donc unique).

**Étape 2.** Nous allons démontrer que la série

$$\sum_{j=1}^{+\infty} \left( \alpha_j^0 e^{-\lambda_j t} + \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right) u_j \quad (8.10)$$

converge dans  $L^2(]0, T[; V) \cap C([0, T]; H)$  et que sa somme, notée  $u(t)$  est solution de (8.7). Considérons la somme partielle à l'ordre  $k$  de cette série

$$w^k(t) = \sum_{j=1}^k \left( \alpha_j^0 e^{-\lambda_j t} + \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right) u_j. \quad (8.11)$$

Clairement  $w^k$  appartient à  $C([0, T]; H)$  puisque chaque  $\alpha_j(t)$  est continu. Montrons que la suite  $w^k$  est de Cauchy dans  $C([0, T]; H)$ . Pour  $l > k$ , en utilisant le caractère



orthonormé des fonctions propres  $u_j$ , on a

$$\begin{aligned}
 \|w^l(t) - w^k(t)\|_H &\leq \left\| \sum_{j=k+1}^l \alpha_j^0 e^{-\lambda_j t} u_j \right\|_H + \left\| \sum_{j=k+1}^l \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds u_j \right\|_H \\
 &\leq \left( \sum_{j=k+1}^l |\alpha_j^0|^2 e^{-2\lambda_j t} \right)^{1/2} + \left( \sum_{j=k+1}^l \left( \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right)^2 \right)^{1/2} \\
 &\leq \left( \sum_{j=k+1}^l |\alpha_j^0|^2 \right)^{1/2} + \left( \sum_{j=k+1}^l \frac{1}{2\lambda_j} \int_0^T |\beta_j(s)|^2 ds \right)^{1/2} \\
 &\leq \left( \sum_{j=k+1}^l |\alpha_j^0|^2 \right)^{1/2} + \frac{1}{\sqrt{2\lambda_1}} \left( \sum_{j=k+1}^l \int_0^T |\beta_j(s)|^2 ds \right)^{1/2},
 \end{aligned}$$

puisque la suite des valeurs propres  $(\lambda_j)$  est croissante et strictement positive. Comme  $u_0 \in H$  et  $f \in L^2(]0, T[; H)$  on a

$$\|u_0\|_H^2 = \sum_{j=1}^{+\infty} |\alpha_j^0|^2 < +\infty, \quad \|f\|_{L^2(]0, T[; H)}^2 = \sum_{j=1}^{+\infty} \int_0^T |\beta_j(s)|^2 ds < +\infty,$$

ce qui entraîne que la suite  $w^k(t)$  est de Cauchy dans  $H$ . Plus précisément, on en déduit que la suite  $w^k$  vérifie

$$\lim_{k, l \rightarrow +\infty} \left( \sup_{0 \leq t \leq T} \|w^l - w^k\|_H \right) = 0,$$

c'est-à-dire qu'elle est de Cauchy dans  $C([0, T]; H)$ .

Montrons que la suite  $w^k$  est aussi de Cauchy dans  $L^2(]0, T[; V)$ . On munit  $V$  du produit scalaire  $a(u, v)$  (équivalent au produit scalaire usuel à cause de la coercivité de  $a$ ). Pour  $l > k$  on a

$$\begin{aligned}
 \|w^l(t) - w^k(t)\|_V^2 &= a(w^l(t) - w^k(t), w^l(t) - w^k(t)) = \sum_{j=k+1}^l \lambda_j |\alpha_j(t)|^2 \\
 &\leq 2 \sum_{j=k+1}^l \lambda_j |\alpha_j^0|^2 e^{-2\lambda_j t} + 2 \sum_{j=k+1}^l \lambda_j \left( \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right)^2
 \end{aligned}$$

Or, par application de l'inégalité de Cauchy-Schwarz on a

$$\begin{aligned}
 \left( \int_0^t \beta_j(s) e^{-\lambda_j(t-s)} ds \right)^2 &\leq \left( \int_0^t |\beta_j(s)|^2 e^{-\lambda_j(t-s)} ds \right) \left( \int_0^t e^{-\lambda_j(t-s)} ds \right) \\
 &\leq \frac{1}{\lambda_j} \left( \int_0^t |\beta_j(s)|^2 e^{-\lambda_j(t-s)} ds \right).
 \end{aligned}$$

Par ailleurs, en vertu du théorème de Fubini

$$\begin{aligned} \int_0^T \left( \int_0^t |\beta_j(s)|^2 e^{-\lambda_j(t-s)} ds \right) dt &= \int_0^T |\beta_j(s)|^2 \left( \int_s^T e^{-\lambda_j(t-s)} dt \right) ds \\ &\leq \frac{1}{\lambda_j} \int_0^T |\beta_j(s)|^2 ds. \end{aligned}$$

Par conséquent, on en déduit que

$$\int_0^T \|w^l(t) - w^k(t)\|_V^2 dt \leq \sum_{j=k+1}^l |\alpha_j^0|^2 + \sum_{j=k+1}^l \frac{2}{\lambda_j} \int_0^T |\beta_j(s)|^2 ds,$$

ce qui implique que la suite  $w^k$  vérifie

$$\lim_{k,l \rightarrow +\infty} \int_0^T \|w^l(t) - w^k(t)\|_V^2 dt = 0,$$

c'est-à-dire qu'elle est de Cauchy dans  $L^2(]0, T[; V)$ .

Comme les deux espaces  $C([0, T]; H)$  et  $L^2(]0, T[; V)$  sont complets, la suite de Cauchy  $w^k$  converge et on peut définir sa limite  $u$

$$\lim_{k \rightarrow +\infty} w^k = u \text{ dans } C([0, T]; H) \cap L^2(]0, T[; V).$$

En particulier, comme  $w^k(0)$  converge vers  $u_0$  dans  $H$ , on en déduit la condition initiale voulue,  $u(0) = u_0$  (qui est une égalité entre fonctions de  $H$ ). D'autre part, il est clair que  $u(t)$ , en tant que somme de la série (8.10) vérifie la formulation variationnelle (8.7) pour chaque fonction test  $v = u_k$ . Comme  $(u_k/\sqrt{\lambda_k})$  est une base hilbertienne de  $V$ ,  $u(t)$  vérifie donc la formulation variationnelle (8.7) pour tout  $v \in V$ , c'est-à-dire que  $u(t)$  est bien la solution recherchée de (8.7).

Pour obtenir l'estimation d'énergie (8.8), il suffit de remarquer que l'on a prouvé les majorations

$$\|w^l(t) - w^k(t)\|_H \leq \|u_0\|_H + \frac{1}{\sqrt{2\lambda_1}} \|f\|_{L^2(]0, T[; H)}$$

et

$$\int_0^T \|w^l(t) - w^k(t)\|_V^2 dt \leq \|u_0\|_H^2 + \frac{2}{\lambda_1} \|f\|_{L^2(]0, T[; H)}^2.$$

En prenant  $k = 0$  et en faisant tendre  $l$  vers l'infini, on obtient immédiatement l'estimation désirée.  $\square$

**Remarque 8.2.6 (délicate)** Pour le lecteur épris de rigueur mathématique, nous revenons sur le sens de la dérivée en temps dans la formulation variationnelle (8.7). Au vu des espaces dans lequel nous cherchons la solution  $u(t)$ , la fonction  $t \rightarrow \langle u(t), v \rangle_H$  n'est pas dérivable

au sens classique : elle appartient seulement à  $L^2(0, T)$  et à  $C[0, T]$ ). On peut néanmoins définir sa dérivée au sens faible du Lemme 4.4.12 (ou au sens des distributions). Plus précisément,  $\frac{d}{dt}\langle u(t), v \rangle_H$  est défini comme un élément de  $H^{-1}(0, T)$  (c'est-à-dire une forme linéaire continue sur  $H_0^1(0, T)$ ) par la formule

$$\left\langle \frac{d}{dt}\langle u(t), v \rangle_H, \phi(t) \right\rangle_{H^{-1}, H_0^1(0, T)} = - \int_0^T \langle u(t), v \rangle_H \frac{d\phi}{dt}(t) dt \quad \forall \phi \in H_0^1(0, T).$$

Par conséquent, dire que l'équation de (8.7) a lieu au sens faible dans  $]0, T[$  est équivalent à dire que

$$- \int_0^T \langle u(t), v \rangle_H \frac{d\phi}{dt}(t) dt + \int_0^T a(u(t), v) \phi(t) dt = \int_0^T \langle f(t), v \rangle_H \phi(t) dt$$

pour tout  $v \in V$  et tout  $\phi \in C_c^\infty(]0, T[)$  puisque  $C_c^\infty(]0, T[)$  est dense dans  $H_0^1(0, T)$ . Pour conclure, rassurons le lecteur : si  $u$  est une solution de (8.7), alors, par l'égalité même qu'est (8.7), la dérivée  $\frac{d}{dt}\langle u(t), v \rangle_H$  appartient à  $L^2(0, T)$  et on peut donc dire que (8.7) a lieu presque partout dans  $]0, T[$ . •

### 8.2.3 Applications

Nous appliquons maintenant le résultat abstrait du Théorème 8.2.3 à l'équation de la chaleur, et nous prouvons que cette approche variationnelle a bien permis de résoudre l'équation aux dérivées partielles d'origine.

**Théorème 8.2.7** *Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . Soit un temps final  $T > 0$ , une donnée initiale  $u_0 \in L^2(\Omega)$ , et un terme source  $f \in L^2(]0, T[; L^2(\Omega))$ . Alors l'équation de la chaleur*

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & p.p. \text{ dans } \Omega \times ]0, T[ \\ u = 0 & p.p. \text{ sur } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & p.p. \text{ dans } \Omega. \end{cases} \quad (8.12)$$

*admet une unique solution  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ . De plus, il existe une constante  $C > 0$  (qui ne dépend que de  $\Omega$ ) telle que, pour tout  $t \in [0, T]$ ,*

$$\int_{\Omega} u(x, t)^2 dx + \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds \leq C \left( \int_{\Omega} u_0(x)^2 dx + \int_0^t \int_{\Omega} f(x, s)^2 dx ds \right). \quad (8.13)$$

**Démonstration.** Nous appliquons le Théorème 8.2.3 à la formulation variationnelle (8.6) de l'équation de la chaleur (8.12) : ses hypothèses sont facilement vérifiées avec  $H = L^2(\Omega)$  et  $V = H_0^1(\Omega)$  (en particulier, comme  $\Omega$  est borné le Théorème 4.3.21 de Rellich affirme que l'injection de  $H$  dans  $V$  est compacte). Il reste à montrer que l'unique solution  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$  de cette formulation variationnelle est bien une solution de (8.12). Tout d'abord, la condition aux limites

de Dirichlet se retrouve par application du Théorème 4.3.13 de trace à  $u(t) \in H_0^1(\Omega)$  pour presque tout  $t \in ]0, T[$ , et la condition initiale est justifiée par la continuité de  $u(t)$  en  $t = 0$  (comme fonction à valeurs dans  $L^2(\Omega)$ ).

Si la solution  $u$  est suffisamment régulière (par exemple, si  $\frac{\partial u}{\partial t}$  et  $\Delta u$  appartiennent à  $L^2([0, T] \times \Omega)$ , ce qui est vrai d'après la Proposition 8.4.6), par intégration par parties, la formulation variationnelle (8.6) est équivalente à

$$\int_{\Omega} \left( \frac{\partial u}{\partial t} - \Delta u - f \right) v \, dx = 0, \quad (8.14)$$

pour toute fonction  $v(x) \in H_0^1(\Omega)$  et presque tout temps  $t \in ]0, T[$ . Par conséquent, on déduit de (8.14) que

$$\frac{\partial u}{\partial t} - \Delta u - f = 0 \quad \text{p.p. dans } ]0, T[ \times \Omega.$$

Si la solution  $u$  n'est pas plus régulière que  $u \in L^2([0, T]; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ , on obtient tout de même cette égalité mais la justification en est légèrement plus délicate. Conformément à la Remarque 8.2.6 le sens précis de (8.6) est

$$-\int_0^T \int_{\Omega} uv \frac{d\phi}{dt} \, dx \, dt + \int_0^T \int_{\Omega} \nabla u \cdot \nabla v \phi \, dx \, dt = \int_0^T \int_{\Omega} f v \phi \, dx \, dt \quad (8.15)$$

pour toute fonction  $v(x) \in C_c^1(\Omega)$  et  $\phi(t) \in C_c^1([0, T])$ . Un résultat classique d'analyse nous dit que l'ensemble des combinaisons linéaires de produit de telles fonctions  $v(x)\phi(t)$  est dense dans  $C_c^1([0, T] \times \Omega)$ . On note  $\sigma = (u, -\nabla u)$  la fonction à valeurs vectorielles dans  $\mathbb{R}^{N+1}$  dont la divergence en "espace-temps" est  $\frac{\partial u}{\partial t} - \Delta u$ . L'identité (8.15) nous dit que cette divergence a bien un sens faible (voir la Définition 4.2.6) et est égale à la fonction  $f$  qui appartient à  $L^2([0, T]; L^2(\Omega))$ , d'où l'égalité presque partout dans  $]0, T[ \times \Omega$ . Il faut cependant faire attention que nous avons montré que la différence  $\frac{\partial u}{\partial t} - \Delta u$  appartient à  $L^2([0, T]; L^2(\Omega))$ , mais pas chaque terme individuellement.  $\square$

**Remarque 8.2.8 L'estimation d'énergie** (8.13) indique que la norme de la solution dans l'espace d'énergie est contrôlée par la norme des données. Il est à noter que cette norme ne correspond pas toujours à la "vraie" énergie physique (dans le cas de la chaleur l'énergie thermique est proportionnelle à  $\int_{\Omega} u(t, x) \, dx$ ). L'inégalité (8.13) a été obtenue comme conséquence de (8.8), ce qui camoufle son origine et son interprétation physique. Les exercices suivants permettent d'obtenir directement (8.13) à partir de l'équation de la chaleur (8.12) en utilisant une **égalité d'énergie** qui ne fait qu'exprimer un bilan physique. En particulier, ces estimations ou égalités d'énergie justifient le choix de l'espace  $L^2([0, T]; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$  pour y chercher des solutions car c'est précisément **l'espace d'énergie** c'est-à-dire l'espace de régularité minimum dans lequel les égalités d'énergie ont un sens.  $\bullet$

**Exercice 8.2.1** On suppose que les hypothèses du Théorème 8.2.7 sont vérifiées.

1. En supposant que la solution  $u$  de (8.12) est assez régulière dans  $]0, T[ \times \Omega$ , montrer que, pour tout  $t \in [0, T]$ , on a l'égalité d'énergie suivante

$$\begin{aligned} \frac{1}{2} \int_{\Omega} u(x, t)^2 dx + \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds &= \frac{1}{2} \int_{\Omega} u_0(x)^2 dx \\ &+ \int_0^t \int_{\Omega} f(x, s) u(x, s) dx ds. \end{aligned} \quad (8.16)$$

2. Démontrer la propriété suivante, appelée "lemme de Gronwall" : si  $z$  est une fonction continue de  $[0, T]$  dans  $\mathbb{R}^+$  telle que

$$z(t) \leq a + b \int_0^t z(s) ds \quad \forall t \in [0, T],$$

où  $a, b$  sont deux constantes positives ou nulles, alors

$$z(t) \leq a e^{bt} \quad \forall t \in [0, T].$$

3. En appliquant le lemme de Gronwall avec  $z(t) = \frac{1}{2} \int_{\Omega} u(x, t)^2 dx$ , déduire de (8.16) que, pour tout  $t \in [0, T]$ ,

$$\begin{aligned} \frac{1}{2} \int_{\Omega} u(x, t)^2 dx + \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds &\leq \frac{e^t}{2} \left( \int_{\Omega} u_0(x)^2 dx \right. \\ &\left. + \int_0^T \int_{\Omega} f(x, s)^2 dx ds \right). \end{aligned} \quad (8.17)$$

**Exercice 8.2.2** Au vu de (8.13), où la constante  $C$  est indépendante de  $T$ , on voit que le terme  $e^t$  n'est certainement pas optimal dans la majoration (8.17). Cette estimation peut être améliorée en raisonnant de la façon suivante, avec une variante du lemme de Gronwall.

1. Soit  $a \in \mathbb{R}^+$  et  $g \in L^2(]0, T[)$  tel que  $g \geq 0$ . Montrer que, si  $z(t)$  est continue de  $[0, T]$  dans  $\mathbb{R}^+$  et vérifie

$$z(t) \leq a + 2 \int_0^t g(s) \sqrt{z(s)} ds \quad \forall t \in [0, T],$$

alors

$$z(t) \leq \left( \sqrt{a} + \int_0^t g(s) ds \right)^2 \quad \forall t \in [0, T].$$

2. Déduire de (8.16) que, pour tout  $t \in [0, T]$ ,

$$\begin{aligned} \int_{\Omega} u(x, t)^2 dx + 2 \int_0^t \int_{\Omega} |\nabla u(x, s)|^2 dx ds &\leq \left( \left( \int_{\Omega} u_0(x)^2 dx \right)^{1/2} \right. \\ &\left. + \int_0^t ds \left( \int_{\Omega} f(x, s)^2 dx \right)^{1/2} \right)^2. \end{aligned} \quad (8.18)$$

L'égalité d'énergie (8.16) n'est pas la seule possible pour l'équation de la chaleur comme le montre l'exercice suivant.

**Exercice 8.2.3** On suppose que les hypothèses du Théorème 8.2.7 sont vérifiées, que  $u_0 \in H_0^1(\Omega)$ , et que la solution  $u$  de (8.12) est assez régulière dans  $]0, T[ \times \Omega$ . Montrer que, pour tout  $t \in [0, T]$ , on a l'égalité d'énergie suivante

$$\begin{aligned} \frac{1}{2} \int_{\Omega} |\nabla u(x, t)|^2 dx + \int_0^t \int_{\Omega} \left| \frac{\partial u}{\partial t}(x, s) \right|^2 dx ds &= \frac{1}{2} \int_{\Omega} |\nabla u_0(x)|^2 dx \\ &+ \int_0^t \int_{\Omega} f(x, s) \frac{\partial u}{\partial t}(x, s) dx ds. \end{aligned} \quad (8.19)$$

Bien sûr, le Théorème d'existence 8.2.7 se généralise facilement au cas d'autres conditions aux limites ou d'un opérateur elliptique général comme le montre les exercices suivants.

**Exercice 8.2.4** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . Soit un temps final  $T > 0$ , une donnée initiale  $u_0 \in L^2(\Omega)$ , et un terme source  $f \in L^2([0, T[; L^2(\Omega)))$ . À l'aide de la Remarque 8.2.5 montrer que l'équation de la chaleur avec condition aux limites de Neumann

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{dans } \Omega \times ]0, T[ \\ \frac{\partial u}{\partial n} = 0 & \text{sur } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{dans } \Omega \end{cases} \quad (8.20)$$

admet une unique solution  $u \in L^2([0, T[; H^1(\Omega))) \cap C([0, T]; L^2(\Omega))$ .

**Exercice 8.2.5** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . Soit  $A(x)$  une fonction de  $\Omega$  dans l'ensemble des matrices symétriques réelles telles qu'il existe deux constantes  $\beta \geq \alpha > 0$  vérifiant

$$\beta |\xi|^2 \geq A(x) \xi \cdot \xi \geq \alpha |\xi|^2 \quad \forall \xi \in \mathbb{R}^N, \text{ p.p. } x \in \Omega.$$

Soit un temps final  $T > 0$ , une donnée initiale  $u_0 \in L^2(\Omega)$ , et un terme source  $f \in L^2([0, T[; L^2(\Omega)))$ . Montrer que le problème aux limites

$$\begin{cases} \frac{\partial u}{\partial t} - \operatorname{div}(A(x) \nabla u) = f & \text{dans } \Omega \times ]0, T[ \\ u = 0 & \text{sur } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{dans } \Omega, \end{cases}$$

admet une unique solution  $u \in L^2([0, T[; H^1(\Omega))) \cap C([0, T]; L^2(\Omega))$ .

On peut étendre le Théorème 8.2.7 aux équations de Stokes instationnaires.

**Théorème 8.2.9** Soit  $\Omega$  un ouvert borné régulier connexe de  $\mathbb{R}^N$ . Soit un temps final  $T > 0$ , une donnée initiale  $u_0 \in L^2(\Omega)^N$  telle que  $\operatorname{div} u_0 = 0$  dans  $\Omega$ , et un terme source  $f \in L^2(]0, T[; L^2(\Omega))^N$ . Alors les équations de Stokes instationnaires

$$\begin{cases} \frac{\partial u}{\partial t} + \nabla p - \mu \Delta u = f & \text{dans } \Omega \times ]0, T[ \\ \operatorname{div} u = 0 & \text{dans } \Omega \times ]0, T[ \\ u = 0 & \text{sur } \partial\Omega \times ]0, T[ \\ u(x, t = 0) = u_0(x) & \text{dans } \Omega \end{cases} \quad (8.21)$$

admettent une unique solution  $u \in L^2(]0, T[; H_0^1(\Omega))^N \cap C([0, T]; L^2(\Omega))^N$ .

**Démonstration.** Pour obtenir une formulation variationnelle de (8.21) nous combinons les arguments de la Sous-section 8.2.1 et de la démonstration du Théorème 5.3.8. On introduit les espaces de Hilbert

$$V = \left\{ v \in H_0^1(\Omega)^N \text{ tel que } \operatorname{div} v = 0 \text{ p.p. dans } \Omega \right\},$$

et

$$H = \left\{ v \in L^2(\Omega)^N \text{ tel que } \operatorname{div} v = 0 \text{ p.p. dans } \Omega \right\},$$

où  $H$  est un sous-espace fermé de  $H(\operatorname{div})$  qu'on peut aussi définir comme l'adhérence de  $V$  dans  $L^2(\Omega)^N$  (voir la Sous-section 4.4.2). Nous obtenons la formulation variationnelle suivante

$$\begin{cases} \frac{d}{dt} \int_{\Omega} u(t) \cdot v \, dx + \mu \int_{\Omega} \nabla u(t) \cdot \nabla v \, dx = \int_{\Omega} f(t) \cdot v \, dx \quad \forall v \in V, 0 < t < T, \\ u(t = 0) = u_0, \end{cases} \quad (8.22)$$

où l'équation de (8.22) a lieu au sens faible dans  $]0, T[$ . On applique le Théorème 8.2.3 à cette formulation variationnelle (8.22) (ses hypothèses sont facilement vérifiées) et on obtient l'existence et l'unicité de sa solution  $u \in L^2(]0, T[; H_0^1(\Omega))^N \cap C([0, T]; L^2(\Omega))^N$ .

Toute la difficulté réside dans la preuve que cette solution de (8.22) est bien une solution de (8.21). La condition aux limites de Dirichlet se retrouve par application du Théorème 4.3.13 de trace à  $u(t) \in H_0^1(\Omega)^N$  pour presque tout  $t \in ]0, T[$ , et la condition initiale est justifiée par la continuité de  $u(t)$  en  $t = 0$  car  $u_0 \in H$ .

Pour retrouver l'équation, on procède comme dans la démonstration du Théorème 8.2.7. Si la solution  $u$  est suffisamment régulière, on obtient

$$\int_{\Omega} \left( \frac{\partial u}{\partial t} - \mu \Delta u - f \right) \cdot v \, dx = 0 \quad (8.23)$$

pour presque tout  $t \in ]0, T[$ , et quel que soit  $v(x) \in C_c^1(\Omega)^N$  tel que  $\operatorname{div} v = 0$  dans  $\Omega$ . Comme pour le problème de Stokes stationnaire (voir la Sous-section 5.3.2), il faut déduire de (8.23) l'existence d'une fonction  $p(t, x)$  telle que l'équation (8.21) ait lieu. Il faut encore utiliser le Théorème de de Rham 5.3.9 (du moins une de ces variantes) qui affirme que "l'orthogonal des vecteurs à divergence nulle est l'ensemble des gradients". Ce point d'analyse est assez délicat (encore plus si la solution  $u$  n'est pas régulière) et nous admettrons simplement l'existence d'une telle pression  $p$  sans préciser même à quel espace elle appartient.  $\square$

**Remarque 8.2.10** Mentionnons brièvement qu'il existe d'autres approches que celle utilisée ici (et qu'on peut qualifier d'approche spectrale) pour obtenir l'existence et l'unicité de solutions de problèmes d'évolution en temps. Il existe une théorie purement variationnelle (voir [31]) ainsi qu'une théorie, dite des semi-groupes (voir [8]). Ces théories sont un peu plus compliquées, mais plus puissantes puisqu'en particulier elles permettent de s'affranchir des hypothèses sur le caractère borné de l'ouvert et sur la symétrie de la forme bilinéaire de la formulation variationnelle. •

## 8.3 Existence et unicité dans le cas hyperbolique

Comme dans la section précédente nous suivons la même démarche en trois étapes. Premièrement (Sous-section 8.3.1), on établit une formulation variationnelle, deuxièmement (Sous-section 8.3.2), on démontre l'existence et l'unicité de la solution de cette formulation variationnelle en utilisant une base hilbertienne de fonctions propres, troisièmement (Sous-section 8.3.3), on montre que cette solution variationnelle vérifie bien le problème aux limites.

### 8.3.1 Formulation variationnelle

L'idée est d'écrire une formulation variationnelle qui ressemble à une **équation différentielle ordinaire** du deuxième ordre, similaire à (7.4). Nous multiplions donc l'équation des ondes (8.3) par une fonction test  $v(x)$  qui **ne dépend pas du temps**  $t$ . A cause de la condition aux limites nous demandons à ce que  $v$  s'annule sur le bord de l'ouvert  $\Omega$ . Un calcul formel conduit à

$$\frac{d^2}{dt^2} \int_{\Omega} u(x, t) v(x) dx + \int_{\Omega} \nabla u(x, t) \cdot \nabla v(x) dx = \int_{\Omega} f(x, t) v(x) dx. \quad (8.24)$$

Il est clair que l'espace "naturel" pour la fonction test  $v$  est  $H_0^1(\Omega)$ . On introduit alors le produit scalaire de  $L^2(\Omega)$  et la forme bilinéaire  $a(w, v)$  définis par

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

Soit un temps final  $T > 0$  (éventuellement égal à  $+\infty$ ), on se donne le terme source  $f \in L^2([0, T]; L^2(\Omega))$ . On se donne aussi des conditions initiales  $u_0 \in H_0^1(\Omega)$  et  $u_1 \in L^2(\Omega)$ . La formulation variationnelle déduite de (8.24) est donc : trouver une solution  $u$  dans  $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  telle que

$$\begin{cases} \frac{d^2}{dt^2} \langle u(t), v \rangle_{L^2(\Omega)} + a(u(t), v) = \langle f(t), v \rangle_{L^2(\Omega)} & \forall v \in H_0^1(\Omega), 0 < t < T, \\ u(t=0) = u_0, \quad \frac{du}{dt}(t=0) = u_1. \end{cases} \quad (8.25)$$

Les données initiales ont bien un sens dans (8.25) grâce au choix de l'**espace d'énergie**  $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  pour la solution  $u$ . Nous justifierons encore ce choix un peu plus loin en établissant son lien avec des égalités d'énergie.



Finalement, la dérivée en temps dans la formulation variationnelle (8.25) doit être prise au sens faible puisqu'a priori la fonction  $t \rightarrow \langle u(t), v \rangle_{L^2(\Omega)}$  n'est qu'une fois dérivable en temps puisqu'elle appartient à  $C^1(0, T)$  (voir le Lemme 4.4.12 et la Remarque 8.2.6 pour plus de précisions).

### 8.3.2 Un résultat général

Pour démontrer l'existence et l'unicité de la solution de la formulation variationnelle (8.25), nous revenons encore au cadre général de la Section 7.3 pour "diagonaliser" l'opérateur Laplacien et nous ramener à la résolution d'une famille de simples équations différentielles ordinaires du deuxième ordre. Soit  $V$  et  $H$  deux espaces de Hilbert tels que  $V \subset H$  avec injection dense et compacte (typiquement  $V = H_0^1(\Omega)$  et  $H = L^2(\Omega)$ ).

**Théorème 8.3.1** *Soient  $V$  et  $H$  deux espaces de Hilbert tels que  $V \subset H$  avec injection compacte et  $V$  est dense dans  $H$ . Soit  $a(u, v)$  une forme bilinéaire symétrique continue et coercive dans  $V$ . Soit un temps final  $T > 0$ , une donnée initiale  $(u_0, u_1) \in V \times H$ , et un terme source  $f \in L^2(]0, T[; H)$ . Alors le problème*

$$\begin{cases} \frac{d^2}{dt^2} \langle u(t), v \rangle_H + a(u(t), v) = \langle f(t), v \rangle_H & \forall v \in V, \ 0 < t < T, \\ u(t=0) = u_0, \quad \frac{du}{dt}(t=0) = u_1, \end{cases} \quad (8.26)$$

(où l'équation de (8.26) a lieu au sens faible dans  $]0, T[$ ) a une unique solution  $u \in C([0, T]; V) \cap C^1([0, T]; H)$ . De plus, il existe une constante  $C > 0$  (qui ne dépend que de  $\Omega$  et de  $T$ ) telle que

$$\|u\|_{C([0, T]; V)} + \|u\|_{C^1([0, T]; H)} \leq C (\|u_0\|_V + \|u_1\|_H + \|f\|_{L^2(]0, T[; H)}). \quad (8.27)$$

**Remarque 8.3.2 L'estimation d'énergie** (8.27) prouve que la solution de (8.26) dépend continûment des données, et donc que le problème hyperbolique (8.26) est bien posé au sens de Hadamard. La Proposition 8.3.5 donnera une interprétation physique importante d'un cas particulier de cette estimation d'énergie. •

**Remarque 8.3.3** Comme dans le cas parabolique (voir la Remarque 8.2.5), on peut affaiblir l'hypothèse du Théorème 8.3.1 sur la coercivité de la forme bilinéaire symétrique  $a(u, v)$ . On obtient les mêmes conclusions en supposant seulement qu'il existe deux constantes positives  $\nu > 0$  et  $\eta > 0$  telles que

$$a(v, v) + \eta \|v\|_H^2 \geq \nu \|v\|_V^2 \text{ pour tout } v \in V.$$

Le changement d'inconnue  $u(t) = e^{\sqrt{\eta}t} w(t)$  transforme l'équation de (8.26) en

$$\frac{d^2}{dt^2} \langle w(t), v \rangle_H + 2\sqrt{\eta} \frac{d}{dt} \langle w(t), v \rangle_H + a(w(t), v) + \eta \langle w(t), v \rangle_H = \langle f(t), v \rangle_H, \quad (8.28)$$

où la forme bilinéaire  $a(w, v) + \eta \langle w, v \rangle_H$  est bien coercive sur  $V$ . L'équation (8.28) est une équation des ondes amorties (voir la Remarque 8.1.1). Il suffit alors de généraliser le Théorème 8.3.1 à de telles équations (ce qui est facile bien que nous ne le fassions pas ici). •

**Démonstration.** La démonstration est très semblable à celle du Théorème 8.2.3, aussi nous ne la détaillons pas autant. Dans une première étape, on montre que toute solution  $u$  est une série de fonctions propres. Dans une deuxième étape, nous démontrons la convergence de cette série dans les espaces  $C([0, T]; V)$  et  $C^1([0, T]; H)$ .

**Étape 1.** Supposons que  $u \in C([0, T]; V) \cap C^1([0, T]; H)$  est solution de (8.26). Introduisons la base hilbertienne  $(u_k)_{k \geq 1}$  de  $H$  composée des fonctions propres de la formulation variationnelle (7.12) qui vérifient

$$u_k \in V, \text{ et } a(u_k, v) = \lambda_k \langle u_k, v \rangle_H \quad \forall v \in V.$$

On écrit  $u(t) = \sum_{k=1}^{+\infty} \alpha_k(t) u_k$  avec  $\alpha_k(t) = \langle u(t), u_k \rangle_H$ . En choisissant  $v = u_k$  dans (8.26), et en notant  $\beta_k(t) = \langle f(t), u_k \rangle_H$ ,  $\alpha_k^0 = \langle u_0, u_k \rangle_H$ , et  $\alpha_k^1 = \langle u_1, u_k \rangle_H$ , on obtient

$$\begin{cases} \frac{d^2 \alpha_k}{dt^2} + \lambda_k \alpha_k = \beta_k \text{ dans } ]0, T[ \\ \alpha_k(t=0) = \alpha_k^0, \quad \frac{d\alpha_k}{dt}(t=0) = \alpha_k^1. \end{cases} \quad (8.29)$$

(Attention à une confusion possible dans les notations : la donnée initiale  $u_1$  n'a rien à voir avec la fonction propre  $u_k$  pour  $k=1$ .) Posant  $\omega_k = \sqrt{\lambda_k}$ , l'unique solution de (8.29) est

$$\alpha_k(t) = \alpha_k^0 \cos(\omega_k t) + \frac{\alpha_k^1}{\omega_k} \sin(\omega_k t) + \frac{1}{\omega_k} \int_0^t \beta_k(s) \sin(\omega_k(t-s)) ds, \quad (8.30)$$

ce qui donne une formule explicite pour la solution  $u$  (qui est donc unique).

**Étape 2.** Pour démontrer que la série

$$\sum_{j=1}^{+\infty} \left( \alpha_j^0 \cos(\omega_j t) + \frac{\alpha_j^1}{\omega_j} \sin(\omega_j t) + \frac{1}{\omega_j} \int_0^t \beta_j(s) \sin(\omega_j(t-s)) ds \right) u_j \quad (8.31)$$

converge dans  $C([0, T]; V) \cap C^1([0, T]; H)$ , on va montrer que la suite  $w^k = \sum_{j=1}^k \alpha_j(t) u_j$  des sommes partielles de cette série est de Cauchy. Dans  $V$  nous considérons le produit scalaire  $a(u, v)$  pour lequel la famille  $(u_j)$  est orthogonale. Par orthogonalité dans  $H$  et dans  $V$  de  $(u_j)$  (voir le Théorème 7.3.2), on obtient, pour  $l > k$ , et pour tout temps  $t$ ,

$$a(w^l - w^k, w^l - w^k) + \left\| \frac{d}{dt} (w^l - w^k) \right\|_H^2 = \sum_{j=k+1}^l \left( \lambda_j |\alpha_j(t)|^2 + \left| \frac{d\alpha_j}{dt}(t) \right|^2 \right).$$

Or, en multipliant (8.29) par  $d\alpha_k/dt$  et en intégrant en temps, on obtient

$$\left| \frac{d\alpha_j}{dt}(t) \right|^2 + \lambda_j |\alpha_j(t)|^2 = |\alpha_j^1|^2 + \lambda_j |\alpha_j^0|^2 + 2 \int_0^t \beta_j(s) \frac{d\alpha_j}{dt}(s) ds.$$

De la formule (8.30) on infère que

$$\left| \frac{d\alpha_j}{dt}(t) \right| \leq \omega_j |\alpha_j^0| + |\alpha_j^1| + \int_0^t |\beta_j(s)| ds.$$

En combinant ces deux résultats on en déduit

$$\left| \frac{d\alpha_j}{dt}(t) \right|^2 + \lambda_j |\alpha_j(t)|^2 \leq 2 |\alpha_j^1|^2 + 2\lambda_j |\alpha_j^0|^2 + 2t \int_0^t |\beta_j(s)|^2 ds. \quad (8.32)$$

Comme  $u_0 \in V$ ,  $u_1 \in H$  et  $f \in L^2([0, T]; H)$ , on a

$$\|u_0\|_V^2 = a(u_0, u_0) = \sum_{j=1}^{+\infty} \lambda_j |\alpha_j^0|^2 < +\infty, \quad \|u_1\|_H^2 = \sum_{j=1}^{+\infty} |\alpha_j^1|^2 < +\infty,$$

$$\|f\|_{L^2([0, T]; H)}^2 = \sum_{j=1}^{+\infty} \int_0^T |\beta_j(s)|^2 ds < +\infty,$$

ce qui implique que la série, dont le terme général est le membre de gauche de (8.32), est convergente, c'est-à-dire que la suite  $w^k$  vérifie

$$\lim_{k, l \rightarrow +\infty} \max_{0 \leq t \leq T} \left( \|w^l(t) - w^k(t)\|_V^2 + \left\| \frac{d}{dt}(w^l(t) - w^k(t)) \right\|_H^2 \right) = 0,$$

autrement dit, elle est de Cauchy dans  $C^1([0, T]; H)$  et dans  $C([0, T]; V)$ . Comme ces espaces sont complets, la suite de Cauchy  $w^k$  converge et on peut définir sa limite  $u$ . En particulier, comme  $(w^k(0), \frac{dw^k}{dt}(0))$  converge vers  $(u_0, u_1)$  dans  $V \times H$ , on obtient les conditions initiales voulues. D'autre part, il est clair que  $u(t)$ , en tant que somme de la série (8.31) vérifie la formulation variationnelle (8.26) pour chaque fonction test  $v = u_k$ . Comme  $(u_k/\sqrt{\lambda_k})$  est une base hilbertienne de  $V$ ,  $u(t)$  vérifie donc la formulation variationnelle (8.26) pour tout  $v \in V$ , c'est-à-dire que  $u(t)$  est bien la solution recherchée de (8.26). Par ailleurs, on a en fait montré que

$$a(w^l - w^k, w^l - w^k) + \left\| \frac{d}{dt}(w^l - w^k) \right\|_H^2 \leq C \left( \|u_0\|_V^2 + \|u_1\|_H^2 + T \|f\|_{L^2([0, T]; H)}^2 \right),$$

et l'estimation d'énergie (8.27) s'obtient alors facilement en prenant  $k = 0$  et en faisant tendre  $l$  vers l'infini.  $\square$

### 8.3.3 Applications

Nous appliquons maintenant le résultat abstrait du Théorème 8.3.1 à l'équation des ondes, et nous prouvons que cette approche variationnelle a bien permis de résoudre l'équation aux dérivées partielles d'origine.

**Théorème 8.3.4** *Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ , et un temps final  $T > 0$ . On considère une donnée initiale  $(u_0, u_1) \in H_0^1(\Omega) \times L^2(\Omega)$  et un terme source  $f \in L^2([0, T]; L^2(\Omega))$ . Alors l'équation des ondes*

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = f & \text{p.p. dans } \Omega \times ]0, T[ \\ u = 0 & \text{p.p. sur } \partial\Omega \times ]0, T[ \\ u(x, 0) = u_0(x) & \text{p.p. dans } \Omega \\ \frac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{p.p. dans } \Omega. \end{cases} \quad (8.33)$$

*admet une unique solution  $u \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ . De plus, il existe une constante  $C > 0$  (qui ne dépend que de  $\Omega$  et de  $T$ ) telle que, pour tout  $t \in [0, T]$ ,*

$$\begin{aligned} \int_{\Omega} \left( \left| \frac{\partial u}{\partial t}(x, t) \right|^2 + |\nabla u(x, t)|^2 \right) dx &\leq C \left( \int_{\Omega} (|u_1(x)|^2 + |\nabla u_0(x)|^2) dx \right. \\ &\quad \left. + \int_0^t \int_{\Omega} |f(x, s)|^2 dx ds \right). \end{aligned} \quad (8.34)$$

**Démonstration.** Nous appliquons le Théorème 8.3.1 à la formulation variationnelle (8.25) de l'équation des ondes obtenue à la Sous-section 8.3.1 (ses hypothèses sont facilement vérifiées avec  $H = L^2(\Omega)$  et  $V = H_0^1(\Omega)$ ). Il reste à montrer que l'unique solution  $u \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  de cette formulation variationnelle est bien une solution de (8.33). Tout d'abord, la condition aux limites de Dirichlet se retrouve par application du Théorème 4.3.13 de trace à  $u(t) \in H_0^1(\Omega)$  pour tout  $t \in [0, T]$ , et la condition initiale est justifiée par la continuité de  $u(t)$  en  $t = 0$  comme fonction à valeurs dans  $H_0^1(\Omega)$  et de  $du/dt(t)$  en  $t = 0$  comme fonction à valeurs dans  $L^2(\Omega)$ .

Si la solution  $u$  est suffisamment régulière, par intégration par parties la formulation variationnelle (8.25) est équivalente à

$$\int_{\Omega} \left( \frac{\partial^2 u}{\partial t^2} - \Delta u - f \right) v dx = 0,$$

pour tout  $v(x) \in C_c^1(\Omega)$  et presque tout  $t \in ]0, T[$ . On en déduit donc l'équation de (8.33). Si la solution  $u$  n'est pas plus régulière que ce qui est donné par le Théorème 8.3.1, on obtient tout de même l'équation au sens "presque partout", en reprenant les arguments de la démonstration du Théorème 8.2.7 (que nous ne détaillons pas). On note  $\sigma = (\frac{\partial u}{\partial t}, -\nabla u)$  la fonction à valeurs vectorielles dans  $\mathbb{R}^{N+1}$ , et on peut montrer

qu'elle admet une divergence faible en "espace-temps" qui est justement  $\frac{\partial^2 u}{\partial t^2} - \Delta u$  qui appartient donc à  $L^2([0, T]; L^2(\Omega))$ .  $\square$

En l'absence de forces,  $f = 0$ , on peut améliorer l'estimation d'énergie (8.34) et obtenir une propriété de **conservation de l'énergie totale** qui est très importante du point de vue des applications. L'énergie totale est ici la somme de deux termes : d'une part l'énergie cinétique  $|\frac{\partial u}{\partial t}|^2$  et d'autre part l'énergie mécanique  $|\nabla u|^2$ .

**Proposition 8.3.5** *On se place sous les hypothèses du Théorème 8.3.4 avec  $f = 0$ . La solution de l'équation des ondes (8.33) vérifie, pour tout  $t \in [0, T]$ , l'égalité de conservation de l'énergie*

$$\int_{\Omega} \left( \left| \frac{\partial u}{\partial t}(x, t) \right|^2 + |\nabla u(x, t)|^2 \right) dx = \int_{\Omega} (|u_1(x)|^2 + |\nabla u_0(x)|^2) dx. \quad (8.35)$$

**Démonstration.** En reprenant la démonstration du Théorème 8.3.1 avec  $f = 0$ , c'est-à-dire  $\beta_k = 0$ , on déduit directement de (8.29) que l'énergie de l'oscillateur harmonique est conservée, c'est-à-dire que

$$\left| \frac{d\alpha_j}{dt}(t) \right|^2 + \lambda_j |\alpha_j(t)|^2 = |\alpha_j^1|^2 + \lambda_j |\alpha_j^0|^2,$$

ce qui donne l'égalité (plutôt que l'inégalité)

$$a(w^l - w^k, w^l - w^k) + \left\| \frac{d}{dt}(w^l - w^k) \right\|_H^2 = \sum_{j=k+1}^l |\alpha_j^1|^2 + \lambda_j |\alpha_j^0|^2,$$

et (8.35) s'obtient en prenant  $k = 0$  et en faisant tendre  $l$  vers l'infini. Si la solution  $u$  est régulière, on peut démontrer plus directement (8.35) en multipliant l'équation des ondes (8.33) par  $\frac{\partial u}{\partial t}$  et en intégrant par parties (voir l'Exercice 8.3.1).  $\square$

Nous revenons maintenant sur l'**estimation d'énergie** (8.34) dans le cas général  $f \neq 0$ . L'exercice suivant montre comment (8.34) peut être obtenue directement à partir de (8.33) à l'aide d'un argument similaire à celui de la Proposition 8.3.5 qui établit une **égalité d'énergie** qui ne fait qu'exprimer un bilan physique. En particulier, ces estimations ou égalités d'énergie justifient le choix de l'espace  $C([0, T]; H_0^1(\Omega))^N \cap C^1([0, T]; L^2(\Omega))^N$  pour y chercher des solutions car c'est précisément l'**espace d'énergie** c'est-à-dire l'espace de régularité minimum dans lequel les égalités d'énergie ont un sens.

**Exercice 8.3.1** On suppose que les hypothèses du Théorème 8.3.4 sont vérifiées.

1. En supposant que la solution  $u$  de (8.33) est assez régulière dans  $]0, T[ \times \Omega$ , montrer que, pour tout  $t \in [0, T]$ , on a l'égalité d'énergie suivante

$$\begin{aligned} \int_{\Omega} \left| \frac{\partial u}{\partial t}(x, t) \right|^2 dx + \int_{\Omega} |\nabla u(x, t)|^2 dx &= \int_{\Omega} |u_1(x)|^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx \\ &\quad + 2 \int_0^t \int_{\Omega} f(x, s) \frac{\partial u}{\partial t}(x, s) dx ds. \end{aligned}$$

2. En déduire qu'il existe une constante  $C(T)$  (indépendante des données autre que  $T$ ) telle que

$$\int_{\Omega} \left| \frac{\partial u}{\partial t}(x, t) \right|^2 dx + \int_{\Omega} |\nabla u(x, t)|^2 dx \leq C(T) \left( \int_{\Omega} u_1(x)^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx + \int_0^t \int_{\Omega} f(x, s)^2 dx ds \right).$$

3. Montrer qu'il existe une constante  $C$  (indépendante de toutes les données y compris  $T$ ) telle que

$$\int_{\Omega} \left| \frac{\partial u}{\partial t}(x, t) \right|^2 dx + \int_{\Omega} |\nabla u(x, t)|^2 dx \leq C \left( \int_{\Omega} u_1(x)^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx + \left( \int_0^t \left( \int_{\Omega} f(x, s)^2 dx \right)^{1/2} ds \right)^2 \right).$$

Il existe d'autres quantités conservées comme l'indique l'exercice suivant.

**Exercice 8.3.2** On suppose que les hypothèses du Théorème 8.3.4 sont vérifiées, que le terme source est nul  $f = 0$  et que la solution  $u$  de (8.33) est régulière dans  $[0, T] \times \Omega$ . Montrer que, pour tout entier  $m \geq 1$ , on a

$$\frac{d}{dt} \int_{\Omega} \left( \left| \frac{\partial^m u}{\partial t^m} \right|^2 + \left| \nabla \frac{\partial^{m-1} u}{\partial t^{m-1}} \right|^2 \right) dx = 0.$$

Bien sûr, le Théorème d'existence 8.3.4 se généralise facilement au cas d'autres conditions aux limites (par exemple, de Neumann), ou au cas d'autres opérateurs que le Laplacien comme

$$\frac{\partial^2 u}{\partial t^2} - \operatorname{div}(A(x)\nabla u) = f.$$

C'est un exercice de généraliser ce résultat aux équations de l'élastodynamique.

**Exercice 8.3.3** Soit  $\Omega$  un ouvert borné régulier connexe de  $\mathbb{R}^N$ . Soit une donnée initiale  $(u_0, u_1) \in H_0^1(\Omega)^N \times L^2(\Omega)^N$ , et un terme source  $f \in L^2([0, T]; L^2(\Omega))^N$ . Montrer qu'il existe une unique solution  $u \in C([0, T]; H_0^1(\Omega))^N \cap C^1([0, T]; L^2(\Omega))^N$  de

$$\begin{cases} \rho \frac{\partial^2 u}{\partial t^2} - \operatorname{div}(2\mu e(u) + \lambda \operatorname{tr}(e(u)) \operatorname{Id}) = f & \text{dans } \Omega \times ]0, T[, \\ u = 0 & \text{sur } \partial\Omega \times ]0, T[, \\ u(t=0) = u_0(x) & \text{dans } \Omega, \\ \frac{\partial u}{\partial t}(t=0) = u_1(x) & \text{dans } \Omega. \end{cases} \quad (8.36)$$

En supposant que la solution  $u$  est assez régulière, montrer que, pour tout  $t \in [0, T]$ , on a l'égalité d'énergie suivante

$$\begin{aligned} \frac{\rho}{2} \int_{\Omega} \left| \frac{\partial u}{\partial t} \right|^2 dx + \mu \int_{\Omega} |e(u)|^2 dx + \frac{\lambda}{2} \int_{\Omega} (\operatorname{div} u)^2 dx &= \frac{\rho}{2} \int_{\Omega} |u_1|^2 dx \\ + \mu \int_{\Omega} |e(u_0)|^2 dx + \frac{\lambda}{2} \int_{\Omega} (\operatorname{div} u_0)^2 dx + \int_0^t \int_{\Omega} f \cdot \frac{\partial u}{\partial t} dx ds. \end{aligned}$$

En déduire une estimation d'énergie.

**Remarque 8.3.6** Comme pour les problèmes paraboliques, l'approche "spectrale" utilisée ici pour obtenir l'existence et l'unicité d'équations aux dérivées partielles hyperboliques n'est pas la seule possible. Citons la théorie purement variationnelle de [31] ainsi que la théorie, dite des semi-groupes (voir [8]). Ces théories sont un peu plus compliquées, mais plus puissantes puisqu'en particulier elles permettent de s'affranchir des hypothèses sur le caractère borné de l'ouvert  $\Omega$  et sur la symétrie de la forme bilinéaire  $a(u, v)$  de la formulation variationnelle. •

## 8.4 Propriétés qualitatives dans le cas parabolique

Nous examinons maintenant les principales propriétés qualitatives de la solution de l'équation de la chaleur, notamment les propriétés de régularité, comportement asymptotiques pour les grandes valeurs de  $t$  et principe du maximum.

### 8.4.1 Comportement asymptotique

Nous étudions le comportement de la solution de l'équation de la chaleur en temps long, c'est-à-dire lorsque  $t$  tend vers  $+\infty$ . Nous allons vérifier que, conformément à l'intuition physique, si le second membre  $f(x)$  est indépendant du temps  $t$ , alors la solution de l'équation de la chaleur tend asymptotiquement vers la solution (stationnaire) du Laplacien. Nous commençons par examiner le cas de l'équation de la chaleur homogène.

**Proposition 8.4.1** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . Soit  $u_0 \in L^2(\Omega)$  et  $u$  la solution du problème

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{dans } ]0, +\infty[ \times \Omega \\ u(x, t) = 0 & \text{sur } ]0, +\infty[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{dans } \Omega. \end{cases} \quad (8.37)$$

Alors,  $u(t)$  converge vers zéro dans  $L^2(\Omega)$  lorsque  $t$  tend vers  $+\infty$

$$\lim_{t \rightarrow +\infty} \|u(t)\|_{L^2(\Omega)} = 0. \quad (8.38)$$

**Démonstration.** On reprend la démonstration du Théorème 8.2.3 dans le cas  $f = 0$ , c'est-à-dire  $\beta_k = 0$ . On obtient facilement que la somme partielle vérifie

$$\|w^l(t) - w^k(t)\|_H^2 = \sum_{j=k+1}^l |\alpha_j^0|^2 e^{-2\lambda_j t},$$

avec  $H = L^2(\Omega)$ , ce qui conduit, en prenant  $k = 0$  et  $l = +\infty$ , et en majorant, à

$$\|u(t)\|_H^2 \leq \|u_0\|_H^2 e^{-2\lambda_1 t}$$

qui tend vers zéro lorsque  $t$  tend vers l'infini puisque  $\lambda_1 > 0$ .  $\square$

Le cas d'un second membre non nul, indépendant du temps, est un simple exercice que nous laissons au lecteur.

**Exercice 8.4.1** On reprend les hypothèses de la Proposition 8.4.1. Soit  $f(x) \in L^2(\Omega)$  et  $u(t, x)$  la solution de

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{dans } ]0, +\infty[ \times \Omega \\ u(x, t) = 0 & \text{sur } ]0, +\infty[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{dans } \Omega. \end{cases}$$

Soit  $v(x) \in H_0^1(\Omega)$  la solution de

$$\begin{cases} -\Delta v = f & \text{dans } \Omega \\ v = 0 & \text{sur } \partial\Omega. \end{cases}$$

Montrer que  $\lim_{t \rightarrow +\infty} \|u(x, t) - v(x)\|_{L^2(\Omega)} = 0$ .

On peut en fait préciser la conclusion de la Proposition 8.4.1 comme le montre l'exercice suivant dont l'interprétation est la suivante

$$u(t, x) \approx \left( \int_{\Omega} u_0 u_1 dx \right) e^{-\lambda_1 t} u_1(x) \quad \text{lorsque } t \rightarrow +\infty,$$

où  $u_1(x)$  est la première fonction propre (normalisée) du Laplacien (7.17). Asymptotiquement, toutes les solutions de l'équation de la chaleur homogène décroissent exponentiellement en temps avec le même profil spatial qui est donné par  $u_1$  (quelle que soit la donnée initiale).

**Exercice 8.4.2** On reprend les hypothèses de la Proposition 8.4.1. Montrer qu'il existe une constante positive  $C$  telle que

$$\|u(t) - \alpha_1^0 e^{-\lambda_1 t} u_1\|_{L^2(\Omega)} \leq C e^{-\lambda_2 t} \quad \forall t > 1, \quad \text{avec } \alpha_1^0 = \int_{\Omega} u_0 u_1 dx, \quad (8.39)$$

où  $\lambda_k$  désigne la  $k$ -ème valeur propre du Laplacien avec condition aux limites de Dirichlet.



### 8.4.2 Principe du maximum

Pour l'équation de la chaleur, le principe du maximum prend une forme voisine de celle que nous avons énoncée au Théorème 5.2.22 pour le Laplacien.

**Proposition 8.4.2** *Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ , et un temps final  $T > 0$ . Soit  $u_0 \in L^2(\Omega)$ ,  $f \in L^2(]0, T[; L^2(\Omega))$ , et  $u \in C([0, T]; L^2(\Omega)) \cap L^2(]0, T[; H_0^1(\Omega))$  l'unique solution de (8.12). Si  $f \geq 0$  presque partout dans  $]0, T[ \times \Omega$  et  $u_0 \geq 0$  presque partout dans  $\Omega$ , alors  $u \geq 0$  presque partout dans  $]0, T[ \times \Omega$ .*

**Démonstration.** Soit  $u^- = \min(u, 0)$  qui appartient bien à  $L^2(]0, T[; H_0^1(\Omega))$  en vertu du Lemme 5.2.24 et qui vérifie, pour  $0 < t < T$ ,

$$\int_{\Omega} \nabla u(t) \cdot \nabla u^-(t) dx = \int_{\Omega} |\nabla u^-(t)|^2 dx. \quad (8.40)$$

Un raisonnement similaire à celui qui a permis de démontrer (8.40) montre que, si  $\frac{\partial u}{\partial t} \in L^2(]0, T[; L^2(\Omega))$ , alors

$$\int_{\Omega} \frac{\partial u}{\partial t}(t) u^-(t) dx = \frac{1}{2} \frac{d}{dt} \left( \int_{\Omega} |u^-(t)|^2 dx \right). \quad (8.41)$$

Nous admettons que l'identité (8.41) reste vraie même si  $\frac{\partial u}{\partial t}$  n'appartient pas à  $L^2(]0, T[; L^2(\Omega))$ . Par conséquent, en prenant  $v = u^-$  dans la formulation variationnelle (8.6) de l'équation de la chaleur on obtient

$$\frac{1}{2} \frac{d}{dt} \left( \int_{\Omega} |u^-|^2 dx \right) + \int_{\Omega} |\nabla u^-|^2 dx = \int_{\Omega} f u^- dx,$$

ce qui donne par intégration en temps

$$\frac{1}{2} \int_{\Omega} |u^-(t)|^2 dx + \int_0^t \int_{\Omega} |\nabla u^-|^2 dx ds = \int_0^t \int_{\Omega} f u^- dx ds + \frac{1}{2} \int_{\Omega} |u^-(0)|^2 dx.$$

Comme  $u^-(0) = (u_0)^- = 0$  on en déduit

$$\frac{1}{2} \int_{\Omega} |u^-(t)|^2 dx + \int_0^t \int_{\Omega} |\nabla u^-|^2 dx ds \leq 0,$$

c'est-à-dire que  $u^- = 0$  presque partout dans  $]0, T[ \times \Omega$ .  $\square$

Comme dans le cas elliptique, le principe du maximum fourni par la Proposition 8.4.2 est conforme à l'intuition physique. Dans le cadre du modèle de thermique décrit au Chapitre 1, si la température initiale  $u_0(x)$  est en tout point supérieure à la valeur 0 à laquelle on maintient la température sur la paroi  $\partial\Omega$  et si le terme source est positif (correspondant à un effet de chauffage), alors il est clair que la température est positive en tout point et à tout instant. Nous n'avons fait que vérifier que le modèle mathématique reproduit bien cette propriété intuitive.

Il est bon de réaliser que ces résultats peuvent également s'énoncer sous plusieurs formes équivalentes. On peut, par exemple, comparer deux solutions de (8.12) : si  $u_0 \leq \tilde{u}_0$  dans  $\Omega$  et  $f \leq \tilde{f}$  dans  $]0, T[ \times \Omega$ , et si  $u$  et  $\tilde{u}$  désignent respectivement les solutions de (8.12) correspondant aux données  $(u_0, f)$  et  $(\tilde{u}_0, \tilde{f})$  respectivement, alors on a  $u \leq \tilde{u}$  dans  $]0, T[ \times \Omega$ .

Les deux exercices suivants illustrent quelques applications intéressantes du principe du maximum.

**Exercice 8.4.3** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . On note  $u_1$  la première fonction propre du Laplacien dans  $\Omega$  avec condition de Dirichlet,  $\lambda_1$  la valeur propre associée. On rappelle que l'on peut choisir  $u_1 > 0$  dans  $\Omega$  (voir le Théorème de Krein-Rutman 7.3.10) et on admettra que l'on a aussi  $\partial u_1 / \partial n > 0$  sur  $\partial\Omega$ . Soit  $f = 0$ ,  $u_0 \in L^2(\Omega)$  et  $u$  l'unique solution (supposée régulière) de (8.12).

Soit  $\epsilon > 0$ . Montrer que l'on peut trouver une constante positive  $K$  telle que

$$-Ku_1(x) \leq u(x, \epsilon) \leq Ku_1(x) \quad \forall x \in \overline{\Omega}, \quad (8.42)$$

et en déduire qu'il existe une constante positive  $C$  telle que

$$\max_{x \in \overline{\Omega}} |u(x, t)| \leq Ce^{-\lambda_1 t} \quad \forall t > \epsilon. \quad (8.43)$$

**Exercice 8.4.4** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . Soit  $u_0 \in L^\infty(\Omega)$ ,  $f \in L^\infty(\mathbb{R}^+ \times \Omega)$ , et  $u \in C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H_0^1(\Omega))$  l'unique solution de (8.12). Montrer que

$$\|u\|_{L^\infty(\mathbb{R}^+ \times \Omega)} \leq \|u_0\|_{L^\infty(\Omega)} + \frac{D^2}{2N} \|f\|_{L^\infty(\mathbb{R}^+ \times \Omega)}, \quad (8.44)$$

où  $D = \sup_{x, y \in \Omega} |x - y|$  est le diamètre de  $\Omega$ . On pourra d'abord considérer le cas plus facile où  $f \equiv 0$ , puis, dans le cas général, utiliser la fonction  $\psi \in H_0^1(\Omega)$  telle que  $-\Delta\psi = 1$  dans  $\Omega$ .

### 8.4.3 Propagation à vitesse infinie

Nous avons déjà évoqué à la Remarque 1.2.9 cette propriété surprenante de l'équation de la chaleur : la chaleur se propage à une vitesse infinie ! Ce résultat découle d'un **principe du maximum fort** que nous énonçons maintenant sans démonstration. Nous le vérifierons plus facilement lorsque le domaine  $\Omega$  est l'espace  $\mathbb{R}^N$  tout entier (voir la Sous-section 8.4.5).

**Proposition 8.4.3** Soit  $\Omega$  un ouvert borné régulier de classe  $C^2$  de  $\mathbb{R}^N$ . Soit un temps final  $T > 0$ . Soit  $u_0 \in L^2(\Omega)$  et  $u$  la solution unique dans  $C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H_0^1(\Omega))$  du problème

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{dans } ]0, T[ \times \Omega \\ u(x, t) = 0 & \text{sur } ]0, T[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{dans } \Omega. \end{cases}$$

On suppose de plus que  $u_0(x) \geq 0$  presque partout dans  $\Omega$  et que  $u_0$  n'est pas identiquement nulle. Alors, pour tout temps  $\epsilon > 0$ , on a

$$u(x, \epsilon) > 0 \quad \forall x \in \Omega. \quad (8.45)$$

C'est l'inégalité **stricte** de (8.45) qui est remarquable (on avait déjà une inégalité large par le principe du maximum de la Proposition 8.4.2). En effet, si  $u_0$  a un support compact dans  $\Omega$  et si on se place en un point  $x \in \Omega$  en dehors du support de  $u_0$ , on trouve que  $u(x, \epsilon) > 0$  bien qu'initialement  $u_0(x) = 0$ . Autrement dit, dans le cadre de la modélisation de l'évolution de la température, même si le point  $x$  est initialement froid ( $u_0(x) = 0$ ) et très loin de la partie chaude initiale (le support de  $u_0$ ), il devient instantanément chaud puisque pour tout temps  $t = \epsilon$  (même très petit), on a  $u(x, \epsilon) > 0$ . Ainsi **la chaleur se propage à vitesse infinie** puisque son effet est immédiat même à grande distance! Il s'agit clairement d'un défaut du modèle mathématique puisque l'on sait que rien ne peut se propager plus vite que la vitesse de la lumière (en fait c'est la loi de Fourier (1.3) qui est en défaut). C'est un modèle, qualitativement et quantitativement correct à bien des égards, comme l'ont démontré d'ailleurs plusieurs résultats précédents, conformes à l'intuition physique, mais ce n'est qu'un modèle idéalisé de la réalité.

**Remarque 8.4.4** La même propriété de “propagation à vitesse infinie” peut aussi être observée pour les équations de Stokes (8.2). Dans ce cadre, on est d'ailleurs peut-être moins surpris par ce paradoxe si on a réalisé que l'hypothèse d'incompressibilité d'un fluide revient à dire que la vitesse du son y est infinie. Autrement dit, en utilisant l'approximation d'incompressibilité, on introduit implicitement dans le modèle la possibilité d'une propagation de l'information à vitesse infinie. •

#### 8.4.4 Régularité et effet régularisant

Dans le cas elliptique nous avons vu que la régularité de la solution est directement liée à celle des données. Dans le cas parabolique, la situation est différente car, si le terme source est nul ( $f = 0$ ), il existe un **effet régularisant** de la condition initiale : de manière surprenante, même si la donnée initiale  $u_0$  est très peu régulière, la solution devient instantanément très régulière.

**Proposition 8.4.5** Soit  $\Omega$  un ouvert borné régulier de classe  $C^\infty$  de  $\mathbb{R}^N$ , et soit un temps final  $T > 0$ . Soit  $u_0 \in L^2(\Omega)$ , et  $u$  l'unique solution dans  $C([0, T]; L^2(\Omega)) \cap L^2([0, T]; H_0^1(\Omega))$  de

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{dans } ]0, T[ \times \Omega \\ u(x, t) = 0 & \text{sur } ]0, T[ \times \partial\Omega \\ u(x, 0) = u_0(x) & \text{dans } \Omega. \end{cases} \quad (8.46)$$

Alors, pour tout  $\epsilon > 0$ ,  $u$  est de classe  $C^\infty$  en  $x$  et  $t$  dans  $\overline{\Omega} \times [\epsilon, T]$ .

**Idée de démonstration.** Plutôt qu'une démonstration rigoureuse (mais assez technique et peu parlante, voir l'Exercice 8.4.5) nous proposons un calcul formel qui montre bien l'idée essentielle derrière ce résultat de régularité (la démonstration est plus facile dans le cas  $\Omega = \mathbb{R}^N$ , voir l'Exercice 8.4.9). Pour  $k \geq 1$  on note  $v = \frac{\partial^k u}{\partial t^k}$  et on dérive (formellement)  $k$  fois l'équation de la chaleur (8.46) par rapport au temps pour obtenir

$$\begin{cases} \frac{\partial v}{\partial t} - \Delta v = 0 & \text{dans } ]0, T[ \times \Omega \\ v(x, t) = 0 & \text{sur } ]0, T[ \times \partial\Omega \\ v(x, 0) = \frac{\partial^k u}{\partial t^k}(0, x) & \text{dans } \Omega, \end{cases} \quad (8.47)$$

qui est encore une équation de la chaleur. Si  $\frac{\partial^k u}{\partial t^k}(0, x)$  appartient à  $L^2(\Omega)$ , on applique le Théorème 8.2.7 d'existence et d'unicité à (8.47) qui nous dit que  $v$  appartient à  $L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ . En particulier,  $u$  est régulier en temps. D'autre part, par égalité,  $v = (\Delta)^k u$  appartient au même espace. Par régularité elliptique (voir le Théorème 5.2.26) on en déduit que  $u$  est régulier en espace. Le point le plus délicat pour donner un sens à ce raisonnement formel est que la donnée initiale de (8.47) n'est pas assez régulière. C'est pour cette raison que la régularité de  $u$  n'est valable que pour les temps  $t > \epsilon > 0$ .  $\square$

En présence de termes sources, le même raisonnement que celui de la démonstration de la Proposition 8.4.5 permet de retrouver un résultat de régularité plus classique qui n'est pas sans rapport avec l'égalité d'énergie (8.19) (en effet, l'espace auquel  $u$  va appartenir est celui qui donne un sens à (8.19)).

**Proposition 8.4.6** *Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ , et un temps final  $T > 0$ . Pour un terme source  $f \in L^2(]0, T[; L^2(\Omega))$  et une donnée initiale régulière  $u_0 \in H_0^1(\Omega)$ , on considère la solution unique  $u \in L^2(]0, T[; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$  de l'équation de la chaleur (8.12). Alors, cette solution est plus régulière au sens où  $\frac{\partial u}{\partial t} \in L^2(]0, T[; L^2(\Omega))$  et  $u \in L^2(]0, T[; H^2(\Omega)) \cap C([0, T]; H_0^1(\Omega))$ .*

**Remarque 8.4.7** On peut bien sûr "monter" en régularité et obtenir que la solution  $u$  de l'équation de la chaleur (8.12) est aussi régulière que l'on veut, pourvu que les données  $u_0$  et  $f$  le soient aussi (voir [8]). Cependant, si l'on veut que la solution  $u$  soit régulière dès l'instant initial, il faut que les données  $u_0$  et  $f$  vérifient des conditions de compatibilité. Ainsi, dans la Proposition 8.4.6 il est demandé à la condition initiale  $u_0$  de vérifier la condition aux limites de Dirichlet (ce qui n'était pas nécessaire pour l'existence d'une solution dans le Théorème 8.2.7). Les autres conditions de compatibilité s'obtiennent en remarquant que les dérivées successives de  $u$  par rapport au temps  $t$  sont aussi solutions d'équations de la chaleur avec conditions aux limites de Dirichlet. Par exemple, la condition initiale pour la dérivée première est  $\frac{\partial u}{\partial t}(0) = f(0) + \Delta u_0$ . Pour que  $\frac{\partial u}{\partial t}$  soit régulier, il faut donc que cette donnée initiale vérifie la condition aux limites de Dirichlet  $f(0) + \Delta u_0 = 0$  sur  $\partial\Omega$ , ce qui est une condition de compatibilité entre  $u_0$  et  $f$ .  $\bullet$

**Exercice 8.4.5 (difficile)** Démontrer rigoureusement la Proposition 8.4.5. Pour cela on introduira, pour tout entier  $m \geq 0$ , l'espace

$$W^{2m}(\Omega) = \{v \in H^{2m}(\Omega), v = \Delta v = \dots \Delta^{m-1} v = 0 \text{ sur } \partial\Omega\}, \quad (8.48)$$

que l'on munit de la norme  $\|v\|_{W^{2m}(\Omega)}^2 = \int_{\Omega} |(\Delta)^m v|^2 dx$ , dont on montrera qu'elle est équivalente à la norme de  $H^{2m}(\Omega)$ . On reprendra la démonstration du Théorème 8.2.3 en montrant que la suite  $(w_k)$  des sommes partielles est de Cauchy dans  $C^\ell([\epsilon, T], W^{2m}(\Omega))$ .

### 8.4.5 Équation de la chaleur dans tout l'espace

Pour terminer cette section, nous indiquons brièvement comment résoudre l'équation de la chaleur posée dans tout l'espace  $\mathbb{R}^N$ . Rappelons que l'approche spectrale suivie dans ce chapitre est limitée au cas des ouverts bornés (cette limitation est en fait artificielle et absolument pas nécessaire pour établir l'existence et l'unicité de la solution d'une équation parabolique). Considérons l'équation de la chaleur homogène dans tout l'espace  $\mathbb{R}^N$ , munie d'une donnée initiale

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = 0 & \text{dans } ]0, +\infty[ \times \mathbb{R}^N \\ u(x, 0) = u_0(x) & \text{dans } \mathbb{R}^N. \end{cases} \quad (8.49)$$

Le résultat classique suivant montre que la solution du problème (8.49) est donnée explicitement comme le produit de convolution de la donnée initiale  $u_0$  avec une Gaussienne dont l'écart-type croît comme  $\sqrt{t}$ .

**Théorème 8.4.8** *On suppose que  $u_0 \in L^2(\mathbb{R}^N)$ . Alors le problème (8.49) a une solution unique  $u \in C(\mathbb{R}^+, L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_+^*, L^2(\mathbb{R}^N))$ , donnée par*

$$u(x, t) = \frac{1}{(4\pi t)^{N/2}} \int_{\mathbb{R}^N} u_0(y) e^{-\frac{|x-y|^2}{4t}} dy. \quad (8.50)$$

**Démonstration.** Pour  $t \geq 0$ , nous introduisons la transformée de Fourier de  $u(t)$  (voir [7]), c'est-à-dire de la fonction  $x \mapsto u(x, t)$ , définie par

$$\hat{u}(k, t) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} u(x, t) e^{ik \cdot x} dx,$$

pour  $k \in \mathbb{R}^N$ . Si  $u \in C(\mathbb{R}^+, L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_+^*; L^2(\mathbb{R}^N))$  vérifie (8.49), on peut appliquer la transformation de Fourier aux deux équations (8.49) pour obtenir

$$\begin{cases} \hat{u} \in C(\mathbb{R}^+, L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_+^*, L^2(\mathbb{R}^N)), \\ \frac{\partial \hat{u}}{\partial t} + |k|^2 \hat{u} = 0 & \text{pour } k \in \mathbb{R}^N, t > 0, \\ \hat{u}(k, 0) = \hat{u}_0(k) & \text{pour } k \in \mathbb{R}^N, \end{cases} \quad (8.51)$$

où  $\hat{u}_0(k) = \frac{1}{(2\pi)^{N/2}} \int_{\mathbb{R}^N} u_0(x) e^{ik \cdot x} dx$  est la transformée de Fourier de  $u_0$ . Le système (8.51) se résout de façon élémentaire puisqu'on a une équation différentielle pour chaque valeur de  $k$ . On obtient

$$\hat{u}(k, t) = \hat{u}_0(k) e^{-|k|^2 t} \quad \text{pour } (k, t) \in \mathbb{R}^N \times \mathbb{R}^+,$$

et il est facile d'en déduire (8.50) par transformation de Fourier inverse (puisque cette transformation convertit un produit de convolution en produit simple).  $\square$

**Remarque 8.4.9** L'utilisation de la transformation de Fourier a permis de "diagonaliser" l'équation de la chaleur (8.49) et de se ramener à la résolution d'une simple équation différentielle ordinaire (8.51). Cette méthode est donc très semblable, dans l'esprit, à l'approche

spectrale précédemment utilisée et qui repose aussi sur un argument de diagonalisation. En d'autres termes,  $|k|^2$  et  $e^{ik \cdot x}$  s'interprètent comme des sortes de valeurs et fonctions propres du Laplacien dans  $\mathbb{R}^N$ . Remarquons que plus le mode de Fourier  $|k|$  est grand, plus la décroissance exponentielle en temps de  $\hat{u}(k, t)$  est rapide : cet amortissement plus rapide pour les petites longueurs d'onde ( $k$  grand) est lié à l'effet régularisant de l'équation de la chaleur (voir l'Exercice 8.4.9 ci-dessous). •

**Remarque 8.4.10** Le résultat (8.50) peut s'interpréter en terme de fonction de Green. En posant  $G(x, t) = \frac{1}{(2\pi t)^{N/2}} e^{-\frac{|x|^2}{4t}}$ , on écrit (8.50) sous la forme  $u(t) = G(t) * u_0$ , c'est-à-dire

$$u(x, t) = \int_{\mathbb{R}^N} G(x - y, t) u_0(y) dy.$$

La fonction de Green est la solution élémentaire de l'équation de la chaleur dans  $\mathbb{R}^N \times \mathbb{R}_*^+$ , ce qui signifie que l'on peut vérifier que

$$\begin{cases} \frac{\partial G}{\partial t} - \Delta G = 0 & \text{dans } ]0, +\infty[ \times \mathbb{R}^N \\ G(x, 0) = \delta_0(x) & \text{dans } \mathbb{R}^N. \end{cases}$$

au sens des distributions, où  $\delta_0$  est la masse de Dirac à l'origine. Ce point de vue peut aussi être développé dans un domaine borné et conduit à une méthode de résolution des équations paraboliques différente de l'approche "spectrale" suivie ici. •

L'exercice suivant permet de résoudre l'équation de la chaleur non homogène.

**Exercice 8.4.6** Pour  $u_0 \in L^2(\mathbb{R}^N)$  et  $t > 0$ , on note  $S(t)u_0$  la fonction donnée par le second membre de (8.50). Vérifier que  $S(t)$  est un opérateur linéaire continu de  $L^2(\mathbb{R}^N)$  dans  $L^2(\mathbb{R}^N)$ . En posant  $S(0) = \text{Id}$  (l'identité de  $L^2(\mathbb{R}^N)$ ), vérifier que  $(S(t))_{t \geq 0}$  est un semi-groupe d'opérateurs qui dépendent continûment de  $t$ , c'est-à-dire qu'ils vérifient  $S(t + t') = S(t)S(t')$  pour  $t, t' \geq 0$ . Soit  $f \in C^1(\mathbb{R}^+; L^2(\mathbb{R}^N))$ . Montrer que le problème

$$\begin{cases} \frac{\partial u}{\partial t} - \Delta u = f & \text{dans } ]0, +\infty[ \times \mathbb{R}^N \\ u(x, 0) = u_0(x) & \text{dans } \mathbb{R}^N. \end{cases}$$

admet une unique solution  $u \in C(\mathbb{R}^+; L^2(\mathbb{R}^N)) \cap C^1(\mathbb{R}_*^+; L^2(\mathbb{R}^N))$ , donnée par

$$u(t) = S(t)u_0 + \int_0^t S(t-s)f(s) ds,$$

c'est-à-dire

$$u(x, t) = \int_{\mathbb{R}^N} u_0(y) e^{-\frac{|x-y|^2}{4t}} \frac{dy}{(2\pi t)^{N/2}} + \int_0^t \int_{\mathbb{R}^N} f(y, s) e^{-\frac{|x-y|^2}{4(t-s)}} \frac{dy ds}{(2\pi(t-s))^{N/2}}.$$

La formule explicite (8.50) permet de retrouver sans difficulté, pour le problème (8.49) posé dans tout l'espace, les propriétés qualitatives étudiées précédemment. C'est l'objet des exercices suivants où l'on notera  $u$  la solution (8.50) du problème (8.49), avec la donnée initiale  $u_0 \in L^2(\mathbb{R}^N)$ .

**Exercice 8.4.7 (égalité d'énergie)** Montrer que, pour tout  $T > 0$ ,

$$\frac{1}{2} \int_{\mathbb{R}^N} u(x, T)^2 dx + \int_0^T \int_{\mathbb{R}^N} |\nabla u(x, t)|^2 dx dt = \frac{1}{2} \int_{\mathbb{R}^N} u_0(x)^2 dx.$$

**Exercice 8.4.8 (principe du maximum)** Montrer que, si  $u_0 \in L^\infty(\mathbb{R}^N)$ , alors  $u(t) \in L^\infty(\mathbb{R}^N)$  et

$$\|u(t)\|_{L^\infty(\mathbb{R}^N)} \leq \|u_0\|_{L^\infty(\mathbb{R}^N)} \quad \forall t > 0.$$

Montrer que, si  $u_0 \geq 0$  presque partout dans  $\mathbb{R}^N$ , alors  $u \geq 0$  dans  $\mathbb{R}^N \times \mathbb{R}^+$ .

**Exercice 8.4.9 (effet régularisant)** Montrer que  $u \in C^\infty(\mathbb{R}^N \times \mathbb{R}_*)$ .

**Exercice 8.4.10 (comportement asymptotique)** Montrer que

$$\lim_{|x| \rightarrow +\infty} u(x, t) = 0 \quad \forall t > 0, \quad \text{et} \quad \lim_{t \rightarrow +\infty} u(x, t) = 0 \quad \forall x \in \mathbb{R}^N.$$

**Exercice 8.4.11 (vitesse de propagation infinie)** Montrer que, si  $u_0 \geq 0$  et  $u_0 \not\equiv 0$ , alors  $u(x, t) > 0$  dans  $\mathbb{R}^N \times \mathbb{R}_*^+$ .

## 8.5 Propriétés qualitatives dans le cas hyperbolique

### 8.5.1 Réversibilité en temps

Nous examinons maintenant les principales propriétés qualitatives de la solution de l'équation des ondes, qui sont très différentes de celles de la solution de l'équation de la chaleur. La propriété la plus frappante, déjà évoquée au Chapitre 1, est la **réversibilité en temps** de cette équation.

**Proposition 8.5.1** Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ , et un temps final  $T > 0$ . Soit  $(v_0, v_1) \in H_0^1(\Omega) \times L^2(\Omega)$ , et un terme source  $f \in L^2([0, T]; L^2(\Omega))$ . Alors l'équation des ondes **rétrograde en temps** (intégrée en remontant le temps à partir de  $T$ )

$$\begin{cases} \frac{\partial^2 v}{\partial t^2} - \Delta v = f & \text{p.p. dans } \Omega \times ]0, T[ \\ v = 0 & \text{p.p. sur } \partial\Omega \times ]0, T[ \\ v(x, T) = v_0(x) & \text{p.p. dans } \Omega \\ \frac{\partial v}{\partial t}(x, T) = v_1(x) & \text{p.p. dans } \Omega \end{cases} \quad (8.52)$$

admet une unique solution  $v \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$ . De plus, si  $u(t, x)$  est la solution de l'équation des ondes (8.3) et si  $v_0(x) = u(x, T)$  dans  $H_0^1(\Omega)$  et  $v_1(x) = \frac{\partial u}{\partial t}(x, T)$  dans  $L^2(\Omega)$ , alors on a  $v(t, x) = u(t, x)$ .

**Démonstration.** On fait le changement d'inconnue  $w(x, t) = v(x, T - t)$  et (8.52) devient une équation des ondes "progressive" avec donnée initiale en  $t = 0$  comme l'équation "usuelle" (8.3) (comme la dérivée en temps est d'ordre 2, il n'y a pas de changement de signe dans l'équation après ce changement d'inconnue). Par application

du Théorème 8.3.4, (8.52) admet donc bien une unique solution. Si  $v_0(x) = u(x, T)$  et  $v_1(x) = \frac{\partial u}{\partial t}(x, T)$ , la solution  $u(t, x)$  de (8.3) est aussi solution de (8.52). Par unicité on en déduit  $v(t, x) = u(t, x)$ .  $\square$

Le caractère réversible en temps de l'équation des ondes a de nombreuses conséquences. La plus importante est qu'il n'y a aucun **effet régularisant** pour l'équation des ondes au contraire de ce qui se passait pour l'équation de la chaleur. En effet, si c'était le cas, en changeant le sens du temps comme dans la Proposition 8.5.1, on obtiendrait un effet "dérégularisant" contradictoire (la solution serait moins régulière que la donnée finale en  $T$ , ce qui n'est pas possible puisque l'effet régularisant doit aussi s'appliquer). Par conséquent, **il n'y a ni gain ni perte de régularité** pour la solution de l'équation des ondes par rapport aux données initiales. On peut tout au plus affirmer, comme dans le cas elliptique, que la régularité de la solution de l'équation des ondes est directement liée à celle des données.

**Proposition 8.5.2** *Soit  $\Omega$  un ouvert borné régulier de  $\mathbb{R}^N$ . Soit un temps final  $T > 0$ , une donnée initiale  $u_0 \in H_0^1(\Omega) \cap H^2(\Omega)$  et  $u_1 \in H_0^1(\Omega)$ , un terme source  $f \in L^2([0, T]; L^2(\Omega))$ , et  $u \in C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  la solution unique de l'équation des ondes (8.33). Alors,  $u$  appartient à  $C([0, T]; H_0^1(\Omega) \cap H^2(\Omega)) \cap C^1([0, T]; H_0^1(\Omega)) \cap C^2([0, T]; L^2(\Omega))$ .*

Nous admettons la Proposition 8.5.2 qui est similaire à la Proposition 8.4.6 que nous avons aussi admise. On peut bien sûr "monter" en régularité à partir de ce résultat et obtenir que la solution  $u$  de l'équation des ondes (8.33) est aussi régulière que l'on veut, pourvu que les données  $u_0$ ,  $u_1$  et  $f$  le soient aussi (avec d'éventuelles conditions de compatibilité des données, voir la Remarque 8.4.7).

### 8.5.2 Comportement asymptotique et équipartition de l'énergie

**Il n'y a pas de principe du maximum** pour l'équation des ondes. En l'absence de terme source ( $f = 0$ ), même si la vitesse initiale est nulle ( $u_1 = 0$ ) et si la donnée initiale est positive ( $u_0 \geq 0$ ), la solution  $u$  peut changer de signe au cours du temps. Cette absence de principe du maximum est conforme à l'intuition physique. Imaginons une corde ou une membrane élastique : si on la déforme initialement dans une position au dessus de son plan de repos, elle va vibrer et passer alternativement en dessus et au dessous de ce plan (autrement dit  $u$  change de signe). Mathématiquement, ce contre-exemple peut s'écrire simplement sous la forme suivante. Soit  $w(x)$  la première fonction propre du Laplacien dans un domaine borné connexe  $\Omega$  avec condition aux limites de Dirichlet. D'après le Théorème 7.3.10, on peut normaliser  $w$  de telle manière que  $w(x) \geq 0$  dans  $\Omega$ . En notant  $\lambda = \omega^2$  la première valeur propre associée à  $w$ , il est facile de vérifier que  $u(t, x) = \cos(\omega t)w(x)$  change de signe au cours du temps tout en étant la solution unique dans  $C([0, T]; H_0^1(\Omega)) \cap C^1([0, T]; L^2(\Omega))$  de l'équation des ondes (8.33) sans terme source et avec les données initiales

$$u(x, 0) = w(x), \quad \frac{\partial u}{\partial t}(x, 0) = 0 \text{ dans } \Omega.$$



Il n'y a donc pas non plus de comportement asymptotique en temps long pour l'équation des ondes en domaine borné. Autrement dit, même si le terme source  $f$  ne dépend pas du temps, la solution  $u$  ne converge pas vers une limite stationnaire lorsque le temps  $t$  tend vers l'infini. En particulier, si  $f = 0$ , l'influence des conditions initiales est la même à tout temps puisque l'énergie est conservée et ne décroît pas (voir l'Exercice 8.3.1). Le même contre-exemple  $u(t, x) = \cos(\omega t)w(x)$  permet de voir qu'il n'y a pas de limite stationnaire mais des oscillations qui perdurent sans amortissement.

Cela n'est évidemment pas le cas pour l'équation des ondes amortie (8.53) comme le montre l'exercice suivant.

**Exercice 8.5.1** Soit  $\eta > 0$ . On considère l'équation des ondes amortie

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} + \eta \frac{\partial u}{\partial t} - \Delta u = f & \text{dans } \Omega \times \mathbb{R}_*^+ \\ u = 0 & \text{sur } \partial\Omega \times \mathbb{R}_*^+ \\ u(x, 0) = u_0(x) & \text{dans } \Omega \\ \frac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{dans } \Omega. \end{cases} \quad (8.53)$$

On suppose que  $u$  est une solution suffisamment régulière de (8.53) et que  $f$  est nul après un temps fini. Montrer, à l'aide d'un lemme de Gronwall (voir l'Exercice 8.2.1), que  $u$  et  $\frac{\partial u}{\partial t}$  décroissent exponentiellement vers zéro lorsque le temps  $t$  tend vers l'infini.

La Proposition 8.3.5 a établi une propriété de conservation de l'énergie totale en l'absence de terme source. Un résultat plus précis, dit **d'équipartition de l'énergie**, affirme que l'énergie totale se partage équitablement en énergie cinétique et énergie mécanique, asymptotiquement pour les temps grands.

**Exercice 8.5.2** Soit  $u(t, x)$  la solution, supposée suffisamment régulière, de l'équation des ondes (8.33). En l'absence de terme source, montrer que

$$\lim_{t \rightarrow +\infty} \frac{1}{t} \int_0^t \int_{\Omega} \left| \frac{\partial u}{\partial t} \right|^2 dx = \lim_{t \rightarrow +\infty} \frac{1}{t} \int_0^t \int_{\Omega} |\nabla u|^2 dx = \frac{1}{2} E_0,$$

avec  $E_0$  l'énergie initiale

$$E_0 = \int_{\Omega} |u_1(x)|^2 dx + \int_{\Omega} |\nabla u_0(x)|^2 dx.$$

Pour cela on multipliera l'équation (8.33) par  $u$  et on intégrera par parties.

### 8.5.3 Vitesse de propagation finie

Une dernière propriété importante de l'équation des ondes est celle, dite de **propagation à vitesse finie**. Nous avons déjà vu au Chapitre 1 (en dimension  $N = 1$  et dans tout l'espace  $\Omega = \mathbb{R}$ ) qu'il existe un cône de lumière (ou domaine de dépendance) qui englobe toute l'information sur la solution de l'équation des ondes (voir

la Figure 8.1). Plus précisément, grâce à la formule explicite (1.19) pour la solution  $u$ , on voit immédiatement que la valeur de  $u$  en  $(t, x)$  ne dépend que des valeurs des données initiales  $u_0$  et  $u_1$  sur le segment  $[x-t, x+t]$ . On en déduit que si les données initiales sont à support compact  $K = [k_{\inf}, k_{\sup}] \subset \mathbb{R}$ , alors la solution au temps  $t$  est à support compact dans  $[k_{\inf} - t, k_{\sup} + t]$ . En termes physiques cela veut dire que des perturbations initiales se propagent à vitesse finie bornée par 1. Cette situation est encore une fois bien différente de ce qui se passe pour l'équation de la chaleur (comparer avec la Proposition 8.4.3). La proposition suivante permet de généraliser cette discussion aux dimensions supérieures, sans utiliser de formule explicite pour la solution.

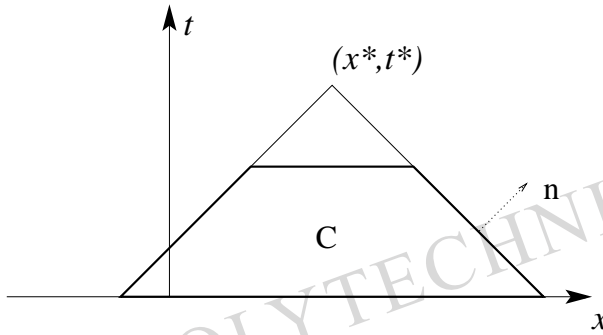


FIGURE 8.1 – Cône de lumière pour l'équation des ondes.

**Proposition 8.5.3** Soit  $u(x, t)$  une solution régulière de l'équation des ondes dans tout l'espace  $\mathbb{R}^N \times \mathbb{R}_+$

$$\frac{\partial^2 u}{\partial t^2} - \Delta u = 0 \quad \text{dans } \mathbb{R}^N \times \mathbb{R}_+. \quad (8.54)$$

Pour chaque temps  $t^* \geq 0$  et point d'espace  $x^* \in \mathbb{R}^N$  on introduit le cône de lumière

$$C^* = \{(x, t) \in \mathbb{R}^N \times \mathbb{R}_+ \text{ tel que } 0 \leq t \leq t^* \text{ et } |x - x^*| \leq t^* - t\},$$

ainsi que sa section au temps  $s$ ,  $C_s^* = \{x \in \mathbb{R}^N \text{ tel que } (x, s) \in C^*\}$ . On définit l'énergie dans  $C_s^*$  par

$$e^*(s) = \frac{1}{2} \int_{C_s^*} \left( \left| \frac{\partial u}{\partial t}(x, s) \right|^2 + |\nabla u(x, s)|^2 \right) dx.$$

Alors la fonction  $s \rightarrow e^*(s)$  est décroissante. En particulier, si les données initiales sont nulles sur la base  $C_0^*$  du cône de lumière, alors la solution est identiquement nulle dans tout le cône  $C^*$ .

**Démonstration.** Par invariance par translation en temps il suffit de montrer que  $e^*(s) \leq e^*(0)$  pour tout  $s \geq 0$ . Etant donné  $s \geq 0$ , on note  $C$  le tronc de cône défini par  $C = \{(x, t) \in C^* \text{ tel que } 0 \leq t \leq s\}$ , avec  $n = (n_t, n_x)$  sa normale extérieure unité (un vecteur de  $\mathbb{R}^{N+1}$ ) et  $d\sigma$  la mesure surfacique sur son bord  $\partial C$  (voir la Figure 8.1). On intègre par parties l'expression

$$\begin{aligned} 0 &= \int_C \frac{\partial u}{\partial t} \left( \frac{\partial^2 u}{\partial t^2} - \Delta u \right) dt dx \\ &= \int_C \left( \frac{1}{2} \frac{\partial}{\partial t} \left| \frac{\partial u}{\partial t} \right|^2 + \nabla u \cdot \nabla \frac{\partial u}{\partial t} \right) dt dx - \int_{\partial C} \frac{\partial u}{\partial t} \nabla u \cdot n_x d\sigma \\ &= \frac{1}{2} \int_C \frac{\partial}{\partial t} \left( \left| \frac{\partial u}{\partial t} \right|^2 + |\nabla u|^2 \right) dt dx - \int_{\partial C} \frac{\partial u}{\partial t} \nabla u \cdot n_x d\sigma \\ &= \frac{1}{2} \int_{\partial C} \left( \left( \left| \frac{\partial u}{\partial t} \right|^2 + |\nabla u|^2 \right) n_t - 2 \frac{\partial u}{\partial t} \nabla u \cdot n_x \right) d\sigma. \end{aligned}$$

Le bord  $\partial C$  du tronc de cône  $C$  est composé de trois parties (voir la Figure 8.1) : sur les deux parties horizontales,  $C_s^*$  et  $C_0^*$ , la normale vérifie  $n_x = 0$  ( $n_t = 1$  sur  $C_s^*$ ,  $n_t = -1$  sur  $C_0^*$ ), tandis que sur la partie latérale, notée  $\partial C_{lat}$  (qui est de pente unité, ou plus généralement égale à l'inverse de la vitesse de propagation) la normale vérifie  $|n_x| = |n_t|$ . Tenant compte de la décomposition  $\partial C = C_0^* \cup C_s^* \cup \partial C_{lat}$ , on a tout d'abord

$$\frac{1}{2} \int_{C_0^* \cup C_s^*} \left( \left( \left| \frac{\partial u}{\partial t} \right|^2 + |\nabla u|^2 \right) n_t - 2 \frac{\partial u}{\partial t} \nabla u \cdot n_x \right) d\sigma = e^*(s) - e^*(0).$$

D'autre part, par application de l'inégalité de Cauchy-Schwarz sur  $\partial C_{lat}$ , on obtient

$$\left| 2 \frac{\partial u}{\partial t} \nabla u \cdot n_x \right| \leq 2 |n_t| \left| \frac{\partial u}{\partial t} \right| |\nabla u| \leq |n_t| \left( \left| \frac{\partial u}{\partial t} \right|^2 + |\nabla u|^2 \right),$$

et donc

$$\frac{1}{2} \int_{\partial C_{lat}} \left( \left( \left| \frac{\partial u}{\partial t} \right|^2 + |\nabla u|^2 \right) n_t - 2 \frac{\partial u}{\partial t} \nabla u \cdot n_x \right) d\sigma \geq 0.$$

On en déduit l'inégalité recherchée,  $e^*(s) \leq e^*(0)$ , pour tout  $s \geq 0$ . Bien évidemment, si les données initiales sont nulles sur la base  $C_0^*$ , alors  $e^*(s) = 0$ , pour tout  $s \geq 0$ , et la solution  $u$  est identiquement nulle dans tout le cône de lumière  $C^*$ .  $\square$

Une conséquence immédiate de la Proposition 8.5.3 est que, si les données initiales  $u_0$  et  $u_1$  sont à support compact  $K \subset \mathbb{R}^N$ , alors, pour tout temps  $t \geq 0$ , la solution  $u(\cdot, t)$  de l'équation des ondes est aussi à support compact dans un domaine un peu plus grand  $K+t$  défini comme  $\{x \in \mathbb{R}^N, d(x, K) \leq t\}$  où  $d$  est la distance euclidienne.

C'est à nouveau une propriété de propagation à vitesse finie bornée par 1. Notons que la Proposition 8.5.3 reste valable pour l'équation des ondes dans un domaine borné  $\Omega$  tant que la base  $C_0^*$  du cône de lumière  $C^*$  reste incluse dans  $\Omega$ .

L'exercice (difficile) suivant donne des formules explicites pour les solutions de l'équation des ondes en dimensions  $N = 2, 3$ .

**Exercice 8.5.3** On considère l'équation des ondes dans tout l'espace  $\mathbb{R}^N$

$$\begin{cases} \frac{\partial^2 u}{\partial t^2} - \Delta u = 0 & \text{dans } \mathbb{R}^N \times \mathbb{R}_+^* \\ u(x, 0) = u_0(x) & \text{dans } x \in \mathbb{R}^N \\ \frac{\partial u}{\partial t}(x, 0) = u_1(x) & \text{dans } x \in \mathbb{R}^N, \end{cases} \quad (8.55)$$

avec une donnée initiale  $(u_0, u_1)$  régulière et à support compact. Montrer que la solution  $u(t, x)$  peut se mettre sous la forme

$$u(x, t) = (Mu_1)(x, t) + \left( \frac{\partial(Mu_0)}{\partial t} \right)(x, t),$$

où  $M$  est un opérateur de moyenne défini par

$$\text{si } N = 1, \quad (Mv)(x, t) = \frac{1}{2} \int_{-t}^{+t} v(x - \xi) d\xi,$$

$$\text{si } N = 2, \quad (Mv)(x, t) = \frac{1}{2\pi} \int_{|\xi| < t} \frac{v(x - \xi)}{\sqrt{t^2 - |\xi|^2}} d\xi,$$

$$\text{si } N = 3, \quad (Mv)(x, t) = \frac{1}{4\pi t} \int_{|\xi|=t} v(x - \xi) ds(\xi),$$

où  $ds(\xi)$  est la mesure surfacique de la sphère. En déduire que la solution  $u$  en  $(t, x)$  ne dépend que des valeurs des données initiales  $u_0$  et  $u_1$  sur la boule  $|x| \leq t$ . (Pour savoir comment on trouve les expressions ci-dessus de l'opérateur  $M$ , nous renvoyons au chapitre 9 de [7].)

**Exercice 8.5.4** On considère l'équation des ondes (8.55) dans un domaine  $\Omega \subset \mathbb{R}^N$  avec des conditions aux limites de Dirichlet ou de Neumann (homogènes), et une donnée initiale  $(u_0, u_1)$  régulière et à support compact dans  $\Omega$ . Vérifier qu'il existe un temps  $T > 0$  tel que sur l'intervalle  $[0, T]$  la solution est encore donnée par les formules de l'Exercice 8.5.3.

L'exercice suivant met en relief la différence essentielle entre la dimension d'espace  $N = 2$  ou 3.

**Exercice 8.5.5 (application musicale)** En admettant que le son se propage selon l'équation des ondes, montrer qu'il n'est pas possible d'écouter de la musique (audible) dans un monde de dimension spatiale  $N = 2$ , alors que c'est (fort heureusement) possible en dimension  $N = 3$ .

## 8.6 Méthodes numériques dans le cas parabolique

Dans cette section, nous montrons comment la méthode des éléments finis (présentée au Chapitre 6) s'adapte à la résolution numérique de l'équation de la chaleur : on utilise des éléments finis pour la discrétisation spatiale, et des différences finies pour la discrétisation temporelle.

### 8.6.1 Semi-discrétisation en espace

Il s'agit de discrétiser **en espace seulement** la formulation variationnelle (8.6) de l'équation de la chaleur (8.1). Pour cela, comme dans le cas des problèmes elliptiques, nous construisons une approximation variationnelle interne en introduisant un sous-espace  $V_{0h}$  de  $H_0^1(\Omega)$ , de dimension finie. Typiquement,  $V_{0h}$  sera un sous-espace d'éléments finis  $P_k$  (ou  $Q_k$ ) sur un maillage triangulaire (ou rectangulaire) comme précisé aux Définitions 6.3.5 et 6.3.25.

La semi-discrétisation de (8.6) est donc l'approximation variationnelle suivante : trouver  $u_h(t)$  fonction de  $]0, T[$  à valeurs dans  $V_{0h}$  telle que

$$\begin{cases} \frac{d}{dt} \langle u_h(t), v_h \rangle_{L^2(\Omega)} + a(u_h(t), v_h) = \langle f(t), v_h \rangle_{L^2(\Omega)} & \forall v_h \in V_{0h}, 0 < t < T, \\ u_h(t=0) = u_{0,h} \end{cases} \quad (8.56)$$

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".

On peut adapter le cadre abstrait du Théorème 8.2.3 pour montrer que (8.56) admet une unique solution, mais il est beaucoup plus simple et parlant de vérifier que (8.56) est en fait un système **d'équations différentielles ordinaires** à coefficients constants dont on calcule facilement l'unique solution. De manière pratique, pour résoudre (8.56) on introduit une base  $(\phi_i)_{1 \leq i \leq n_{dl}}$  de  $V_{0h}$  (typiquement, la base des éléments finis), et on cherche  $u_h(t)$  sous la forme

$$u_h(t) = \sum_{i=1}^{n_{dl}} U_i^h(t) \phi_i, \quad (8.57)$$

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 (8.57) les fonctions de base  $\phi_i$  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 (8.56) devient, pour tout  $1 \leq j \leq n_{dl}$ ,

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

Introduisant (comme dans la démonstration du Lemme 7.4.1) 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 (8.56) est équivalente au système linéaire **d'équations différentielles ordinaires** à coefficients constants

$$\begin{cases} \mathcal{M}_h \frac{dU^h}{dt}(t) + \mathcal{K}_h U^h(t) = b^h(t), & 0 < t < T, \\ U^h(t=0) = U^{0,h} \end{cases} \quad (8.58)$$

avec  $b_i^h(t) = \langle f(t), \phi_i \rangle_{L^2(\Omega)}$ . L'existence et l'unicité, ainsi qu'une formule explicite, de la solution de (8.58) s'obtiennent classiquement par simple diagonalisation simultanée de  $\mathcal{M}_h$  et  $\mathcal{K}_h$  (voir la Sous-section 7.4.1 à ce sujet). Comme il est difficile et coûteux de diagonaliser (8.58), en pratique on résout numériquement (8.58) 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. Avant cela, nous énonçons un résultat de convergence des solutions "semi-discrètes" de (8.56) vers la solution exacte de (8.6).

**Proposition 8.6.1** *Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages triangulaires réguliers de  $\Omega$ . Soit  $V_{0h}$  le sous-espace de  $H_0^1(\Omega)$ , défini par la méthode des éléments finis  $P_k$ , de dimension  $n_{dl}$ . Soit  $f(t) \in L^2([0, T]; L^2(\Omega))$ ,  $u_0 \in H_0^1(\Omega)$ , et  $u \in L^2([0, T]; H_0^1(\Omega)) \cap C([0, T]; L^2(\Omega))$ , la solution unique de l'équation de la chaleur (8.12). Soit  $u_h$  la solution unique de l'approximation variationnelle (8.56) dans  $V_{0h}$ . Si  $\lim_{h \rightarrow 0} \|u_{0,h} - u_0\|_{L^2(\Omega)} = 0$ , alors on a*

$$\lim_{h \rightarrow 0} \|u - u_h\|_{L^2([0, T]; H_0^1(\Omega))} = \lim_{h \rightarrow 0} \|u - u_h\|_{C([0, T]; L^2(\Omega))} = 0.$$

La démonstration de la Proposition 8.6.1 est tout à fait similaire à celle des résultats précédents d'approximation variationnelle. Elle peut se trouver dans l'ouvrage [36]. On peut aussi obtenir des estimations d'erreurs et des vitesses de convergence explicites, mais cela nous entraînerait trop loin...

## 8.6.2 Discrétisation totale en espace-temps

Après avoir discrétisé l'équation de la chaleur 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 le système d'équations différentielles ordinaires (8.58) issu 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. Pour simplifier les notations, nous réécrivons le système (8.58) sans mentionner la dépendance par rapport au paramètre  $h$  du maillage spatial

$$\begin{cases} \mathcal{M} \frac{dU}{dt}(t) + \mathcal{K}U(t) = b(t) \\ U(t=0) = U^0 \end{cases} \quad (8.59)$$

Pour simplifier l'analyse on va supposer que  $b(t)$  est continu sur  $[0, T]$ . 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 (8.59) le schéma le plus simple et le plus utilisé est le  $\theta$ -schéma (déjà vu, voir (2.5))

$$\mathcal{M} \frac{U^{n+1} - U^n}{\Delta t} + \mathcal{K}(\theta U^{n+1} + (1-\theta)U^n) = \theta b(t_{n+1}) + (1-\theta)b(t_n). \quad (8.60)$$

Lorsque  $\theta = 0$ , on appelle (8.60) **schéma explicite**, lorsque  $\theta = 1$ , **schéma implicite**, et pour  $\theta = 1/2$ , **schéma de Crank-Nicholson**. On peut réécrire (8.60) sous la forme

$$(\mathcal{M} + \theta \Delta t \mathcal{K})U^{n+1} = (\mathcal{M} - (1-\theta)\Delta t \mathcal{K})U^n + \Delta t(\theta b(t_{n+1}) + (1-\theta)b(t_n)). \quad (8.61)$$

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 la Remarque 7.4.3). Évidemment, on peut construire une foule de schémas en s'inspirant de ceux du Chapitre 2. Citons juste pour mémoire un exemple de schéma à 3 niveaux de temps, le **schéma de Gear**,

$$\mathcal{M} \frac{3U^{n+1} - 4U^n + U^{n-1}}{2\Delta t} + \mathcal{K}U^{n+1} = b(t_{n+1}). \quad (8.62)$$

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 8.6.1** Montrer que le schéma de Crank-Nicholson et celui de Gear sont d'ordre 2 (en temps), tandis que le  $\theta$ -schéma pour  $\theta \neq 1/2$  est d'ordre 1.

**Remarque 8.6.2** Il est possible de mettre en oeuvre une méthode d'éléments finis **en espace et en temps**, mais cela ne présente pas d'intérêt particulier sauf dans le cas où le domaine  $\Omega(t)$  varie en fonction du temps. •

Nous donnons maintenant une définition de la stabilité de ces schémas qui est une variante de la Définition 2.2.8.

**Définition 8.6.3** Un schéma aux différences finies pour (8.59) 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  $\Delta t$  et de la dimension du système  $n_d$  (donc du pas du maillage  $h$ ), mais peut dépendre de la donnée initiale  $U^0$ , du second membre  $b$ , et de  $T$ .

**Remarque 8.6.4** Le choix de la norme  $\sqrt{\mathcal{M}U \cdot U}$  dans la Définition 8.6.3 s'explique par le fait que  $\mathcal{M}U \cdot U = \int_{\Omega} |u|^2 dx$  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 pondérerait par  $\Delta x$  la norme euclidienne de  $U$  pour retrouver aussi l'analogie avec la norme de  $u$  dans  $L^2(\Omega)$ . •

**Lemme 8.6.5** Si  $1/2 \leq \theta \leq 1$ , le  $\theta$ -schéma (8.60) est inconditionnellement stable, tandis que, si  $0 \leq \theta < 1/2$ , il est stable sous la condition CFL

$$\max_i \lambda_i \Delta t \leq \frac{2}{1 - 2\theta}, \quad (8.63)$$

où les  $\lambda_i$  sont les valeurs propres de  $KU = \lambda \mathcal{M}U$  (voir (7.23)).

**Remarque 8.6.6** On ne reconnaît pas immédiatement dans (8.63) 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, on peut montrer que, si le maillage  $\mathcal{T}_h$  est uniformément régulier au sens où toute maille contient une boule de rayon  $Ch$  (avec  $C > 0$  indépendant de la maille), alors on a effectivement  $\max_i \lambda_i = \mathcal{O}(h^{-2})$ . On propose au lecteur de vérifier ce fait en dimension  $N = 1$  lors de l'Exercice 8.6.4 ci-dessous. En pratique on n'utilise pas le  $\theta$ -schéma pour  $\theta < 1/2$  car la condition de stabilité (8.63) 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 (8.61) dans la base orthonormale pour  $\mathcal{M}$  et diagonale pour  $\mathcal{K}$  (voir la démonstration du Lemme 7.4.1)

$$(\text{Id} + \theta \Delta t \text{diag}(\lambda_i)) \tilde{U}^{n+1} = (\text{Id} - (1 - \theta) \Delta t \text{diag}(\lambda_i)) \tilde{U}^n + \Delta t \tilde{b}^n, \quad (8.64)$$

avec  $\mathcal{M} = PP^*$ ,  $\mathcal{K} = P \text{diag}(\lambda_i) P^*$ ,  $\tilde{U}^n = P^* U^n$ , et  $\tilde{b}^n = P^{-1}(\theta b(t_{n+1}) + (1 - \theta)b(t_n))$ . On déduit de (8.64) que les composantes  $\tilde{U}_i^n$  de  $\tilde{U}^n$  vérifient

$$\tilde{U}_i^n = (\rho_i)^n \tilde{U}_i^0 + \frac{\Delta t}{1 + \theta \Delta t \lambda_i} \sum_{k=1}^n (\rho_i)^{k-1} \tilde{b}_i^{n-k}. \quad (8.65)$$

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 (8.63) si  $0 \leq \theta < 1/2$ , et qui est toujours satisfaite si  $\theta \geq 1/2$ .  $\square$

**Remarque 8.6.7** Il est clair dans l'estimation (8.65) que plus  $\theta$  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  $\theta$ -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.  $\bullet$

**Remarque 8.6.8** Le système d'équations différentielles ordinaires (8.59) est dit "raide" car, ses solutions faisant intervenir des termes du type  $\exp(-\lambda_i t)$ , évoluent sur des échelles de temps très différentes (rappelons que  $\min_i \lambda_i = \mathcal{O}(1)$  et  $\max_i \lambda_i = \mathcal{O}(h^{-2})$ ). Il existe des méthodes numériques d'intégration d'équations différentielles ordinaires raides plus efficaces et plus compliquées que le  $\theta$ -schéma. Citons pour mémoire les méthodes de Runge-Kutta (voir, par exemple, le chapitre XX de [18]). Néanmoins, il y a un compromis à trouver entre l'utilisation d'une méthode numérique d'intégration en temps robuste, mais chère, et la taille très importante des systèmes à résoudre (rappelons que la taille est proportionnelle au nombre de mailles).  $\bullet$

On peut utiliser le caractère variationnel de la discrétisation par éléments finis qui a conduit à (8.59) pour montrer autrement la stabilité inconditionnelle du  $\theta$ -schéma.

**Exercice 8.6.2** On considère le  $\theta$ -schéma (8.60) avec  $1/2 \leq \theta \leq 1$ . On note  $\|U\|_{\mathcal{M}} = \sqrt{\mathcal{M}U \cdot U}$ . Démontrer l'équivalent discret suivant de l'inégalité d'énergie (8.17)

$$\|U^{n_0}\|_{\mathcal{M}}^2 + \Delta t \sum_{n=0}^{n_0} \mathcal{K} \hat{U}^n \cdot \hat{U}^n \leq C \left( \|U^0\|_{\mathcal{M}}^2 + \int_0^T \|f(t)\|_{L^2(\Omega)}^2 dt + \mathcal{O}(1) \right).$$

Pour cela, on prendra le produit scalaire de (8.60) avec  $\hat{U}^n = \theta U^{n+1} + (1 - \theta)U^n$ .

**Exercice 8.6.3** Montrer que le schéma de Gear (8.62) est inconditionnellement stable.

**Exercice 8.6.4** On résout par éléments finis  $P_1$  et schéma explicite en temps l'équation de la chaleur (8.12) en dimension  $N = 1$ . On utilise une formule de quadrature qui rend la matrice  $\mathcal{M}$  diagonale (voir la Remarque 7.4.3 et l'Exercice 7.4.1). On rappelle que la matrice  $\mathcal{K}$  est donnée par (6.12) et qu'on a calculé ses valeurs propres lors de l'Exercice 13.1.3. Montrer que dans ce cas la condition CFL (8.63) est bien du type  $\Delta t \leq Ch^2$ .

Finalement, nous pouvons énoncer un résultat de convergence de cette méthode de discrétisation que nous nous garderons bien de démontrer (la démonstration de la Proposition 8.6.9 est dans l'esprit des démonstrations précédentes). Pour plus de détails, ainsi que pour des estimations d'erreurs et des vitesses de convergence explicites, nous renvoyons à [36].

**Proposition 8.6.9** *Soit  $u$  la solution “suffisamment régulière” de l'équation de la chaleur (8.12). Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages triangulaires réguliers de  $\Omega$ . Soit  $V_{0h}$  le sous-espace de  $H_0^1(\Omega)$ , défini par la méthode des éléments finis  $P_k$ . Soit  $(\Delta t)$  une suite de pas de temps qui tend vers zéro. Soit  $u_h^n \in V_{0h}$  la fonction dont les coordonnées  $U^n$  dans la base des éléments finis de  $V_{0h}$  sont calculées par le  $\theta$ -schéma. Si  $\lim_{h \rightarrow 0} u_h^0 = u_0$  dans  $L^2(\Omega)$ , et si  $h$  et  $\Delta t$  tendent vers 0 en respectant la condition de stabilité (8.63), alors on a*

$$\lim_{h \rightarrow 0, \Delta t \rightarrow 0} \max_{0 \leq n \leq n_0} \|u(t_n) - u_h^n\|_{L^2(\Omega)} = 0.$$

Pour terminer nous illustrons notre propos en reprenant l'exemple de la Sous-section 6.3 sur la diffusion dans l'atmosphère d'un polluant émis par une source localisée (voir le maillage à la Figure 6.7 et le second membre  $f$  à la Figure 6.14). La donnée initiale  $u_0$  est prise nulle dans le domaine. La Figure 8.2 présente les résultats pour 4 temps différents. Comme le terme source est indépendant du temps, la solution converge bien vers un régime stationnaire lorsque le temps tend vers l'infini.

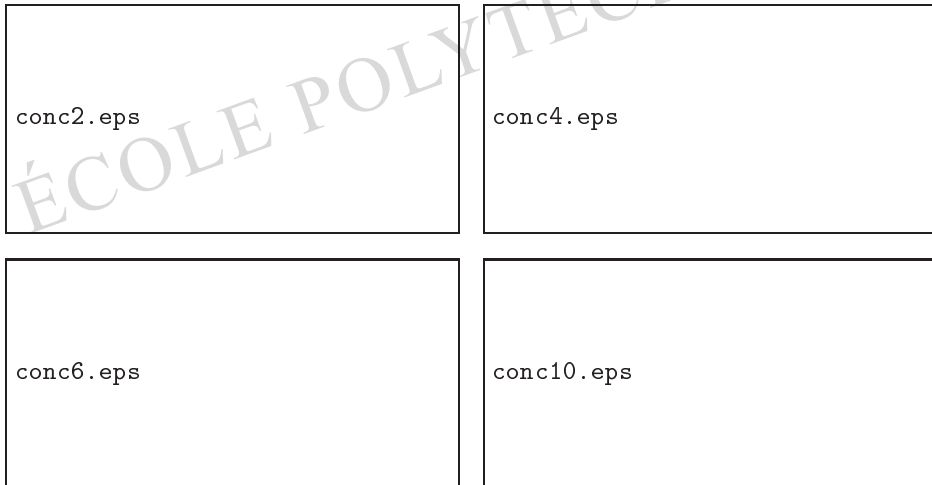


FIGURE 8.2 – Instantanés aux temps  $t = 1, 2, 3, 5$  de la concentration.

## 8.7 Méthodes numériques dans le cas hyperbolique

Les méthodes numériques pour résoudre l'équation des ondes sont très semblables (dans le principe, mais pas toujours dans la pratique) à celles que nous avons vues pour l'équation de la chaleur.

### 8.7.1 Semi-discrétisation en espace

On discrétise **en espace seulement** la formulation variationnelle (8.25) de l'équation des ondes (8.3). Pour cela, on construit une approximation variationnelle interne en introduisant un sous-espace  $V_{0h}$  de  $H_0^1(\Omega)$ , de dimension finie (typiquement, un sous-espace d'éléments finis). La semi-discrétisation de (8.25) est donc l'approximation variationnelle suivante : trouver  $u_h(t)$  fonction de  $]0, T[$  à valeurs dans  $V_{0h}$  telle que

$$\begin{cases} \frac{d^2}{dt^2} \langle u_h(t), v_h \rangle_{L^2(\Omega)} + a(u_h(t), v_h) = \langle f(t), v_h \rangle_{L^2(\Omega)} & \forall v_h \in V_{0h}, \quad 0 < t < T, \\ u_h(t=0) = u_{0,h}, \quad \frac{\partial u_h}{\partial t}(t=0) = u_{1,h} \end{cases} \quad (8.66)$$

où  $u_{0,h} \in V_{0h}$  et  $u_{1,h} \in V_{0h}$  sont des approximations des données initiales  $u_0$  et  $u_1$ .

Pour montrer que (8.66) admet une unique solution et la calculer de manière pratique, on introduit une base  $(\phi_i)_{1 \leq i \leq n_{dl}}$  de  $V_{0h}$  (qui ne dépend pas du temps), et on cherche  $u_h(t)$  sous la forme

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

avec  $U^h = (U_i^h)_{1 \leq i \leq n_{dl}}$  le vecteur des coordonnées de  $u_h$ . En posant

$$u_{0,h} = \sum_{i=1}^{n_{dl}} U_i^{0,h} \phi_i, \quad u_{1,h} = \sum_{i=1}^{n_{dl}} U_i^{1,h} \phi_i, \quad b_i^h(t) = \langle f(t), \phi_i \rangle_{L^2(\Omega)}, \quad 1 \leq i \leq n_{dl},$$

l'approximation variationnelle (8.66) est équivalente au système linéaire **d'équations différentielles ordinaires** d'ordre 2 à coefficients constants

$$\begin{cases} \mathcal{M}_h \frac{d^2 U^h}{dt^2}(t) + \mathcal{K}_h U^h(t) = b^h(t), & 0 < t < T, \\ U^h(t=0) = U^{0,h}, \quad \frac{dU^h}{dt}(t=0) = U^{1,h}, \end{cases} \quad (8.67)$$

où on retrouve les mêmes **matrices de masse  $\mathcal{M}_h$  et de rigidité  $\mathcal{K}_h$**  que pour l'équation de la chaleur

$$(\mathcal{M}_h)_{ij} = \langle \phi_i, \phi_j \rangle_{L^2(\Omega)}, \quad (\mathcal{K}_h)_{ij} = a(\phi_i, \phi_j) \quad 1 \leq i, j \leq n_{dl}.$$

L'existence et l'unicité, ainsi qu'une formule explicite, de la solution de (8.67) s'obtiennent facilement par simple diagonalisation simultanée des matrices  $\mathcal{M}_h$  et  $\mathcal{K}_h$ . Comme il est difficile et coûteux de diagonaliser (8.67), en pratique on résout numériquement (8.67) par discrétisation et marche en temps.

**Exercice 8.7.1** Écrire le système linéaire d'équations différentielles ordinaires obtenu par semi-discrétisation de l'équation des ondes amortie (8.53).

### 8.7.2 Discrétisation totale en espace-temps

On utilise une méthode de **différences finies en temps** pour résoudre le système d'équations différentielles ordinaires (8.67). Pour simplifier les notations, nous réécrivons le système (8.67) sans mentionner la dépendance spatiale en  $h$

$$\begin{cases} \mathcal{M} \frac{d^2 U}{dt^2}(t) + \mathcal{K}U(t) = b(t) \\ U(t=0) = U_0, \quad \frac{dU}{dt}(t=0) = U_1, \end{cases} \quad (8.68)$$

où on suppose que  $b(t)$  est continu sur  $[0, T]$ . On découpe l'intervalle de temps  $[0, T]$  en  $n_0$  pas de temps  $\Delta t = T/n_0$ , on pose  $t_n = n\Delta t$   $0 \leq n \leq n_0$ , et on note  $U^n$  l'approximation de  $U(t_n)$  calculé par un schéma. Pour  $0 \leq \theta \leq 1/2$  on propose le  **$\theta$ -schéma**

$$\begin{aligned} \mathcal{M} \frac{U^{n+1} - 2U^n + U^{n-1}}{(\Delta t)^2} + \mathcal{K}(\theta U^{n+1} + (1-2\theta)U^n + \theta U^{n-1}) \\ = \theta b(t_{n+1}) + (1-2\theta)b(t_n) + \theta b(t_{n-1}). \end{aligned} \quad (8.69)$$

Lorsque  $\theta = 0$ , on appelle (8.69) **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

$$U^0 = U_0 \quad \text{et} \quad \frac{U^1 - U^0}{\Delta t} = U_1.$$

Un schéma plus fréquemment utilisé car plus général est le **schéma de Newmark**. Pour résoudre le système "amorti"

$$\mathcal{M} \frac{d^2 U}{dt^2}(t) + \mathcal{C} \frac{dU}{dt}(t) + \mathcal{K}U(t) = b(t)$$

on approche  $U(t)$ ,  $dU/dt(t)$ ,  $d^2U/dt^2(t)$  par trois suites  $U^n, \dot{U}^n, \ddot{U}^n$

$$\begin{cases} \mathcal{M}\ddot{U}^{n+1} + \mathcal{C}\dot{U}^{n+1} + \mathcal{K}U^{n+1} = b(t_{n+1}) \\ \dot{U}^{n+1} = \dot{U}^n + \Delta t(\delta\ddot{U}^{n+1} + (1-\delta)\ddot{U}^n) \\ U^{n+1} = U^n + \Delta t\dot{U}^n + \frac{(\Delta t)^2}{2}(2\theta\ddot{U}^{n+1} + (1-2\theta)\ddot{U}^n) \end{cases} \quad (8.70)$$

avec  $0 \leq \delta \leq 1$  et  $0 \leq \theta \leq 1/2$ . Lorsque la matrice d'amortissement est nulle ( $\mathcal{C} = 0$ ), on peut éliminer les suites  $\dot{U}^n, \ddot{U}^n$  et (8.70) est équivalent à

$$\begin{aligned} \mathcal{M} \frac{U^{n+1} - 2U^n + U^{n-1}}{(\Delta t)^2} + \mathcal{K} \left( \theta U^{n+1} + \left( \frac{1}{2} + \delta - 2\theta \right) U^n + \left( \frac{1}{2} - \delta + \theta \right) U^{n-1} \right) \\ = \theta b(t_{n+1}) + \left( \frac{1}{2} + \delta - 2\theta \right) b(t_n) + \left( \frac{1}{2} - \delta + \theta \right) b(t_{n-1}). \end{aligned} \quad (8.71)$$

Remarquons que pour  $\delta = 1/2$  le schéma de Newmark redonne le  $\theta$ -schéma. En pratique, plus  $\delta$  est grand, plus le schéma est dissipatif et robuste (les erreurs numériques passées s'amortissent plus vite), même s'il est moins précis.

**Exercice 8.7.2** Montrer que le schéma de Newmark est d'ordre 1 (en temps) pour  $\delta \neq 1/2$ , d'ordre 2 pour  $\delta = 1/2$  et  $\theta \neq 1/12$ , et d'ordre 4 si  $\delta = 1/2$  et  $\theta = 1/12$  (on se limitera à l'équation sans amortissement).

On étudie la stabilité de ces schémas au sens de la Définition 8.6.3. Pour éviter des calculs trop lourds, on se contente d'étudier la condition nécessaire de stabilité de Von Neumann (voir la Remarque 2.2.24). Le résultat suivant est dans le même esprit que le Lemme 2.3.6.

**Lemme 8.7.1** *On considère le schéma de Newmark (8.71). Si  $\delta < 1/2$ , il est toujours instable. Supposons désormais que  $\delta \geq 1/2$ . La condition nécessaire de stabilité de Von Neumann est toujours vérifiée si  $\delta \leq 2\theta \leq 1$ , tandis que, si  $0 \leq 2\theta < \delta$  elle n'est satisfaite que sous la condition CFL*

$$\max_i \lambda_i (\Delta t)^2 < \frac{2}{\delta - 2\theta}, \quad (8.72)$$

où les  $\lambda_i$  sont les valeurs propres de  $KU = \lambda MU$  (voir (7.23)).

**Remarque 8.7.2** On ne reconnaît pas immédiatement dans (8.72) la condition CFL (Courant-Friedrichs-Lewy) usuelle  $\Delta t \leq Ch$  pour l'équation des ondes (voir le Lemme 2.3.6). En fait, on peut montrer que, si le maillage  $\mathcal{T}_h$  est uniformément régulier au sens où toute maille contient une boule de rayon  $Ch$  (avec  $C > 0$  indépendant de la maille), alors on a effectivement  $\max_i \lambda_i = \mathcal{O}(h^{-2})$ . Le lecteur a pu vérifier ce fait en dimension  $N = 1$  lors de l'Exercice 8.6.4. Contrairement au cas parabolique, la condition CFL (8.72) n'est pas trop sévère puisqu'on peut prendre des pas de temps  $\Delta t$  de l'ordre du pas d'espace  $h$ . Cependant, quitte à inverser un système linéaire (pour pouvoir calculer  $U^{n+1}$  en fonction de  $U^n, U^{n-1}$  et du second membre), autant utiliser le schéma de Newmark pour des valeurs de  $\delta$  et  $\theta$  telles qu'il soit inconditionnellement stable. Le seul cas intéressant pour un schéma stable sous condition CFL est le cas où il est **explicite**, c'est-à-dire qu'il n'y a pas de système linéaire à résoudre à chaque pas de temps. En effet, un schéma explicite nécessite très peu d'opérations par pas de temps et conduit donc à des calculs peu coûteux. La seule possibilité pour le schéma de Newmark (8.71) d'être explicite est que  $\theta = 0$  et que la matrice de masse  $\mathcal{M}$  soit diagonale grâce à une formule d'intégration approchée (voir la Remarque 7.4.3). Ce schéma explicite est souvent utilisé en pratique avec  $\delta = 1/2$ . •

**Démonstration.** Elle est très semblable à celle du Lemme 8.6.5 mais avec des complications techniques qui rappellent le Lemme 2.3.6. On décompose  $U^n$  et le second membre de (8.71) dans la base orthonormale pour  $\mathcal{M}$  et orthogonale pour  $\mathcal{K}$ . Par

conséquent, (8.71) est équivalent, composante par composante, à

$$\frac{U_i^{n+1} - 2U_i^n + U_i^{n-1}}{(\Delta t)^2} + \lambda_i \left( \theta U_i^{n+1} + \left(\frac{1}{2} + \delta - 2\theta\right) U_i^n + \left(\frac{1}{2} - \delta + \theta\right) U_i^{n-1} \right) = b_i^n, \quad (8.73)$$

avec des notations évidentes (les  $\lambda_i$  sont les valeurs propres du système matriciel  $KV_i = \lambda_i M(V_i)$ ). Comme le schéma (8.73) est à trois niveaux (voir la Sous-section 2.2.5), on introduit une matrice d'itérations  $A_i$  telle que

$$\begin{pmatrix} U_i^{n+1} \\ U_i^n \end{pmatrix} = A_i \begin{pmatrix} U_i^n \\ U_i^{n-1} \end{pmatrix} + \frac{(\Delta t)^2}{1 + \theta \lambda_i (\Delta t)^2} \begin{pmatrix} b_i^n \\ 0 \end{pmatrix} \quad \text{avec} \quad A_i = \begin{pmatrix} a_{11} & a_{12} \\ 1 & 0 \end{pmatrix},$$

$$a_{11} = \frac{2 - \lambda_i (\Delta t)^2 (\frac{1}{2} + \delta - 2\theta)}{1 + \theta \lambda_i (\Delta t)^2}, \quad a_{12} = -\frac{1 + \lambda_i (\Delta t)^2 (\frac{1}{2} - \delta + \theta)}{1 + \theta \lambda_i (\Delta t)^2}.$$

On en déduit que

$$\begin{pmatrix} U_i^{n+1} \\ U_i^n \end{pmatrix} = A_i^n \begin{pmatrix} U_i^1 \\ U_i^0 \end{pmatrix} U_i^0 + \frac{(\Delta t)^2}{1 + \theta \lambda_i (\Delta t)^2} \sum_{p=0}^{n-1} A_i^p \begin{pmatrix} b_i^{n-p} \\ 0 \end{pmatrix}. \quad (8.74)$$

La condition nécessaire de stabilité de Von Neumann est  $\rho(A_i) \leq 1$ . On calcule donc les valeurs propres de  $A_i$  qui sont les racines du polynôme en  $\mu$  suivant

$$\mu^2 - a_{11}\mu - a_{12} = 0$$

dont le discriminant est

$$\Delta = \frac{-4\lambda_i (\Delta t)^2 + \lambda_i^2 (\Delta t)^4 \left( (\frac{1}{2} + \delta)^2 - 4\theta \right)}{(1 + \theta \lambda_i (\Delta t)^2)^2}.$$

On vérifie aisément que les racines de ce polynôme sont de module inférieur ou égal à 1 si et seulement si on est dans l'un des deux cas suivants : soit  $\Delta \leq 0$  et  $a_{12} \geq -1$ , soit  $\Delta > 0$  et  $1 - a_{12} \geq |a_{11}|$ . Un calcul fastidieux mais simple dans le principe conduit alors à la condition (8.72) (voir si nécessaire le théorème 6, section 3, chapitre XX dans [18]).  $\square$

**Exercice 8.7.3** On considère le cas limite du Lemme 8.7.1, c'est-à-dire  $\delta = 1/2$  et  $\lambda_i (\Delta t)^2 = \frac{4}{1-4\theta}$ . Montrer que le schéma de Newmark est instable dans ce cas en vérifiant que

$$A_i = \begin{pmatrix} -2 & -1 \\ 1 & 0 \end{pmatrix}, \quad \text{et} \quad A_i^n = (-1)^n \begin{pmatrix} n+1 & n \\ -n & 1-n \end{pmatrix}.$$

Remarquez qu'il s'agit d'une instabilité "faible" puisque la croissance de  $A_i^n$  est linéaire et non exponentielle.

**Remarque 8.7.3** En l'absence de termes sources les solutions du système d'équations différentielles ordinaires (8.68) sont des fonctions oscillantes du type  $\cos(\omega_i t)$ , avec  $\omega_i^2 = \lambda_i$ . Comme les schémas numériques d'intégration de (8.68) ont tendance à amortir un peu ces oscillations (on dit qu'ils sont diffusifs ou dissipatifs; voir la discussion après la Définition 2.3.5), on peut leur préférer une autre méthode de résolution de (8.68) basée sur une décomposition de la solution sur les vecteurs propres  $V_i$  du système matriciel  $\mathcal{K}V_i = \lambda_i \mathcal{M}V_i$ . On cherche la solution  $U(t)$  de (8.68) sous la forme

$$U(t) = \sum_{i \in I} x_i(t) V_i,$$

où  $I$  est une collection d'indices de modes propres choisis pour représenter correctement la solution, et  $x_i(t)$  est la solution d'une équation différentielle ordinaire scalaire

$$\frac{d^2 x_i}{dt^2}(t) + \lambda_i x_i(t) = b_i(t)$$

qu'on peut intégrer facilement, rapidement, et précisément par la méthode numérique de son choix. Cette méthode, dite de **superpositions de modes**, est très utilisée en mécanique vibratoire. Elle a l'avantage d'être peu dissipative, c'est-à-dire d'amortir très peu les oscillations. Bien sûr, la qualité des résultats dépend pour une grande part du choix des modes entrant dans la décomposition, choix souvent motivé par des considérations physiques. •

Finalement, nous pouvons énoncer un résultat de convergence de cette méthode de discrétisation que nous nous garderons bien de démontrer (la démonstration de la Proposition 8.7.4 est dans l'esprit des démonstrations précédentes). Pour plus de détails, ainsi que pour des estimations d'erreurs et des vitesses de convergence explicites, nous renvoyons à [36].

**Proposition 8.7.4** Soit  $u$  la solution "suffisamment régulière" de l'équation des ondes (8.33). Soit  $(\mathcal{T}_h)_{h>0}$  une suite de maillages triangulaires réguliers de  $\Omega$ . Soit  $V_{0h}$  le sous-espace de  $H_0^1(\Omega)$ , défini par la méthode des éléments finis  $P_k$ . Soit  $(\Delta t)$  une suite de pas de temps qui tend vers zéro. Soit  $u_h^n \in V_{0h}$  la fonction dont les coordonnées  $U^n$  dans la base des éléments finis de  $V_{0h}$  sont calculées par le schéma de Newmark. Si  $\lim_{h \rightarrow 0} u_h^0 = u_0$  dans  $L^2(\Omega)$ ,  $\lim_{h \rightarrow 0} u_h^1 = u_1$  dans  $L^2(\Omega)$ , et si  $h$  et  $\Delta t$  tendent vers 0 en respectant la condition de stabilité (8.72), alors on a

$$\lim_{h \rightarrow 0, \Delta t \rightarrow 0} \max_{0 \leq n \leq n_0} \|u(t_n) - u_h^n\|_{L^2(\Omega)} = 0.$$

Pour terminer nous illustrons notre propos en simulant la propagation d'une onde sphérique dans une cavité carrée sur les parois de laquelle elle se réfléchit. On résout donc l'équation des ondes dans  $\Omega = ]0, 1[^2$  avec une condition aux limites de Neumann, une vitesse initiale nulle et un déplacement initial à support compact et symétrie sphérique centré sur le point (0.3, 0.4). Dans la Figure 8.3 on trace le module de la déformation  $\nabla u$  à l'instant initial et à 5 instants ultérieurs (ce type d'image, appelé diagramme de Schlieren, représente ce que l'on pourrait voir dans une expérience réelle).

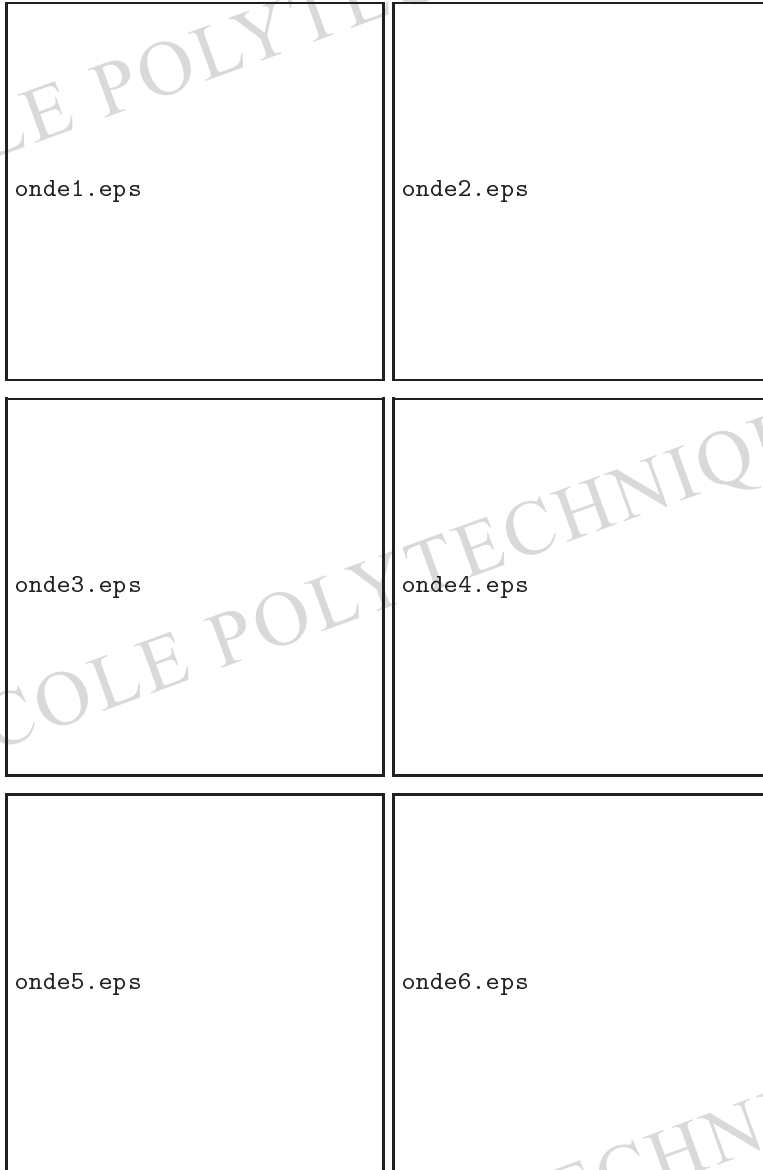


FIGURE 8.3 – Instantanés aux temps  $t = 0.0, 0.1, 0.2, 0.3, 0.4, 0.5$  de  $|\nabla u|$ .



## Chapitre 9

# INTRODUCTION À L'OPTIMISATION

### 9.1 Motivation et exemples

#### 9.1.1 Introduction

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.

Par la suite nous mélangerons donc les exemples de problèmes d'optimisation

où les modèles sont de nature très différente. Dans le cas le plus simple, le modèle sera une simple équation algébrique et il s'agira simplement d'optimiser une fonction définie sur un espace de dimension finie (disons  $\mathbb{R}^n$ ). Typiquement, c'est la situation la plus fréquente en recherche opérationnelle. Une deuxième catégorie de problèmes correspond au cas où la fonction à optimiser dépend de la solution d'une équation différentielle ordinaire (autrement dit, cette fonction est définie sur un espace de dimension infinie, disons  $C[0, T]$ ). On parle alors de commande optimale, et les applications sont très nombreuses en automatique et robotique. La troisième et dernière catégorie correspond à l'optimisation de fonctions de la solution d'une équation aux dérivées partielles. Il s'agit alors de la théorie du contrôle optimal des systèmes distribués qui a de nombreuses applications, par exemple en conception optimale ou pour la stabilisation de structures mécaniques. La prochaine sous-section fait un panorama de quelques exemples typiques de ces problèmes d'optimisation. Remarquons que ces catégories ne sont pas hermétiquement cloisonnées puisqu'après discrétisation spatiale et/ou temporelle une équation différentielle ordinaire ou aux dérivées partielles se ramène à un système d'équations algébriques.

On peut aussi séparer l'optimisation en deux grandes branches aux méthodes fort différentes selon que les variables sont continues ou discrètes. Typiquement, si l'on minimise une fonction  $f(x)$  avec  $x \in \mathbb{R}^n$ , il s'agit d'**optimisation en variables continues**, tandis que si  $x \in \mathbb{Z}^n$  on a affaire à de l'**optimisation combinatoire** ou en variables discrètes. Malgré les apparences, l'optimisation en variables continues est souvent plus "facile" que l'optimisation en variables discrètes car on peut utiliser la notion de dérivée qui est fort utile tant du point de vue théorique qu'algorithmique. L'optimisation combinatoire est naturelle et essentielle dans de nombreux problèmes de la recherche opérationnelle. C'est un domaine où, à côté de résultats théoriques rigoureux, fleurissent de nombreuses "heuristiques" essentielles pour obtenir de bonnes performances algorithmiques.

Pour finir cette brève introduction nous indiquons le plan de la suite du cours. Ce chapitre va principalement porter sur la question de l'existence et de l'unicité en optimisation continue, que ce soit en dimension finie ou infinie. En particulier, nous verrons le rôle crucial de la **convexité** pour obtenir des résultats d'existence en dimension infinie. Le Chapitre 10 développera les conditions d'optimalité et les algorithmes numériques qui en découlent. Le Chapitre 11 constitue une introduction aux méthodes de la recherche opérationnelle, incluant la programmation linéaire et des méthodes combinatoires. Pour plus de détails sur l'optimisation nous renvoyons le lecteur aux ouvrages [6], [10], [13], [15], [21], [23], [34].

### 9.1.2 Exemples

Passons en revue quelques problèmes typiques d'optimisation, d'importance pratique ou théorique inégale, mais qui permettent de faire le tour des différentes "branches" de l'optimisation.

Commençons par quelques exemples en **recherche opérationnelle**, c'est-à-dire en optimisation de la gestion ou de la programmation des ressources.

**Exemple 9.1.1 (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

$$v_{ij} \geq 0, \sum_{j=1}^N v_{ij} \leq s_i, \sum_{i=1}^M v_{ij} = r_j \quad \text{pour } 1 \leq i \leq M, 1 \leq j \leq N.$$

•

**Exemple 9.1.2 (problème d'affectation)** Il s'agit d'un exemple d'optimisation combinatoire ou en variables entières. Imaginez vous à la tête d'une agence matrimoniale... Soit  $N$  femmes, indicées par  $1 \leq i \leq N$ , et  $N$  hommes, indicés par  $1 \leq j \leq N$ . Si la femme  $i$  et l'homme  $j$  sont d'accord pour se marier leur variable d'accord  $a_{ij}$  vaut 1; dans le cas contraire elle vaut 0. Restons classiques : seuls les mariages hétérosexuels sont autorisés et la polygamie n'est pas admise... (On verra dans la Sous-section 11.3.7 que cette hypothèse de monogamie n'est pas nécessaire!) Le but du jeu est de maximiser le nombre de mariages "satisfaisants" entres ces  $N$  femmes et  $N$  hommes. Autrement dit, on cherche une permutation  $\sigma$  dans l'ensemble des permutations  $\mathcal{S}_N$  de  $\{1, \dots, N\}$  qui réalise le maximum de

$$\max_{\sigma \in \mathcal{S}_N} \sum_{i=1}^N a_{i\sigma(i)}.$$

Une variante consiste à autoriser des valeurs de  $a_{ij}$  entre 0 et 1. Ce type de problèmes est appelé problème d'affectation (il intervient dans des contextes industriels plus sérieux comme l'affectation des équipages et des avions dans une compagnie aérienne). Bien que ce ne soit pas forcément la meilleure manière de poser le problème, on peut l'écrire sous une forme voisine de l'Exemple 9.1.1. Les variables de décision sont notées  $v_{ij}$  qui vaut 1 s'il y a mariage entre la femme  $i$  et l'homme  $j$  et 0 sinon. On veut maximiser

$$\sup_{(v_{ij})} \left( \sum_{i=1}^N \sum_{j=1}^N a_{ij} v_{ij} \right)$$

sous les contraintes

$$v_{ij} = 0 \text{ ou } 1, \sum_{j=1}^N v_{ij} \leq 1, \sum_{i=1}^M v_{ij} \leq 1 \text{ pour } 1 \leq i, j \leq N.$$

On pourrait croire que ce problème d'affectation est simple puisqu'il y a un nombre fini de possibilités qu'il "suffit" d'énumérer pour trouver l'optimum. Il s'agit bien sûr d'un leurre car la caractéristique des problèmes combinatoires est leur très grand nombre de combinaisons possibles qui empêche toute énumération exhaustive en pratique. Néanmoins pour ce problème il existe des techniques de résolution efficaces (voir la Sous-section 11.3.7). •

**Exemple 9.1.3 (problème du sac-à-dos)** Un problème classique est celui du sac-à-dos. Soit  $n$  objets de poids respectifs  $p_1, \dots, p_n \in \mathbb{R}$ , et d'utilités respectives  $u_1, \dots, u_n \in \mathbb{R}$ , et  $P \in \mathbb{R}$  un poids maximal que l'on est disposé à porter. On pose  $x_i = 1$  si on met le  $i$ -ème objet dans le sac-à-dos, et  $x_i = 0$  sinon. On veut maximiser l'utilité du sac à dos sous contrainte de poids :

$$\max_{\substack{x \in \{0,1\}^n \\ \sum_{1 \leq i \leq n} x_i p_i \leq P}} \sum_{1 \leq i \leq n} x_i u_i.$$

Encore une fois la difficulté vient de ce que les variables d'optimisation  $x_i$  sont discrètes (voir l'Exercice 11.4.5 pour une méthode de résolution). •

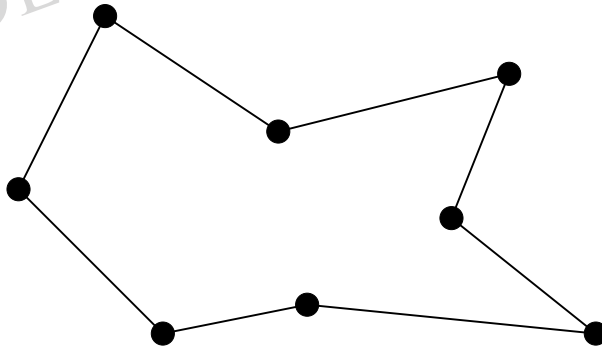


FIGURE 9.1 – Tournée pour un voyageur de commerce dans l'Exemple 9.1.4.

**Exemple 9.1.4 (tourn e du voyageur de commerce)** Un exemple c l bre en optimisation combinatoire est le probl me du voyageur de commerce. Un repr sentant (ou VRP) doit visiter  $n$  villes successivement et revenir   son point de d part en un temps minimum. On note  $t_{ij}$  le temps pour rejoindre la ville  $i$  de la ville  $j$  ( ventuellement diff rent de  $t_{ji}$ ). On trace alors le graphe orient  des  $n$  villes reli es entre elles

par des arcs pondérés par les  $(t_{ij})$  (voir la Figure 9.1). Il faut alors trouver un cycle dans ce graphe qui passe une et une seule fois par toutes les villes. Il est possible de mettre ce problème sous une forme de type programme linéaire en variables entières (voir la Sous-section 11.6.1 pour une méthode de résolution). •

**Exemple 9.1.5 (chemin de coût minimum)** Soit un graphe orienté  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ , où  $\mathcal{N}$  est l'ensemble des nœuds et  $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$  est un ensemble d'arcs reliant ces nœuds. On associe à chaque arc  $(k, m) \in \mathcal{A}$  un coût  $w(k, m)$ . On fixe un nœud d'origine  $i$  et un nœud de destination  $j$ . Le problème du chemin de coût minimum consiste à trouver un chemin du graphe allant de  $i$  à  $j$  en un nombre quelconque d'étapes qui soit de coût total minimum. Autrement dit, on cherche une suite de nœuds  $i = \ell_0, \dots, \ell_T = j$  telle que  $(\ell_r, \ell_{r+1}) \in \mathcal{A}$  pour tout  $r = 0, 1, \dots, T-1$  et  $w(\ell_0, \ell_1) + \dots + w(\ell_{T-1}, \ell_T)$  est minimal.

Lorsque les nœuds sont des villes, que  $\mathcal{A}$  est l'ensemble des routes directes d'une ville à l'autre, et que  $w(k, m)$  est la distance routière entre les villes  $k$  et  $m$ , on retrouve le problème classique du plus court chemin. Malgré son aspect combinatoire, ce problème est facile à résoudre y compris pour des instances de grande taille. •

Voici un exemple algébrique très simple qui provient, par exemple, de la discrétisation par éléments finis des équations de Stokes (voir la Sous-section 6.3.4).

**Exemple 9.1.6 (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

$$\inf_{x \in \text{Ker} B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}.$$

La contrainte d'appartenance à  $\text{Ker} B$  rend cette minimisation non évidente (voir la Sous-section 10.2.2 pour sa résolution). •

Un autre exemple algébrique simple est celui du quotient de Rayleigh qui permet de calculer les valeurs et vecteurs propres d'une matrice symétrique.

**Exemple 9.1.7 (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 10.2.2). •

Considérons un exemple classique en économie.

**Exemple 9.1.8 (consommation des ménages)** 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 10.3.2 pour la résolution). •

Passons à un exemple d'optimisation d'un système modélisé par une équation différentielle ordinaire, c'est-à-dire à un problème de **commande optimale**.

**Exemple 9.1.9 (commande optimale)** On considère un système différentiel linéaire avec critère quadratique. Le but est de guider un robot (ou un engin spatial, un véhicule, etc.) afin qu'il suive "au plus près" une trajectoire prédéfinie. L'état du robot à l'instant  $t$  est représenté par une fonction  $y(t)$  à valeurs dans  $\mathbb{R}^N$  (typiquement, la position et la vitesse). On agit sur le robot par l'intermédiaire d'une commande  $v(t)$  à valeurs dans  $\mathbb{R}^M$  (typiquement, la puissance du moteur, la direction des roues, etc.). En présence de forces  $f(t) \in \mathbb{R}^N$  les lois de la mécanique conduisent à un système d'équations différentielles ordinaires (supposées linéaires pour simplifier)

$$\begin{cases} \frac{dy}{dt} = Ay + Bv + f \text{ pour } 0 \leq t \leq T \\ y(0) = y_0 \end{cases} \quad (9.1)$$

où  $y_0 \in \mathbb{R}^N$  est l'état initial du système,  $A$  et  $B$  sont deux matrices constantes de dimensions respectives  $N \times N$  et  $N \times M$ . On note  $z(t)$  une trajectoire "cible" et  $z_T$  une position finale "cible". Pour approcher au mieux ces cibles et pour minimiser le coût du contrôle, on introduit trois matrices symétriques positives  $R, Q, D$  dont seule  $R$  est supposée en plus être définie positive. On définit alors un critère quadratique

$$J(v) = \int_0^T Rv(t) \cdot v(t) dt + \int_0^T Q(y-z)(t) \cdot (y-z)(t) dt + D(y(T) - z_T) \cdot (y(T) - z_T).$$

Remarquons que la fonction  $y(t)$  dépend de la variable  $v$  à travers (9.1). Comme les commandes admissibles sont éventuellement limitées (la puissance d'un moteur est souvent bornée...), on introduit un convexe fermé non vide  $K$  de  $\mathbb{R}^M$  qui représente l'ensemble des commandes admissibles. Le problème est donc de résoudre

$$\inf_{v(t) \in K, 0 \leq t \leq T} J(v).$$

Il faudra, bien sûr, préciser dans quels espaces fonctionnels on minimise  $J(v)$  et on définit la solution  $y$  de (9.1) (voir la Sous-section 10.4.2 pour la résolution). •

**Exemple 9.1.10 (minimisation d'une énergie mécanique)** Il s'agit de minimiser l'énergie mécanique d'une membrane ou bien l'énergie électrostatique d'un conducteur. Nous renvoyons aux Chapitres 1 et 5 pour plus de détails sur la modélisation et sur les notations mathématiques. Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$  et  $f \in L^2(\Omega)$ . D'après la Proposition 5.2.7, pour résoudre le problème de Dirichlet pour le Laplacien, il faut minimiser l'énergie  $J(v)$  définie pour  $v \in H_0^1(\Omega)$  par

$$J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx.$$

Autrement dit, on veut résoudre

$$\inf_{v \in H_0^1(\Omega)} J(v).$$

On peut se poser le même problème pour des équations plus compliquées que le Laplacien comme les équations de Stokes. Selon l'Exercice 5.3.10, la résolution des équations de Stokes est équivalente à la minimisation

$$\inf_{v \in H_0^1(\Omega)^N \text{ tel que } \operatorname{div} v = 0} \left\{ J(v) = \frac{\mu}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f \cdot v dx \right\}.$$

Remarquons qu'il y a une contrainte "d'incompressibilité" dans cette minimisation, et que la pression est absente de l'énergie. Nous verrons que ces deux faits sont étroitement liés. •

Donnons maintenant un exemple issu du calcul des variations. Historiquement, il s'agit d'un des plus vieux problèmes d'optimisation, résolu par Zénodore environ deux siècles avant notre ère et dont la démonstration complète est due à Weierstrass vers la fin du dix-neuvième siècle.

**Exemple 9.1.11 (problème de Didon)** Virgile raconte dans l'Énéide que lorsque la reine Didon fonda la ville de Carthage, il ne lui fut alloué comme superficie que "ce que pourrait contenir une peau de boeuf". Elle découpa alors cette peau en fines lanières et encercla de cette cordelette sa future ville situé au bord de la mer. La question était donc de trouver la plus grande surface possible s'appuyant sur une droite (le rivage) et de frontière terrestre de longueur donnée. La réponse est bien sûr un demi disque (voir l'Exercice 10.2.10). En termes mathématiques un peu simplifiés, le problème est de trouver la courbe plane de longueur fixée  $l \geq 0$  qui enclot avec le segment reliant ses deux extrémités l'aire maximum. Autrement dit, on résout

$$\sup \int_0^\xi y(x) dx,$$

avec les contraintes

$$\xi \geq 0, \quad y(0) = 0, \quad \int_0^\xi \sqrt{1 + y'(x)^2} dx = l,$$

où  $\xi$  est l'extrémité du segment et  $y(x)$  la position de la courbe au dessus du point  $x$  du segment. •

Venons en à l'optimisation d'un **système distribué**, c'est-à-dire modélisé par une équation aux dérivées partielles.

**Exemple 9.1.12 (contrôle d'une membrane)** On considère une membrane élastique, fixée sur son contour, et se déformant sous l'action d'une force  $f$ . Comme nous l'avons vu à la Sous-section 1.3.3, 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  $\omega$  du domaine  $\Omega$  avec une intensité limitée. On définit donc l'ensemble des contrôles admissibles

$$K = \{v(x) \text{ tel que } v_{\min}(x) \leq v(x) \leq v_{\max}(x) \text{ dans } \omega \text{ et } v = 0 \text{ dans } \Omega \setminus \omega\},$$

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

$$J(v) = \frac{1}{2} \int_{\Omega} (|u - u_0|^2 + c|v|^2) dx,$$

avec  $c > 0$ . Le problème de contrôle s'écrit

$$\inf_{v \in K} J(v).$$

Il reste bien sûr à préciser le choix des espaces fonctionnels pour  $v$  et les autres données de ce problème (voir la Sous-section 10.4.3 pour la résolution). •

### 9.1.3 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  $\|v\|$ . Dans la Sous-section 9.1.4  $V$  sera l'espace  $\mathbb{R}^N$ , tandis que dans la section suivante  $V$  sera un espace de Hilbert réel (on pourrait également considérer le cas, plus général, d'un espace de Banach, c'est-à-dire un espace vectoriel normé complet). 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

$$\inf_{v \in K \subset V} J(v). \quad (9.2)$$

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  $\bar{v} \in K$  tel que

$$J(\bar{v}) = \inf_{v \in K \subset V} J(v).$$

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 9.1.1** *On dit que  $u$  est un minimum (ou un point de minimum) local de  $J$  sur  $K$  si et seulement si*

$$u \in K \quad \text{et} \quad \exists \delta > 0, \forall v \in K, \|v - u\| < \delta \implies J(v) \geq J(u).$$

*On dit que  $u$  est un minimum (ou un point de minimum) global de  $J$  sur  $K$  si et seulement si*

$$u \in K \quad \text{et} \quad J(v) \geq J(u) \quad \forall v \in K.$$

**Définition 9.1.2** *On appelle infimum de  $J$  sur  $K$  (ou, plus couramment, valeur minimum), que l'on désigne par la notation (9.2), 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*

$$u^n \in K \quad \forall n \quad \text{et} \quad \lim_{n \rightarrow +\infty} J(u^n) = \inf_{v \in K} J(v).$$

Par la définition même de l'infimum de  $J$  sur  $K$  il existe toujours des suites minimisantes.

### 9.1.4 Optimisation en dimension finie

Intéressons nous maintenant à la question de l'**existence de minima** pour des problèmes d'optimisation posés en dimension finie. Nous supposons dans cette sous-section (sans perte de généralité) 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 9.1.3 (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",

$$\forall (u^n)_{n \geq 0} \text{ suite dans } K, \quad \lim_{n \rightarrow +\infty} \|u^n\| = +\infty \implies \lim_{n \rightarrow +\infty} J(u^n) = +\infty. \quad (9.3)$$

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 (9.3) 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 9.1.2.  $\square$

**Remarque 9.1.4** Notons que la propriété (9.3), 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 9.1.1** 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 9.1.2** Montrer que l'on peut remplacer la propriété "infinie à l'infini" (9.3) par la condition plus faible

$$\inf_{v \in K} J(v) < \lim_{R \rightarrow +\infty} \left( \inf_{\|v\| \geq R} J(v) \right).$$

**Exercice 9.1.3** Montrer que l'on peut remplacer la continuité de  $J$  par la semi-continuité inférieure de  $J$  définie par

$$\forall (u^n)_{n \geq 0} \text{ suite dans } K, \quad \lim_{n \rightarrow +\infty} u^n = u \implies \liminf_{n \rightarrow +\infty} J(u^n) \geq J(u).$$

**Exercice 9.1.4** Montrer qu'il existe un minimum pour les Exemples 9.1.1, 9.1.6 et 9.1.7.

**Exercice 9.1.5** Soit  $a$  et  $b$  deux réels avec  $0 < a < b$ , et pour  $n \in \mathbb{N}^*$ , soit  $\mathcal{P}_n$  l'ensemble des polynômes  $P$  de degré inférieur ou égal à  $n$  tels que  $P(0) = 1$ . Pour  $P \in \mathcal{P}_n$ , on note  $\|P\| = \max_{x \in [a,b]} |P(x)|$ .

1. Montrer que le problème

$$\inf_{P \in \mathcal{P}_n} \|P\| \quad (9.4)$$

a une solution.

2. On rappelle que les polynômes de Tchebycheff  $T_n(X)$  sont définis par les relations

$$T_0(X) = 1, T_1(X) = X, T_{n+1}(X) = 2XT_n(X) - T_{n-1}(X).$$

Montrer que le degré de  $T_n$  est égal à  $n$  et que pour tout  $\theta \in \mathbb{R}$ ,  $T_n(\cos \theta) = \cos(n\theta)$ . En déduire l'existence de  $n+1$  réels  $\xi_0^n = 1 > \xi_1^n > \xi_2^n > \dots > \xi_n^n = -1$  tels que  $T_n(\xi_k^n) = (-1)^k$  pour  $0 \leq k \leq n$  et que  $\max_{-1 \leq x \leq 1} |T_n(x)| = 1$ .

3. Montrer que l'unique solution de (9.4) est le polynôme

$$P(X) = \frac{1}{T_n\left(\frac{b+a}{b-a}\right)} T_n\left(\frac{\frac{b+a}{2} - X}{\frac{b-a}{2}}\right).$$

## 9.2 Existence d'un minimum en dimension infinie

### 9.2.1 Exemples de non-existence

Cette sous-section est consacrée à deux exemples montrant que 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 9.1.3. Cette difficulté est intimement liée au fait qu'en dimension infinie les fermés bornés ne sont pas compacts!

Commençons par donner un exemple abstrait qui explique bien le mécanisme de "fuite à l'infini" qui empêche l'existence d'un minimum.

**Exemple 9.2.1** 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  $\langle x, y \rangle = \sum_{i=1}^{+\infty} x_i y_i$ . On considère la fonction  $J$  définie sur  $\ell_2(\mathbb{R})$  par

$$J(x) = (\|x\|^2 - 1)^2 + \sum_{i=1}^{+\infty} \frac{x_i^2}{i}.$$

Prenant  $K = \ell_2(\mathbb{R})$ , on considère le problème

$$\inf_{x \in \ell_2(\mathbb{R})} J(x), \quad (9.5)$$

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  $x_i^n = \delta_{in}$  pour tout  $i \geq 1$ . On vérifie aisément que

$$J(x^n) = \frac{1}{n} \rightarrow 0 \text{ quand } n \rightarrow +\infty.$$

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  $\bar{x} \in \ell_2(\mathbb{R})$  tel que  $J(\bar{x}) = 0$ . Par conséquent, il n'existe pas de point de minimum pour (9.5). On voit dans cet exemple que la suite minimisante  $x^n$  "part à l'infini" et n'est pas compacte dans  $\ell_2(\mathbb{R})$  (bien qu'elle soit bornée). •

Voici maintenant un exemple modèle qui n'est pas sans ressemblance avec les problèmes de minimisation d'énergie que nous avons rencontrés lors de la résolution d'équations aux dérivées partielles (voir, par exemple, la Proposition 5.2.7). Malgré son caractère simplifié, cet exemple est très représentatif de problèmes réalistes et pratiques de minimisation d'énergies de transition de phases en science des matériaux.

**Exemple 9.2.2** On considère l'espace de Sobolev  $V = H^1(0,1)$  muni de la norme  $\|v\| = \left( \int_0^1 (v'(x)^2 + v(x)^2) dx \right)^{1/2}$  (qui est un espace de Hilbert de dimension infinie, voir le Chapitre 4). On pose  $K = V$  et, pour  $1 \geq h > 0$ , on considère

$$J_h(v) = \int_0^1 \left( (|v'(x)| - h)^2 + v(x)^2 \right) dx.$$

L'application  $J$  est continue sur  $V$ , et la condition (9.3) est vérifiée puisque

$$J_h(v) = \|v\|^2 - 2h \int_0^1 |v'(x)| dx + h^2 \geq \|v\|^2 - \frac{1}{2} \int_0^1 v'(x)^2 dx - h^2 \geq \frac{\|v\|^2}{2} - h^2.$$

Montrons que

$$\inf_{v \in V} J_h(v) = 0, \quad (9.6)$$

ce qui impliquera qu'il n'existe pas de minimum de  $J_h$  sur  $V$  : en effet, si (9.6) a lieu et si  $u$  était un minimum de  $J_h$  sur  $V$ , on devrait avoir  $J_h(u) = 0$ , d'où  $u \equiv 0$  et  $|u'| \equiv h > 0$  (presque partout) sur  $(0,1)$ , ce qui est impossible.

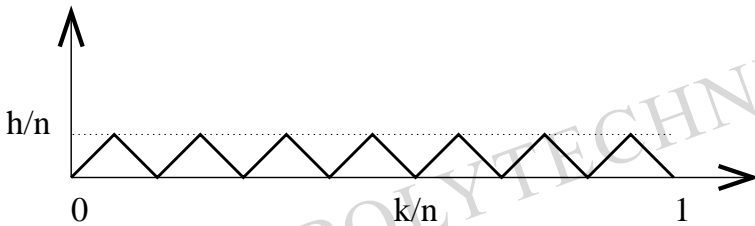


FIGURE 9.2 – Suite minimisante  $u^n$  pour l'Exemple 9.2.2.

Pour obtenir (9.6), on construit une suite minimisante  $(u^n)$  définie pour  $n \geq 1$  par

$$u^n(x) = \begin{cases} h(x - \frac{k}{n}) & \text{si } \frac{k}{n} \leq x \leq \frac{2k+1}{2n}, \\ h(\frac{k+1}{n} - x) & \text{si } \frac{2k+1}{2n} \leq x \leq \frac{k+1}{n}, \end{cases} \quad \text{pour } 0 \leq k \leq n-1,$$

comme le montre la Figure 9.2. On voit facilement que  $u^n \in V$  et que la dérivée  $(u^n)'(x)$  ne prend que deux valeurs :  $+h$  et  $-h$ . Par conséquent,  $J_h(u^n) = \int_0^1 u^n(x)^2 dx = \frac{h^2}{4n}$ , ce qui prouve (9.6), c'est-à-dire que  $J_h$  n'admet pas de point de minimum sur  $V$ . Et pourtant, si  $h = 0$ , il est clair que  $J_0$  admet un unique point de minimum  $v \equiv 0$  !

**Exercice 9.2.1** Modifier la construction précédente pour montrer qu'il n'existe pas non plus de minimum de  $J$  sur  $C^1[0, 1]$ .

A la lumière de ces contre-exemples, examinons la difficulté qui se présente en dimension infinie et sous quelles hypothèses nous pouvons espérer obtenir un résultat d'existence pour un problème de minimisation posé dans un espace de Hilbert de dimension infinie.

Soit  $V$  un espace vectoriel de norme  $\|v\|$ . Soit  $J$  une fonction définie sur une partie  $K$  de  $V$  à valeurs dans  $\mathbb{R}$ , vérifiant la condition (9.3) (infinie à l'infini). Alors, toute suite minimisante  $(u^n)$  du problème

$$\inf_{v \in K} J(v) \tag{9.7}$$

est bornée. En dimension finie (si  $V = \mathbb{R}^N$ ), on conclut aisément comme dans la Sous-section 9.1.4 en utilisant la compacité des fermés bornés (et en supposant que  $K$  est fermé et que  $J$  est continue ou semi-continue inférieurement). Malheureusement, un tel résultat est faux en dimension infinie, comme nous venons de le constater. De manière générale, on peut conclure si le triplet  $(V, K, J)$  vérifie la condition suivante : pour toute suite  $(u^n)_{n \geq 1}$  dans  $K$  telle que  $\sup_{n \in \mathbb{N}} \|u^n\| < +\infty$  on a

$$\lim_{n \rightarrow +\infty} J(u^n) = \ell < +\infty \implies \exists u \in K \text{ tel que } J(u) \leq \ell. \tag{9.8}$$

Ainsi, sous les conditions (9.3) et (9.8), le problème (9.7) admet une solution.

Malheureusement, la condition (9.8) est inutilisable car invérifiable en général ! On peut cependant la vérifier pour une classe particulière de problèmes, très importants en théorie comme en pratique : les problèmes de minimisation **convexe**. Comme nous le verrons dans la Sous-section 9.2.3, si  $V$  est un espace de Hilbert,  $K$  un **convexe** fermé de  $V$ , et que  $J$  est une fonction **convexe** et continue sur  $K$ , alors (9.8) a lieu et le problème (9.7) admet une solution. Les motivations pour introduire ces conditions sont, d'une part, que les hypothèses de convexité sont souvent naturelles dans beaucoup d'applications, et d'autre part, qu'il s'agit d'une des rares classes de problèmes pour lesquels la théorie est suffisamment générale et complète. Mais ceci ne

signifie pas que ces conditions sont les seules qui assurent l'existence d'un minimum ! Néanmoins, en dehors du cadre convexe développé dans les sous-sections suivantes, des difficultés du type de celles rencontrées dans les contre-exemples précédents peuvent survenir.

### 9.2.2 Analyse convexe

Dans tout ce qui suit, nous supposons que  $V$  est un espace de Hilbert 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 (voir la Définition 12.1.9). Donnons quelques propriétés des fonctions convexes.

**Définition 9.2.1** 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

$$J(\theta u + (1 - \theta)v) \leq \theta J(u) + (1 - \theta)J(v) \quad \forall u, v \in K, \quad \forall \theta \in [0, 1]. \quad (9.9)$$

De plus,  $J$  est dite strictement convexe si l'inégalité (9.9) est stricte lorsque  $u \neq v$  et  $\theta \in ]0, 1[$ .

**Remarque 9.2.2** Si  $J$  est une application définie sur  $K$  à valeurs dans  $\mathbb{R}$ , on appelle **épigraphe** de  $J$  l'ensemble  $\text{Epi}(J) = \{(\lambda, v) \in \mathbb{R} \times K, \lambda \geq J(v)\}$ . Alors  $J$  est convexe si et seulement si  $\text{Epi}(J)$  est une partie convexe de  $\mathbb{R} \times V$ . •

**Exercice 9.2.2** Soient  $J_1$  et  $J_2$  deux fonctions convexes sur  $V$ ,  $\lambda > 0$ , et  $\varphi$  une fonction convexe croissante sur un intervalle de  $\mathbb{R}$  contenant l'ensemble  $J_1(V)$ . Montrer que  $J_1 + J_2$ ,  $\max(J_1, J_2)$ ,  $\lambda J_1$  et  $\varphi \circ J_1$  sont convexes.

**Exercice 9.2.3** 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 9.2.3** 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 9.1.1, nous pouvons écrire

$$\exists \delta > 0, \forall w \in K, \|w - u\| < \delta \implies J(w) \geq J(u). \quad (9.10)$$

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 (9.10), 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  $u_1 = u_2$ .  $\square$

Nous nous servirons par la suite d'une notion de "forte convexité" **plus restrictive** que la stricte convexité.

**Définition 9.2.4** On dit qu'une fonction  $J$  définie sur un ensemble convexe  $K$  est *fortement convexe* si et seulement si il existe  $\alpha > 0$  tel que, pour tout  $u, v \in K$ ,

$$J\left(\frac{u+v}{2}\right) \leq \frac{J(u) + J(v)}{2} - \frac{\alpha}{8}\|u - v\|^2. \quad (9.11)$$

On dit aussi dans ce cas que  $J$  est  $\alpha$ -convexe.

Dans la Définition 9.2.4, 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, l'Exercice 9.2.7 montrera que les fonctions convexes à valeurs finies sont automatiquement continues.

**Exercice 9.2.4** Si  $J$  est continue et  $\alpha$ -convexe, montrer que, pour tout  $\theta \in [0, 1]$ ,

$$J(\theta u + (1 - \theta)v) \leq \theta J(u) + (1 - \theta)J(v) - \frac{\alpha\theta(1 - \theta)}{2}\|u - v\|^2. \quad (9.12)$$

**Exercice 9.2.5** 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 convexe et trouver la meilleure constante  $\alpha$ .

**Exercice 9.2.6** Soit  $\Omega$  un ouvert de  $\mathbb{R}^N$  et  $H^1(\Omega)$  l'espace de Sobolev associé (voir la Définition 4.3.1). Soit la fonction  $J$  définie sur  $\Omega$  par

$$J(v) = \frac{1}{2} \int_{\Omega} (|\nabla v(x)|^2 + v(x)^2) dx - \int_{\Omega} f(x)v(x) dx,$$

avec  $f \in L^2(\Omega)$ . Montrer que  $J$  est fortement convexe sur  $H^1(\Omega)$ .

Le résultat suivant sera essentiel dans l'obtention d'un résultat d'existence d'un minimum en dimension infinie. En particulier, il permet de conclure qu'une fonction  $J$  fortement convexe et continue sur un ensemble  $K$  convexe fermé non vide est "infinie à l'infini" dans  $K$ , c'est-à-dire vérifie la propriété (9.3).

**Proposition 9.2.5** *Si  $J$  est convexe continue sur un ensemble  $K$  convexe fermé non vide, alors il existe une forme linéaire continue  $L \in V'$  et une constante  $\delta \in \mathbb{R}$  telles que*

$$J(v) \geq L(v) + \delta \quad \forall v \in K. \quad (9.13)$$

*Si de plus  $J$  est fortement convexe sur  $K$ , alors il existe deux constantes  $\gamma > 0$  et  $\eta \in \mathbb{R}$  telles que*

$$J(v) \geq \gamma \|v\|^2 + \eta \quad \forall v \in K. \quad (9.14)$$

**Démonstration.** Prouvons d'abord (9.13). Si  $J$  est convexe continue (ou simplement semi-continue inférieurement) sur un ensemble  $K$  convexe fermé non vide, alors son épigraphe  $Epi(J)$  (défini dans la Remarque 9.2.2) est convexe fermé non vide. Soit  $v_0 \in K$  et  $\lambda_0 < J(v_0)$ . Puisque  $(\lambda_0, v_0) \notin Epi(J)$ , nous déduisons du Théorème 12.1.19 de séparation d'un point et d'un convexe l'existence de  $\alpha, \beta \in \mathbb{R}$  et d'une forme linéaire continue  $L \in V'$  tels que

$$\beta \lambda + L(v) > \alpha > \beta \lambda_0 + L(v_0) \quad \forall (\lambda, v) \in Epi(J). \quad (9.15)$$

Comme, pour  $v$  fixé, on peut prendre  $\lambda$  arbitrairement grand dans le membre de gauche de (9.15), il est clair que  $\beta \geq 0$ ; de plus, comme on peut prendre  $v = v_0$  dans le membre de gauche de (9.15),  $\beta$  ne peut être nul. On a donc  $\beta > 0$ . On déduit alors de (9.15), en choisissant  $\lambda = J(v)$ , que  $J(v) + L(v)/\beta > \alpha/\beta$  pour tout  $v \in K$ , ce qui prouve (9.13).

Prouvons maintenant (9.14). Soit encore  $v_0 \in K$  fixé. Pour tout  $v \in K$ , (9.11) et (9.13) impliquent que

$$\frac{J(v)}{2} + \frac{J(v_0)}{2} \geq J\left(\frac{v+v_0}{2}\right) + \frac{\alpha}{8} \|v - v_0\|^2 \geq \frac{L(v) + L(v_0)}{2} + \frac{\alpha}{8} \|v - v_0\|^2 + \delta.$$

On en déduit

$$J(v) \geq \frac{\alpha}{4} \|v\|^2 - \frac{\alpha}{2} \langle v, v_0 \rangle + L(v) + C_1,$$

avec  $C_1 = (\alpha/4)\|v_0\|^2 + L(v_0) - J(v_0) + 2\delta$ . D'après l'inégalité de Cauchy-Schwarz appliqué à  $\langle v, v_0 \rangle$  et la continuité de  $L$ , i.e.  $|L(v)| \leq \|L\|_{V'} \|v\|$  (voir la Définition 12.1.17), il vient

$$J(v) \geq \frac{\alpha}{4} \|v\|^2 - \left( \|L\|_{V'} + \frac{\alpha \|v_0\|}{2} \right) \|v\| + C_1 \geq \frac{\alpha}{8} \|v\|^2 + \eta,$$

pour  $\eta \in \mathbb{R}$  bien choisi. □

Terminons cette sous-section par une propriété agréable des fonctions convexes : les fonctions convexes "propres" (c'est-à-dire qui ne prennent pas la valeur  $+\infty$ ) sont continues.



**Exercice 9.2.7** 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.

### 9.2.3 Résultats d'existence

Nous pouvons maintenant énoncer un premier résultat d'existence de minimum dans le cas particulier où  $J$  est fortement convexe ( $\alpha$ -convexe).

**Théorème 9.2.6 (Existence d'un minimum, cas fortement convexe)** Soit  $K$  un convexe fermé non vide d'un Hilbert  $V$  et  $J$  une fonction  $\alpha$ -convexe continue sur  $K$ . Alors, il existe un unique minimum  $u$  de  $J$  sur  $K$  et on a

$$\|v - u\|^2 \leq \frac{4}{\alpha} [J(v) - J(u)] \quad \forall v \in K. \quad (9.16)$$

En particulier, toute suite minimisante de  $J$  sur l'ensemble  $K$  converge vers  $u$ .

**Démonstration.** Soit  $(u^n)$  une suite minimisante de  $J$  sur  $K$ . D'après (9.14),  $J(v) \geq \delta$  pour tout  $v \in K$ , c'est-à-dire que  $J$  est minorée sur  $K$ , donc  $\inf_{v \in K} J(v)$  est une valeur finie. Pour  $n, m \in \mathbb{N}$  la propriété (9.11) de forte convexité entraîne que

$$\begin{aligned} \frac{\alpha}{8} \|u^n - u^m\|^2 &\leq \frac{\alpha}{8} \|u^n - u^m\|^2 + J\left(\frac{u^n + u^m}{2}\right) - \inf_{v \in K} J(v) \\ &\leq \frac{1}{2} \left( J(u^n) - \inf_{v \in K} J(v) \right) + \frac{1}{2} \left( J(u^m) - \inf_{v \in K} J(v) \right), \end{aligned}$$

ce qui montre que la suite  $(u^n)$  est de Cauchy, et donc converge vers une limite  $u$ , qui est nécessairement un minimum de  $J$  sur  $K$  puisque  $J$  est continue et  $K$  fermé. L'unicité du point de minimum a été montrée dans la Proposition 9.2.3. Enfin, si  $v \in K$ ,  $(u + v)/2 \in K$  car  $K$  est convexe, d'où, toujours grâce à (9.11),

$$\frac{\alpha}{8} \|u - v\|^2 \leq \frac{J(u)}{2} + \frac{J(v)}{2} - J\left(\frac{u + v}{2}\right) \leq \frac{J(v) - J(u)}{2},$$

$$\text{car } J\left(\frac{u + v}{2}\right) \geq J(u). \quad \square$$

**Exercice 9.2.8** Montrer que le Théorème 9.2.6 s'applique à l'Exemple 9.1.10 (utiliser l'inégalité de Poincaré dans  $H_0^1(\Omega)$ ).

**Exercice 9.2.9** Généraliser l'Exercice 9.2.8 aux différents modèles rencontrés au Chapitre 5 : Laplacien avec conditions aux limites de Neumann (voir la Proposition 5.2.16), élasticité (voir l'Exercice 5.3.3), Stokes (voir l'Exercice 5.3.10).

Il est possible de généraliser en grande partie le Théorème 9.2.6 au cas de fonctions  $J$  qui sont seulement convexes (et non pas fortement convexes). Cependant, autant la démonstration du Théorème 9.2.6 est élémentaire, autant celle du théorème suivant est délicate. Elle repose en particulier sur la notion de convergence faible que l'on peut considérer comme "hors-programme" dans le cadre de ce cours.

**Théorème 9.2.7 (Existence d'un minimum, cas convexe)** *Soit  $K$  un convexe fermé non vide d'un espace de Hilbert  $V$ , et  $J$  une fonction convexe continue sur  $K$ , qui est "infinie à l'infini" dans  $K$ , c'est-à-dire qui vérifie la condition (9.3), à savoir,*

$$\forall (u^n)_{n \geq 0} \text{ suite dans } K, \quad \lim_{n \rightarrow +\infty} \|u^n\| = +\infty \implies \lim_{n \rightarrow +\infty} J(u^n) = +\infty.$$

*Alors il existe un minimum de  $J$  sur  $K$ .*

**Remarque 9.2.8** Le Théorème 9.2.7 donne l'existence d'un minimum comme le précédent Théorème 9.2.6, mais ne dit rien sur l'unicité ni sur l'estimation d'erreur (9.16). Remarquons au passage que (9.16) sera fort utile pour l'étude d'algorithmes numériques de minimisation puisqu'elle fournit une estimation de la vitesse de convergence d'une suite minimisante  $(u^n)$  vers le point de minimum  $u$ . •

**Remarque 9.2.9** Le Théorème 9.2.7 reste vrai si l'on suppose simplement que  $V$  est un espace de Banach réflexif (i.e. que le dual de  $V'$  est  $V$ ). •

Nous indiquons brièvement comment on peut démontrer le Théorème 9.2.7 dans le cas d'un espace de Hilbert séparable (c'est-à-dire qui admet une base hilbertienne dénombrable, voir la Proposition 12.1.15). On définit la notion de **convergence faible** dans  $V$ .

**Définition 9.2.10** *On dit qu'une suite  $(u^n)$  de  $V$  converge faiblement vers  $u \in V$  si*

$$\forall v \in V, \quad \lim_{n \rightarrow +\infty} \langle u^n, v \rangle = \langle u, v \rangle.$$

Soit  $(e^i)_{i \geq 1}$  une base hilbertienne de  $V$ . Si on note  $u_i^n = \langle u^n, e^i \rangle$  les composantes dans cette base d'une suite  $u^n$ , uniformément bornée dans  $V$ , il est facile de vérifier que la Définition 9.2.10 de la convergence faible est équivalente à la **convergence de toutes les suites de composantes**  $(u_i^n)_{n \geq 1}$  pour  $i \geq 1$ .

Comme son nom l'indique la convergence faible est une notion "plus faible" que la convergence usuelle dans  $V$ , puisque  $\lim_{n \rightarrow +\infty} \|u^n - u\| = 0$  implique que  $u^n$  converge faiblement vers  $u$ . Réciproquement, en dimension infinie il existe des suites qui convergent faiblement mais pas au sens usuel (que l'on appelle parfois "convergence forte" par opposition). Par exemple, la suite  $u^n = e^n$  converge faiblement vers zéro, mais pas fortement puisqu'elle est de norme constante égale à 1. L'intérêt de la convergence faible vient du résultat suivant.

**Lemme 9.2.11** *De toute suite  $u^n$  bornée dans  $V$  on peut extraire une sous-suite qui converge faiblement.*

**Démonstration.** Comme la suite  $u^n$  est bornée, chaque suite d'une composante  $u_i^n$  est bornée dans  $\mathbb{R}$ . Pour chaque  $i$ , il existe donc une sous-suite, notée  $u_i^{n'}$ , qui converge vers une limite  $u_i$ . Par un procédé d'extraction diagonale de suites, on obtient alors une sous-suite commune  $n'$  telle que, pour tout  $i$ ,  $u_i^{n'}$  converge vers  $u_i$ . Ce qui prouve que  $u^{n'}$  converge faiblement vers  $u$  (on vérifie que  $u \in V$ ).  $\square$

Si on appelle "demi-espace fermé" de  $V$  tout ensemble de la forme  $\{v \in V, L(v) \leq \alpha\}$ , où  $L$  est une forme linéaire continue non identiquement nulle sur  $V$  et  $\alpha \in \mathbb{R}$ , on peut caractériser de façon commode les ensembles convexes fermés.

**Lemme 9.2.12** *Une partie convexe fermée  $K$  de  $V$  est l'intersection des demi-espaces fermés qui contiennent  $K$ .*

**Démonstration.** Il est clair que  $K$  est inclus dans l'intersection des demi-espaces fermés qui le contiennent. Réciproquement, supposons qu'il existe un point  $u_0$  de cette intersection qui n'appartiennent pas à  $K$ . On peut alors appliquer le Théorème 12.1.19 de séparation d'un point et d'un convexe et construire ainsi un demi-espace fermé qui contient  $K$  mais pas  $u_0$ . Ceci est une contradiction avec la définition de  $u_0$ , donc  $u_0 \in K$ .  $\square$

**Lemme 9.2.13** *Soit  $K$  un ensemble convexe fermé non vide de  $V$ . Alors  $K$  est fermé pour la convergence faible.*

*De plus, si  $J$  est convexe et semi-continue inférieurement sur  $K$  (voir l'Exercice 9.1.3 pour cette notion), alors  $J$  est aussi semi-continue inférieurement sur  $K$  pour la convergence faible.*

**Démonstration.** Par définition, si  $u^n$  converge faiblement vers  $u$ , alors  $L(u^n)$  converge vers  $L(u)$ . Par conséquent, un demi-espace fermé de  $V$  est fermé pour la convergence faible. Le Lemme 9.2.12 permet d'obtenir la même conclusion pour  $K$ .

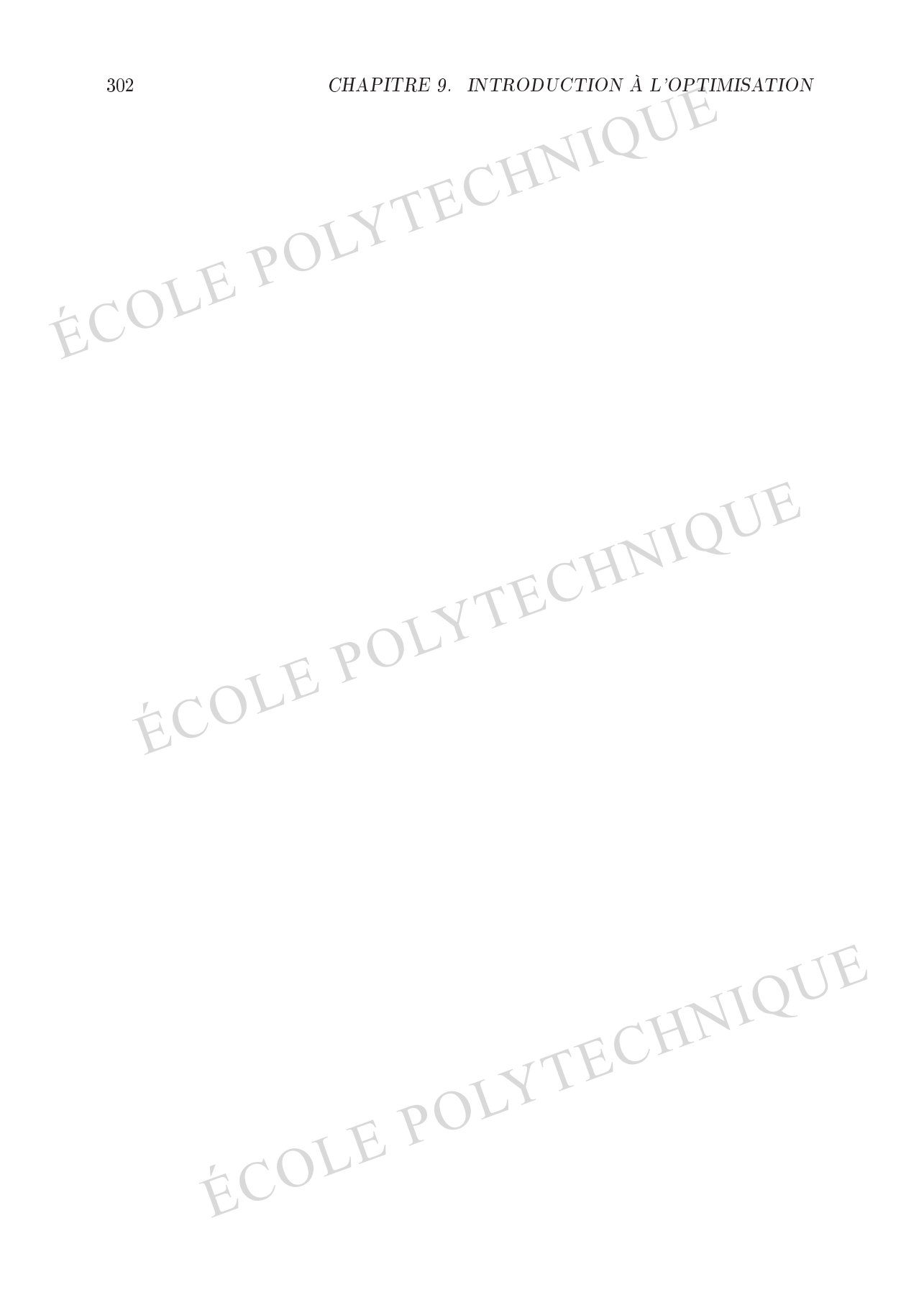
D'après les hypothèses sur  $J$ , l'ensemble  $Epi(J)$  (défini à la Remarque 9.2.2) est un convexe fermé de  $\mathbb{R} \times V$ , donc il est aussi fermé pour la convergence faible. On en déduit alors facilement le résultat : si la suite  $(v^n)$  tend faiblement vers  $v$  dans  $K$ , alors  $\liminf_{n \rightarrow +\infty} J(v^n) \geq J(v)$ .  $\square$

Nous avons maintenant tous les ingrédients pour finir.

**Démonstration du Théorème 9.2.7.** D'après (9.3), toute suite minimisante  $(u^n)$  est bornée. On déduit alors du Lemme 9.2.11 qu'il existe une sous-suite  $(u^{n'})$  convergeant faiblement vers une limite  $u \in V$ . Mais, d'après le Lemme 9.2.13,  $u \in K$  et

$$J(u) \leq \liminf_k J(u^{n_k}) = \inf_{v \in K} J(v).$$

Le point  $u$  est donc bien un minimum de  $J$  sur  $K$ .  $\square$



## Chapitre 10

# CONDITIONS D'OPTIMALITÉ ET ALGORITHMES

### 10.1 Généralités

#### 10.1.1 Introduction

Dans le Chapitre 9 nous nous sommes intéressés aux questions d'existence de minimum aux problèmes d'optimisation. Dans ce chapitre, nous allons chercher à obtenir des conditions nécessaires et parfois suffisantes de minimalité. L'objectif est d'une certaine manière beaucoup plus pratique, puisque ces conditions d'optimalité seront le plus souvent utilisées pour tenter de **calculer un minimum** (parfois même sans avoir su démontrer son existence!). L'idée générale des conditions d'optimalité est la même que celle qui, lorsque l'on calcule l'extremum d'une fonction sur  $\mathbb{R}$ , 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.

Ces conditions d'optimalité généralisent la remarque élémentaire suivante : si  $x_0$  est un point de minimum local de  $J$  sur l'intervalle  $[a, b] \subset \mathbb{R}$  ( $J$  étant une fonction dérivable sur  $[a, b]$ ), alors on a

$$J'(x_0) \geq 0 \text{ si } x_0 = a, \quad J'(x_0) = 0 \text{ si } x_0 \in ]a, b[, \quad J'(x_0) \leq 0 \text{ si } x_0 = b.$$

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  $J(x_0) + \frac{h^2}{2}J''(x_0) + o(h^2) \geq J(x_0)$  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 !

Le plan de ce chapitre est le suivant. Le reste de cette section est consacrée à préciser quelques notations et à rappeler des notions élémentaires de dérivabilité. La Section 10.2 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 **multiplicateurs de Lagrange**. La Section 10.3 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**. La Section 10.4 explore trois applications de l'optimisation à des systèmes modélisés par des équations différentielles ordinaires ou aux dérivées partielles. Finalement, la Section 10.5 traite des **algorithmes numériques d'optimisation**. On étudie principalement les algorithmes de **gradient** qui sont les plus importants en pratique.

### 10.1.2 Différentiabilité

Désormais (et nous ne le rappellerons plus systématiquement), nous supposons que  $V$  est un espace de Hilbert réel, et que le critère  $J$  est une fonction continue à valeurs dans  $\mathbb{R}$ . Le produit scalaire dans  $V$  est toujours noté  $\langle u, v \rangle$  et la norme associée  $\|u\|$ .

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 10.1.1** *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  $V$ ,  $L \in V'$ , telle que*

$$J(u + w) = J(u) + L(w) + o(w) \quad , \quad \text{avec} \quad \lim_{w \rightarrow 0} \frac{|o(w)|}{\|w\|} = 0 . \quad (10.1)$$

On appelle  $L$  la dérivée (ou la différentielle, ou le gradient) de  $J$  en  $u$  et on note  $L = J'(u)$ .

**Remarque 10.1.2** La Définition 10.1.1 est en fait valable si  $V$  est seulement un espace de Banach (on n'utilise pas de produit scalaire dans (10.1)). Cependant, si  $V$  est un espace de Hilbert, on peut préciser la relation (10.1) en identifiant  $V$  et son dual  $V'$  grâce au Théorème de représentation de Riesz 12.1.18. En effet, il existe un unique  $p \in V$  tel que  $\langle p, w \rangle = L(w)$ , donc (10.1) devient

$$J(u + w) = J(u) + \langle p, w \rangle + o(w) \quad , \quad \text{avec} \quad \lim_{w \rightarrow 0} \frac{|o(w)|}{\|w\|} = 0 . \quad (10.2)$$

On note aussi parfois  $p = J'(u)$ , ce qui peut prêter à confusion... La formule (10.2) est souvent plus “naturelle” que (10.1), notamment si  $V = \mathbb{R}^n$  ou  $V = L^2(\Omega)$ . Par contre, elle peut être d'interprétation un peu plus délicate pour des espaces de Hilbert plus “compliqués” comme  $V = H^1(\Omega)$  (voir l'Exercice 10.1.5). •

Dans la plupart des applications, il suffit souvent de déterminer la forme linéaire continue  $L = J'(u) \in V'$  car on n'a pas besoin de l'expression explicite de  $p = J'(u) \in V$  lorsque  $V'$  est identifié à  $V$ . En pratique, il est plus facile de trouver l'expression explicite de  $L$  que celle de  $p$ , comme le montrent les exercices suivants.

**Exercice 10.1.1** Montrer que (10.1) implique la continuité de  $J$  en  $u$ . Montrer aussi que, si deux formes linéaires continues  $L_1, L_2$  vérifient

$$\begin{cases} J(u + w) \geq J(u) + L_1(w) + o(w) , \\ J(u + w) \leq J(u) + L_2(w) + o(w) , \end{cases} \quad (10.3)$$

alors  $J$  est dérivable et  $L_1 = L_2 = J'(u)$ .

**Exercice 10.1.2 (essentiel !)** Soit  $a$  une forme bilinéaire symétrique continue sur  $V \times V$ . 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  $\langle J'(u), w \rangle = a(u, w) - L(w)$  pour tout  $u, w \in V$ .

**Exercice 10.1.3** 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 10.1.4** On reprend l'Exercice 10.1.2 avec  $V = L^2(\Omega)$  ( $\Omega$  étant un ouvert de  $\mathbb{R}^N$ ),  $a(u, v) = \int_{\Omega} uv \, dx$ , et  $L(u) = \int_{\Omega} fu \, dx$  avec  $f \in L^2(\Omega)$ . En identifiant  $V$  et  $V'$ , montrer que  $J'(u) = u - f$ .

**Exercice 10.1.5** On reprend l'Exercice 10.1.2 avec  $V = H_0^1(\Omega)$  ( $\Omega$  étant un ouvert de  $\mathbb{R}^N$ ) que l'on munit du produit scalaire

$$\langle u, v \rangle = \int_{\Omega} (\nabla u \cdot \nabla v + uv) \, dx.$$

On pose  $a(u, v) = \int_{\Omega} \nabla u \cdot \nabla v \, dx$ , et  $L(u) = \int_{\Omega} f u \, dx$  avec  $f \in L^2(\Omega)$ . Montrer (au moins formellement) que  $J'(u) = -\Delta u - f$  dans  $V' = H^{-1}(\Omega)$ . Montrer que, si on identifie  $V$  et  $V'$ , alors  $J'(u) = u_0$  où  $u_0$  est l'unique solution dans  $H_0^1(\Omega)$  de

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

**Exercice 10.1.6** Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$  (on pourra se restreindre au cas où  $N = 1$  avec  $\Omega = ]0, 1[$ ). Soit  $L = L(p, t, x)$  une fonction continue sur  $\mathbb{R}^N \times \mathbb{R} \times \overline{\Omega}$ , dérivable par rapport à  $p$  et  $t$  sur cet ensemble, de dérivées partielles  $\frac{\partial L}{\partial p}$  et  $\frac{\partial L}{\partial t}$  Lipschitziennes sur cet ensemble. On pose  $V = H_0^1(\Omega)$  et  $J(v) = \int_{\Omega} L(\nabla v(x), v(x), x) \, dx$ .

1. Montrer que  $J$  est dérivable sur  $H_0^1(\Omega)$  et que

$$\langle J'(u), w \rangle = \int_{\Omega} \left( \frac{\partial L}{\partial p}(\nabla u(x), u(x), x) \cdot \nabla w(x) + \frac{\partial L}{\partial t}(\nabla u(x), u(x), x) w(x) \right) dx.$$

2. Si  $N = 1$  et  $\Omega = ]0, 1[$ , montrer que, si  $u \in H_0^1(0, 1)$  satisfait  $J'(u) = 0$ , alors  $u$  vérifie

$$\frac{d}{dx} \left( \frac{\partial L}{\partial p}(u'(x), u(x), x) \right) - \frac{\partial L}{\partial t}(u'(x), u(x), x) = 0, \quad (10.4)$$

presque partout dans l'intervalle  $]0, 1[$ .

3. Si  $L$  ne dépend pas de  $x$  (i.e.  $L = L(p, t)$ ) et si  $u \in C^2(]0, 1[)$  est une solution de l'équation différentielle (10.4), montrer que la quantité

$$L(u'(x), u(x)) - u'(x) \frac{\partial L}{\partial p}(u'(x), u(x))$$

est constante sur l'intervalle  $[0, 1]$ .

**Remarque 10.1.3** 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  $L \in V'$  tel que

$$\forall w \in V, \quad \lim_{\delta \rightarrow 0, \delta > 0} \frac{J(u + \delta w) - J(u)}{\delta} = L(w). \quad (10.5)$$

On parle aussi de différentiabilité directionnelle et  $w$  est la direction de dérivation dans (10.5). L'intérêt de cette notion est que la vérification de (10.5) est plus aisée que celle de (10.1). Cependant, si une fonction dérivable au sens de Fréchet l'est aussi au sens de Gâteaux, la réciproque est fautive, même en dimension finie, comme le montre l'exemple suivant dans  $\mathbb{R}^2$

$$J(x, y) = \frac{x^6}{(y - x^2)^2 + x^8} \quad \text{pour } (x, y) \neq (0, 0) \quad , \quad J(0, 0) = 0.$$



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 10.1.4** *Soit  $J$  une application différentiable de  $V$  dans  $\mathbb{R}$ . Les assertions suivantes sont équivalentes*

$$J \text{ est convexe sur } V, \quad (10.6)$$

$$J(v) \geq J(u) + \langle J'(u), v - u \rangle \quad \forall u, v \in V, \quad (10.7)$$

$$\langle J'(u) - J'(v), u - v \rangle \geq 0 \quad \forall u, v \in V. \quad (10.8)$$

**Proposition 10.1.5** *Soit  $J$  une application différentiable de  $V$  dans  $\mathbb{R}$  et  $\alpha > 0$ . Les assertions suivantes sont équivalentes*

$$J \text{ est } \alpha\text{-convexe sur } V, \quad (10.9)$$

$$J(v) \geq J(u) + \langle J'(u), v - u \rangle + \frac{\alpha}{2} \|v - u\|^2 \quad \forall u, v \in V, \quad (10.10)$$

$$\langle J'(u) - J'(v), u - v \rangle \geq \alpha \|u - v\|^2 \quad \forall u, v \in V. \quad (10.11)$$

**Remarque 10.1.6** Les conditions (10.10) et (10.7) ont une interprétation géométrique simple. Elles signifient que la fonction convexe  $J(v)$  est toujours au dessus de son plan tangent en  $u$  (considéré comme une fonction affine de  $v$ ). Les conditions (10.11) et (10.8) sont des hypothèses de monotonie ou de croissance de  $J'$ . Par ailleurs, si  $J(u) = \frac{1}{2}a(u, u) - L(u)$  avec  $a$  une forme bilinéaire symétrique continue sur  $V$  et  $L$  une forme linéaire continue sur  $V$ , alors l'Exercice 10.1.2 montre que (10.11) est exactement la définition de la coercivité de  $a$ . •

**Démonstration.** Il suffit de démontrer la Proposition 10.1.5 en observant que le cas  $\alpha = 0$  donne la Proposition 10.1.4. Montrons que (10.9) implique (10.10). Comme  $J$  est  $\alpha$ -convexe, on vérifie aisément (par récurrence) que, pour tout  $k \geq 1$ ,

$$J\left(\left(1 - \frac{1}{2^k}\right)u + \frac{1}{2^k}v\right) \leq \left(1 - \frac{1}{2^k}\right)J(u) + \frac{1}{2^k}J(v) - \frac{\alpha}{2^{k+1}}\left(1 - \frac{1}{2^k}\right)\|u - v\|^2,$$

d'où

$$2^k \left[ J\left(u + \frac{1}{2^k}(v - u)\right) - J(u) \right] \leq J(v) - J(u) - \frac{\alpha}{2}\left(1 - \frac{1}{2^k}\right)\|u - v\|^2.$$

En faisant tendre  $k$  vers  $+\infty$ , on trouve (10.10). Pour obtenir (10.11) il suffit d'additionner (10.10) avec lui-même en échangeant  $u$  et  $v$ .

Montrons que (10.11) implique (10.9). Pour  $u, v \in V$  et  $t \in \mathbb{R}$ , on pose  $\varphi(t) = J(u + t(v-u))$ . Alors  $\varphi$  est dérivable et donc continue sur  $\mathbb{R}$ , et  $\varphi'(t) = \langle J'(u + t(v-u)), v-u \rangle$ , de sorte que, d'après (10.11)

$$\varphi'(t) - \varphi'(s) \geq \alpha(t-s)\|v-u\|^2 \quad \text{si } t \geq s. \quad (10.12)$$

Soit  $\theta \in ]0, 1[$ . En intégrant l'inégalité (10.12) de  $t = \theta$  à  $t = 1$  et de  $s = 0$  à  $s = \theta$ , on obtient

$$\theta\varphi(1) + (1-\theta)\varphi(0) - \varphi(\theta) \geq \frac{\alpha\theta(1-\theta)}{2}\|v-u\|^2,$$

c'est-à-dire (10.9).  $\square$

**Exercice 10.1.7** Montrer qu'une fonction  $J$  dérivable sur  $V$  est strictement convexe si et seulement si

$$J(v) > J(u) + \langle J'(u), v-u \rangle \quad \forall u, v \in V \quad \text{avec } u \neq v,$$

ou encore

$$\langle J'(u) - J'(v), u-v \rangle > 0 \quad \forall u, v \in V \quad \text{avec } u \neq v.$$

Terminons cette sous-section en définissant la **dérivée seconde** de  $J$ . Remarquons tout d'abord qu'il est très facile de généraliser la Définition 10.1.1 de différentiabilité au cas d'une fonction  $f$  définie sur  $V$  à valeurs dans un autre espace de Hilbert  $W$  (et non pas seulement dans  $\mathbb{R}$ ). On dira que  $f$  est différentiable (au sens de Fréchet) en  $u$  s'il existe une application linéaire continue  $L$  de  $V$  dans  $W$  telle que

$$f(u+w) = f(u) + L(w) + o(w) \quad , \quad \text{avec} \quad \lim_{w \rightarrow 0} \frac{\|o(w)\|_W}{\|w\|_V} = 0. \quad (10.13)$$

On appelle  $L = f'(u)$  la différentielle de  $f$  en  $u$ . La définition (10.13) est utile pour définir la dérivée de  $f(u) = J'(u)$  qui est une application de  $V$  dans son dual  $V'$ .

**Définition 10.1.7** Soit  $J$  une fonction de  $V$  dans  $\mathbb{R}$ . On dit que  $J$  est deux fois dérivable en  $u \in V$  si  $J$  est dérivable dans un voisinage de  $u$  et si sa dérivée  $J'(u)$  est dérivable en  $u$ . On note  $J''(u)$  la dérivée seconde de  $J$  en  $u$  qui vérifie

$$J'(u+w) = J'(u) + J''(u)w + o(w) \quad , \quad \text{avec} \quad \lim_{w \rightarrow 0} \frac{\|o(w)\|_{V'}}{\|w\|_V} = 0.$$

Telle qu'elle est définie la dérivée seconde est difficile à évaluer en pratique car  $J''(u)w$  est un élément de  $V'$ . Heureusement, en la faisant agir sur  $v \in V$  on obtient une brave forme bilinéaire continue sur  $V \times V$  que l'on notera  $J''(u)(w, v)$  en lieu et place de  $(J''(u)w)v$ . Nous laissons au lecteur le soin de prouver le résultat élémentaire suivant.

**Lemme 10.1.8** Si  $J$  est une fonction deux fois dérivable de  $V$  dans  $\mathbb{R}$ , elle vérifie

$$J(u+w) = J(u) + J'(u)w + \frac{1}{2}J''(u)(w, w) + o(\|w\|^2), \text{ avec } \lim_{w \rightarrow 0} \frac{o(\|w\|^2)}{\|w\|^2} = 0, \quad (10.14)$$

où  $J''(u)$  est identifiée à une forme bilinéaire continue sur  $V \times V$ .

En pratique c'est  $J''(u)(w, w)$  que l'on calcule.

**Exercice 10.1.8** Soit  $a$  une forme bilinéaire symétrique continue sur  $V \times V$ . Soit  $L$  une forme linéaire continue sur  $V$ . On pose  $J(u) = \frac{1}{2}a(u, u) - L(u)$ . Montrer que  $J$  est deux fois dérivable sur  $V$  et que  $J''(u)(v, w) = a(v, w)$  pour tout  $u, v, w \in V$ . Appliquer ce résultat aux exemples des Exercices 10.1.3, 10.1.4, 10.1.5.

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 10.1.9** Montrer que si  $J$  est deux fois dérivable sur  $V$  les conditions des Propositions 10.1.4 et 10.1.5 sont respectivement équivalentes à

$$J''(u)(w, w) \geq 0 \quad \text{et} \quad J''(u)(w, w) \geq \alpha \|w\|^2 \quad \forall u, w \in V. \quad (10.15)$$

## 10.2 Conditions d'optimalité

### 10.2.1 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 supposons 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 10.2.1 (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

$$\langle J'(u), v-u \rangle \geq 0 \quad \forall v \in K. \quad (10.16)$$

Si  $u \in K$  vérifie (10.16) et si  $J$  est convexe, alors  $u$  est un minimum global de  $J$  sur  $K$ .

**Remarque 10.2.2** On appelle (10.16), "inéquation d'Euler". Il s'agit d'une condition **nécessaire** d'optimalité qui devient **nécessaire et suffisante** si  $J$  est convexe. La condition (10.16) 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  $K$ . Il faut aussi remarquer

que, dans deux cas importants, (10.16) **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 (10.16) 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

$$\frac{J(u + h(v - u)) - J(u)}{h} \geq 0. \quad (10.17)$$

On en déduit (10.16) en faisant tendre  $h$  vers 0. Le caractère suffisant de (10.16) pour une fonction convexe découle immédiatement de la propriété de convexité (10.7). □

**Exercice 10.2.1** Soit  $K$  un convexe fermé non vide de  $V$ . Pour  $x \in V$ , on cherche la projection  $x_K \in K$  de  $x$  sur  $K$  (voir le Théorème 12.1.10)

$$\|x - x_K\| = \min_{y \in K} \|x - y\|.$$

Montrer que la condition nécessaire et suffisante (10.16) se ramène exactement à (12.1).

**Exercice 10.2.2** Soit  $A$  une matrice réelle d'ordre  $p \times n$  et  $b \in \mathbb{R}^p$ . On considère le problème "aux moindres carrés"

$$\inf_{x \in \mathbb{R}^n} \|Ax - b\|^2.$$

Montrer que ce problème admet toujours une solution et écrire l'équation d'Euler correspondante.

**Exercice 10.2.3** On reprend l'Exemple 9.1.6

$$\inf_{x \in \text{Ker} B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\}$$

avec  $A$  matrice symétrique carrée d'ordre  $n$ , et  $B$  de taille  $m \times n$  ( $m \leq n$ ). Montrer qu'il existe une solution si  $A$  est positive et  $b \in (\text{Ker} A \cap \text{Ker} B)^\perp$ , et qu'elle est unique si  $A$  est définie positive. Montrer que tout point de minimum  $\bar{x} \in \mathbb{R}^n$  vérifie

$$A\bar{x} - b = B^*p \quad \text{avec } p \in \mathbb{R}^m.$$

**Exercice 10.2.4** On reprend l'Exemple 9.1.10. Montrer que l'équation d'Euler vérifiée par le point de minimum  $u \in H_0^1(\Omega)$  de

$$\inf_{v \in H_0^1(\Omega)} \left\{ J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right\}$$

est précisément la formulation variationnelle

$$\int_{\Omega} \nabla u \cdot \nabla v dx = \int_{\Omega} f v dx \quad \forall v \in H_0^1(\Omega).$$

(On retrouve ainsi un résultat de la Proposition 5.2.7.)

**Exercice 10.2.5** 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  $V$ . Montrer que  $J(v) = \frac{1}{2}a(v, v) - L(v)$  admet un unique point de minimum dans  $K$ , noté  $u$ . Montrer que  $u$  est aussi l'unique solution du problème (appelé inéquation variationnelle)

$$u \in K \quad \text{et} \quad a(u, v - u) \geq L(v - u) \quad \forall v \in K .$$

**Exercice 10.2.6** 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

$$\langle J'_1(u), v - u \rangle + J_2(v) - J_2(u) \geq 0 \quad \forall v \in K .$$

Les remarques suivantes, qui sont des applications simples du Théorème 10.2.1, vont nous donner l'intuition de la notion de "multiplicateur de Lagrange" qui sera développée à la Sous-section suivante.

**Remarque 10.2.3** Examinons le cas particulier où  $K$  est un sous-espace affine fermé de  $V$ . On suppose donc que  $K = u_0 + \mathcal{P}$ , où  $u_0 \in V$  et où  $\mathcal{P}$  est un sous-espace vectoriel fermé de  $V$ . Alors, lorsque  $v$  décrit  $K$ ,  $v - u$  est un élément quelconque de  $\mathcal{P}$  si bien que (10.16) équivaut à

$$\langle J'(u), w \rangle = 0 \quad \forall w \in \mathcal{P} ,$$

c'est-à-dire  $J'(u) \in \mathcal{P}^\perp$ . En particulier, si  $\mathcal{P}$  est une intersection finie d'hyperplans, c'est-à-dire si

$$\mathcal{P} = \{v \in V \quad , \quad \langle a_i, v \rangle = 0 \quad \text{pour} \quad 1 \leq i \leq M\} ,$$

où  $a_1, \dots, a_m$  sont donnés dans  $V$ , alors (le lecteur vérifiera que)  $\mathcal{P}^\perp$  est l'espace vectoriel engendré par la famille  $(a_i)_{1 \leq i \leq M}$ . La condition d'optimalité s'écrit donc sous la forme :

$$u \in K \quad \text{et} \quad \exists \lambda_1, \dots, \lambda_M \in \mathbb{R} \quad , \quad J'(u) + \sum_{i=1}^M \lambda_i a_i = 0 , \quad (10.18)$$

et les réels  $\lambda_i$  sont appelés **multiplicateurs de Lagrange**. Nous verrons au Théorème 10.2.8 leur rôle plus général et fondamental. •

**Remarque 10.2.4** Supposons maintenant que  $K$  est un cône convexe fermé, ce qui signifie que  $K$  est un ensemble convexe fermé tel que  $\lambda v \in K$  pour tout  $v \in K$  et tout  $\lambda \geq 0$ . En prenant  $v = 0$  puis  $v = 2u$  dans (10.16), on obtient que

$$\langle J'(u), u \rangle = 0 . \quad (10.19)$$

Par conséquent, (10.16) implique que

$$\langle J'(u), w \rangle \geq 0 \quad \forall w \in K . \quad (10.20)$$

En fait, (10.16) est équivalent à (10.19) et (10.20). Dans le cas où

$$K = \{v \in V \text{ , } \langle a_i, v \rangle \leq 0 \text{ pour } 1 \leq i \leq M\} \text{ ,}$$

où  $a_1, \dots, a_M$  sont donnés dans  $V$ , le Lemme de Farkas 10.2.17 (voir ci-dessous) montre alors que

$$u \in K \text{ et } \exists \lambda_1, \dots, \lambda_M \geq 0 \text{ , } J'(u) + \sum_{i=1}^M \lambda_i a_i = 0 \text{ ,} \quad (10.21)$$

et l'égalité (10.19) montre que  $\lambda_i = 0$  si  $\langle a_i, u \rangle < 0$ . Les réels positifs ou nuls  $\lambda_i$  sont encore appelés **multiplicateurs de Lagrange**. Nous verrons au Théorème 10.2.15 leur rôle plus général et fondamental. •

Terminons cette sous-section en donnant une **condition d'optimalité du deuxième ordre**.

**Proposition 10.2.5** *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*

$$J'(u) = 0 \text{ et } J''(u)(w, w) \geq 0 \quad \forall w \in V \text{ .} \quad (10.22)$$

Réciproquement, si, pour tout  $v$  dans un voisinage de  $u$ ,

$$J'(u) = 0 \text{ et } J''(v)(w, w) \geq 0 \quad \forall w \in V \text{ ,} \quad (10.23)$$

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 (10.14) nous donne (10.22). Réciproquement, si  $u$  vérifie (10.23), 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 9.1.1). □

## 10.2.2 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 10.2.6** *En tout point  $v \in K$ , l'ensemble*

$$K(v) = \left\{ w \in V \text{ , } \exists (v^n) \in K^{\mathbb{N}} \text{ , } \exists (\varepsilon^n) \in (\mathbb{R}_+^*)^{\mathbb{N}} \text{ , } \lim_{n \rightarrow +\infty} v^n = v \text{ , } \lim_{n \rightarrow +\infty} \varepsilon^n = 0 \text{ , } \lim_{n \rightarrow +\infty} \frac{v^n - v}{\varepsilon^n} = w \right\}$$

*est appelé le cône des directions admissibles au point  $v$ .*

En termes plus imagés, on peut dire aussi que  $K(v)$  est l'ensemble de tous les vecteurs qui sont tangents en  $v$  à une courbe contenue dans  $K$  et passant par  $v$  (si  $K$  est une variété régulière,  $K(v)$  est simplement 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  $K$ .

En posant  $w^n = (v^n - v)/\varepsilon^n$ , on peut aussi dire de façon équivalente que  $w \in K(v)$  si et seulement si il existe une suite  $w^n$  dans  $V$  et une suite  $\varepsilon^n$  dans  $\mathbb{R}_+^*$  telles que

$$\lim_{n \rightarrow +\infty} w^n = w, \quad \lim_{n \rightarrow +\infty} \varepsilon^n = 0 \text{ et } v + \varepsilon^n w^n \in K \quad \forall n.$$

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 10.2.7** 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. La démonstration, très simple, est laissée au lecteur.

**Proposition 10.2.7 (Inéquation d'Euler, cas général)** Soit  $u$  un minimum local de  $J$  sur  $K$ . Si  $J$  est différentiable en  $u$ , on a

$$\langle J'(u), w \rangle \geq 0 \quad \forall w \in K(u).$$

Nous allons maintenant préciser la condition nécessaire de la Proposition 10.2.7 dans le cas où  $K$  est donné par des **contraintes d'égalité** ou **d'inégalité**. Les résultats que nous obtiendrons vont généraliser ceux des Remarques 10.2.3 et 10.2.4.

### Contraintes d'égalité

Dans ce premier cas on suppose que  $K$  est donné par

$$K = \{v \in V, \quad F(v) = 0\}, \quad (10.24)$$

où  $F(v) = (F_1(v), \dots, F_M(v))$  est une application de  $V$  dans  $\mathbb{R}^M$ , avec  $M \geq 1$ . La condition **nécessaire** d'optimalité prend alors la forme suivante.

**Théorème 10.2.8** Soit  $u \in K$  où  $K$  est donné par (10.24). 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  $(F'_i(u))_{1 \leq i \leq M}$  sont linéairement indépendants. Alors, si  $u$  est un minimum local de  $J$  sur  $K$ , il existe  $\lambda_1, \dots, \lambda_M \in \mathbb{R}$ , appelés **multiplicateurs de Lagrange**, tels que

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0. \quad (10.25)$$

**Démonstration.** Comme les vecteurs  $(F'_i(u))_{1 \leq i \leq M}$  sont linéairement indépendants, le théorème des fonctions implicites permet de montrer que

$$K(u) = \{w \in V, \langle F'_i(u), w \rangle = 0 \text{ pour } i = 1, \dots, M\}, \quad (10.26)$$

ou de façon équivalente

$$K(u) = \bigcap_{i=1}^M [F'_i(u)]^\perp. \quad (10.27)$$

Nous ne détaillons pas la preuve de (10.26) qui est un résultat classique de calcul différentiel (voir par exemple [5]) et qui requiert que chaque fonction  $F_i$  soit continûment différentiable (c'est-à-dire que la fonction  $u \rightarrow F'_i(u)$  soit continue). En fait,  $K(u)$  est l'espace tangent à la variété  $K$  au point  $u$ . Comme  $K(u)$  est un espace vectoriel, on peut prendre successivement  $w$  et  $-w$  dans la Proposition 10.2.7, 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 que  $J'(u)$  est engendré par les  $(F'_i(u))_{1 \leq i \leq M}$  (notons que les multiplicateurs de Lagrange sont définis de manière unique). Une autre démonstration (plus géométrique) est proposé dans la preuve de la Proposition 10.2.11.  $\square$

**Remarque 10.2.9** Lorsque les vecteurs  $(F'_i(u))_{1 \leq i \leq M}$  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 10.2.8 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  $J'(u) = 1$ , (10.25) n'a pas lieu.  $\bullet$

Pour bien comprendre la portée du Théorème 10.2.8, nous l'appliquons sur l'Exemple 9.1.6

$$\min_{x \in \text{Ker } B} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\},$$

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 (10.25). Il existe donc un multiplicateur de Lagrange  $p \in \mathbb{R}^m$  tel que un point de minimum  $\bar{x}$  vérifie

$$A\bar{x} - b = \sum_{i=1}^m p_i b_i = B^* p.$$



Comme  $A$  est inversible, on en déduit la valeur  $\bar{x} = A^{-1}(b + B^*p)$ . Par ailleurs  $B\bar{x} = 0$  et, comme  $B$  est de rang maximal, la matrice  $BA^{-1}B^*$  est inversible, ce qui conduit à

$$p = -(BA^{-1}B^*)^{-1}BA^{-1}b \quad \text{et} \quad \bar{x} = A^{-1} \left( \text{Id} - B^*(BA^{-1}B^*)^{-1}BA^{-1} \right) b.$$

Notons que le multiplicateur de Lagrange  $p$  est unique. Si  $B$  n'est pas de rang  $m$ , l'Exercice 10.2.3 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 10.2.8** Généraliser les résultats ci-dessus pour cette variante de l'Exemple 9.1.6

$$\min_{Bx=c} \left\{ J(x) = \frac{1}{2}Ax \cdot x - b \cdot x \right\},$$

où  $c \in \mathbb{R}^m$  est un vecteur donné.

**Exercice 10.2.9** Appliquer le Théorème 10.2.8 à l'Exemple 9.1.7 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 10.2.10** En utilisant les résultats précédents et ceux de l'Exercice 10.1.6, montrer que la solution du problème de Didon (Exemple 9.1.11) est nécessairement un arc de cercle.

**Exercice 10.2.11** On étudie la première valeur propre du Laplacien dans un domaine borné  $\Omega$  (voir la Section 7.3). Pour cela on introduit le problème de minimisation sur  $K = \{v \in H_0^1(\Omega), \int_{\Omega} v^2 dx = 1\}$

$$\min_{v \in K} \left\{ J(v) = \int_{\Omega} |\nabla v|^2 dx \right\}.$$

Montrer que ce problème admet un minimum (on montrera que  $K$  est compact pour les suites minimisantes à l'aide du Théorème de Rellich 4.3.21). Écrire l'équation d'Euler de ce problème et en déduire que la valeur du minimum est bien la première valeur propre et que les points de minimum sont des vecteurs propres associés.

**Exercice 10.2.12** 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 10.2.8 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  $x \in \mathbb{R}^n$ . Dédurre de la question précédente que, si  $A \geq B$ , alors  $B^{-1} \geq A^{-1}$ .

**Exercice 10.2.13** 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  $\rho$ , sa vitesse  $u$ , et sa température  $T$ , se retrouvent grâce aux moments de la fonction  $f(v)$

$$\rho = \int_{\mathbb{R}^N} f(v) dv, \quad \rho u = \int_{\mathbb{R}^N} v f(v) dv, \quad \frac{1}{2} \rho u^2 + \frac{N}{2} \rho T = \frac{1}{2} \int_{\mathbb{R}^N} |v|^2 f(v) dv. \quad (10.28)$$

Boltzmann a introduit l'entropie cinétique  $H(f)$  définie par

$$H(f) = \int_{\mathbb{R}^N} f(v) \log(f(v)) dv.$$

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 (10.28), et on admettra qu'il existe un unique point de minimum  $M(v)$ . Montrer que ce point de minimum est une Maxwellienne définie par

$$M(v) = \frac{\rho}{(2\pi T)^{N/2}} \exp\left(-\frac{|v-u|^2}{2T}\right).$$

**Remarque 10.2.10** Pour obtenir un nouvel éclairage sur le Théorème 10.2.8 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 10.2.8 nous dit alors que, dans le cas régulier, il existe  $\lambda \in \mathbb{R}^M$  tel que

$$\frac{\partial \mathcal{L}}{\partial v}(u, \lambda) = 0, \quad \frac{\partial \mathcal{L}}{\partial \mu}(u, \lambda) = 0,$$

puisque  $\frac{\partial \mathcal{L}}{\partial \mu}(u, \lambda) = F(u) = 0$  si  $u \in K$  et  $\frac{\partial \mathcal{L}}{\partial v}(u, \lambda) = J'(u) + \lambda F'(u) = 0$  d'après (10.25). 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

$$\inf_{v \in V, F(v)=0} J(v) = \inf_{v \in V} \sup_{\mu \in \mathbb{R}^M} \mathcal{L}(v, \mu).$$

•

Nous donnons maintenant une condition **nécessaire** d'optimalité du deuxième ordre.

**Proposition 10.2.11** *On se place sous les hypothèses du Théorème 10.2.8 et on suppose que les fonctions  $J$  et  $F_1, \dots, F_M$  sont deux fois continûment dérivables et que les vecteurs  $(F'_i(u))_{1 \leq i \leq M}$  sont linéairement indépendants. Soit  $\lambda \in \mathbb{R}^M$  le multiplicateur de Lagrange défini par le Théorème 10.2.8. Alors tout minimum local  $u$  de  $J$  sur  $K$  vérifie*

$$\left( J''(u) + \sum_{i=1}^M \lambda_i F''_i(u) \right) (w, w) \geq 0 \quad \forall w \in K(u) = \bigcap_{i=1}^M [F'_i(u)]^\perp. \quad (10.29)$$

**Démonstration.** Supposons qu'il existe un chemin admissible de classe  $C^2$ , c'est-à-dire une fonction  $t \rightarrow u(t)$  de  $[0, 1]$  dans  $V$  telle que  $u(0) = u$  et  $F(u(t)) = 0$  pour tout  $t \in [0, 1]$ . Par définition, la dérivée  $u'(0)$  appartient au cône des directions admissibles  $K(u)$ . On pose

$$j(t) = J(u(t)) \quad \text{et} \quad f_i(t) = F_i(u(t)) \quad \text{pour } 1 \leq i \leq M.$$

En dérivant on obtient

$$j'(t) = \langle J'(u(t)), u'(t) \rangle \quad \text{et} \quad f'_i(t) = \langle F'_i(u(t)), u'(t) \rangle \quad \text{pour } 1 \leq i \leq M,$$

et

$$j''(t) = J''(u(t))(u'(t), u'(t)) + \langle J'(u(t)), u''(t) \rangle$$

$$f''_i(t) = F''_i(u(t))(u'(t), u'(t)) + \langle F'_i(u(t)), u''(t) \rangle \quad \text{pour } 1 \leq i \leq M.$$

Comme  $f_i(t) = 0$  pour tout  $t$  et puisque 0 est un minimum de  $j(t)$ , on en déduit  $j'(0) = 0$ ,  $j''(0) \geq 0$ , et  $f'_i(0) = f''_i(0) = 0$ . Les conditions  $f'_i(0) = 0$  nous disent que  $u'(0)$  est orthogonal au sous-espace engendré par  $(F'_i(u))_{1 \leq i \leq M}$  (qui est égal à  $K(u)$  quand cette famille est libre), tandis que  $j'(0) = 0$  signifie que  $J'(u)$  est orthogonal à  $u'(0)$ . Si  $u'(0)$  décrit tout  $K(u)$  lorsque on fait varier les chemins admissibles, on en déduit que  $J'(u)$  et les  $F'_i(u)$  appartiennent au même sous-espace (l'orthogonal de  $K(u)$ ). On retrouve ainsi la condition du premier ordre : il existe  $\lambda \in \mathbb{R}^M$  tel que

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0. \quad (10.30)$$

Les conditions  $f_i''(0) = 0$  impliquent que

$$0 = \sum_{i=1}^M \lambda_i \left( F_i''(u)(u'(0), u'(0)) + \langle F_i'(u), u''(0) \rangle \right),$$

tandis que  $j''(0) \geq 0$  donne

$$J''(u)(u'(0), u'(0)) + \langle J'(u), u''(0) \rangle \geq 0.$$

Grâce à (10.30) on peut éliminer les dérivées premières et  $u''(0)$  pour obtenir (en sommant les deux dernières équations)

$$\left( \sum_{i=1}^M \lambda_i F_i''(u) + J''(u) \right) (u'(0), u'(0)) \geq 0,$$

qui n'est rien d'autre que (10.29) lorsque  $u'(0)$  parcourt  $K(u)$ .

L'existence de tels chemins admissibles  $u(t)$  et le fait que l'ensemble des  $u'(0)$  décrit la totalité du cône des directions admissibles  $K(u)$  est une conséquence du théorème des fonctions implicites que l'on peut appliquer grâce à l'hypothèse que la famille  $(F_i'(u))_{1 \leq i \leq M}$  est libre (voir par exemple [5]).  $\square$

**Exercice 10.2.14** Calculer la condition nécessaire d'optimalité du second ordre pour les Exemples 9.1.6 et 9.1.7.

### Contraintes d'inégalité

Dans ce deuxième cas on suppose que  $K$  est donné par

$$K = \{v \in V, \quad F_i(v) \leq 0 \quad \text{pour} \quad 1 \leq i \leq M\}, \quad (10.31)$$

où  $F_1, \dots, F_M$  sont des fonctions continues 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 (10.31) 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  $\varepsilon$  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 (10.31) 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). Nous allons donner une définition dont le principe est de regarder sur le problème **linéarisé** s'il est possible de faire des variations respectant les contraintes linéarisées. Ces considérations de "calcul des variations" motivent les définitions suivantes.

**Définition 10.2.12** Soit  $u \in K$ . L'ensemble  $I(u) = \{i \in \{1, \dots, M\} \mid F_i(u) = 0\}$  est appelé l'ensemble des contraintes **actives** en  $u$ .

**Définition 10.2.13** On dit que les contraintes (10.31) sont **qualifiées** en  $u \in K$  si et seulement si il existe une direction  $\bar{w} \in V$  telle que l'on ait pour tout  $i \in I(u)$

$$\begin{aligned} &\text{ou bien } \langle F'_i(u), \bar{w} \rangle < 0, \\ &\text{ou bien } \langle F'_i(u), \bar{w} \rangle = 0 \quad \text{et } F_i \text{ est affine.} \end{aligned} \quad (10.32)$$

**Remarque 10.2.14** La direction  $\bar{w}$  est en quelque sorte une “direction rentrante” puisque on déduit de (10.32) que  $u + \varepsilon \bar{w} \in K$  pour tout  $\varepsilon \geq 0$  suffisamment petit. Bien sûr, si toutes les fonctions  $F_i$  sont affines, on peut prendre  $\bar{w} = 0$  et les contraintes sont automatiquement qualifiées. Le fait de distinguer les contraintes affines dans la Définition 10.2.13 est justifié non seulement parce que celles-ci sont qualifiées sous des conditions moins strictes, mais surtout en regard de l'importance des contraintes affines dans les applications (comme le montre les exemples du Chapitre 9). •

Nous pouvons alors énoncer les conditions **nécessaires** d'optimalité sur l'ensemble (10.31).

**Théorème 10.2.15** On suppose que  $K$  est donné par (10.31), que les fonctions  $J$  et  $F_1, \dots, F_M$  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  $\lambda_1, \dots, \lambda_M \geq 0$ , appelés *multiplicateurs de Lagrange*, tels que

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0, \quad \lambda_i \geq 0, \quad \lambda_i = 0 \text{ si } F_i(u) < 0 \quad \forall i \in \{1, \dots, M\}. \quad (10.33)$$

**Remarque 10.2.16** On peut réécrire la condition (10.33) sous la forme suivante

$$J'(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0, \quad \lambda \geq 0, \quad \lambda \cdot F(u) = 0,$$

où  $\lambda \geq 0$  signifie que chacune des composantes du vecteur  $\lambda = (\lambda_1, \dots, \lambda_M)$  est positive, puisque, pour tout indice  $i \in \{1, \dots, M\}$ , on a soit  $F_i(u) = 0$ , soit  $\lambda_i = 0$ . Le fait que  $\lambda \cdot F(u) = 0$  est appelée condition des écarts complémentaires. •

**Démonstration.** Considérons tout d'abord l'ensemble

$$\tilde{K}(u) = \{w \in V \mid \langle F'_i(u), w \rangle \leq 0 \quad \forall i \in I(u)\}. \quad (10.34)$$

(On peut montrer que  $\tilde{K}(u)$  n'est autre que le cône  $K(u)$  des directions admissibles, voir [5]). Soit  $\bar{w}$  une direction admissible satisfaisant (10.32),  $w \in \tilde{K}(u)$ , et un réel  $\delta > 0$ . Nous allons montrer que  $u + \varepsilon(w + \delta \bar{w}) \in K$  pour tout réel  $\varepsilon > 0$  assez petit. Il faut examiner trois cas de figure.

1. Si  $i \notin I(u)$ , on a  $F_i(u) < 0$  et  $F_i(u + \varepsilon(w + \delta\bar{w})) < 0$  par continuité de  $F_i$  si  $\varepsilon$  est assez petit.
2. Si  $i \in I(u)$  et  $\langle F'_i(u), \bar{w} \rangle < 0$ , alors

$$\begin{aligned} F_i(u + \varepsilon(w + \delta\bar{w})) &= F_i(u) + \varepsilon \langle F'_i(u), w + \delta\bar{w} \rangle + o(\varepsilon) \\ &\leq \varepsilon \delta \langle F'_i(u), \bar{w} \rangle + o(\varepsilon) < 0, \end{aligned} \quad (10.35)$$

pour  $\varepsilon > 0$  assez petit.

3. Enfin, si  $i \in I(u)$  et  $\langle F'_i(u), \bar{w} \rangle = 0$ , alors  $F_i$  est affine et

$$F_i(u + \varepsilon(w + \delta\bar{w})) = F_i(u) + \varepsilon \langle F'_i(u), w + \delta\bar{w} \rangle = \varepsilon \langle F'_i(u), w \rangle \leq 0. \quad (10.36)$$

Finalement, si  $u$  est un minimum local de  $J$  sur  $K$ , on déduit de ce qui précède que

$$\langle J'(u), w + \delta\bar{w} \rangle \geq 0 \quad \forall w \in \tilde{K}(u) \quad , \quad \forall \delta \in \mathbb{R}_+^* .$$

En faisant tendre  $\delta$  vers 0, on obtient  $\langle J'(u), w \rangle \geq 0$  pour toute direction  $w \in \tilde{K}(u)$  et on termine la démonstration grâce au Lemme de Farkas 10.2.17 ci-dessous.  $\square$

**Lemme 10.2.17 (de Farkas)** Soient  $a_1, \dots, a_M$  des éléments fixés de  $V$ . On considère les ensembles

$$\mathcal{K} = \left\{ w \in V, \langle a_i, w \rangle \leq 0 \text{ pour } 1 \leq i \leq M \right\},$$

et

$$\hat{\mathcal{K}} = \left\{ q \in V, \exists \lambda_1, \dots, \lambda_M \geq 0, q = - \sum_{i=1}^M \lambda_i a_i \right\}.$$

Alors pour tout  $p \in V$ , on a l'implication

$$\langle p, w \rangle \geq 0 \quad \forall w \in \mathcal{K} \implies p \in \hat{\mathcal{K}}.$$

(La réciproque étant évidente, il s'agit en fait d'une équivalence.)

**Démonstration.** Commençons par montrer que  $\hat{\mathcal{K}}$  est fermé. Supposons d'abord que les vecteurs  $(a_i)_{1 \leq i \leq M}$  sont linéairement indépendants. Soit  $(q^n) = \left( - \sum_{i=1}^M \lambda_i^n a_i \right)$  une suite d'éléments de  $\hat{\mathcal{K}}$  (donc avec  $\lambda_i^n \geq 0 \quad \forall i \forall n$ ), convergeant vers une limite  $q \in V$ . Alors il est clair que chaque suite  $(\lambda_i^n)$  converge dans  $\mathbb{R}_+$  vers une limite  $\lambda_i \geq 0$  (pour  $1 \leq i \leq M$ ) puisque les vecteurs  $(a_i)_{1 \leq i \leq M}$  forment une base de l'espace qu'ils engendrent. On a donc  $q = - \sum_{i=1}^M \lambda_i a_i \in \hat{\mathcal{K}}$ , qui est donc fermé.

Si les vecteurs  $(a_i)_{1 \leq i \leq M}$  sont linéairement dépendants, nous procédons par récurrence sur leur nombre  $M$ . La propriété est évidente lorsque  $M = 1$ , et nous supposons qu'elle est vraie lorsque le nombre de vecteurs  $a_i$  est inférieur à  $M$ . Comme les vecteurs  $(a_i)_{1 \leq i \leq M}$  sont liés, il existe une relation de la forme  $\sum_{i=1}^M \mu_i a_i = 0$ , avec au moins un

des coefficients  $\mu_i$  qui est strictement positif. Soit alors  $q = -\sum_{i=1}^M \lambda_i a_i$  un élément de  $\hat{\mathcal{K}}$ . Pour tout  $t \leq 0$ , on peut aussi écrire  $q = -\sum_{i=1}^M (\lambda_i + t\mu_i) a_i$ , et on peut choisir  $t \leq 0$  pour que

$$\lambda_i + t\mu_i \geq 0 \quad \forall i \in \{1, \dots, M\} \quad \text{et} \quad \exists i_0 \in \{1, \dots, M\}, \quad \lambda_{i_0} + t\mu_{i_0} = 0.$$

Ce raisonnement montre que

$$\hat{\mathcal{K}} = \bigcup_{i_0=1}^M \left\{ q \in V, \exists \lambda_1, \dots, \lambda_M \geq 0, q = -\sum_{i \neq i_0} \lambda_i a_i \right\}. \quad (10.37)$$

Par notre hypothèse de récurrence, chacun des ensembles apparaissant dans le membre de droite de (10.37) est fermé, et il en est donc de même de  $\hat{\mathcal{K}}$ .

Raisonnons maintenant par l'absurde : supposons que  $\langle p, w \rangle \geq 0 \quad \forall w \in \mathcal{K}$  et que  $p \notin \hat{\mathcal{K}}$ . On peut alors utiliser le Théorème 12.1.19 de séparation d'un point et d'un convexe pour séparer  $p$  et  $\hat{\mathcal{K}}$  qui est fermé et, à l'évidence, convexe et non vide. Il existe donc  $w \neq 0$  dans  $V$  et  $\alpha \in \mathbb{R}$  tels que

$$\langle p, w \rangle < \alpha < \langle q, w \rangle \quad \forall q \in \hat{\mathcal{K}}. \quad (10.38)$$

Mais alors, on doit avoir  $\alpha < 0$  puisque  $0 \in \hat{\mathcal{K}}$ ; d'autre part, pour tout  $i \in \{1, \dots, M\}$  nous pouvons choisir dans (10.38)  $q = -\lambda a_i$  avec  $\lambda$  arbitrairement grand, ce qui montre que  $\langle a_i, w \rangle \leq 0$ . On obtient donc que  $w \in \mathcal{K}$  et que  $\langle p, w \rangle < \alpha < 0$ , ce qui est impossible.  $\square$

**Exercice 10.2.15** Soit  $A$  une matrice symétrique définie positive d'ordre  $n$ , et  $B$  une matrice de taille  $m \times n$  avec  $m \leq n$  et de rang  $m$ . On considère le problème de minimisation

$$\min_{x \in \mathbb{R}^n, Bx \leq c} \left\{ J(x) = \frac{1}{2} Ax \cdot x - b \cdot x \right\},$$

Appliquer le Théorème 10.2.15 pour obtenir l'existence d'un multiplicateur de Lagrange  $p \in \mathbb{R}^m$  tel qu'un point de minimum  $\bar{x}$  vérifie

$$A\bar{x} - b + B^*p = 0, \quad p \geq 0, \quad p \cdot (B\bar{x} - c) = 0.$$

**Exercice 10.2.16** Soit  $f \in L^2(\Omega)$  une fonction définie sur un ouvert borné  $\Omega$ . Pour  $\epsilon > 0$  on considère le problème de régularisation suivant

$$\min_{u \in H_0^1(\Omega), \|u-f\|_{L^2(\Omega)} \leq \epsilon} \int_{\Omega} |\nabla u|^2 dx.$$

Montrer que ce problème admet une unique solution  $u_\epsilon$ . Montrer que, soit  $u_\epsilon = f$ , soit  $u_\epsilon = 0$ , soit il existe  $\lambda > 0$  tel que  $u_\epsilon$  est solution de

$$\begin{cases} -\Delta u_\epsilon + \lambda(u_\epsilon - f) = 0 & \text{dans } \Omega, \\ u_\epsilon = 0 & \text{sur } \partial\Omega. \end{cases}$$

### Contraintes d'égalité et d'inégalité

On peut bien sûr mélanger les deux types de contraintes. On suppose donc que  $K$  est donné par

$$K = \{v \in V \quad , \quad G(v) = 0 \quad , \quad F(v) \leq 0\} \quad , \quad (10.39)$$

où  $G(v) = (G_1(v), \dots, G_N(v))$  et  $F(v) = (F_1(v), \dots, F_M(v))$  sont deux applications de  $V$  dans  $\mathbb{R}^N$  et  $\mathbb{R}^M$ . Dans ce nouveau contexte, il faut donner une définition adéquate de la qualification des contraintes. On note toujours  $I(u) = \{i \in \{1, \dots, M\} \quad , \quad F_i(u) = 0\}$  l'ensemble des contraintes d'inégalité actives en  $u \in K$ .

**Définition 10.2.18** *On dit que les contraintes (10.39) sont **qualifiées** en  $u \in K$  si et seulement si les vecteurs  $(G'_i(u))_{1 \leq i \leq N}$  sont linéairement indépendants et il existe une direction  $\bar{w} \in \bigcap_{i=1}^N [G'_i(u)]^\perp$  telle que l'on ait pour tout  $i \in I(u)$*

$$\langle F'_i(u), \bar{w} \rangle < 0 \quad . \quad (10.40)$$

Nous pouvons alors énoncer les conditions **nécessaires** d'optimalité sur l'ensemble (10.39).

**Théorème 10.2.19** *Soit  $u \in K$  où  $K$  est donné par (10.39). On suppose que  $J$  et  $F$  sont dérivables en  $u$ , que  $G$  est dérivable dans un voisinage de  $u$ , et que les contraintes sont qualifiées en  $u$  (au sens de la Définition 10.2.18). Alors, si  $u$  est un minimum local de  $J$  sur  $K$ , il existe des multiplicateurs de Lagrange  $\mu_1, \dots, \mu_N$ , et  $\lambda_1, \dots, \lambda_M \geq 0$ , tels que*

$$J'(u) + \sum_{i=1}^N \mu_i G'_i(u) + \sum_{i=1}^M \lambda_i F'_i(u) = 0 \quad , \quad \lambda \geq 0 \quad , \quad F(u) \leq 0 \quad , \quad \lambda \cdot F(u) = 0 \quad . \quad (10.41)$$

La démonstration du Théorème 10.2.19 est une simple adaptation de celles des Théorèmes 10.2.8 et 10.2.15, que nous laissons au lecteur en guise d'exercice.

### Autres formes des conditions de qualification

Les conditions de qualification sont des conditions **suffisantes** de type "géométrique" qui permettent de faire des variations internes à l'ensemble  $K$  à partir d'un point  $u \in K$ . La condition de qualification de la Définition 10.2.13 est assez générale (quoique loin d'être nécessaire), mais parfois difficile à vérifier dans les applications. C'est pourquoi les remarques qui suivent donnent des conditions de qualifications plus simples (donc plus faciles à vérifier en pratique) mais moins générales (i.e. moins souvent vérifiées).

**Remarque 10.2.20** Dans le cas des contraintes d'inégalité, on peut s'inspirer de la notion de cas régulier (introduite à la Remarque 10.2.9 pour les contraintes d'égalité) afin de donner une condition très simple qui entraîne la condition de qualification de



la Définition 10.2.13. En effet, pour  $u \in K$  les contraintes inactives ne “jouent” pas et seules sont à prendre en compte les contraintes actives  $i \in I(u)$  qui sont justement des contraintes d'égalité en ce point ! On vérifie alors sans peine que la condition suivante (qui dit que  $u$  est un point régulier pour les contraintes d'égalité  $F_i(u) = 0$  pour  $i \in I(u)$ )

$$(F'_i(u))_{i \in I(u)} \text{ est une famille libre} \quad (10.42)$$

entraîne (10.32), c'est-à-dire que les contraintes sont qualifiées. En effet, il suffit de prendre  $\bar{w} = \sum_{i \in I(u)} \alpha_i F'_i(u)$  tel que  $\langle F'_j(u), \bar{w} \rangle = -1$  pour tout  $j \in I(u)$  (l'existence des coefficients  $\alpha_i$  découle de l'inversibilité de la matrice  $(\langle F'_i(u), F'_j(u) \rangle)_{ij}$ ). Il est clair cependant que (10.32) n'implique pas (10.42). •

**Remarque 10.2.21** Dans le cas des contraintes combinées d'égalité et d'inégalité, on peut aussi s'inspirer de la notion de cas régulier pour donner une condition plus simple qui implique la condition de qualification de la Définition 10.2.18. Cette condition “forte” (c'est-à-dire moins souvent vérifiée) de qualification est

$$(G'_i(u))_{1 \leq i \leq N} \cup (F'_i(u))_{i \in I(u)} \text{ est une famille libre.} \quad (10.43)$$

On vérifie facilement que (10.43) entraîne (10.40), c'est-à-dire que les contraintes sont qualifiées. •

**Remarque 10.2.22** Revenant au cas des contraintes d'inégalité, supposées convexes, une autre condition de qualification possible est la suivante. On suppose qu'il existe  $\bar{v} \in V$  tel que l'on ait, pour tout  $i \in \{1, \dots, M\}$ ,

$$\begin{aligned} &\text{les fonctions } F_i \text{ sont convexes et,} \\ &\text{ou bien } F_i(\bar{v}) < 0, \\ &\text{ou bien } F_i(\bar{v}) = 0 \text{ et } F_i \text{ est affine.} \end{aligned} \quad (10.44)$$

L'hypothèse (10.44) entraîne que les contraintes sont qualifiées en  $u \in K$  au sens de la Définition 10.2.13. En effet, si  $i \in I(u)$  et si  $F_i(\bar{v}) < 0$ , alors, d'après la condition de convexité (10.7)

$$\langle F'_i(u), \bar{v} - u \rangle = F_i(u) + \langle F'_i(u), \bar{v} - u \rangle \leq F_i(\bar{v}) < 0.$$

D'autre part, si  $i \in I(u)$  et si  $F_i(\bar{v}) = 0$ , alors  $F_i$  est affine et

$$\langle F'_i(u), \bar{v} - u \rangle = F_i(\bar{v}) - F_i(u) = 0,$$

et la Définition 10.2.13 de qualification des contraintes est satisfaite avec  $\bar{w} = \bar{v} - u$ . L'avantage de l'hypothèse (10.44) est de ne pas nécessiter de connaître le point de minimum  $u$  ni de calculer les dérivées des fonctions  $F_1, \dots, F_M$ . •

### 10.3 Point-selle, théorème de Kuhn et Tucker, dualité

Nous avons vu dans la Remarque 10.2.10 comment il est possible d'interpréter le couple  $(u, \lambda)$  (point de minimum, multiplicateur de Lagrange) comme **point stationnaire du 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 10.5. Signalons enfin que la notion de point-selle joue un rôle fondamental dans la **théorie des jeux**.

#### 10.3.1 Point-selle

De manière abstraite,  $V$  et  $Q$  étant deux espaces de Hilbert réels, 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 du Théorème 10.2.8 sur les contraintes d'égalité (ou plutôt de la Remarque 10.2.10), nous avons  $U = V$ ,  $P = Q = \mathbb{R}^M$  et  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ . La situation est un peu différente dans le cadre du Théorème 10.2.15 sur les contraintes d'inégalité, où pour le même Lagrangien  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$  il faut prendre  $U = V$ ,  $Q = \mathbb{R}^M$  et  $P = (\mathbb{R}_+)^M$ .

Donnons maintenant la définition d'un point-selle, souvent appelé également min-max ou col.

**Définition 10.3.1** *On dit que  $(u, p) \in U \times P$  est un point-selle de  $\mathcal{L}$  sur  $U \times P$  si*

$$\forall q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in U. \quad (10.45)$$

Le résultat suivant montre le lien entre cette notion de point-selle et les problèmes de minimisation avec contraintes d'égalité (10.24) ou contraintes d'inégalité (10.31) 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 10.3.2** *On suppose que les fonctions  $J, F_1, \dots, F_M$  sont continues sur  $V$ , et que l'ensemble  $K$  est défini par (10.24) ou (10.31). On note  $P = \mathbb{R}^M$  dans le cas de contraintes d'égalité (10.24) et  $P = (\mathbb{R}_+)^M$  dans le cas de contraintes d'inégalité (10.31). Soit  $U$  un ouvert de  $V$  contenant  $K$ . Pour  $(v, q) \in U \times P$ , on pose  $\mathcal{L}(v, q) = J(v) + q \cdot F(v)$ .*

*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  $F_1, \dots, F_M$  sont dérivables en  $u$ , on a*

$$J'(u) + \sum_{i=1}^M p_i F'_i(u) = 0. \quad (10.46)$$

**Démonstration.** Écrivons la condition de point-selle

$$\forall q \in P \quad J(u) + q \cdot F(u) \leq J(u) + p \cdot F(u) \leq J(v) + p \cdot F(v) \quad \forall v \in U. \quad (10.47)$$

Examinons d'abord le cas de contraintes d'égalité. Puisque  $P = \mathbb{R}^M$ , la première inégalité dans (10.47) 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 (10.47) 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  $F_1, \dots, F_M$  sont dérivables en  $u$ , la deuxième inégalité de (10.47) 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 10.2.2).  $\square$

### 10.3.2 Théorème de Kuhn et Tucker

Nous revenons au problème de minimisation sous contraintes d'inégalité pour lequel l'ensemble  $K$  est donné par (10.31), c'est-à-dire

$$K = \{v \in V, \quad F_i(v) \leq 0 \quad \text{pour} \quad 1 \leq i \leq m\}. \quad (10.48)$$

Le Théorème 10.2.15 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 10.3.2 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  $J, F_1, \dots, F_M$  sont convexes continues sur  $V$ .

**Remarque 10.3.3** Comme  $J, F_1, \dots, F_M$  sont convexes continues,  $K$  est convexe fermé et l'existence d'un minimum global de  $J$  sur  $K$  est assuré par le Théorème 9.2.7 dès que  $K$  est non vide et que la condition "infinie à l'infini" (9.11) est vérifiée.  $\bullet$

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 10.2.15 est en fait une condition **nécessaire et suffisante**.

**Théorème 10.3.4 (de Kuhn et Tucker)** *On suppose que les fonctions  $J, F_1, \dots, F_M$  sont convexes continues sur  $V$  et dérivables sur l'ensemble  $K$  (10.48). 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 10.2.13. 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

$$F(u) \leq 0, \quad p \geq 0, \quad p \cdot F(u) = 0, \quad J'(u) + \sum_{i=1}^M p_i F'_i(u) = 0. \quad (10.49)$$

**Démonstration.** Si  $u$  est un minimum de  $J$  sur  $K$ , on peut appliquer le Théorème 10.2.15, qui donne exactement la condition d'optimalité (10.49), d'où l'on déduit facilement que  $(u, p)$  est point-selle de  $\mathcal{L}$  sur  $V \times (\mathbb{R}_+)^M$  (en utilisant le fait que  $J(v) + p \cdot F(v)$  est convexe). Réciproquement, si  $(u, p)$  est point-selle, on a déjà montré à la Proposition 10.3.2 que  $u$  est un minimum global de  $J$  sur  $K$ .  $\square$

**Remarque 10.3.5** Le Théorème 10.3.4 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 10.3.4 de Kuhn et Tucker à un problème de minimisation avec contraintes d'égalité affines.  $\bullet$

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 10.3.1** On considère le problème d'optimisation, dit perturbé

$$\inf_{F_i(v) \leq u_i, 1 \leq i \leq m} J(v), \quad (10.50)$$

avec  $u = (u_1, \dots, u_m) \in \mathbb{R}^m$ . On se place sous les hypothèses du Théorème 10.3.4 de Kuhn et Tucker. On note  $m^*(u)$  la valeur minimale du problème perturbé (10.50).

1. Montrer que si  $p \in \mathbb{R}^m$  est le multiplicateur de Lagrange pour le problème non perturbé (c'est-à-dire (10.50) avec  $u = 0$ ), alors

$$m^*(u) \geq m^*(0) - p \cdot u. \quad (10.51)$$

2. Dédurre de (10.51) que si  $u \mapsto m^*(u)$  est dérivable, alors

$$p_i = -\frac{\partial m^*}{\partial u_i}(0).$$

Interpréter ce résultat (cf. l'Exemple 9.1.8 en économie).

### 10.3.3 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 10.3.1.

**Définition 10.3.6** Soit  $V$  et  $Q$  deux espaces de Hilbert réels, et  $\mathcal{L}$  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  $U \times P$

$$\forall q \in P \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in U. \quad (10.52)$$

Pour  $v \in U$  et  $q \in P$ , posons

$$\mathcal{J}(v) = \sup_{q \in P} \mathcal{L}(v, q) \quad \mathcal{G}(q) = \inf_{v \in U} \mathcal{L}(v, q). \quad (10.53)$$

On appelle **problème primal** le problème de minimisation

$$\inf_{v \in U} \mathcal{J}(v), \quad (10.54)$$

et **problème dual** le problème de maximisation

$$\sup_{q \in P} \mathcal{G}(q). \quad (10.55)$$

**Remarque 10.3.7** 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 10.3.6 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 bien définies) ne sont pas vides, puisque (10.52) 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 10.3.8 (de dualité)** Le couple  $(u, p)$  est un point-selle de  $\mathcal{L}$  sur  $U \times P$  si et seulement si

$$\mathcal{J}(u) = \min_{v \in U} \mathcal{J}(v) = \max_{q \in P} \mathcal{G}(q) = \mathcal{G}(p). \quad (10.56)$$

**Remarque 10.3.9** Par la Définition (10.53) de  $\mathcal{J}$  et  $\mathcal{G}$ , (10.56) est équivalent à

$$\mathcal{J}(u) = \min_{v \in U} \left( \sup_{q \in P} \mathcal{L}(v, q) \right) = \max_{q \in P} \left( \inf_{v \in U} \mathcal{L}(v, q) \right) = \mathcal{G}(p). \quad (10.57)$$

Si le sup et l'inf sont atteints dans (10.57) (c'est-à-dire qu'on peut les écrire max et min, respectivement), on voit alors que (10.57) 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 (10.53) que  $\mathcal{J}(v) \geq \mathcal{L}(v, p)$ , d'où  $\mathcal{J}(v) \geq \mathcal{L}^*$  d'après (10.52). 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 (10.56) a lieu et posons  $\mathcal{L}^* = \mathcal{J}(u)$ . La définition (10.53) de  $\mathcal{J}$  montre que

$$\mathcal{L}(u, q) \leq \mathcal{J}(u) = \mathcal{L}^* \quad \forall q \in P. \quad (10.58)$$

De même, on a aussi :

$$\mathcal{L}(v, p) \geq \mathcal{G}(p) = \mathcal{L}^* \quad \forall v \in U, \quad (10.59)$$

et on déduit facilement de (10.58)-(10.59) que  $\mathcal{L}(u, p) = \mathcal{L}^*$ , ce qui montre que  $(u, p)$  est point-selle.  $\square$

**Remarque 10.3.10** 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**

$$\inf_{v \in U} \left( \sup_{q \in P} \mathcal{L}(v, q) \right) \geq \sup_{q \in P} \left( \inf_{v \in U} \mathcal{L}(v, q) \right). \quad (10.60)$$

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  $v \in U$ ,  $\inf_{v \in U} \sup_{q \in P} \mathcal{L}(v, q) \geq \sup_{q \in P} \inf_{v' \in U} \mathcal{L}(v', q)$ , ce qui donne (10.60). La différence (positive) entre les deux membres de l'inégalité (10.60) est appelée **saut de dualité**.  $\bullet$

**Exercice 10.3.2** Donner un exemple de Lagrangien pour lequel l'inégalité (10.60) est stricte avec ses deux membres finis.

**Exercice 10.3.3** Soit  $U$  (respectivement  $P$ ) un convexe compact non vide de  $V$  (respectivement  $Q$ ). On suppose que le Lagrangien est tel que  $v \rightarrow \mathcal{L}(v, q)$  est convexe sur  $U$  pour tout  $q \in P$ , et  $q \rightarrow \mathcal{L}(v, q)$  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

$$\inf_{v \in V, F(v) \leq 0} J(v) \quad (10.61)$$

avec  $J$  et  $F = (F_1, \dots, F_M)$  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  $v \in V$ ,

$$\mathcal{J}(v) = \sup_{q \in (\mathbb{R}_+)^M} \mathcal{L}(v, q) = \begin{cases} J(v) & \text{si } F(v) \leq 0 \\ +\infty & \text{sinon,} \end{cases} \quad (10.62)$$

ce qui montre que le problème primal  $\inf_{v \in V} \mathcal{J}(v)$  est exactement le problème d'origine (10.61) ! D'autre part, la fonction  $\mathcal{G}(q)$  du problème dual est bien définie par (10.53), car

(10.53) 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 9.2.3). 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 (10.61) 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 10.3.4 et de dualité 10.3.8 nous donne le résultat suivant.

**Corollaire 10.3.11** *On suppose que les fonctions  $J, F_1, \dots, F_M$  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 10.2.13. 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

$$F(u) \leq 0, \quad p \geq 0, \quad p \cdot F(u) = 0, \quad J'(u) + p \cdot F'(u) = 0. \quad (10.63)$$

L'application la plus courante du Corollaire 10.3.11 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'unicité des solutions pour tous ces problèmes. Précisons aussi que pour obtenir l'existence du minimum  $u$  dans le Corollaire 10.3.11 il suffit d'ajouter une hypothèse de forte convexité ou de comportement infini à l'infini sur  $J$ .

**Remarque 10.3.12** Pour illustrer le Corollaire 10.3.11 et l'intérêt de la dualité, nous considérons un problème de minimisation quadratique dans  $\mathbb{R}^N$  avec contraintes d'inégalité affines

$$\min_{v \in \mathbb{R}^N, F(v) = Bv - c \leq 0} \left\{ J(v) = \frac{1}{2} Av \cdot v - b \cdot v \right\}, \quad (10.64)$$

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

$$\mathcal{L}(v, q) = \frac{1}{2} Av \cdot v - b \cdot v + q \cdot (Bv - c) \quad \forall (v, q) \in \mathbb{R}^N \times (\mathbb{R}_+)^M. \quad (10.65)$$

Nous avons déjà fait dans (10.62) le calcul de  $\mathcal{J}$ , et dit que le problème primal est exactement (10.64). 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  $v \rightarrow \mathcal{L}(v, q)$  est une fonction fortement convexe. 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

$$\mathcal{G}(q) = \mathcal{L}(A^{-1}(b - B^*q), q),$$

et le problème dual s'écrit finalement

$$\sup_{q \geq 0} \left( -\frac{1}{2}q \cdot BA^{-1}B^*q + (BA^{-1}b - c) \cdot q - \frac{1}{2}A^{-1}b \cdot b \right). \quad (10.66)$$

Certes, la fonctionnelle à maximiser dans (10.66) n'a pas une allure particulièrement sympathique. Il s'agit encore d'un problème avec fonctionnelle quadratique et contraintes affines. Cependant, le Corollaire 10.3.11 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 (10.66) 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 10.5.3 que cet avantage peut être utilisé pour mettre au point un algorithme de calcul de la solution du problème primal. •

Terminons par un exercice récréatif qui montre la relation entre les problèmes de point-selle ou min-max et la théorie des jeux.

**Exercice 10.3.4** Soit une matrice rectangulaire

$$A = \begin{pmatrix} 1 & 0 & 4 & 2 & 3 & 5 \\ -3 & 2 & -1 & 2 & -5 & 2 \\ -4 & 2 & -2 & 0 & -1 & 2 \\ -2 & 4 & -1 & 6 & -2 & 2 \\ -1 & 2 & -6 & 3 & -1 & 1 \end{pmatrix}.$$

On suppose que deux joueurs choisissent l'un une ligne  $i$ , l'autre une colonne  $j$ , sans qu'ils ne connaissent le choix de l'autre. Une fois révélé leurs choix, le gain (ou la perte, selon le signe) du premier joueur est déterminé par le coefficient  $a_{ij}$  de la matrice  $A$  (l'autre joueur recevant ou payant  $-a_{ij}$ ). Montrer que la stratégie optimale de minimisation du risque conduit à un problème de min-max que l'on résoudra. Le jeu est-il équitable avec cette matrice  $A$ ?

## 10.4 Applications

Dans cette section nous étudions quelques applications des résultats des sections précédentes. Signalons qu'une autre application, la programmation linéaire, sera traitée à part au prochain chapitre en raison de son importance en recherche opérationnelle.

### 10.4.1 Énergie duale ou complémentaire

Au Chapitre 5 nous avons vu que la résolution du problème aux limites suivant

$$\begin{cases} -\Delta u = f & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (10.67)$$



où  $\Omega$  est un ouvert borné de  $\mathbb{R}^N$  et  $f \in L^2(\Omega)$ , est équivalent à la minimisation d'une énergie

$$\min_{v \in H_0^1(\Omega)} \left\{ J(v) = \frac{1}{2} \int_{\Omega} |\nabla v|^2 dx - \int_{\Omega} f v dx \right\} \quad (10.68)$$

(voir la Proposition 5.2.7). Nous avons vu à l'Exercice 9.2.8 que (10.68) admet un unique minimum et, à l'Exercice 10.2.4, que son équation d'Euler est la formulation variationnelle de (10.67). La signification physique de l'énergie (10.68) est évidente. Par exemple, si (10.67) modélise la déformation d'une membrane élastique ( $u$  est le déplacement normal sous l'action des forces  $f$ ), la solution est le déplacement qui minimise la somme de l'énergie élastique de déformation et de l'énergie potentielle des forces extérieures. On se propose de montrer que la théorie de la dualité permet d'associer à (10.67) un **deuxième principe de minimisation** mettant en jeu une énergie, dite **complémentaire** en mécanique (ou duale), dont la signification physique est tout aussi importante que celle de (10.68).

Nous allons introduire un Lagrangien associé à l'énergie primale (10.68) bien que celle-ci ne présente pas de contraintes. Pour ce faire nous utilisons une petite astuce en introduisant une variable intermédiaire  $e \in L^2(\Omega)^N$  et une contrainte  $e = \nabla v$ . Alors (10.68) est équivalent à

$$\min_{\substack{v \in H_0^1(\Omega), \\ e = \nabla v}} \left\{ \tilde{J}(v, e) = \frac{1}{2} \int_{\Omega} |e|^2 dx - \int_{\Omega} f v dx \right\}.$$

On introduit un Lagrangien intermédiaire pour ce problème

$$\mathcal{M}(e, v, \tau) = \tilde{J}(v, e) + \int_{\Omega} \tau \cdot (\nabla v - e) dx,$$

avec un multiplicateur de Lagrange  $\tau \in L^2(\Omega)^N$ . On élimine maintenant  $e$  pour obtenir le Lagrangien recherché

$$\mathcal{L}(v, \tau) = \min_{e \in L^2(\Omega)^N} \mathcal{M}(e, v, \tau).$$

Comme  $e \rightarrow \mathcal{M}(e, v, \tau)$  est fortement convexe, il existe un unique point de minimum, et un calcul facile montre que

$$\mathcal{L}(v, \tau) = -\frac{1}{2} \int_{\Omega} |\tau|^2 dx - \int_{\Omega} f v dx + \int_{\Omega} \tau \cdot \nabla v dx. \quad (10.69)$$

On vérifie sans peine que le problème primal associé au Lagrangien (10.69) est bien (10.68)

$$\left( \max_{\tau \in L^2(\Omega)^N} \mathcal{L}(v, \tau) \right) = J(v),$$

et que le problème dual est

$$\left( \min_{v \in H_0^1(\Omega)} \mathcal{L}(v, \tau) \right) = G(\tau) = \begin{cases} -\frac{1}{2} \int_{\Omega} |\tau|^2 dx & \text{si } -\operatorname{div} \tau = f \text{ dans } \Omega \\ -\infty & \text{sinon.} \end{cases} \quad (10.70)$$

On peut maintenant énoncer le résultat principal.

**Théorème 10.4.1** *Il existe un unique point selle  $(u, \sigma)$  du Lagrangien  $\mathcal{L}(v, \tau)$  sur  $H_0^1(\Omega) \times L^2(\Omega)^N$*

$$\mathcal{L}(u, \sigma) = \max_{\tau \in L^2(\Omega)^N} \min_{v \in H_0^1(\Omega)} \mathcal{L}(v, \tau) = \min_{v \in H_0^1(\Omega)} \max_{\tau \in L^2(\Omega)^N} \mathcal{L}(v, \tau).$$

Autrement dit,  $u$  est l'unique point de minimum de  $J(v)$  dans  $H_0^1(\Omega)$ ,  $\sigma$  est l'unique point de maximum de  $G(\tau)$  dans  $L^2(\Omega)^N$ ,

$$J(u) = \min_{v \in H_0^1(\Omega)} J(v) = \max_{\tau \in L^2(\Omega)^N} G(\tau) = G(\sigma),$$

et ils sont reliés par la relation  $\sigma = \nabla u$ .

**Remarque 10.4.2** Le problème dual (10.70) a une interprétation physique claire. Comme  $\max G(\tau) = -\min(-G(\tau))$ , il s'agit de minimiser l'énergie (dite complémentaire) des contraintes mécaniques  $\frac{1}{2} \int_{\Omega} |\tau|^2 dx$  parmi l'ensemble des champs de contraintes **statiquement admissibles**, c'est-à-dire vérifiant l'équilibre des forces  $-\operatorname{div} \tau = f$  dans  $\Omega$ . Dans cette formulation duale, le déplacement  $v$  apparaît comme le multiplicateur de Lagrange de la contrainte d'équilibre  $-\operatorname{div} \tau = f$ . Une autre conséquence du Théorème 10.4.1 est qu'on a toujours  $G(\tau) \leq J(v)$  ce qui permet d'obtenir des bornes sur les énergies primale ou duale. Remarquons que l'on a aussi

$$J(u) = G(\sigma) = \frac{1}{2} \int_{\Omega} f u \, dx$$

qui n'est rien d'autre que la moitié du travail des forces extérieures. •

**Démonstration.** La preuve pourrait être une conséquence immédiate du Corollaire 10.3.11 (en remarquant qu'une contrainte d'égalité affine s'écrit comme deux inégalités affines opposées) si ce théorème n'était pas restreint (ici) à un nombre fini de contraintes. Or, dans le problème dual (10.70) il y a une infinité de contraintes puisque la contrainte  $-\operatorname{div} \tau = f$  a lieu pour presque tout point  $x \in \Omega$ . Néanmoins, le résultat est vrai et il est facile de voir pourquoi. En effet, par construction on a

$$G(\tau) \leq \mathcal{L}(v, \tau) \leq J(v),$$

et on sait que les problèmes primal et dual admettent un unique point de minimum  $u$  et  $\sigma$ , respectivement. Or, comme  $u$  est solution de (10.67), une simple intégration par parties montre que

$$J(u) = \frac{1}{2} \int_{\Omega} |\nabla u|^2 dx - \int_{\Omega} f u \, dx = -\frac{1}{2} \int_{\Omega} |\nabla u|^2 dx = -\frac{1}{2} \int_{\Omega} f u \, dx.$$

Si on définit  $\sigma = \nabla u$ , on déduit de (10.67) que  $-\operatorname{div} \sigma = f$ , et on obtient donc

$$G(\tau) \leq J(u) = G(\sigma),$$

c'est-à-dire que  $\sigma$  est le point de maximum de  $G$ , donc  $(u, \sigma)$  est le point selle de  $\mathcal{L}(v, \tau)$ . □

## 10.4.2 Commande optimale

On résout ici l'Exemple 9.1.9 d'un problème de commande optimale appelé système linéaire-quadratique. On considère le système différentiel linéaire dont l'inconnue (l'état du système)  $y(t)$  est à valeurs dans  $\mathbb{R}^N$

$$\begin{cases} \frac{dy}{dt} = Ay + Bv + f \text{ pour } 0 \leq t \leq T \\ y(0) = y_0 \end{cases} \quad (10.71)$$

où  $y_0 \in \mathbb{R}^N$  est l'état initial du système,  $f(t) \in \mathbb{R}^N$  est un terme source,  $v(t) \in \mathbb{R}^M$  est la commande qui permet d'agir sur le système, et  $A$  et  $B$  sont deux matrices constantes de dimensions respectives  $N \times N$  et  $N \times M$ .

On veut choisir la commande  $v$  de manière à minimiser un critère quadratique

$$J(v) = \frac{1}{2} \int_0^T Rv(t) \cdot v(t) dt + \frac{1}{2} \int_0^T Q(y-z)(t) \cdot (y-z)(t) dt + \frac{1}{2} D(y(T) - z_T) \cdot (y(T) - z_T),$$

où  $z(t)$  une trajectoire "cible",  $z_T$  une position finale "cible", et  $R, Q, D$  trois matrices symétriques positives dont seule  $R$  est supposée définie positive. Remarquons que la fonction  $y(t)$  dépend de la variable  $v$  à travers (10.71).

Pour pouvoir appliquer les résultats d'optimisation précédents, nous choisissons de chercher  $v$  dans l'espace de Hilbert  $L^2([0, T[; \mathbb{R}^M)$  des fonctions de  $[0, T[$  dans  $\mathbb{R}^M$  de carré intégrable. (L'espace "plus naturel" des fonctions continues n'est malheureusement pas un espace de Hilbert.) Pour tenir compte d'éventuelles contraintes sur la commande, on introduit un convexe fermé non vide  $K$  de  $\mathbb{R}^M$  qui représente l'ensemble des commandes admissibles. Le problème de minimisation est donc

$$\inf_{v(t) \in L^2([0, T[; K)} J(v). \quad (10.72)$$

Commençons par vérifier que le système (10.71) est bien posé.

**Lemme 10.4.3** *On suppose que  $f(t) \in L^2([0, T[; \mathbb{R}^N)$  et  $v(t) \in L^2([0, T[; K)$ . Alors (10.71) admet une unique solution  $y(t) \in H^1([0, T[; \mathbb{R}^N)$  qui est, de plus, continue sur  $[0, T]$ .*

**Démonstration.** Ce résultat d'existence et d'unicité est bien connu si  $f$  et  $v$  sont continues. Il n'est pas plus difficile dans le cadre  $L^2$ . On utilise la formule explicite de représentation de la solution

$$y(t) = \exp(tA)y_0 + \int_0^t \exp((t-s)A)(Bv + f)(s) ds$$

qui permet de vérifier l'existence et l'unicité de  $y$  dans  $H^1([0, T[; \mathbb{R}^N)$ . Le Lemme 4.3.3 nous dit enfin que  $y$  est continue sur  $[0, T]$ .  $\square$

On peut alors montrer l'existence et l'unicité de la commande optimale.

**Proposition 10.4.4** *Il existe un unique  $u \in L^2([0, T[; K)$  qui minimise (10.72). Cette commande optimale  $u$  est caractérisée par*

$$\begin{aligned} \int_0^T Q(y_u - z) \cdot (y_v - y_u) dt &+ \int_0^T Ru \cdot (v - u) dt \\ &+ D(y_u(T) - z_T) \cdot (y_v(T) - y_u(T)) \geq 0, \end{aligned} \quad (10.73)$$

pour tout  $v \in L^2([0, T[; K)$ , où  $y_v$  désigne la solution de (10.71) associée à la commande  $v$ .

**Démonstration.** On commence par remarquer que  $v \rightarrow y$  est une fonction affine. En effet, par linéarité de (10.71) on a  $y_v = \tilde{y}_v + \tilde{y}$ , où  $\tilde{y}_v$  est solution de

$$\begin{cases} \frac{d\tilde{y}_v}{dt} = A\tilde{y}_v + Bv \text{ pour } 0 \leq t \leq T \\ \tilde{y}_v(0) = 0 \end{cases} \quad (10.74)$$

et  $\hat{y}$  est solution de

$$\begin{cases} \frac{d\hat{y}}{dt} = A\hat{y} + f \text{ pour } 0 \leq t \leq T \\ \hat{y}(0) = y_0 \end{cases}$$

Il est clair que  $\hat{y}$  ne dépend pas de  $v$  et  $v \rightarrow \tilde{y}_v$  est linéaire continue de  $L^2([0, T]; K)$  dans  $H^1([0, T]; \mathbb{R}^N)$ . Par conséquent,  $v \rightarrow J(v)$  est une fonction quadratique positive de  $v$  (plus précisément, la somme d'une forme quadratique et d'une fonction affine), donc  $J$  est convexe, et même fortement convexe car la matrice  $R$  est définie positive. Comme  $L^2([0, T]; K)$  est un convexe fermé non vide, le Théorème 9.2.6 permet de conclure à l'existence et à l'unicité du point de minimum  $u$  de (10.72). D'autre part, la condition d'optimalité nécessaire et suffisante du Théorème 10.2.1 est  $\langle J'(u), v - u \rangle \geq 0$ . Pour calculer le gradient, la méthode la plus sûre et la plus simple est de calculer

$$\lim_{\epsilon \rightarrow 0} \frac{J(u + \epsilon w) - J(u)}{\epsilon} = \langle J'(u), w \rangle.$$

Comme  $J(v)$  est quadratique le calcul est très simple puisque  $y_{u+\epsilon w} = y_u + \epsilon \tilde{y}_w$ . On obtient aisément (10.73) en remarquant que  $y_u - y_v = \tilde{y}_u - \tilde{y}_v$ .  $\square$

**Remarque 10.4.5** En explicitant la condition d'optimalité de (10.72) on a en fait calculé le gradient  $J'(w)$  pour tout  $w \in L^2[0, T[$  (et pas seulement pour le minimum  $u$ ), ce qui est utile pour les méthodes numériques de minimisation (voir la Section 10.5). On a obtenu

$$\begin{aligned} \int_0^T J'(w)v \, dt &= \int_0^T R w \cdot v \, dt + \int_0^T Q(y_w - z) \cdot \tilde{y}_v \, dt \\ &\quad + D(y_w(T) - z_T) \cdot \tilde{y}_v(T), \end{aligned} \quad (10.75)$$

où  $v$  est une fonction quelconque de  $L^2[0, T[$ .  $\bullet$

La condition nécessaire et suffisante d'optimalité (10.73) est en fait inexploitable! En effet, pour tester l'optimalité de  $u$  il est nécessaire pour chaque fonction test  $v$  de calculer l'état correspondant  $y_v$ . Une autre façon de voir cette difficulté est l'impossibilité d'obtenir une expression explicite de  $J'(u)$  à partir de (10.75). Pour contourner cette difficulté on a recouru à la notion d'état adjoint qui est une des idées les plus profondes de la théorie du contrôle optimal. Montrons comment procéder sur l'exemple étudié dans cette sous-section (nous donnerons l'idée générale dans la Remarque 10.4.8 ci-dessous). Pour le problème (10.72) on définit l'état adjoint  $p$  comme la solution unique de

$$\begin{cases} \frac{dp}{dt} = -A^*p - Q(y - z) \text{ pour } 0 \leq t \leq T \\ p(T) = D(y(T) - z_T) \end{cases} \quad (10.76)$$

où  $y$  est la solution de (10.71) pour la commande  $u$ . Le nom d'état adjoint vient de ce que c'est la matrice adjointe  $A^*$  qui apparaît dans (10.76). L'intérêt de l'état adjoint est qu'il permet d'obtenir une expression explicite de  $J'(u)$ .

**Théorème 10.4.6** La dérivée de  $J$  en  $u$  est donnée par

$$J'(u) = B^*p + Ru. \quad (10.77)$$

En particulier, la condition d'optimalité nécessaire et suffisante du problème (10.72) est

$$\int_0^T (B^*p + Ru) \cdot (v - u) dt \geq 0 \quad \forall v \in L^2[0, T[; K). \quad (10.78)$$

**Remarque 10.4.7** La formule (10.77) se généralise pour tout  $w \in L^2[0, T[$  en  $J'(w) = B^*p + Rw$ , quitte à calculer  $p$  par (10.76) en utilisant l'état  $y$  correspondant à la commande  $w$ . Le Théorème 10.4.6 donne une expression explicite du gradient au prix de la résolution supplémentaire du système adjoint (10.76). C'est une différence fondamentale avec la formule (10.75) qui, pour chaque fonction test  $v$ , nécessitait la résolution du système (10.71) avec la commande  $v$ . •

**Démonstration.** Soit  $p$  la solution de (10.76) et  $\tilde{y}_v$  celle de (10.74). L'idée est de multiplier (10.76) par  $\tilde{y}_v$  et (10.74) par  $p$ , d'intégrer par parties et de comparer les résultats. Plus précisément, on calcule la quantité suivante de deux manières différentes. Tout d'abord, en intégrant et en tenant compte des conditions initiales  $\tilde{y}_v(0) = 0$  et  $p(T) = D(y(T) - z_T)$ , on a

$$\int_0^T \left( \frac{dp}{dt} \cdot \tilde{y}_v + p \cdot \frac{d\tilde{y}_v}{dt} \right) dt = D(y(T) - z_T) \cdot \tilde{y}_v(T). \quad (10.79)$$

D'autre part, en utilisant les équations on obtient

$$\int_0^T \left( \frac{dp}{dt} \cdot \tilde{y}_v + p \cdot \frac{d\tilde{y}_v}{dt} \right) dt = - \int_0^T Q(y - z) \cdot \tilde{y}_v dt + \int_0^T Bv \cdot p dt. \quad (10.80)$$

On déduit de l'égalité entre (10.79) et (10.80) une simplification de l'expression (10.75) de la dérivée

$$\int_0^T J'(u)v dt = \int_0^T Ru \cdot v dt + \int_0^T Bv \cdot p dt,$$

ce qui donne les résultats (10.77) et (10.78). □

**Remarque 10.4.8** Comment a-t-on bien pu deviner le problème (10.76) qui définit l'état adjoint afin de simplifier l'expression de  $J'(v)$ ? Encore une fois, l'idée directrice est l'introduction d'un Lagrangien associé au problème de minimisation (10.72). On considère l'équation d'état (10.71) comme une contrainte entre deux variables indépendantes  $v$  et  $y$  et on définit le Lagrangien comme la somme de  $J(v)$  et de l'équation d'état multipliée par  $p$ , c'est-à-dire

$$\begin{aligned} \mathcal{L}(v, y, p) &= \int_0^T Rv(t) \cdot v(t) dt + \int_0^T Q(y - z)(t) \cdot (y - z)(t) dt \\ &\quad + D(y(T) - z_T) \cdot (y(T) - z_T) + \int_0^T p \cdot \left( -\frac{dy}{dt} + Ay + Bv + f \right) dt \\ &\quad - p(0) \cdot (y(0) - y_0), \end{aligned}$$

où  $p$  est le multiplicateur de Lagrange pour la contrainte (10.71) entre  $v$  et  $y$ . Formellement, les conditions d'optimalité de (10.72) s'obtiennent en disant que le Lagrangien est stationnaire, c'est-à-dire que

$$\frac{\partial \mathcal{L}}{\partial v} = \frac{\partial \mathcal{L}}{\partial y} = \frac{\partial \mathcal{L}}{\partial p} = 0.$$

La première dérivée donne la condition d'optimalité (10.77), la seconde donne l'équation vérifiée par l'état adjoint  $p$ , et la troisième l'équation vérifiée par l'état  $y$ . Insistons sur le fait que ce calcul est purement formel, mais qu'en général il donne la "bonne" équation de l'état adjoint. •

A partir du Théorème 10.4.6 on peut soit obtenir des propriétés qualitatives de la solution  $y$  et de la commande optimale  $u$  (voir l'Exercice 10.4.1), soit construire une méthode numérique de minimisation de (10.72) par un algorithme de type gradient. En l'absence de contraintes sur la commande, c'est-à-dire si  $K = \mathbb{R}^M$ , on peut aller encore plus loin dans l'analyse et trouver une "loi de commande" qui donne l'état adjoint  $p$  (et donc la commande optimale  $u = -R^{-1}B^*p$  en vertu de (10.78)).

**Proposition 10.4.9** *On suppose que  $K = \mathbb{R}^M$ ,  $f = 0$ ,  $z = 0$ , et  $z_T = 0$ . Soit  $P(t)$  la fonction de  $[0, T]$  à valeurs matricielles (d'ordre  $N$ ) qui est solution unique de*

$$\begin{cases} \frac{dP}{dt} = -A^*P - PA + PBR^{-1}B^*P - Q \text{ pour } 0 \leq t \leq T \\ P(T) = D \end{cases} \quad (10.81)$$

Alors  $P(t)$  est symétrique positive pour tout  $t \in [0, T]$  et on a  $p(t) = P(t)y(t)$ .

**Démonstration.** Pour tout  $t \in [0, T]$ , l'application qui, à  $y_0 \in \mathbb{R}^N$ , fait correspondre  $(y, p, u)(t)$  (où  $u = -R^{-1}B^*p$  est la commande optimale) est clairement linéaire. Elle est injective en vertu de l'Exercice 10.4.1, donc bijective de  $\mathbb{R}^N$  dans  $\mathbb{R}^N$ . Par conséquent, l'application  $y(t) \rightarrow p(t)$  est linéaire de  $\mathbb{R}^N$  dans  $\mathbb{R}^N$ , et il existe une matrice  $P(t)$  d'ordre  $N$  telle que  $p(t) = P(t)y(t)$ . En dérivant cette expression et en utilisant les équations (10.71) et (10.76), on obtient

$$\frac{dP}{dt}y = -A^*Py - PAy + PBR^{-1}B^*Py - Qy$$

pour tout  $y(t)$  (qui est quelconque puisque  $y_0$  l'est). On en déduit l'équation de (10.81). On obtient la condition finale  $P(T) = D$  car  $p(T) = Dy(T)$  et  $y(T)$  est quelconque dans  $\mathbb{R}^N$ . On admet l'unicité de la solution de (10.81). □

**Exercice 10.4.1** On suppose que  $K = \mathbb{R}^M$ ,  $f = 0$ ,  $z = 0$ , et  $z_T = 0$ . Montrer que, pour tout  $t \in [0, T]$ ,

$$p(t) \cdot y(t) = Dy(t) \cdot y(t) + \int_t^T Qy(s) \cdot y(s) ds + \int_t^T R^{-1}B^*p(s) \cdot B^*p(s) ds.$$

En déduire que s'il existe  $t_0 \in [0, T]$  tel que  $y(t_0) = 0$ , alors  $y(t) = p(t) = 0$  pour tout  $t \in [0, T]$ . Interpréter ce résultat.

**Exercice 10.4.2** Obtenir l'équivalent de la Proposition 10.4.4 et du Théorème 10.4.6 pour le système parabolique

$$\begin{cases} \frac{\partial y}{\partial t} - \Delta y = v + f \text{ dans } ]0, T[ \times \Omega \\ y = 0 \text{ sur } ]0, T[ \times \partial\Omega \\ y(0) = y_0 \text{ dans } \Omega \end{cases}$$

où  $y_0 \in L^2(\Omega)$ ,  $f \in L^2(]0, T[ \times \Omega)$ ,  $v \in L^2(]0, T[ \times \Omega)$  est la commande, et on minimise

$$\inf_{v \in L^2(]0, T[ \times \Omega)} J(v) = \int_0^T \int_{\Omega} v^2 dt dx + \int_0^T \int_{\Omega} |y - z|^2 dt dx + \int_{\Omega} |y(T) - z_T|^2 dx,$$

où  $z \in L^2(]0, T[ \times \Omega)$  et  $z_T \in L^2(\Omega)$ .

**Exercice 10.4.3** Généraliser l'exercice précédent à l'équation des ondes.

### 10.4.3 Optimisation des systèmes distribués

On résout ici l'Exemple 9.1.12 à propos du contrôle d'une membrane élastique déformée par une force extérieure  $f$  et fixée sur son contour. Le comportement de la membrane est modélisé par

$$\begin{cases} -\Delta u = f + v & \text{dans } \Omega \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (10.82)$$

où  $u$  est le déplacement vertical de la membrane et  $v$  est une force de contrôle qui sera la variable d'optimisation. On se donne un ouvert  $\omega \subset \Omega$  sur lequel agit le contrôle et deux fonctions limitatives  $v_{\min} \leq v_{\max}$  dans  $L^2(\omega)$ . Il est entendu dans tout ce qui suit que les fonctions de  $L^2(\omega)$  sont étendues par zéro dans  $\Omega \setminus \omega$ . On définit alors l'ensemble des contrôles admissibles

$$K = \{v \in L^2(\omega) \text{ tel que } v_{\min}(x) \leq v(x) \leq v_{\max}(x) \text{ dans } \omega \text{ et } v = 0 \text{ dans } \Omega \setminus \omega\}. \quad (10.83)$$

Si  $f \in L^2(\Omega)$ , le Théorème 5.2.2 nous dit qu'il existe une unique solution  $u \in H_0^1(\Omega)$ . On cherche à contrôler la membrane pour qu'elle adopte un déplacement  $u_0 \in L^2(\Omega)$ . On définit une fonction coût

$$J(v) = \frac{1}{2} \int_{\Omega} (|u - u_0|^2 + c|v|^2) dx, \quad (10.84)$$

où  $u$  est la solution de (10.82) (et donc dépend de  $v$ ) et  $c > 0$ . Le problème d'optimisation s'écrit

$$\inf_{v \in K} J(v). \quad (10.85)$$

**Proposition 10.4.10** *Il existe un unique contrôle optimal  $\bar{v} \in K$  pour le problème (10.85).*

**Démonstration.** On remarque que la fonction  $v \rightarrow u$  est affine de  $L^2(\Omega)$  dans  $H_0^1(\Omega)$ . Par conséquent,  $J(v)$  est une fonction quadratique positive de  $v$ , donc elle est convexe. Elle est même fortement convexe puisque  $J(v) \geq c\|v\|_{L^2(\Omega)}^2$ . D'autre part,  $K$  est un convexe fermé non vide de  $L^2(\Omega)$ . Par conséquent, le Théorème 9.2.6 permet de conclure à l'existence et à l'unicité du point de minimum de (10.85).  $\square$

Pour obtenir une condition nécessaire d'optimalité qui soit exploitable, on introduit, comme dans la Sous-section 10.4.2, un état adjoint  $p$  défini comme l'unique solution dans  $H_0^1(\Omega)$  de

$$\begin{cases} -\Delta p = u - u_0 & \text{dans } \Omega \\ p = 0 & \text{sur } \partial\Omega. \end{cases} \quad (10.86)$$

**Proposition 10.4.11** *La fonction coût  $J(v)$  est dérivable sur  $K$  et on a*

$$J'(v) = p + cv,$$

où  $p$  (qui dépend de  $v$ ) est donné par (10.86). Par conséquent, la condition nécessaire et suffisante d'optimalité pour le contrôle optimal  $\bar{v}$  est

$$-\Delta \bar{u} = f + \bar{v} \text{ dans } \Omega, \quad \bar{u} \in H_0^1(\Omega), \quad (10.87)$$

$$-\Delta \bar{p} = \bar{u} - u_0 \text{ dans } \Omega, \quad \bar{p} \in H_0^1(\Omega), \quad (10.88)$$

$$\bar{v} = \mathbb{I}_\omega P_{[v_{\min}(x), v_{\max}(x)]} \left( -\frac{\bar{p}}{c} \right), \quad (10.89)$$

où  $\mathbb{I}_\omega$  est la fonction indicatrice de  $\omega$  (c'est-à-dire qui vaut 1 dans  $\omega$  et 0 dans  $\Omega \setminus \omega$ ) et  $P_{[v_{\min}(x), v_{\max}(x)]}$  est l'opérateur de projection orthogonal sur le segment  $[v_{\min}(x), v_{\max}(x)]$  défini par  $P_{[v_{\min}(x), v_{\max}(x)]} w = \min \left( v_{\max}(x), \max \left( v_{\min}(x), w(x) \right) \right)$ .

**Démonstration.** Comme dans la Proposition 10.4.4, la méthode la plus sûre et la plus simple de calculer le gradient est

$$\lim_{\epsilon \rightarrow 0} \frac{J(v + \epsilon w) - J(v)}{\epsilon} = \int_{\Omega} J'(v) w \, dx.$$

Comme  $J(v)$  est quadratique le calcul est très simple et on obtient

$$\int_{\Omega} J'(v) w \, dx = \int_{\Omega} ((u - u_0) \tilde{u}_w + cvw) \, dx,$$

où  $\tilde{u}_w$  est donné par

$$\begin{cases} -\Delta \tilde{u}_w = w & \text{dans } \Omega \\ \tilde{u}_w = 0 & \text{sur } \partial\Omega. \end{cases} \quad (10.90)$$

Pour simplifier l'expression du gradient on utilise l'état adjoint pour cela : on multiplie (10.90) par  $p$  et (10.86) par  $\tilde{u}_w$  et on intègre par parties

$$\begin{aligned} \int_{\Omega} \nabla p \cdot \nabla \tilde{u}_w \, dx &= \int_{\Omega} (u - u_0) \tilde{u}_w \, dx \\ \int_{\Omega} \nabla \tilde{u}_w \cdot \nabla p \, dx &= \int_{\Omega} wp \, dx \end{aligned}$$

Par comparaison de ces deux égalités on en déduit que

$$\int_{\Omega} J'(v) w \, dx = \int_{\Omega} (p + cv) w \, dx,$$

d'où l'expression du gradient. La condition nécessaire et suffisante d'optimalité donnée par le Théorème 10.2.1 est

$$\int_{\Omega} (\bar{p} + c\bar{v}) (w - \bar{v}) \, dx \geq 0 \quad \forall w \in K. \quad (10.91)$$

En prenant  $w$  égal à  $\bar{v}$  partout sauf sur une petite boule dans  $\omega$ , puis en faisant tendre le rayon de cette boule vers zéro, on peut "localiser" (10.91) en (presque) tout point  $x$  de  $\omega$

$$(\bar{p}(x) + c\bar{v}(x)) (w(x) - \bar{v}(x)) \geq 0 \quad \forall w(x) \in [v_{\min}(x), v_{\max}(x)].$$



Cette dernière condition n'est que la définition de  $\bar{v}(x)$  comme projection orthogonale de  $-\bar{p}(x)/c$  sur le segment  $[v_{\min}(x), v_{\max}(x)]$  (voir le Théorème 12.1.10). Finalement, on obtient (10.89) en remarquant que le support des fonctions de  $K$  est restreint à  $\omega$ .  $\square$

**Remarque 10.4.12** Comme dans la Remarque 10.4.8 nous expliquons comment trouver la forme de (10.86) qui définit l'état adjoint. On introduit le Lagrangien associé au problème de minimisation (10.85) sous la contrainte que l'équation d'état (10.82) (qui relie les deux variables indépendantes  $v$  et  $u$ ) soit satisfaite

$$\mathcal{L}(v, u, p) = \frac{1}{2} \int_{\Omega} (|u - u_0|^2 + c|v|^2) dx + \int_{\Omega} p(\Delta u + f + v) dx,$$

où  $p$  est le multiplicateur de Lagrange pour la contrainte (10.82) entre  $v$  et  $u$ . Formellement, les conditions d'optimalité s'obtiennent en disant que le Lagrangien est stationnaire, c'est-à-dire que

$$\frac{\partial \mathcal{L}}{\partial v} = \frac{\partial \mathcal{L}}{\partial u} = \frac{\partial \mathcal{L}}{\partial p} = 0.$$

La première dérivée donne la condition d'optimalité (10.89), la seconde donne l'équation vérifiée par l'état adjoint  $p$ , et la troisième l'équation vérifiée par l'état  $u$ .  $\bullet$

## 10.5 Algorithmes numériques

### 10.5.1 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.

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 supposons que la fonction objectif à minimiser  $J$  est  $\alpha$ -convexe différentiable. Cette hypothèse d' $\alpha$ -convexité est assez forte, mais nous verrons plus loin 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, sans parler des **grosses** difficultés qui apparaissent lorsque l'on cherche à approcher le minimum d'une fonction non convexe ! Typiquement, ces algorithmes peuvent ne pas converger et osciller entre plusieurs points de minimum, ou bien pire ils peuvent converger vers un minimum local, très loin d'un minimum global (dans le cas non convexe, cf. la Proposition 9.2.3).

**Remarque 10.5.1** 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). •

### 10.5.2 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  $\alpha$ -convexe différentiable définie sur l'espace de Hilbert réel  $V$ , on considère le problème sans contrainte

$$\inf_{v \in V} J(v) . \quad (10.92)$$

D'après le Théorème 9.2.6 il existe une unique solution  $u$ , caractérisée d'après la Remarque 10.2.2 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

$$u^{n+1} = u^n - \mu^n w^n , \quad (10.93)$$

avec  $\mu^n > 0$  petit et  $w^n$  unitaire dans  $V$ , c'est avec le choix de la direction  $w_n = \frac{J'(u^n)}{\|J'(u^n)\|}$  que l'on peut espérer trouver la plus petite valeur de  $J(u^{n+1})$  (en l'absence d'autres informations comme les dérivées supérieures ou les itérées antérieures).

Cette remarque simple nous conduit, parmi les méthodes du type (10.93) qui sont appelées “méthodes de descente”, à l'algorithme de **gradient à pas optimal**, dans lequel on résout une succession de problème de minimisation à une seule variable réelle (même si  $V$  n'est pas de dimension finie). A partir de  $u^0$  quelconque dans  $V$ , on construit la suite  $(u^n)$  définie par

$$u^{n+1} = u^n - \mu^n J'(u^n) , \quad (10.94)$$

où  $\mu^n \in \mathbb{R}$  est choisi à chaque étape tel que

$$J(u^{n+1}) = \inf_{\mu \in \mathbb{R}} J(u^n - \mu J'(u^n)) . \quad (10.95)$$

Cet algorithme converge comme l'indique le résultat suivant.

**Théorème 10.5.2** *On suppose que  $J$  est  $\alpha$ -convexe différentiable et que  $J'$  est Lipschitzien sur tout borné de  $V$ , c'est-à-dire que*

$$\forall M > 0, \exists C_M > 0, \|v\| + \|w\| \leq M \Rightarrow \|J'(v) - J'(w)\| \leq C_M \|v - w\|. \quad (10.96)$$

*Alors l'algorithme de gradient à pas optimal converge : quel que soit  $u^0$ , la suite  $(u^n)$  définie par (10.94) et (10.95) converge vers la solution  $u$  de (10.92).*

**Démonstration.** La fonction  $f(\mu) = J(u^n - \mu J'(u^n))$  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 (10.95) a donc bien une solution unique, caractérisée par la condition  $f'(\mu) = 0$ , ce qui s'écrit aussi

$$\langle J'(u^{n+1}), J'(u^n) \rangle = 0. \quad (10.97)$$

Ceci montre que deux “directions de descente” consécutives sont orthogonales.

Puisque (10.97) implique que  $\langle J'(u^{n+1}), u^{n+1} - u^n \rangle = 0$ , on déduit de l' $\alpha$ -convexité de  $J$  que

$$J(u^n) - J(u^{n+1}) \geq \frac{\alpha}{2} \|u^n - u^{n+1}\|^2, \quad (10.98)$$

ce qui prouve que la suite  $J(u^n)$  est décroissante. Comme elle est minorée par  $J(u)$ , elle converge et (10.98) montre que  $u^{n+1} - u^n$  tend vers 0. D'autre part, l' $\alpha$ -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 (10.96) pour  $v = u^n$  et  $w = u^{n+1}$  et utilisant (10.97), 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' $\alpha$ -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 de l'algorithme.

□

**Remarque 10.5.3** Il est utile de noter l'intérêt pratique de la dernière inégalité de cette démonstration : outre la preuve de la convergence, elle donne une majoration aisément calculable de l'erreur  $u^n - u$ . •

### 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

$$u^{n+1} = u^n - \mu J'(u^n), \quad (10.99)$$

où  $\mu$  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 (10.95). Le résultat suivant montre sous quelles hypothèses on peut choisir le paramètre  $\mu$  pour assurer la convergence.

**Théorème 10.5.4** *On suppose que  $J$  est  $\alpha$ -convexe différentiable et que  $J'$  est Lipschitzien sur  $V$ , c'est-à-dire qu'il existe une constante  $C > 0$  telle que*

$$\|J'(v) - J'(w)\| \leq C\|v - w\| \quad \forall v, w \in V. \quad (10.100)$$

*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 (10.95) converge vers la solution  $u$  de (10.92).*

**Démonstration.** Posons  $v^n = u^n - u$ . Comme  $J'(u) = 0$ , on a  $v^{n+1} = v^n - \mu(J'(u^n) - J'(u))$ , d'où il vient

$$\begin{aligned} \|v^{n+1}\|^2 &= \|v^n\|^2 - 2\mu\langle J'(u^n) - J'(u), u^n - u \rangle + \mu^2 \|J'(u^n) - J'(u)\|^2 \\ &\leq (1 - 2\alpha\mu + C^2\mu^2) \|v^n\|^2, \end{aligned} \quad (10.101)$$

d'après (10.100) et l' $\alpha$ -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 (10.101). 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  $u^n$ .  $\square$

**Remarque 10.5.5** Une adaptation simple de la démonstration précédente, laissée au lecteur en guise d'exercice, permet de montrer la convergence en remplaçant (10.100) par l'hypothèse plus faible (10.96). Il faut noter aussi que, pour l'algorithme de gradient à pas fixe, à la différence du gradient à pas optimal, la suite  $J(u^n)$  n'est pas nécessairement monotone.  $\bullet$

**Remarque 10.5.6** Il existe de nombreux autres algorithmes de descente du type (10.93) 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 13.1.5, pour le cas particulier d'une fonctionnelle quadratique du type  $\frac{1}{2}Ax \cdot x - b \cdot x$ .  $\bullet$

**Remarque 10.5.7** Comment choisir entre les deux algorithmes de gradient que nous venons de voir, et plus généralement entre les différentes méthodes de minimisation numérique qui existent ? Un premier critère de choix concerne le coût de chaque itération. Par exemple, nous l'avons dit, chaque itération de l'algorithme de gradient à pas fixe est moins chère qu'une itération de gradient à pas optimal. Évidemment,

si l'on part de la même itérée  $u^n$ , une itération du gradient à pas optimal décroît plus la fonction coût qu'une itération du gradient à pas fixe. On en arrive donc au deuxième critère de choix, souvent plus déterminant, qui est celui de la **vitesse de convergence** de l'algorithme, qui fixe le nombre d'itérations nécessaires pour rendre l'erreur  $\|u^n - u\|$  inférieure à une tolérance  $\epsilon$  fixée a priori.

Par exemple, l'inégalité (10.101) montre que la convergence de l'algorithme de gradient à pas fixe est au moins géométrique, puisque

$$\|u^n - u\| \leq \gamma^n \|u^0 - u\| \quad \text{avec} \quad \gamma = \sqrt{1 - 2\alpha\mu + \mu^2 C^2}.$$

Cette remarque conduit d'ailleurs, sous réserve d'une analyse plus poussée, à préférer pour le paramètre  $\mu$  la valeur médiane  $\alpha/C^2$  dans l'intervalle  $]0, 2\alpha/C^2[$ , valeur qui minimise celle de  $\gamma$ . En fait, on peut montrer que la convergence des deux algorithmes étudiés ci-dessus est effectivement géométrique dans certains cas particulier (ce qui signifie que la quantité  $\|u^n - u\|^{1/n}$  a une limite finie, comprise strictement entre 0 et 1, lorsque  $n$  tend vers  $+\infty$ ). •

**Exercice 10.5.1** Pour  $V = \mathbb{R}^2$  et  $J(x, y) = ax^2 + by^2$  avec  $a, b > 0$ , montrer que l'algorithme de gradient à pas optimal converge en une seule itération si  $a = b$  ou si  $x^0 y^0 = 0$ , et que la convergence est géométrique dans les autres cas. Étudier aussi la convergence de l'algorithme de gradient à pas fixe : pour quelles valeurs du paramètre  $\mu$  la convergence se produit-elle, pour quelle valeur est-elle la plus rapide ?

### 10.5.3 Algorithmes de type gradient (cas avec contraintes)

On étudie maintenant la résolution de problèmes d'optimisation avec contraintes

$$\inf_{v \in K} J(v), \quad (10.102)$$

où  $J$  est une fonction  $\alpha$ -convexe différentiable définie sur  $K$ , sous-ensemble convexe fermé non vide de l'espace de Hilbert réel  $V$ . Le Théorème 9.2.6 assure alors l'existence et l'unicité de la solution  $u$  de (10.102), caractérisée d'après le Théorème 10.2.1 par la condition

$$\langle J'(u), v - u \rangle \geq 0 \quad \forall v \in K. \quad (10.103)$$

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 (10.102) avec contraintes à partir de la remarque suivante. Pour tout réel  $\mu > 0$ , (10.103) s'écrit

$$\langle u - (u - \mu J'(u)), v - u \rangle \geq 0 \quad \forall v \in K. \quad (10.104)$$

Notons  $P_K$  l'opérateur de projection sur l'ensemble convexe  $K$ , défini au Théorème 12.1.10 de projection sur un convexe (voir la Remarque 12.1.11). Alors, d'après ce théorème, (10.104) n'est rien d'autre que la caractérisation de  $u$  comme la projection orthogonale de  $u - \mu J'(u)$  sur  $K$ . Autrement dit,

$$u = P_K(u - \mu J'(u)) \quad \forall \mu > 0. \quad (10.105)$$

Il est facile de voir que (10.105) est en fait équivalent à (10.103), et caractérise donc la solution  $u$  de (10.102). L'algorithme de **gradient à pas fixe avec projection** (ou plus simplement de gradient projeté) est alors défini par l'itération

$$u^{n+1} = P_K(u^n - \mu J'(u^n)), \quad (10.106)$$

où  $\mu$  est un paramètre positif fixé.

**Théorème 10.5.8** *On suppose que  $J$  est  $\alpha$ -convexe différentiable et que  $J'$  est Lipschitzien sur  $V$  (de constante  $C$ , voir (10.100)). 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 (10.106) converge vers la solution  $u$  de (10.102).*

**Démonstration.** La démonstration reprend celle du Théorème 10.5.4 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

$$\exists \gamma \in ]0, 1[, \quad \|(v - \mu J'(v)) - (w - \mu J'(w))\| \leq \gamma \|v - w\|.$$

Puisque la projection  $P_K$  est faiblement contractante d'après (12.2), l'application  $v \mapsto P_K(v - \mu J'(v))$  est strictement contractante, ce qui prouve la convergence de la suite  $(u^n)$  définie par (10.106) vers la solution  $u$  de (10.102).  $\square$

**Exercice 10.5.2** 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 (10.105) 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, en dimension finie (pour  $V = \mathbb{R}^M$ ), les sous-ensembles  $K$  de la forme

$$K = \prod_{i=1}^M [a_i, b_i] \quad (10.107)$$

(avec éventuellement  $a_i = -\infty$  ou  $b_i = +\infty$  pour certains indices  $i$ ). En effet, il est alors facile de voir que, si  $x = (x_1, x_2, \dots, x_M) \in \mathbb{R}^M$ ,  $y = P_K(x)$  a pour composantes

$$y_i = \min(\max(a_i, x_i), b_i) \quad \text{pour } 1 \leq i \leq M, \quad (10.108)$$

autrement dit, il suffit juste de “tronquer” les composantes de  $x$ . Cette propriété simple, jointes aux remarques sur la dualité énoncée dans la Section 10.3, 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 (10.107), 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

$$\inf_{F(v) \leq 0} J(v), \quad (10.109)$$

où  $J$  est une fonctionnelle convexe définie sur  $V$  et  $F$  une fonction convexe de  $V$  sur  $\mathbb{R}^M$ . Sous les hypothèses du Théorème de Kuhn et Tucker 10.3.4, la résolution de (10.109) revient à trouver un point-selle  $(u, p)$  du Lagrangien

$$\mathcal{L}(v, q) = J(v) + q \cdot F(v), \quad (10.110)$$

sur  $V \times (\mathbb{R}_+)^M$ . À partir de la Définition 10.3.1 du point-selle

$$\forall q \in (\mathbb{R}_+)^M \quad \mathcal{L}(u, q) \leq \mathcal{L}(u, p) \leq \mathcal{L}(v, p) \quad \forall v \in V, \quad (10.111)$$

on déduit que  $(p - q) \cdot F(u) \geq 0$  pour tout  $q \in (\mathbb{R}_+)^M$ , d'où on tire, pour tout réel  $\mu > 0$ ,

$$(p - q) \cdot (p - (p + \mu F(u))) \leq 0 \quad \forall q \in (\mathbb{R}_+)^M,$$

ce qui, d'après (12.1), montre que

$$p = P_{\mathbb{R}_+^M}(p + \mu F(u)) \quad \forall \mu > 0, \quad (10.112)$$

$P_{\mathbb{R}_+^M}$  désignant la projection de  $\mathbb{R}^M$  sur  $(\mathbb{R}_+)^M$ .

Au vu de cette propriété et de la seconde inégalité dans (10.111), 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

$$\begin{aligned} \mathcal{L}(u^n, p^n) &= \inf_{v \in V} \mathcal{L}(v, p^n), \\ p^{n+1} &= P_{\mathbb{R}_+^M}(p^n + \mu F(u^n)), \end{aligned} \quad (10.113)$$

$\mu$  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 10.5.9** *On suppose que  $J$  est  $\alpha$ -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*

$$\|F(v) - F(w)\| \leq C\|v - w\| \quad \forall v, w \in V, \quad (10.114)$$

*et qu'il existe un point-selle  $(u, p)$  du Lagrangien (10.110) 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 (10.113) converge vers la solution  $u$  du problème (10.109).*

**Démonstration.** Rappelons d'abord que l'existence d'une solution  $u$  de (10.109) découle de celle du point-selle  $(u, p)$  (voir la Proposition 10.3.2), alors que son unicité est une conséquence de l' $\alpha$ -convexité de  $J$ . De même,  $p^n$  étant fixé, le problème de minimisation dans (10.113) a bien une solution unique  $u^n$ . D'après l'Exercice 10.2.6, les inéquations d'Euler-satisfaites par  $u$  et  $u^n$  s'écrivent

$$\langle J'(u), v - u \rangle + p \cdot (F(v) - F(u)) \geq 0 \quad \forall v \in V, \quad (10.115)$$

$$\langle J'(u^n), v - u^n \rangle + p^n \cdot (F(v) - F(u^n)) \geq 0 \quad \forall v \in V. \quad (10.116)$$

Prenant successivement  $v = u^n$  dans (10.115) et  $v = u$  dans (10.116) et additionnant, on obtient

$$\langle J'(u) - J'(u^n), u^n - u \rangle + (p - p^n) \cdot (F(u^n) - F(u)) \geq 0,$$

d'où en utilisant l' $\alpha$ -convexité de  $J$  et en posant  $r^n = p^n - p$

$$r^n \cdot (F(u^n) - F(u)) \leq -\alpha\|u^n - u\|^2. \quad (10.117)$$

D'autre part, la projection  $P_{\mathbb{R}_+^M}$  étant faiblement contractante d'après (12.2), en soustrayant (10.112) à (10.113) on obtient

$$\|r^{n+1}\| \leq \|r^n + \mu(F(u^n) - F(u))\|,$$

soit

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + 2\mu r^n \cdot (F(u^n) - F(u)) + \mu^2 \|F(u^n) - F(u)\|^2.$$

Utilisant (10.114) et (10.117), il vient

$$\|r^{n+1}\|^2 \leq \|r^n\|^2 + (C^2\mu^2 - 2\mu\alpha)\|u^n - u\|^2.$$



Si  $0 < \mu < 2\alpha/C^2$ , on peut trouver  $\beta > 0$  tel que  $C^2\mu^2 - 2\mu\alpha < -\beta$ , d'où

$$\beta \|u^n - u\|^2 \leq \|r^n\|^2 - \|r^{n+1}\|^2. \quad (10.118)$$

Ceci montre alors que la suite  $\|r^n\|^2$  est décroissante : le membre de droite de (10.118) tend donc vers 0, ce qui entraîne que  $u^n$  tend vers  $u$ .  $\square$

Ainsi, l'algorithme d'Uzawa permet d'approcher la solution de (10.109) en remplaçant ce problème avec contraintes par une suite de problèmes de minimisation sans contraintes (10.113). A chaque itération, la détermination de  $p^n$  est élémentaire, puisque d'après (10.108) l'opérateur de projection  $P_{\mathbb{R}_+^M}$  est une simple troncature à zéro des composantes négatives. Il faut aussi noter que le Théorème 10.5.9 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 10.3.12 et l'Exercice 10.5.3 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 (10.109) s'écrit

$$\sup_{q \geq 0} \mathcal{G}(q), \quad (10.119)$$

où, par définition

$$\mathcal{G}(q) = \inf_{v \in V} \mathcal{L}(v, q), \quad (10.120)$$

et que le multiplicateur de Lagrange  $p$  est une solution du problème dual (10.119). 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 (10.120). 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 (10.113) peut s'écrire  $p^{n+1} = P_{\mathbb{R}_+^M}(p^n + \mu \mathcal{G}'(p^n))$  (le changement de signe par rapport à (10.106) vient du fait que le problème dual (10.119) 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 10.5.3** Appliquer l'algorithme d'Uzawa au problème de la Remarque 10.3.12 (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 10.3.12, montrer que la suite  $p^n$  converge vers  $p$ .

### 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 de "méthode" ou "d'algorithme" car la pénalisation des contraintes n'est pas une méthode à proprement parler. La résolution effective des problèmes sans contraintes que nous allons construire doit être réalisée à l'aide de l'un des algorithmes de la Sous-section 10.5.2. Cette résolution peut d'ailleurs soulever des difficultés, car le problème "pénalisé" (10.122) est souvent "mal conditionné" (voir la Sous-section 13.1.2).

Nous nous plaçons pour simplifier dans le cas où  $V = \mathbb{R}^N$ , et nous considérons de nouveau le problème de minimisation convexe

$$\inf_{F(v) \leq 0} J(v), \quad (10.121)$$

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  $\mathbb{R}^M$ .

Pour  $\varepsilon > 0$ , nous introduisons alors le problème sans contraintes

$$\inf_{v \in \mathbb{R}^N} \left( J(v) + \frac{1}{\varepsilon} \sum_{i=1}^M [\max(F_i(v), 0)]^2 \right), \quad (10.122)$$

dans lequel on dit que les contraintes  $F_i(v) \leq 0$  sont "pénalisées". On peut alors énoncer le résultat suivant, qui montre que, pour  $\varepsilon$  petit, le problème (10.122) "approche bien" le problème (10.121).

**Proposition 10.5.10** *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*

$$K = \{v \in \mathbb{R}^N, \quad F_i(v) \leq 0 \quad \forall i \in \{1, \dots, M\}\}$$

*est non vide. En notant  $u$  l'unique solution de (10.121) et, pour  $\varepsilon > 0$ ,  $u_\varepsilon$  l'unique solution de (10.122), on a alors*

$$\lim_{\varepsilon \rightarrow 0} u_\varepsilon = u.$$

**Démonstration.** L'ensemble  $K$  étant convexe fermé, l'existence et l'unicité de  $u$  découlent du Théorème 9.1.3 et de la stricte convexité de  $J$ . De plus, la fonction  $G(v) = \sum_{i=1}^M [\max(F_i(v), 0)]^2$  est continue et convexe puisque la fonction de  $\mathbb{R}$  dans  $\mathbb{R}$  qui à  $x$  associe  $\max(x, 0)^2$  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  $u_\varepsilon$ . Comme  $G(u) = 0$ , on peut écrire

$$J_\varepsilon(u_\varepsilon) = J(u_\varepsilon) + \frac{G(u_\varepsilon)}{\varepsilon} \leq J_\varepsilon(u) = J(u). \quad (10.123)$$

Ceci montre que

$$J(u_\varepsilon) \leq J_\varepsilon(u_\varepsilon) \leq J(u), \quad (10.124)$$

et donc que  $u_\varepsilon$  est borné d'après la condition "infinie à l'infini". On peut donc extraire de la famille  $(u_\varepsilon)$  une suite  $(u_{\varepsilon_k})$  qui converge vers une limite  $u_*$  lorsque  $\varepsilon_k$  tend vers 0. On a alors  $0 \leq G(u_{\varepsilon_k}) \leq \varepsilon_k(J(u) - J(u_{\varepsilon_k}))$  d'après (10.123). Passant à la limite, on obtient  $G(u_*) = 0$ , qui montre que  $u_* \in K$ . Comme (10.124) implique que  $J(u_*) \leq J(u)$ , on a alors  $u_* = u$ , ce qui conclut la démonstration, toutes les suites extraites  $(u_{\varepsilon_k})$  convergeant vers la même limite  $u$ .  $\square$

**Exercice 10.5.4** En plus des hypothèses de la Proposition 10.5.10, on suppose que les fonctions  $J$  et  $F_1, \dots, F_M$  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 10.2.13. Enfin, on suppose que les vecteurs  $(F'_i(u))_{i \in I(u)}$  sont linéairement indépendants, ce qui assure l'unicité des multiplicateurs de Lagrange  $\lambda_1, \dots, \lambda_M$  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  $i \in \{1, \dots, M\}$

$$\lim_{\varepsilon \rightarrow 0} \left[ \frac{2}{\varepsilon} \max(F_i(u_\varepsilon), 0) \right] = \lambda_i.$$

**Remarque 10.5.11** Nous verrons à la Sous-section 11.2.3 une autre méthode de pénalisation par introduction de fonctions "barrières".  $\bullet$

### 10.5.4 Méthode de Newton

On se place 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

$$F(u) = F(v) + F'(v)(u - v) + \mathcal{O}(\|u - v\|^2),$$

c'est-à-dire

$$u = v - (F'(v))^{-1} F(v) + \mathcal{O}(\|v - u\|^2).$$

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

$$u^{n+1} = u^n - (F'(u^n))^{-1} F(u^n) \quad \text{pour } n \geq 0. \quad (10.125)$$

Rappelons que l'on ne calcule pas l'inverse de la matrice  $F'(u^n)$  dans (10.125) mais que l'on résout un système linéaire par l'une des méthodes exposées à la Section 13.1.

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

$$J(w) = J(v) + J'(v) \cdot (w - v) + \frac{1}{2} J''(v)(w - v) \cdot (w - v) + \mathcal{O}(\|w - v\|^3), \quad (10.126)$$

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 (10.126) est donné par  $w = v - (J''(v))^{-1} J'(v)$  si la matrice  $J''(v)$  est définie positive. On retrouve alors la formule itérative (10.125).

L'avantage principal de la méthode de Newton est sa convergence bien plus rapide que les méthodes précédentes.

**Proposition 10.5.12** *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 (10.125) converge, c'est-à-dire que la suite  $(u^n)$  converge vers  $u$ , et il existe une constante  $C > 0$  telle que*

$$\|u^{n+1} - u\| \leq C \|u^n - u\|^2. \quad (10.127)$$

**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  $\epsilon$ . 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 (10.125)

$$u^{n+1} - u = u^n - u - (F'(u^n))^{-1} (F(u^n) - F(u))$$

qui, par développement de Taylor autour de  $u^n$ , devient

$$u^{n+1} - u = (F'(u^n))^{-1} \mathcal{O}(\|u^n - u\|^2).$$

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  $\epsilon$ ) telle que

$$\|u^{n+1} - u\| \leq C \|u^n - u\|^2. \quad (10.128)$$

Si  $\epsilon$  est suffisamment petit de manière à ce que  $C\epsilon \leq 1$ , on déduit de (10.128) que  $u^{n+1}$  reste dans la boule de centre  $u$  et de rayon  $\epsilon$ . Cela permet de vérifier par récurrence l'hypothèse que  $\|u - u^n\| \leq \epsilon$  pour tout  $n \geq 0$ , et (10.128) est bien la conclusion désirée.  $\square$

**Remarque 10.5.13** Bien sûr, il faut conserver à l'esprit que chaque itération de la méthode de Newton (10.125) nécessite la résolution d'un système linéaire, ce qui est coûteux. De plus, la convergence rapide (dite "quadratique") donnée par (10.127) 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 avons 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 (10.127) 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 (10.127) la rend cependant particulièrement intéressante. •

**Remarque 10.5.14** 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 (10.125) par

$$u^{n+1} = u^n - S^n F(u^n) \quad \text{pour } n \geq 0.$$

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 [6] et [15]. •

On peut adapter la méthode de Newton à la minimisation d'une fonction  $J$  avec des contraintes d'égalité. Soit  $J$  une fonction de classe  $C^3$  de  $\mathbb{R}^N$  dans  $\mathbb{R}$ ,  $G = (G_1, \dots, G_M)$  une fonction de classe  $C^3$  de  $\mathbb{R}^N$  dans  $\mathbb{R}^M$  (avec  $M \leq N$ ), et soit  $u$  un minimum local de

$$\min_{v \in \mathbb{R}^N, G(v)=0} J(v). \quad (10.129)$$

Si les vecteurs  $(G'_1(u), \dots, G'_M(u))$  sont linéairement indépendants, la condition nécessaire d'optimalité du Théorème 10.2.8 est

$$J'(u) + \sum_{i=1}^M \lambda_i G'_i(u) = 0, \quad G_i(u) = 0 \quad 1 \leq i \leq M. \quad (10.130)$$

où les  $\lambda_1, \dots, \lambda_M \in \mathbb{R}$  sont les multiplicateurs de Lagrange. On peut alors résoudre le système (10.130) de  $(N+M)$  équations à  $(N+M)$  inconnues  $(u, \lambda) \in \mathbb{R}^{N+M}$  par une méthode de Newton. On pose donc

$$F(u, \lambda) = \begin{pmatrix} J'(u) + \lambda \cdot G'(u) \\ G(u) \end{pmatrix},$$

dont la matrice dérivée est

$$F'(u, \lambda) = \begin{pmatrix} J''(u) + \lambda \cdot G''(u) & -(G'(u))^* \\ G'(u) & 0 \end{pmatrix}.$$

On peut alors appliquer l'algorithme de Newton (10.125) à cette fonction  $F(u, \lambda)$  si la matrice  $F'(u, \lambda)$  est inversible. Nous allons voir que cette condition est "naturelle" au sens où elle correspond à une version un peu plus forte de la condition d'optimalité d'ordre 2 de la Proposition 10.2.11. La matrice  $F'(u, \lambda)$  est inversible si elle est injective. Soit  $(w, \mu)$  un élément de son noyau

$$\begin{cases} J''(u)w + \lambda \cdot G''(u)w + (G'(u))^* \mu = 0 \\ G'_i(u) \cdot w = 0 \text{ pour } 1 \leq i \leq M \end{cases}$$

On en déduit que  $w \in \text{Ker} G'(u) = \bigcap_{i=1}^M \text{Ker} G'_i(u)$  et  $(J''(u) + \lambda \cdot G''(u))w \in \text{Im}(G'(u))^*$ . Or  $\text{Im}(G'(u))^* = [\text{Ker} G'(u)]^\perp$ . Par conséquent, si on suppose que

$$(J''(u) + \lambda \cdot G''(u))(w, w) > 0 \quad \forall w \in \text{Ker} G'(u), w \neq 0, \quad (10.131)$$

la matrice  $F'(u, \lambda)$  est inversible. On remarque que (10.131) est l'inégalité stricte dans la condition d'optimalité d'ordre 2 de la Proposition 10.2.11. Il est donc naturel de faire l'hypothèse (10.131) qui permet d'utiliser l'algorithme de Newton. On peut ainsi démontrer la convergence de cette méthode (voir [6]). Il est intéressant d'interpréter cet algorithme comme une méthode de minimisation. On introduit le Lagrangien  $\mathcal{L}(v, \mu) = J(v) + \mu \cdot G(v)$ , ses dérivées par rapport à  $v$ ,  $\mathcal{L}'$  et  $\mathcal{L}''$ , et on vérifie que l'équation

$$(u^{n+1}, \lambda^{n+1}) = (u^n, \lambda^n) - (F'(u^n, \lambda^n))^{-1} F(u^n, \lambda^n)$$

est la condition d'optimalité pour que  $u^{n+1}$  soit un point de minimum du problème quadratique à contraintes affines

$$\min_{\substack{w \in \mathbb{R}^N \\ G(u^n) + G'(u^n) \cdot (w - u^n) = 0}} Q^n(w), \quad (10.132)$$

avec

$$Q^n(w) = \left( \mathcal{L}(u^n, \lambda^n) + \mathcal{L}'(u^n, \lambda^n) \cdot (w - u^n) + \frac{1}{2} \mathcal{L}''(u^n, \lambda^n)(w - u^n) \cdot (w - u^n) \right),$$

et  $\lambda^{n+1}$  est le multiplicateur de Lagrange associé au point de minimum de (10.132). On remarque que dans (10.132) on a effectué un développement de Taylor à l'ordre deux en  $w$  sur le Lagrangien  $\mathcal{L}(w, \lambda^n)$  et on a linéarisé la contrainte  $G(w)$  autour du point  $u^n$ .

**Remarque 10.5.15** Dans (10.132) on a utilisé une approximation quadratique du Lagrangien et non pas de la fonction  $J$ . On pourrait essayer de se contenter d'une méthode itérative de résolution de l'approximation quadratique à contraintes affines suivante

$$\min_{\substack{w \in \mathbb{R}^N \\ G(v) + G'(v) \cdot (w - v) = 0}} \left( J(v) + J'(v) \cdot (w - v) + \frac{1}{2} J''(v)(w - v) \cdot (w - v) \right). \quad (10.133)$$

Malheureusement la méthode basée sur (10.133) peut ne pas converger ! En particulier, il n'est pas évident que le Hessien  $J''(v)$  soit définie positif sur l'espace des contraintes (c'est le Hessien du Lagrangien qui est positif comme l'affirme la condition d'optimalité d'ordre 2 de la Proposition 10.2.11). •

## Chapitre 11

# MÉTHODES DE LA RECHERCHE OPÉRATIONNELLE

(Rédigé en collaboration avec Stéphane Gaubert)

### 11.1 Introduction

Dans ce chapitre, nous présentons divers outils de la RO (recherche opérationnelle). Dans RO, le mot “opérationnel” s’est d’abord entendu au sens propre : la RO est née, en grande partie, des problèmes de planification qui se sont posés pendant la seconde guerre mondiale et peu après. Ainsi G. Dantzig, l’inventeur de l’algorithme du simplexe, était conseiller pour l’armée de l’air américaine, et la planification du pont aérien sur Berlin en 1948 est une application célèbre de la programmation linéaire (voir [34] pour plus de détails). Le domaine s’est depuis considérablement développé et civilisé... Les problèmes de RO abondent dans l’industrie et les services : on peut citer par exemple les problèmes d’emploi du temps (pour les équipages de compagnies aériennes, pour les employés d’un centre d’appel, etc.), de tournées de véhicules, de routage dans les réseaux, de localisation d’entrepôts, de gestion de stock, d’ordonnancement d’atelier, ... La RO emprunte des outils à plusieurs domaines scientifiques : optimisation continue et optimisation combinatoire, mais aussi aux mathématiques discrètes et en particulier à l’algorithmique des graphes ; à l’informatique, d’une part via la théorie de la complexité, qui permet de discerner les problèmes “faciles”, c’est-à-dire résolubles en temps polynomial en la taille du problème, des problèmes difficiles, et d’autre part via la programmation par contraintes, ou “PPC” (qui traite de l’art d’énumérer intelligemment des solutions). Quelques questions de RO sont aussi reliées à la théorie des probabilités (par exemple pour la compréhension d’algorithmes d’optimisation stochastiques tels que le recuit simulé), à l’automatique ou à la théorie des

jeux (pour les problèmes de décision dynamiques). Une part importante de l'activité en RO relève par ailleurs de l'art du praticien (modélisation, conception d'heuristiques, etc.). Notre propos dans ce chapitre n'est pas de présenter la RO, mais plutôt d'initier à une partie mathématisée du domaine, l'optimisation combinatoire, qu'unissent des liens profonds à l'optimisation continue traitée dans les chapitres précédents : les méthodes les plus efficaces pour la résolution exacte de problèmes combinatoire s'appuient souvent sur la programmation convexe, et en retour, considérer des problèmes et objets discrets (problèmes de flots, problèmes d'affectation, points extrémaux de polyèdres, réseaux électriques, Laplaciens de graphes, etc.), permet souvent de mieux comprendre les analogues de ces problèmes et objets qui interviennent en Analyse.

Nous présenterons successivement dans ce chapitre cinq grandes méthodes. La **programmation linéaire**, qui fait l'objet de la Section 11.2, permet de résoudre efficacement les problèmes d'optimisation continue où les contraintes et le critère s'expriment linéairement, qui abondent en RO. Parfois, il est indispensable de trouver une solution entière. Nous verrons dans la Section 11.3, à l'aide de la notion de **polyèdre entier**, que cela peut se faire sans augmenter la complexité dans certains cas spéciaux, qui incluent la classe importante des problèmes de flots. La Section 11.4 présente une autre méthode, la **programmation dynamique**, naturellement adaptée aux problèmes de plus court chemin, et qui sert aussi comme auxiliaire dans des problèmes combinatoires plus difficiles. La Section 11.5 donne un exemple d'**algorithme glouton** : les algorithmes gloutons ne sont optimaux que pour des problèmes très particuliers, mais ils sont souvent utiles pour produire des heuristiques. Nous verrons enfin dans la Section 11.6, que lorsque les outils précédents ne s'appliquent pas directement, il est souvent possible d'obtenir une solution optimale par **séparation et évaluation** ("branch and bound"), c'est-à-dire, par une exploration arborescente combinée à une approximation du problème.

## 11.2 Programmation linéaire

Nous n'avons encore rien dit sur l'Exemple 9.1.1 qui est typique d'une classe très large de problèmes, dits de programmation linéaire. Au regard de l'importance pratique de ces problèmes nous leur réservons maintenant une section entière.

### 11.2.1 Définitions et propriétés

On veut résoudre le problème suivant, dit **programme linéaire sous forme standard**,

$$\inf_{x \in \mathbb{R}^n \text{ tel que } Ax=b, x \geq 0} c \cdot x, \quad (11.1)$$

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  $\text{rg}(A) < m$ , 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 (11.1) 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 (11.1) 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)

$$\inf_{x \in \mathbb{R}^n \text{ tel que } Ax \geq b} c \cdot x. \quad (11.2)$$

Dans (11.2) 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 (11.2) est équivalent à

$$\inf_{(x, \lambda) \in \mathbb{R}^{(n+m)} \text{ tel que } Ax = b + \lambda, \lambda \geq 0} c \cdot x. \quad (11.3)$$

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 (11.2) est équivalent à

$$\inf_{(x^+, x^-, \lambda) \in \mathbb{R}^{(2n+m)} \text{ tel que } Ax^+ - Ax^- = b + \lambda, x^+ \geq 0, x^- \geq 0, \lambda \geq 0} c \cdot (x^+ - x^-). \quad (11.4)$$

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 (11.1).

Nous avons déjà donné une motivation concrète de la programmation linéaire au début du Chapitre 9 (voir l'Exemple 9.1.1). Considérons pour l'instant un exemple simple qui va nous permettre de comprendre quelques aspects essentiels d'un programme linéaire

$$\min_{\substack{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0 \\ 2x_1 + x_2 + 3x_3 = 6}} x_1 + 4x_2 + 2x_3. \quad (11.5)$$

Sur la Figure 11.1 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 (11.5) 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

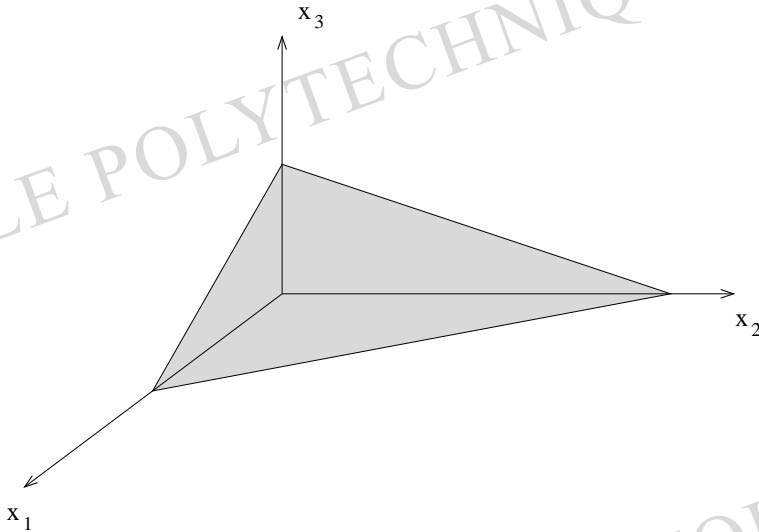


FIGURE 11.1 – Ensemble admissible pour l'exemple (11.5).

dans  $T$  donc ses extréma se trouvent sur le bord de  $T$ . Pour l'exemple (11.5) 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 : c'est précisément ce que fait (de manière intelligente) l'algorithme du simplexe que nous verrons dans la prochaine sous-section.

Pour établir cette propriété en toute généralité pour le programme linéaire standard (11.1), nous avons besoin de quelques définitions qui permettent de préciser le vocabulaire.

**Définition 11.2.1** *L'ensemble  $X_{ad}$  des vecteurs de  $\mathbb{R}^n$  qui satisfont les contraintes de (11.1), 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  $\bar{x} \in X_{ad}$  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  $\bar{x} = \theta y + (1 - \theta)z$ , alors  $y = z = \bar{x}$ .*

**Remarque 11.2.2** 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 (11.1) 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 11.2.3** *Il existe au moins une solution optimale (ou point de minimum) du programme linéaire standard (11.1) 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 (11.1). On introduit la matrice  $\mathcal{A}$  définie par

$$\mathcal{A} = \begin{pmatrix} c^* \\ A \end{pmatrix}.$$

La suite  $\mathcal{A}x^k$  appartient au cône suivant

$$C = \left\{ \sum_{i=1}^n x_i \mathcal{A}_i \text{ avec } x_i \geq 0 \right\},$$

où les  $\mathcal{A}_i$  sont les colonnes de la matrice  $\mathcal{A}$ . D'après le Lemme de Farkas 10.2.17 le cône  $C$  est fermé, ce qui implique que

$$\lim_{k \rightarrow +\infty} \mathcal{A}x^k = \begin{pmatrix} z_0 \\ b \end{pmatrix} \in C,$$

donc il existe  $\bar{x} \geq 0$  tel que

$$\begin{pmatrix} z_0 \\ b \end{pmatrix} = \begin{pmatrix} c \cdot \bar{x} \\ A\bar{x} \end{pmatrix},$$

et le minimum est atteint en  $\bar{x}$ . □

**Définition 11.2.4** *On appelle **base** associée à (11.1) 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  $(x_B, x_N)$  de sorte qu'on a*

$$Ax = Bx_B + Nx_N.$$

Les composantes du vecteur  $x_B$  sont appelées **variables de base** et celles de  $x_N$  **variables hors base**. Une **solution basique** (ou de base) est un vecteur  $x \in X_{ad}$  tel que  $x_N = 0$ . Si en plus l'une des composantes de  $x_B$  est nulle, on dit que la solution basique est **dégénérée**.

La notion de solution basique correspond à celle de sommet de  $X_{ad}$ .

**Lemme 11.2.5** *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  $x = (x_1, \dots, x_m, 0, \dots, 0)$ ,  $A = (B, N)$  avec  $B = (b_1, \dots, b_m)$ , 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  $(a_1, \dots, a_k)$  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  $(y_{k+1}, \dots, y_n) = 0$ . Comme les composantes  $(x_1, \dots, x_k)$  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.  $\square$

Le résultat fondamental suivant nous dit qu'il est suffisant de chercher une solution optimale parmi les sommets du polyèdre  $X_{ad}$ .

**Proposition 11.2.6** *S'il existe une solution optimale du programme linéaire standard (11.1), alors il existe une solution optimale basique.*

**Démonstration.** La démonstration est très similaire à celle du Lemme 11.2.5. Soit  $x \in X_{ad}$  une solution optimale de (11.1). 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  $(a_1, \dots, a_k)$  est libre dans  $\mathbb{R}^m$ , alors  $x$  est une solution optimale basique. Si  $(a_1, \dots, a_k)$  est lié, alors il existe  $y \neq 0$  tel que

$$\sum_{i=1}^k y_i a_i = 0 \text{ et } (y_{k+1}, \dots, y_n) = 0.$$

Comme les composantes  $(x_1, \dots, x_k)$  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  $\epsilon$ . En partant de la valeur  $\epsilon = 0$ , si on augmente ou on diminue  $\epsilon$  on reste dans l'ensemble  $X_{ad}$  jusqu'à une valeur  $\epsilon_0$  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.  $\square$

**Remarque 11.2.7** En appliquant la Proposition 11.2.6 lorsque  $c = 0$  (toute solution admissible est alors optimale), on voit grâce au Lemme 11.2.5 que dès que  $X_{ad}$  est non-vide,  $X_{ad}$  a au moins un sommet. Cette propriété n'a pas lieu pour des polyèdres généraux (considérer un demi-plan de  $\mathbb{R}^2$ ).  $\bullet$

**Exercice 11.2.1** Résoudre le programme linéaire suivant

$$\max_{x_1 \geq 0, x_2 \geq 0} x_1 + 2x_2$$

sous les contraintes

$$\begin{cases} -3x_1 + 2x_2 & \leq 2, \\ -x_1 + 2x_2 & \leq 4, \\ x_1 + x_2 & \leq 5. \end{cases}$$

En pratique le nombre de sommets du polyèdre  $X_{ad}$  est gigantesque car il peut être exponentiel par rapport au nombre de variables. On le vérifie sur un exemple dans l'exercice suivant.

**Exercice 11.2.2** Montrer que l'on peut choisir la matrice  $A$  de taille  $m \times n$  et le vecteur  $b \in \mathbb{R}^m$  de telle façon que  $X_{ad}$  soit le cube unité  $[0, 1]^{n-m}$  dans le sous-espace affine de dimension  $n - m$  défini par  $Ax = b$ . En déduire que le nombre de sommets de  $X_{ad}$  est alors  $2^{n-m}$ .

## 11.2.2 Algorithme du simplexe

L'algorithme du simplexe est dû à G. Dantzig dans les années 1940. Il consiste à parcourir les sommets du polyèdre des solutions admissibles jusqu'à ce qu'on trouve une solution optimale (ce qui est garanti si le programme linéaire admet effectivement une solution optimale). L'algorithme du simplexe ne se contente pas d'énumérer tous les sommets, il décroît la valeur de la fonction  $c \cdot x$  en passant d'un sommet au suivant.

On considère le programme linéaire standard (11.1). Rappelons qu'un sommet (ou solution basique) de l'ensemble des solutions admissibles  $X_{ad}$  est caractérisé par une base  $B$  ( $m$  colonnes libres de  $A$ ). Après permutation de ses colonnes, on peut écrire

$$A = (B, N) \text{ et } x = (x_B, x_N),$$

de sorte qu'on a  $Ax = Bx_B + Nx_N$ . Toute solution admissible peut s'écrire  $x_B = B^{-1}(b - Nx_N) \geq 0$  et  $x_N \geq 0$ . Le sommet associé à  $B$  est défini (s'il existe) par

$\bar{x}_N = 0$  et  $\bar{x}_B = B^{-1}b \geq 0$ . Si on décompose aussi  $c = (c_B, c_N)$  dans cette base, alors on peut comparer le coût d'une solution admissible quelconque  $x$  avec celui de la solution basique  $\bar{x}$

$$c \cdot x - c \cdot \bar{x} = c_B \cdot B^{-1}(b - Nx_N) + c_N \cdot x_N - c_B \cdot B^{-1}b = (c_N - N^*(B^{-1})^*c_B) \cdot x_N. \quad (11.6)$$

On en déduit la condition d'optimalité suivante.

**Proposition 11.2.8** *Supposons que la solution basique associée à  $B$  est non dégénérée, c'est-à-dire que  $B^{-1}b > 0$ . Une condition nécessaire et suffisante pour que cette solution basique associée à  $B$  soit optimale est que*

$$\tilde{c}_N = c_N - N^*(B^{-1})^*c_B \geq 0. \quad (11.7)$$

Le vecteur  $\tilde{c}_N$  est appelé **vecteur des coûts réduits**.

**Démonstration.** Soit  $\bar{x}$  une solution basique non dégénérée associée à  $B$ . Si  $\tilde{c}_N \geq 0$ , alors pour toute solution admissible  $x$  (11.6) implique que

$$c \cdot x - c \cdot \bar{x} = \tilde{c}_N \cdot x_N \geq 0,$$

puisque  $x_N \geq 0$ . Donc la condition (11.7) est suffisante pour que  $\bar{x}$  soit optimal. Réciproquement, supposons qu'il existe une composante  $i$  de  $\tilde{c}_N$  qui soit strictement négative,  $(\tilde{c}_N \cdot e_i) < 0$ . Pour  $\epsilon > 0$  on définit alors un vecteur  $x(\epsilon)$  par  $x_N(\epsilon) = \epsilon e_i$  et  $x_B(\epsilon) = B^{-1}(b - Nx_N(\epsilon))$ . Par construction  $Ax(\epsilon) = b$  et, comme  $B^{-1}b > 0$ , pour des valeurs suffisamment petites de  $\epsilon$  on a  $x(\epsilon) \geq 0$ , donc  $x(\epsilon) \in X_{ad}$ . D'autre part,  $x(0) = \bar{x}$  et, comme  $\epsilon > 0$ , on a

$$c \cdot x(\epsilon) = c \cdot x(0) + \epsilon(\tilde{c}_N \cdot e_i) < c \cdot \bar{x},$$

ce qui montre que  $\bar{x}$  n'est pas optimal. Donc la condition (11.7) est nécessaire.  $\square$

**Remarque 11.2.9** Dans le cadre de la Proposition 11.2.8, si la solution basique considérée est dégénérée, la condition (11.7) reste suffisante mais n'est plus nécessaire.

•

On déduit de la Proposition 11.2.8 une méthode pratique pour décroître la valeur de la fonction coût  $c \cdot x$  à partir d'une solution basique  $\bar{x}$  (non dégénérée et non optimale). Comme  $\bar{x}$  est non-optimale, il existe une composante du vecteur des coûts réduits  $\tilde{c}_N$  telle que  $\tilde{c}_N \cdot e_i < 0$ . On définit alors  $x(\epsilon)$  comme ci-dessus. Puisque le coût décroît linéairement avec  $\epsilon$ , on a intérêt à prendre la plus grande valeur possible de  $\epsilon$  telle que l'on reste dans  $X_{ad}$ . C'est le principe de l'algorithme du simplexe que nous présentons maintenant.

### Algorithme du simplexe

- Initialisation (phase I) : on cherche une base initiale  $B^0$  telle que la solution basique associée  $x^0$  soit admissible

$$x^0 = \begin{pmatrix} (B^0)^{-1}b \\ 0 \end{pmatrix} \geq 0.$$

- Itérations (phase II) : à l'étape  $k \geq 0$ , on dispose d'une base  $B^k$  et d'une solution basique admissible  $x^k$ . On calcule le coût réduit  $\tilde{c}_N^k = c_N^k - (N^k)^*(B^k)^{-1}*c_B^k$ . Si  $\tilde{c}_N^k \geq 0$ , alors  $x^k$  est optimal et l'algorithme est fini. Sinon, il existe une variable hors-base d'indice  $i$  telle que  $(\tilde{c}_N \cdot e_i) < 0$ , et on note  $a_i$  la colonne correspondante de  $A$ . On pose

$$x^k(\epsilon) = (x_B^k(\epsilon), x_N^k(\epsilon)) \text{ avec } x_N^k(\epsilon) = \epsilon e_i, \quad x_B^k(\epsilon) = (B^k)^{-1}(b - \epsilon a_i).$$

- Soit on peut choisir  $\epsilon > 0$  aussi grand que l'on veut avec  $x^k(\epsilon) \in X_{ad}$ . Dans ce cas, le minimum du programme linéaire est  $-\infty$ .
- Soit il existe une valeur maximale  $\epsilon^k \geq 0$  et un indice  $j$  tels que la  $j$ -ème composante de  $x^k(\epsilon^k)$  s'annule. On obtient ainsi une nouvelle solution admissible basique

$$x^{k+1} = x^k(\epsilon^k),$$

correspondant à une nouvelle base  $B^{k+1}$  déduite de  $B^k$  en remplaçant sa  $j$ -ème colonne par la colonne  $a_i$ . La solution admissible  $x^{k+1}$  a un coût inférieur ou égal à celui de  $x^k$ .

Il reste un certain nombre de points pratiques à préciser dans l'algorithme du simplexe. Nous les passons rapidement en revue.

### Dégénérescence et cyclage

On a toujours  $c \cdot x^{k+1} \leq c \cdot x^k$ , mais il peut y avoir égalité si la solution admissible basique  $x^k$  est dégénérée, auquel cas on trouve que  $\epsilon^k = 0$  (si  $x^k$  n'est pas dégénérée, la démonstration de la Proposition 11.2.8 garantit une inégalité stricte). On a donc changé de base sans améliorer le coût : c'est le phénomène du cyclage qui peut empêcher l'algorithme de converger. Il existe des moyens de s'en prémunir, mais en pratique le cyclage n'apparaît jamais.

En l'absence de cyclage, l'algorithme du simplexe parcourt un sous-ensemble des sommets de  $X_{ad}$  en diminuant de façon stricte le coût. Comme il y a un nombre fini de sommets, l'algorithme doit nécessairement trouver un sommet optimal de coût minimal. On a donc démontré le résultat suivant.

**Lemme 11.2.10** *Si toutes solutions admissibles basiques  $x^k$  produites par l'algorithme du simplexe sont non dégénérées, alors l'algorithme converge en un nombre fini d'étapes.*

A priori le nombre d'itérations de l'algorithme du simplexe peut être aussi grand que le nombre de sommets (qui est exponentiel par rapport au nombre de variables  $n$ ; voir l'Exercice 11.2.2). Bien qu'il existe des exemples (académiques) où c'est effectivement le cas, en pratique cet algorithme converge en un nombre d'étapes qui est une fonction polynomiale de  $n$ .

### Choix du changement de base

S'il y a plusieurs composantes du vecteur coût réduit  $\bar{c}_N^k$  strictement négatives, il faut faire un choix dans l'algorithme. Plusieurs stratégies sont possibles, mais en général on choisit la plus négative.

### Initialisation

Comment trouver une solution admissible basique lors de l'initialisation? (Rappelons que la condition d'admissibilité  $x_B = B^{-1}b \geq 0$  n'est pas évidente en général.) Soit on en connaît une à cause de la structure du problème. Par exemple, pour le problème (11.4) qui possède  $m$  variables d'écart,  $-\text{Id}_m$  est une base de la matrice "globale" des contraintes d'égalité de (11.4). Si de plus  $b \leq 0$ , le vecteur  $(x_+^0, x_-^0, \lambda^0) = (0, 0, -b)$  est alors une solution admissible basique pour (11.4).

Dans le cas général, on introduit une nouvelle variable  $y \in \mathbb{R}^m$ , un nouveau vecteur coût  $k = (1, \dots, 1)$  et un nouveau programme linéaire

$$\inf_{\substack{x \geq 0, \quad y \geq 0 \\ Ax + y = b}} k \cdot y, \quad (11.8)$$

où on a préalablement multiplié par  $-1$  toutes les contraintes d'égalité correspondant à des composantes négatives de  $b$  de telles sortes que  $b \geq 0$ . Le vecteur  $(x^0, y^0) = (0, b)$  est une solution admissible basique pour ce problème. S'il existe une solution admissible du programme linéaire original (11.1), alors il existe au moins une solution optimale de (11.8) et toutes les solutions optimales  $(x, y)$  vérifient nécessairement  $y = 0$  et  $x$  est solution admissible de (11.1). En appliquant l'algorithme du simplexe à (11.8), on trouve ainsi une solution admissible basique pour (11.1) s'il en existe une. S'il n'en existe pas (c'est-à-dire si  $X_{ad} = \emptyset$ ), on le détecte car le minimum de (11.8) est atteint par un vecteur  $(x, y)$  avec  $y \neq 0$ .

### Inversion de la base

Tel que nous l'avons décrit l'algorithme du simplexe demande l'inversion à chaque étape de la base  $B^k$ , ce qui peut être très coûteux pour les problèmes de grande taille (avec beaucoup de contraintes puisque l'ordre de  $B^k$  est égal au nombre de contraintes). On peut tirer parti du fait que  $B^{k+1}$  ne diffère de  $B^k$  que par une colonne pour mettre au point une meilleure stratégie. En effet, si c'est la  $j$ -ème colonne qui



change, on a

$$B^{k+1} = B^k E^k \text{ avec } E^k = \begin{pmatrix} 1 & & l_1 & & \\ & \ddots & \vdots & & 0 \\ & & 1 & & \\ & & & l_j & \\ & & & \vdots & 1 \\ & 0 & & \vdots & \ddots \\ & & l_n & & & 1 \end{pmatrix},$$

et  $E^k$  est facile à inverser

$$(E^k)^{-1} = \frac{1}{l_j} \begin{pmatrix} 1 & & -l_1 & & \\ & \ddots & \vdots & & 0 \\ & & 1 & -l_{j-1} & \\ & & & 1 & \\ & & & -l_{j+1} & 1 \\ 0 & & & \vdots & \ddots \\ & & -l_n & & & 1 \end{pmatrix}.$$

On utilise donc la formule, sous forme factorisée,

$$(B^k)^{-1} = (E^{k-1})^{-1} (E^{k-2})^{-1} \dots (E^0)^{-1} (B^0)^{-1}.$$

**Exercice 11.2.3** Résoudre par l'algorithme du simplexe le programme linéaire

$$\min_{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0, x_5 \geq 0} x_1 + 2x_2$$

sous les contraintes

$$\begin{cases} -3x_1 + 2x_2 + x_3 = 2, \\ -x_1 + 2x_2 + x_4 = 4, \\ x_1 + x_2 + x_5 = 5. \end{cases}$$

**Exercice 11.2.4** Résoudre par l'algorithme du simplexe le programme linéaire

$$\min_{x_1 \geq 0, x_2 \geq 0} 2x_1 - x_2$$

sous les contraintes  $x_1 + x_2 \leq 1$  et  $x_2 - x_1 \leq 1/2$  (on pourra s'aider d'un dessin et introduire des variables d'écart).

**Exercice 11.2.5** Résoudre par l'algorithme du simplexe le programme linéaire

$$\min_{x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0} 3x_3 - x_4$$

sous les contraintes

$$\begin{cases} x_1 - 3x_3 + 3x_4 = 6, \\ x_2 - 8x_3 + 4x_4 = 4. \end{cases}$$

### 11.2.3 Algorithmes de points intérieurs

Depuis les travaux de Khachian et Karmarkar au début des années 1980, une nouvelle classe d'algorithmes, dits de points intérieurs, est apparu pour résoudre des programmes linéaires. Le nom de cette classe d'algorithmes vient de ce qu'au contraire de la méthode du simplexe (qui, parcourant les sommets, reste sur le bord du polyèdre  $X_{ad}$ ) ces algorithmes de points intérieurs évoluent à l'intérieur de  $X_{ad}$  et ne rejoignent son bord qu'à convergence. Nous allons décrire ici un de ces algorithmes que l'on appelle aussi **algorithme de trajectoire centrale**. Il y a deux idées nouvelles dans cette méthode : premièrement, on pénalise certaines contraintes à l'aide de potentiels ou fonctions "barrières"; deuxièmement, on utilise une méthode de Newton pour passer d'une itérée à la suivante.

Décrivons cette méthode sur le programme linéaire standard

$$\inf_{x \in \mathbb{R}^n \text{ tel que } Ax=b, x \geq 0} c \cdot x. \quad (11.9)$$

On définit un potentiel logarithmique pour  $x > 0$

$$\pi(x) = - \sum_{i=1}^n \log x_i. \quad (11.10)$$

Pour un paramètre de pénalisation  $\mu > 0$ , on introduit le problème strictement convexe

$$\min_{x \in \mathbb{R}^n \text{ tel que } Ax=b, x > 0} \mu \pi(x) + c \cdot x. \quad (11.11)$$

Remarquons qu'en pratique la contrainte  $x > 0$  n'en est pas une car elle n'est jamais active : quand on minimise (11.11) on ne peut pas "s'approcher" du bord de  $x > 0$  sous peine de faire "exploser" le potentiel  $\pi(x)$  vers  $+\infty$ .

Le principe de l'algorithme de trajectoire centrale est de minimiser (11.11) par une méthode de Newton pour des valeurs de plus en plus petites de  $\mu$ . En effet, lorsque  $\mu$  tend vers zéro, le problème pénalisé (11.11) tend vers le programme linéaire (11.9).

**Exercice 11.2.6** Montrer que, si  $X_{ad}$  est borné non vide, (11.11) admet une unique solution optimale  $x^\mu$ . Écrire les conditions d'optimalité et en déduire que, si (11.9) admet une unique solution optimale  $x^0$ , alors  $x^\mu$  converge vers  $x^0$  lorsque  $\mu$  tend vers zéro.

### 11.2.4 Dualité

La théorie de la dualité (déjà évoquée lors de la Sous-section 10.3.3) est très utile en programmation linéaire. Considérons à nouveau le programme linéaire standard

que nous appellerons primal (par opposition au dual)

$$\inf_{x \in \mathbb{R}^n \text{ tel que } Ax=b, x \geq 0} c \cdot x, \quad (11.12)$$

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 (11.12)

$$L(x, p) = c \cdot x + p \cdot (b - Ax), \quad (11.13)$$

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

$$G(p) = \begin{cases} p \cdot b & \text{si } A^*p - c \leq 0 \\ -\infty & \text{sinon.} \end{cases} \quad (11.14)$$

Le problème dual de (11.12) est donc

$$\sup_{p \in \mathbb{R}^m \text{ tel que } A^*p - c \leq 0} p \cdot b. \quad (11.15)$$

L'espace de solutions admissibles du problème dual (11.15) est noté

$$P_{ad} = \{p \in \mathbb{R}^m \text{ tel que } A^*p - c \leq 0\}.$$

Rappelons que l'espace de solutions admissibles de (11.12) est

$$X_{ad} = \{x \in \mathbb{R}^n \text{ tel que } Ax = b, x \geq 0\}.$$

Les programmes linéaires (11.12) et (11.15) sont dits en **dualité**. L'intérêt de cette notion vient du résultat suivant qui est un cas particulier du Théorème de dualité 10.3.11.

**Théorème 11.2.11** *Si (11.12) ou (11.15) a une valeur optimale finie, alors il existe  $\bar{x} \in X_{ad}$  solution optimale de (11.12) et  $\bar{p} \in P_{ad}$  solution optimale de (11.15) qui vérifient*

$$\left( \min_{x \in \mathbb{R}^n \text{ tel que } Ax=b, x \geq 0} c \cdot x \right) = c \cdot \bar{x} = \bar{p} \cdot b = \left( \max_{p \in \mathbb{R}^m \text{ tel que } A^*p - c \leq 0} p \cdot b \right) \quad (11.16)$$

*De plus,  $\bar{x}$  et  $\bar{p}$  sont solutions optimales de (11.12) et (11.15) si et seulement si elles vérifient les conditions d'optimalité de Kuhn et Tucker*

$$A\bar{x} = b, \bar{x} \geq 0, A^*\bar{p} - c \leq 0, \bar{x} \cdot (c - A^*\bar{p}) = 0. \quad (11.17)$$

*Si (11.12) ou (11.15) a une valeur optimale infinie, alors l'ensemble des solutions admissibles de l'autre problème est vide.*

**Remarque 11.2.12** Une conséquence immédiate du Théorème 11.2.11 de dualité est que, si  $x \in X_{ad}$  et  $p \in P_{ad}$  sont deux solutions admissibles de (11.12) et (11.15), respectivement, elles vérifient

$$c \cdot x \geq b \cdot p.$$

De même, si  $\bar{x} \in X_{ad}$  et  $\bar{p} \in P_{ad}$  vérifient

$$c \cdot \bar{x} = b \cdot \bar{p}$$

alors  $\bar{x}$  est solution optimale de (11.12) et  $\bar{p}$  de (11.15). Ces deux propriétés permettent de trouver facilement des bornes pour les valeurs optimales de (11.12) et (11.15), et de tester si un couple  $(\bar{x}, \bar{p})$  est optimal. •

**Démonstration.** Supposons que  $X_{ad}$  et  $P_{ad}$  sont non vides. 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 11.2.3. L'égalité (11.16) et la condition d'optimalité (11.17) sont alors une conséquence du Théorème de dualité 10.3.11.

Supposons maintenant que l'un des deux problèmes primal ou dual admet une valeur optimale finie. Pour fixer les idées, admettons qu'il s'agisse du problème dual (un argument symétrique fonctionne pour le problème primal). Alors, le Lemme 11.2.3 affirme qu'il existe une solution optimale  $\bar{p}$  de (11.15). Si  $X_{ad}$  n'est pas vide, on se retrouve dans la situation précédente ce qui finit la démonstration. Montrons donc que  $X_{ad}$  n'est pas vide en utilisant encore le Lemme de Farkas 10.2.17. Pour  $p \in \mathbb{R}^m$ , on introduit les vecteurs de  $\mathbb{R}^{m+1}$

$$\tilde{b} = \begin{pmatrix} b \\ -b \cdot \bar{p} \end{pmatrix} \quad \text{et} \quad \tilde{p} = \begin{pmatrix} p \\ 1 \end{pmatrix}.$$

On vérifie que  $\tilde{b} \cdot \tilde{p} = b \cdot p - b \cdot \bar{p} \leq 0$ , pour tout  $p \in P_{ad}$ . D'autre part, la condition  $p \in P_{ad}$  peut se réécrire

$$\tilde{p} \in C = \left\{ \tilde{p} \in \mathbb{R}^{m+1} \text{ tel que } \tilde{p}_{m+1} = 1, \tilde{A}^* \tilde{p} \leq 0 \right\} \quad \text{avec} \quad \tilde{A} = \begin{pmatrix} A \\ -c^* \end{pmatrix}.$$

Comme  $\tilde{b} \cdot \tilde{p} \leq 0$  pour tout  $\tilde{p} \in C$ , le Lemme de Farkas 10.2.17 nous dit qu'il existe  $\tilde{x} \in \mathbb{R}^n$  tel que  $\tilde{x} \geq 0$  et  $\tilde{b} = \tilde{A}\tilde{x}$ , c'est-à-dire que  $\tilde{x} \in X_{ad}$  qui n'est donc pas vide.

Finalement, supposons que la valeur optimale du problème primal est (11.12)  $-\infty$ . Si  $P_{ad}$  n'est pas vide, pour tout  $x \in X_{ad}$  et tout  $p \in P_{ad}$ , on a  $c \cdot x \geq b \cdot p$ . En prenant une suite minimisante dans  $X_{ad}$  on obtient  $b \cdot p = -\infty$ , ce qui absurde. Donc  $P_{ad}$  est vide. Un raisonnement similaire montre que, si la valeur optimale de (11.12) est infinie, alors  $X_{ad}$  est vide. □

L'intérêt de la dualité pour résoudre le programme linéaire (11.12) est multiple. D'une part, selon l'algorithme choisi, il peut être plus facile de résoudre le problème dual (11.15) (qui a  $m$  variables et  $n$  contraintes d'inégalités) que le problème primal (11.12) (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 (11.12) qui utilisent les deux formes primale et duale du programme linéaire.

**Exercice 11.2.7** 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 11.2.8** Trouver le problème dual de (11.12) 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 (11.15) 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 11.2.9** Vérifier que le problème dual de (11.15) est à nouveau (11.12).

**Exercice 11.2.10** Soit  $v \in \mathbb{R}^n$ ,  $c \in \mathbb{R}^n$ ,  $A$  une matrice  $m \times n$  et  $b \in \mathbb{R}^m$ . On considère le programme linéaire

$$\inf_{\substack{v \geq 0 \\ Av \leq b}} c \cdot v. \quad (11.18)$$

Montrer que le problème dual peut se mettre sous la forme suivante, avec  $q \in \mathbb{R}^m$

$$\sup_{\substack{q \geq 0 \\ A^*q \leq c}} b \cdot q. \quad (11.19)$$

Soient  $v$  et  $q$  des solutions admissibles de (11.18) et (11.19), respectivement. Montrer que  $v$  et  $q$  sont des solutions optimales si, et seulement si,

$$(c - A^*q) \cdot v = 0 \quad \text{et} \quad (b - Ae) \cdot q = 0. \quad (11.20)$$

Les deux égalités de (11.20) sont appelées **conditions des écarts complémentaires** (primales et duales, respectivement). Généraliser au cas où le problème primal comprend en outre des contraintes égalités.

### 11.3 Polyèdres entiers

Nous avons jusqu'ici traité de problèmes d'**optimisation continue** : la fonction à minimiser était différentiable, et l'ensemble des solutions admissibles était défini par l'intersection d'un nombre fini de contraintes inégalités, elles mêmes différentiables. L'**optimisation combinatoire**, au contraire, traite de problèmes pour lesquels l'ensemble des solutions admissibles est **discret**. Ainsi, dans le cas du problème d'affectation qui a fait l'objet de l'Exemple 9.1.2, l'ensemble des solutions admissibles était l'ensemble des permutations de  $n$  éléments. La difficulté des problèmes combinatoires est d'une part, que l'on ne peut énumérer l'ensemble des solutions admissibles, qui est trop gros (de cardinal  $n!$  dans le cas du problème d'affectation), et d'autre part, que la nature discrète de l'espace des solutions ne permet pas d'écrire directement des conditions d'optimalité à l'aide du calcul différentiel.

Nous allons cependant voir dans la suite de ce chapitre que, malgré les apparences, les méthodes de l'optimisation continue sont utiles en optimisation combinatoire. Considérons en effet le problème combinatoire typique

$$\sup_{x \in P \cap \mathbb{Z}^n} c \cdot x, \quad (11.21)$$

où  $P$  est un polyèdre de  $\mathbb{R}^n$ , c'est-à-dire une intersection d'un nombre fini de demi-espaces, et  $c \in \mathbb{R}^n$ . La formulation (11.21) montre bien la différence entre un problème combinatoire et un problème continu : si nous oublions la contrainte d'intégrité dans (11.21), nous obtenons

$$\sup_{x \in P} c \cdot x, \quad (11.22)$$

ce qui est un problème de programmation linéaire, parfois qualifié de problème continu **relâché** de (11.21). (De manière générale, on parle de problème relâché, ou relâché, quand on oublie certaines contraintes.) Le problème relâché (11.22) peut se traiter efficacement par les méthodes des sections précédentes : toute la difficulté de (11.21) vient de ce que nous nous restreignons aux points entiers du polyèdre  $P$  (par **point entier**, nous entendons point à coordonnées entières). Nous allons maintenant essentiellement caractériser les cas où la résolution du problème discret (11.21) est équivalente à celle de son relâché continu (11.22). Ces cas, qui peuvent sembler exceptionnels, sont en fait d'une grande importance pratique, car ils apparaissent naturellement dans un certain nombre de problèmes combinatoires concrets : plus courts chemins, affectation, et plus généralement problèmes de flots à coût minimum.

#### 11.3.1 Points extrémaux de compacts convexes

La notion qui va permettre de relier problèmes combinatoires et problèmes discrets est celle de **point extrémal**, notion déjà rencontrée dans la Définition 11.2.1 : un point extrémal d'un convexe  $K$  est un point  $x$  tel que  $x = (y+z)/2$  et  $y, z \in K$  entraîne  $y = z = x$ . On note  $\text{extr } K$  l'ensemble des points extrémaux de  $K$ . Rappelons aussi que si  $X$  est un sous-ensemble de  $\mathbb{R}^n$ , on appelle **enveloppe convexe** de  $X$ , et l'on

note  $\text{co } X$ , le plus petit convexe contenant  $X$ , dont on vérifie qu'il est égal à l'ensemble des barycentres d'un nombre fini d'éléments de  $X$ . L'**enveloppe convexe fermée** de  $X$ , notée  $\overline{\text{co}} X$ , est le plus petit convexe fermé contenant  $X$ . Il est égal à la fermeture de  $\text{co } X$ . Le résultat suivant est fondamental.

**Théorème 11.3.1 (Minkowski)** *Un compact convexe de  $\mathbb{R}^n$  est enveloppe convexe de l'ensemble de ses points extrémaux.*

Ce théorème affirme donc que  $K = \text{coextr } K$  lorsque  $K$  est un convexe compact de  $\mathbb{R}^n$ . A fortiori,  $K = \overline{\text{co}} \text{extr } K$  puisque  $K$  est fermé. La preuve du théorème de Minkowski repose sur la notion d'hyperplan d'appui, introduite dans l'annexe sur les espaces de Hilbert : un hyperplan affine  $H = \{y \in \mathbb{R}^n \mid c \cdot y = \alpha\}$ , avec  $c \in \mathbb{R}^n$ ,  $c \neq 0$ , et  $\alpha \in \mathbb{R}$  est un **hyperplan d'appui** d'un convexe  $K$ , au point  $x \in K$ , si  $\alpha = c \cdot x \leq c \cdot y$ , pour tout  $y \in K$ . Nous utiliserons l'observation suivante.

**Lemme 11.3.2** *Si  $H$  est un hyperplan d'appui d'un convexe  $K \subset \mathbb{R}^n$ , alors tout point extrémal de  $H \cap K$  est point extrémal de  $K$ .*

**Démonstration.** Soit  $H = \{y \in \mathbb{R}^n \mid c \cdot y = \alpha\}$  avec  $c \in \mathbb{R}^n$ ,  $c \neq 0$ , et  $\alpha \in \mathbb{R}$ , un hyperplan d'appui de  $K$ . Si  $x = (y + z)/2$  avec  $y, z \in K$ , et si  $x \in K \cap H$ , il vient  $\alpha = c \cdot x = (c \cdot y + c \cdot z)/2$ , et comme  $\alpha \leq c \cdot y$  et  $\alpha \leq c \cdot z$ , on a nécessairement  $\alpha = c \cdot y = c \cdot z$ , donc  $y, z \in K \cap H$ . Si l'on suppose que  $x$  est un point extrémal de  $K \cap H$ , il vient donc  $x = y = z$ , ce qui montre que  $x$  est un point extrémal de  $K$ .  $\square$

**Démonstration du théorème de Minkowski 11.3.1.** On suppose évidemment que  $K \neq \emptyset$  (sinon, le résultat est trivial). Rappelons que la dimension d'un convexe non-vidé est par définition la dimension de l'espace affine qu'il engendre. On va montrer le théorème par récurrence sur la dimension de  $K$ . Quitte à remplacer  $\mathbb{R}^n$  par un sous-espace affine, on peut supposer que  $K$  est de dimension  $n$ . Si  $n = 0$ ,  $K$  est réduit à un point, et le théorème est vérifié. Supposons donc le théorème démontré pour les compacts convexes de dimension au plus  $n - 1$ , et montrons que tout point  $x$  de  $K$  est barycentre d'un nombre fini de point extrémaux de  $K$ . Si  $x$  est un point frontière de  $K$ , le Corollaire 12.1.20 fournit un hyperplan d'appui  $H$  de  $K$  en  $x$ . Comme  $K \cap H$  est un compact convexe de dimension au plus  $n - 1$ , par hypothèse de récurrence,  $x$  est barycentre d'un nombre fini de points extrémaux de  $K \cap H$ , qui sont aussi des points extrémaux de  $K$  d'après le Lemme 11.3.2. Prenons maintenant un point quelconque  $x$  de  $K$ , et soit  $D$  une droite affine passant par  $x$ . L'ensemble  $D \cap K$  est un segment de la forme  $[y, z]$ , où les points  $y, z$  sont des points frontières de  $K$ . D'après ce qui précède,  $y$  et  $z$  sont barycentres d'un nombre fini de points extrémaux de  $K$ . Comme  $x$  est lui même barycentre de  $y$  et  $z$ , le théorème est démontré.  $\square$

**Remarque 11.3.3** Le théorème de Minkowski est un cas particulier en dimension finie d'un résultat d'analyse fonctionnelle, le théorème de Krein-Milman, qui affirme qu'un compact convexe est enveloppe convexe fermée de l'ensemble des points extrémaux (ce résultat, qui est une conséquence du théorème de Hahn-Banach, a lieu dans des espaces très généraux et en

particulier dans les espaces de Banach). On notera qu'en dimension infinie, c'est l'enveloppe convexe fermée, et non l'enveloppe convexe, qui intervient dans l'énoncé du théorème. •

Nous appliquons maintenant le théorème de Minkowski au problème d'optimisation combinatoire (11.21). Dans ce cas, la fonction coût  $J(x) = c \cdot x$  est linéaire, mais il sera plus clair de considérer plus généralement la **maximisation** de fonctions convexes, qui a des propriétés très différentes de la **minimisation** de fonctions convexes traitée aux Chapitres 9 et 10. Nous considérerons aussi un ensemble  $X$  arbitraire, au lieu de  $P \cap \mathbb{Z}^n$ .

**Proposition 11.3.4 (Maximisation de fonction convexes)** *Pour toute fonction convexe  $J : \mathbb{R}^n \rightarrow \mathbb{R}$ , et pour tout sous-ensemble  $X \subset \mathbb{R}^n$ ,*

$$\sup_{x \in X} J(x) = \sup_{x \in \text{co } X} J(x) = \sup_{x \in \overline{\text{co}} X} J(x) , \quad (11.23)$$

*et si  $X$  est borné,*

$$\sup_{x \in X} J(x) = \sup_{x \in \text{extr } \overline{\text{co}} X} J(x) . \quad (11.24)$$

**Démonstration.** Si  $y \in \text{co } X$ , on peut écrire  $y = \sum_{1 \leq i \leq k} \alpha_i x_i$ , avec  $x_i \in X$ ,  $\alpha_i \geq 0$ , et  $\sum_{1 \leq j \leq k} \alpha_j = 1$ . Puisque  $J$  est convexe, on a  $J(y) \leq \sum_{1 \leq j \leq k} \alpha_j J(x_j) \leq \max_{1 \leq j \leq k} J(x_j) \leq \sup_{x \in X} J(x)$ , et puisque ceci est vrai pour tout  $y \in \text{co } X$ , on a  $\sup_{x \in \text{co } X} J(x) \leq \sup_{x \in X} J(x)$ . Par ailleurs, pour tout  $z \in \overline{\text{co}} X$ , on peut écrire  $z = \lim_{k \rightarrow \infty} y_k$ , avec  $y_k \in \text{co } X$ . Comme une fonction convexe  $\mathbb{R}^n \rightarrow \mathbb{R}$  est nécessairement continue (cf. Exercice 9.2.7), on a  $J(z) = \lim_{k \rightarrow \infty} J(y_k) \leq \sup_{x \in \text{co } X} J(x)$ , et puisque ceci est vrai pour tout  $z \in \overline{\text{co}} X$ , on a  $\sup_{x \in \overline{\text{co}} X} J(x) \leq \sup_{x \in \text{co } X} J(x)$ . Les autres inégalités étant triviales, on a montré (11.23). Lorsque  $X$  est borné,  $\overline{\text{co}} X$  qui est aussi borné, est compact. D'après le théorème de Minkowski 11.3.1,  $\overline{\text{co}} X = \text{co extr } \overline{\text{co}} X$ , et en appliquant (11.23),  $\sup_{x \in \text{extr } \overline{\text{co}} X} J(x) = \sup_{x \in \text{co extr } \overline{\text{co}} X} J(x) = \sup_{x \in \overline{\text{co}} X} J(x) = \sup_{x \in X} J(x)$ , ce qui prouve (11.24).  $\square$

La Proposition 11.3.4 nous suggère de considérer l'enveloppe convexe de l'ensemble admissible  $X = P \cap \mathbb{Z}^n$  de notre problème initial (11.21).

**Définition 11.3.5** *On appelle **enveloppe entière** d'un polyèdre  $P \subset \mathbb{R}^n$ , l'enveloppe convexe de l'ensemble des points entiers de  $P$ , que l'on note  $P_e = \text{co}(P \cap \mathbb{Z}^n)$ .*

Le terme “enveloppe entière” est traditionnel mais légèrement trompeur : d'ordinaire, une enveloppe est un objet plus gros, alors qu'ici  $P_e \subset P$ .

**Corollaire 11.3.6** *Si  $J : \mathbb{R}^n \rightarrow \mathbb{R}$  est convexe, et si  $P \subset \mathbb{R}^n$  est un polyèdre, alors*

$$\sup_{x \in P \cap \mathbb{Z}^n} J(x) = \sup_{x \in P_e} J(x) . \quad (11.25)$$



Ainsi, on peut toujours remplacer le problème discret (11.21) par un problème dont l'ensemble admissible est un convexe. Lorsque  $J$  est linéaire, le problème à droite de (11.25) est un programme linéaire classique : on a ainsi concentré la difficulté dans le calcul, ou l'approximation, du polyèdre  $P_e$ . Il y a un cas où tout devient facile.

**Définition 11.3.7** *On dit qu'un polyèdre  $P$  est un polyèdre entier si  $P = P_e$ .*

Nous allons maintenant donner des conditions suffisantes (précises) pour qu'un polyèdre soit entier.

### 11.3.2 Matrices totalement unimodulaires

Un polyèdre quelconque peut s'écrire

$$P = \{x \in \mathbb{R}^n \mid Ax \leq b\} \quad (11.26)$$

avec  $A \in \mathbb{R}^{m \times n}$  et  $b \in \mathbb{R}^m$ . Il convient de noter qu'un tel polyèdre est plus général que le polyèdre  $X_{ad}$  des solutions admissibles du programme linéaire standard (cf. Définition 11.2.1) : en effet,  $X_{ad}$  est par définition inclus dans le cône positif de  $\mathbb{R}^n$ , et nous avons en outre déjà noté que  $X_{ad}$  s'il est non-vide, a toujours des points extrémaux ce qui n'est pas le cas pour un polyèdre quelconque (cf. Remarque 11.2.7). La caractérisation des points extrémaux de  $X_{ad}$  (Lemme 11.2.5) s'étend cependant de la manière suivante.

**Lemme 11.3.8** *Un point extrémal du polyèdre  $P$  défini par (11.26) est nécessairement solution d'un système  $A'x = b'$ , où  $A'$  est une sous-matrice inversible formée de  $n$  lignes de  $A$ , et  $b'$  est le vecteur formé des lignes correspondantes de  $b$ .*

**Démonstration.** Soit  $x$  un point extrémal de  $P$ , et soit  $I(x) = \{1 \leq i \leq m \mid A_i \cdot x = b_i\}$  (l'ensemble des contraintes actives en  $x$ ), où  $A_i$  désigne la  $i$ -ème ligne de  $A$ . Si la famille  $\{A_i\}_{i \in I(x)}$ , n'est pas de rang  $n$ , on peut trouver un vecteur non nul  $y$  tel que  $A_i \cdot y = 0$  pour tout  $i \in I(x)$ . Comme  $x$  est le milieu des points  $x - \epsilon y$  et  $x + \epsilon y$ , qui sont bien des éléments de  $P$  si  $\epsilon$  est assez petit, on contredit l'extrémalité de  $x$ . Ainsi, on peut trouver un sous ensemble  $I' \subset I(x)$  de cardinal  $n$  tel que la matrice  $n \times n$  dont les lignes sont les  $A_i$ , avec  $i \in I'$ , est inversible. Le système  $A_i \cdot x = b_i, i \in I'$ , caractérise alors  $x$ .  $\square$

Le Lemme 11.3.8 montre en particulier qu'un polyèdre n'a qu'un nombre fini de points extrémaux.

**Exercice 11.3.1** Montrer réciproquement que si  $x$  est un point de  $P$  vérifiant  $A'x = b'$ , avec  $A'$  et  $b'$  comme dans le Lemme 11.3.8, alors  $x$  est un point extrémal.

Le Lemme 11.3.8 suggère d'étudier les cas où la solution d'un système linéaire est entière.

**Proposition 11.3.9** Soit  $A \in \mathbb{Z}^{n \times n}$  une matrice inversible. Les assertions suivantes sont équivalentes :

1.  $\det A = \pm 1$  ;
2. pour tout  $b \in \mathbb{Z}^n$ , on a  $A^{-1}b \in \mathbb{Z}^n$ .

**Démonstration.** L'implication  $1 \Rightarrow 2$  résulte aussitôt des formules de Cramer. Réciproquement, supposons que  $A$  vérifie l'assertion 2. Montrons d'abord que  $A^{-1}$  est à coefficients entiers. En prenant pour  $b$  le  $i$ -ème vecteur de la base canonique de  $\mathbb{R}^n$ , on voit que la  $i$ -ème colonne de  $A^{-1}$ , qui coïncide avec  $A^{-1}b$ , est à coefficients entiers. Comme ceci est vrai pour tout  $1 \leq i \leq n$ , on a  $A^{-1} \in \mathbb{Z}^{n \times n}$ . Donc  $\det A^{-1} \in \mathbb{Z}$ , et  $1 = \det A \det A^{-1}$  montre que  $\det A$  divise 1, c'est-à-dire que  $\det A = \pm 1$ .  $\square$

**Définition 11.3.10** On dit qu'une matrice  $A \in \mathbb{Z}^{n \times n}$  est **unimodulaire** quand  $\det A = \pm 1$ , et qu'une matrice  $B \in \mathbb{Z}^{m \times n}$  est **totalement unimodulaire** quand toute sous-matrice carrée extraite de  $B$  est de déterminant  $\pm 1$  ou 0.

En prenant des sous-matrices  $1 \times 1$ , on voit en particulier que les coefficients d'une matrice totalement unimodulaire valent nécessairement  $\pm 1$  ou 0. L'introduction des matrices totalement unimodulaires est motivée par le résultat suivant.

**Corollaire 11.3.11 (Optimalité des solutions entières)** Soient  $D \in \mathbb{Z}^{m \times n}$  une matrice totalement unimodulaire,  $f \in (\mathbb{Z} \cup \{+\infty\})^m$ ,  $f' \in (\mathbb{Z} \cup \{-\infty\})^m$ ,  $g \in (\mathbb{Z} \cup \{+\infty\})^n$ , et  $g' \in (\mathbb{Z} \cup \{-\infty\})^n$ . Alors, les points extrémaux du polyèdre

$$Q = \{x \in \mathbb{R}^n \mid f' \leq Dx \leq f, \quad g' \leq x \leq g\} \quad (11.27)$$

sont nécessairement entiers. En particulier, si  $Q$  est borné, on a  $Q = Q_e$ , et pour toute fonction convexe  $J$  de  $\mathbb{R}^n$  dans  $\mathbb{R}$ , on a

$$\sup_{x \in Q} J(x) = \sup_{x \in Q \cap \mathbb{Z}^n} J(x) . \quad (11.28)$$

**Démonstration.** On peut écrire

$$Q = \{x \in \mathbb{R}^n \mid Ax \leq b\} , \quad (11.29)$$

où  $b$  est un vecteur entier fini et  $A$  est une matrice dont chaque ligne est soit de la forme  $\pm D_i$ , avec  $D_i$  une ligne quelconque de  $D$ , soit de la forme  $\pm e_j$ , où  $e_j$  est le  $j$ -ème vecteur de la base canonique de  $\mathbb{R}^n$ , pour un indice quelconque  $1 \leq j \leq n$ .

On montre d'abord que  $A$  est totalement unimodulaire. Soit donc  $M$  une sous-matrice  $k \times k$  extraite de  $A$ . Montrons par récurrence sur  $k$  que  $\det M \in \{\pm 1, 0\}$ . Si  $k = 1$ , cela résulte aussitôt de la totale unimodularité de  $D$ . Supposons maintenant le résultat prouvé pour toutes les sous-matrices carrées de  $A$  de dimension au plus  $k - 1$ , et montrons le pour  $M$ . Si  $M$  contient une ligne égale à un vecteur  $\pm e_j$ , on développe  $\det M$  par rapport à cette ligne, et par récurrence, le résultat est prouvé. Si  $M$  contient deux lignes égales au signe près,  $\det M = 0$ , et le résultat est encore

prouvé. Sinon,  $M$  coïncide, au changement du signe de certaines lignes près, avec une sous-matrice de  $D$ , et comme  $D$  est totalement unimodulaire,  $\det M \in \{\pm 1, 0\}$ , ce qui achève la preuve de la totale unimodularité de  $A$ .

Comme  $Q$  est donné par (11.29), avec  $b$  entier et  $A$  totalement unimodulaire, il résulte du Lemme 11.3.8 et de la Proposition 11.3.9 que les points extrémaux de  $Q$ , s'ils existent, sont entiers.

Si l'on suppose en outre que  $Q$  est borné,  $Q$  est compact, et d'après le Théorème de Minkowski 11.3.1,  $Q = \text{co extr } Q$ . Comme  $\text{extr } Q$  est formé de vecteurs entiers,  $Q_e = \text{co}(Q \cap \mathbb{Z}^n) \supset \text{co extr } Q = Q$ , et par ailleurs l'inclusion  $Q_e \subset Q$  est triviale. L'égalité (11.28) est alors obtenue en appliquant le Corollaire 11.3.6.  $\square$

**Exercice 11.3.2** Il s'agit d'établir une réciproque au Corollaire 11.3.11. Commençons par examiner le cas spécial du polyèdre  $X_{ad} = \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0\}$  des solutions admissibles du programme linéaire standard, avec  $A \in \mathbb{Z}^{m \times n}$  de rang  $m$ . Montrer que les deux propriétés suivantes sont équivalentes :

1. pour tout  $b \in \mathbb{Z}^m$ , les points extrémaux de  $X_{ad}$  sont entiers ;
2. toutes les sous-matrices  $m \times m$  de  $A$  sont de déterminant  $\pm 1$  ou 0.

Soit maintenant  $D \in \mathbb{Z}^{m \times n}$ , et considérons  $Q = \{x \in \mathbb{R}^n \mid Dx \leq b, x \geq 0\}$ . Dédurre de l'équivalence qui précède l'équivalence des deux propriétés suivantes (théorème de Hoffman et Kruskal) :

3. pour tout  $b \in \mathbb{Z}^m$ , les points extrémaux de  $Q$  sont entiers ;
4.  $D$  est totalement unimodulaire.

**Remarque 11.3.12** On notera que le Corollaire 11.3.11 ne pose aucune condition sur  $J$ , hormis la convexité. En particulier, si  $J(x) = c \cdot x$  est linéaire, le caractère entier des solutions optimales n'est pas directement relié au caractère entier du vecteur de coût  $c$ .  $\bullet$

**Remarque 11.3.13** Le Corollaire 11.3.11 ne dit surtout pas que toutes les solutions optimales sont entières. D'ailleurs, lorsque  $J$  est linéaire, il ne peut en être ainsi à moins que la solution optimale ne soit unique, car tout barycentre de solutions optimales d'un programme linéaire est solution optimale.  $\bullet$

**Remarque 11.3.14** La condition que  $Q$  soit borné n'est pas nécessaire pour affirmer que  $Q = Q_e$  dans le Corollaire 11.3.11 : nous nous sommes limités aux polyèdres bornés, qui sont suffisants en pratique pour modéliser la plupart des problèmes combinatoires, uniquement pour simplifier l'exposé. Voir [38] pour plus de détails.  $\bullet$

Il existe de nombreux résultats sur les matrices totalement unimodulaires. Nous nous bornons ici à donner une condition suffisante très utile.

**Proposition 11.3.15 (Poincaré)** Si  $A$  est une matrice à coefficients  $\pm 1$  ou 0, avec au plus un coefficient 1 par colonne, et au plus un coefficient  $-1$  par colonne, alors  $A$  est totalement unimodulaire.

**Démonstration.** Comme la propriété que vérifie  $A$  passe aux sous-matrices, il suffit de voir que si  $A$  est carrée,  $\det A \in \{\pm 1, 0\}$ . Si  $A$  a une colonne nulle,  $\det A = 0$ . Si  $A$  a une colonne avec seulement un coefficient non-nul, on développe le déterminant par rapport à cette colonne, et l'on conclut par récurrence que  $\det A \in \{\pm 1, 0\}$ . Il ne reste qu'à considérer le cas où chaque colonne de  $A$  a exactement un coefficient 1 et exactement un coefficient  $-1$  : alors, chaque colonne est de somme nulle, donc  $\det A = 0$ .  $\square$

Nous appliquerons la Proposition 11.3.15 aux problèmes de flots dans la sous-section suivante. Donnons pour l'instant en exercice un cas où l'on peut conclure à la main à la totale unimodularité.

**Exercice 11.3.3 (Problème de couverture)** Un centre d'appel téléphonique a une courbe de charge :  $c_t$  est le nombre de clients devant être servis à l'instant discret  $t \in \{1, \dots, T\}$ . Un certain nombre de conseillers de clientèle répondent aux appels. On simplifie le problème en supposant que tous les appels sont de même type. On supposera qu'il y a  $k$  horaires de travail possibles, l'horaire  $i$  étant caractérisé par un intervalle  $[\alpha_i, \beta_i]$ , avec  $1 \leq \alpha_i \leq \beta_i \leq T$ , ce qui revient à faire abstraction des pauses. On note  $S_i$  le salaire à verser à un conseiller de clientèle travaillant de l'instant  $\alpha_i$  à l'instant  $\beta_i$ . On pose  $u_{it} = 1$  si  $\alpha_i \leq t \leq \beta_i$ , et  $u_{it} = 0$  sinon. Justifier le problème

$$\inf_{\substack{x \in \mathbb{N}^k \\ \sum_{1 \leq i \leq k} x_i u_{it} \geq c_t, \forall 1 \leq t \leq T}} \sum_{1 \leq i \leq k} x_i S_i . \quad (11.30)$$

Montrer que l'ensemble des solutions admissibles de ce problème peut s'écrire comme l'ensemble des points entiers d'un polyèdre de la forme (11.27), où la matrice  $D$  est une matrice d'intervalles, c'est-à-dire une matrice à coefficients 0, 1 telle que les 1 apparaissent consécutivement sur une colonne. Montrer qu'une matrice d'intervalles est totalement unimodulaire. Conclure.

### 11.3.3 Problèmes de flots

Afin de définir les problèmes de flots, considérons un graphe orienté  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  :  $\mathcal{N}$  est l'ensemble des nœuds, et  $\mathcal{A} \subset \mathcal{N} \times \mathcal{N}$  est l'ensemble des **arcs**. Un arc allant du nœud  $i$  au nœud  $j$  est ainsi noté  $(i, j)$ . On munit chaque arc  $(i, j) \in \mathcal{A}$  d'une **capacité**  $u_{ij} \in \mathbb{R}_+ \cup \{+\infty\}$  et d'un **coût**  $c_{ij} \in \mathbb{R}$ . (Le “ $u$ ” dans  $u_{ij}$  est pour “upper bound”.) On se donne aussi en chaque nœud du graphe un flot entrant exogène  $b_i \in \mathbb{R}$  (si  $b_i < 0$ , il s'agit d'un flot sortant, compté algébriquement). On appelle **flot** une fonction  $x \in \mathbb{R}^{\mathcal{A}}$ ,  $(i, j) \mapsto x_{ij}$ , vérifiant la **loi des nœuds de Kirchoff**

$$b_i + \sum_{j \in \mathcal{N}, (j, i) \in \mathcal{A}} x_{ji} = \sum_{j \in \mathcal{N}, (i, j) \in \mathcal{A}} x_{ij}, \quad \forall i \in \mathcal{N}, \quad (11.31)$$

ainsi que la contrainte de positivité

$$0 \leq x_{ij}, \quad \forall (i, j) \in \mathcal{A} . \quad (11.32)$$

En sommant les lois des nœuds (11.31), on voit qu'une condition nécessaire pour l'existence d'un flot est que la somme des flots exogènes entrants soit nulle

$$\sum_{i \in \mathcal{N}} b_i = 0 . \quad (11.33)$$

Nous supposons toujours que la condition (11.33) est vérifiée. Un flot est dit **admissible** s'il satisfait les contraintes de capacité

$$x_{ij} \leq u_{ij}, \quad \forall (i, j) \in \mathcal{A} . \quad (11.34)$$

**Définition 11.3.16** On appelle **problème de flot à coût minimum** le programme linéaire

$$\min \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \text{ sous les contraintes (11.31), (11.32), (11.34).} \quad (11.35)$$

Le problème de flot à coût minimum admet plusieurs sous-problèmes importants, comme le problème de transport de l'Exemple 9.1.1, ou le problème d'affectation de l'Exemple 9.1.2.

**Exercice 11.3.4** Expliciter le problème de flot à coût minimum correspondant au problème de transport de l'Exemple 9.1.1. (On dessinera le graphe.)

Un cas particulier fondamental du problème de flot à coût minimum est le problème du flot maximal, ou **problème de flot** proprement dit, qui concerne seulement les capacités (et non les coûts). Il sera commode de supposer que  $\mathcal{G}$  a deux nœuds distingués,  $s$  et  $p$ , appelés respectivement **source** et **puits**, tels que  $s$  n'a pas de prédécesseur ( $\{i \in \mathcal{N} \mid (i, s) \in \mathcal{A}\} = \emptyset$ ), et  $p$  n'a pas de successeur ( $\{i \in \mathcal{N} \mid (p, i) \in \mathcal{A}\} = \emptyset$ ). Soit  $v \in \mathbb{R}_+$ . On appelle **flot admissible de  $s$  à  $p$  de valeur  $v$**  une solution  $x$  de (11.31), (11.32), (11.34), avec

$$b_i = \begin{cases} v & \text{si } i = s, \\ -v & \text{si } i = p, \\ 0 & \text{sinon.} \end{cases}$$

**Définition 11.3.17** Le **problème du flot maximal** consiste à trouver un **flot admissible de  $s$  à  $p$  de valeur maximale**.

**Exercice 11.3.5** Montrer que le problème du flot maximal est effectivement un cas particulier de problème de flot à coût minimal. (Indication : on pourra rajouter un arc au graphe intervenant dans la définition du problème du flot maximal.)

En pratique, on cherche souvent des solutions entières d'un problème de flot : par exemple, pour le problème de transport de l'Exemple 9.1.1, les marchandises à livrer peuvent être des colis, et livrer un demi-colis peut ne pas avoir de sens. Il est donc

naturel de se demander si un problème de flot à coût minimum a automatiquement des solutions optimales entières. Afin d'appliquer le Corollaire 11.3.11, notons que la loi des nœuds de Kirchoff (11.31) peut s'écrire  $Ax = b$ , où la matrice  $A \in \mathbb{R}^{\mathcal{N} \times \mathcal{A}}$ , appelée **matrice d'incidence nœuds-arcs** de  $\mathcal{G}$ , est définie par

$$A_{i,(j,k)} = \begin{cases} -1 & \text{si } i = k, \\ 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases}$$

La matrice  $A$  est bien définie, sauf dans le cas dégénéré où le graphe a une boucle, c'est-à-dire un arc  $(j, k)$  tel que  $j = k$ . Un flot circulant sur une boucle a une contribution qui se simplifie dans la loi des nœuds de Kirchoff (11.31), aussi n'y a-t-il aucune perte de généralité à supposer le graphe sans boucle, ce que nous ferons dans la suite de la section.

Nous pouvons maintenant écrire l'ensemble des flots admissibles

$$\{x \in \mathbb{R}^{\mathcal{A}} \mid Ax = b, 0 \leq x \leq u\}. \quad (11.36)$$

**Proposition 11.3.18** *La matrice d'incidence nœuds-arcs d'un graphe est totalement unimodulaire.*

**Démonstration.** C'est une conséquence immédiate de la Proposition 11.3.15.  $\square$

**Corollaire 11.3.19 (Optimalité des flots entiers)** *Si les flots entrants exogènes  $b_i$  sont entiers, et si les capacités  $u_{ij}$  sont entières ou infinies, alors, les points extrémaux de l'ensemble (11.36) des solutions admissibles d'un problème de flot à coût minimal sont entiers. En particulier, si l'ensemble des flots admissibles est borné et non-vide, il existe un flot entier optimal.*

**Démonstration.** C'est une conséquence immédiate du Corollaire 11.3.11 et de la Proposition 11.3.18.  $\square$

**Exercice 11.3.6** Montrer que le problème de couverture (Exercice 11.3.3) peut se modéliser par un problème de flot à coût minimum, et retrouver ainsi la conclusion de l'Exercice 11.3.3.

**Exercice 11.3.7** Reprenons le problème d'affectation, introduit dans l'Exemple 9.1.2. On considère toujours  $n$  garçons et  $n$  filles, mais ici,  $a_{ij}$  est un réel qui représente le bonheur du couple  $(i, j)$ , et on cherche une permutation  $\sigma \in \mathcal{S}_n$ , solution optimale de

$$\max_{\sigma \in \mathcal{S}_n} \sum_{1 \leq i \leq n} a_{i\sigma(i)}. \quad (11.37)$$

1. Montrer que ce problème est équivalent au problème linéaire en nombres entiers

$$\max_{x \in \mathcal{B}_n \cap \mathbb{Z}^{n \times n}} \sum_{1 \leq i, j \leq n} a_{ij} x_{ij}, \quad (11.38)$$

où  $\mathcal{B}_n$  désigne l'ensemble des matrices bistochastiques, c'est-à-dire l'ensemble des matrices réelles  $x = (x_{ij})$  de taille  $n \times n$  telles que

$$\begin{aligned} \forall i \in \{1, \dots, n\}, \quad 1 &= \sum_{1 \leq k \leq n} x_{ik}, \\ \forall j \in \{1, \dots, n\}, \quad 1 &= \sum_{1 \leq k \leq n} x_{kj}, \\ \forall i, j \in \{1, \dots, n\}, \quad 0 &\leq x_{ij} \end{aligned}$$

2. Montrer que le problème (11.38) est un problème de flot entier à coût minimum (on dessinera le graphe). En déduire que le polyèdre  $\mathcal{B}_n$  est entier. Que peut-on en conclure quant à la difficulté du problème d'affectation ?

3. Déduire de ce qui précède que toute matrice bistochastique est barycentre d'un nombre fini de matrices de permutations (ce théorème est dû à Birkhoff).

4. Déduire de ce qui précède que si à un bal, il y a  $n$  garçons et  $n$  filles, chaque garçon ayant déjà rencontré exactement  $r$  filles, et chaque fille ayant aussi rencontré exactement  $r$  garçons, avec  $r \geq 1$ , il est possible de former  $n$  couples de danseurs de sorte que les danseurs d'un même couple se soient déjà rencontrés précédemment (ce théorème est dû à König).

**Exercice 11.3.8** Cet exercice présente un algorithme fondamental en théorie des flots, dû à Ford et Fulkerson. On considère le problème du flot maximal d'une source  $s$  à un puits  $p$  dans un graphe  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  (voir la Définition 11.3.17). Pour simplifier, on supposera qu'aucune des capacités  $u_{ij}$  ne prend la valeur  $+\infty$ . Pour tous  $I, J \subset \mathcal{N}$ , et pour tout  $x = (x_{ij}) \in \mathbb{R}^{\mathcal{A}}$ , on pose  $x(I, J) = \sum_{i \in I, j \in J, (i, j) \in \mathcal{A}} x_{ij}$ . On dit qu'une partition de  $\mathcal{N}$  en deux sous-ensembles  $S$  et  $\bar{S}$  est une *coupe séparant  $s$  de  $p$*  si  $s \in S$  et  $p \in \bar{S}$ , et l'on dit que  $u(S, \bar{S})$  est la *capacité* de cette coupe.

1. Montrer que pour toute coupe  $(S, \bar{S})$  séparant  $s$  de  $p$ , et pour tout flot admissible  $x$  de  $s$  à  $p$  de valeur  $v$ ,

$$x(S, \bar{S}) - x(\bar{S}, S) = v.$$

En déduire que la valeur de tout flot admissible de  $s$  à  $p$  est majorée par la capacité de toute coupe séparant  $s$  de  $p$ .

2. Étant donné un flot admissible  $x$  de  $s$  à  $p$ , on définit le *graphe résiduel*  $\mathcal{G}_r(x) = (\mathcal{N}, \mathcal{A}_r(x))$  où  $\mathcal{A}_r(x)$  désigne l'ensemble des couples  $(i, j)$  tels que ou bien  $(i, j) \in \mathcal{A}$  et  $x_{ij} < u_{ij}$ , ou bien  $(j, i) \in \mathcal{A}$  et  $x_{ji} > 0$ . Montrer que s'il existe un chemin  $\gamma$  de  $s$  à  $p$  dans le graphe résiduel  $\mathcal{G}_r(x)$ , il est possible de construire un nouveau flot admissible  $x'$  de  $s$  à  $p$  de valeur strictement supérieure à celle de  $x$ , en modifiant seulement les valeurs  $x_{ij}$  lorsque  $(i, j)$  ou  $(j, i)$  est un arc du chemin  $\gamma$ . Observer en outre que si  $x$  est entier et si les capacités sont entières, on peut choisir  $x'$  entier.

3. On suppose maintenant qu'il n'y a pas de chemin de  $s$  à  $p$  dans  $\mathcal{G}_r(x)$ . Soit  $S$  l'ensemble des nœuds accessibles depuis  $s$  dans  $\mathcal{G}_r(x)$ , et  $\bar{S}$  le complémentaire de  $S$  dans  $\mathcal{N}$ . Montrer que  $(S, \bar{S})$  est une coupe séparant  $s$  de  $p$  dont la capacité est égale à la valeur du flot  $x$ .

4. Conclure que la valeur maximale du flot de  $s$  à  $p$  est égale à la capacité minimale

d'une coupe séparant  $s$  de  $p$ . (C'est le théorème "flot-maximale=coupe minimale", ou "max-flow=min-cut", de Ford et Fulkerson.)

5. En déduire un algorithme permettant, lorsque les capacités sont entières, de calculer un flot maximal de  $s$  à  $p$  en un temps  $O(v^*|\mathcal{A}|)$ , où  $v^*$  est la valeur maximale d'un flot de  $s$  à  $p$ , et  $|\mathcal{A}|$  est le nombre d'arcs.

**Remarque 11.3.20** Le théorème "flot maximal=coupe minimale" de Ford et Fulkerson (Question 4 de l'Exercice 11.3.8) peut être obtenu comme conséquence du théorème de dualité en programmation linéaire. Voir à ce sujet *Combinatorial Optimization : Polyhedra and Efficiency*, Volume A, Alexander Schrijver, Springer-Verlag Berlin and Heidelberg, 2002. •

**Remarque 11.3.21** L'algorithme de Ford et Fulkerson, présenté dans l'Exercice 11.3.8, n'est que le plus simple des algorithmes de flots. Une variante de cet algorithme, appelé algorithme d'accroissement du flot sur le chemin le plus court, ou algorithme d'Edmonds et Karp, aussi dû à Dinits, peut s'implémenter en un temps  $O(nm^2)$  indépendant du caractère entier des capacités, où  $n = |\mathcal{N}|$  désigne le nombre nœuds, et  $m = |\mathcal{A}|$  désigne le nombre d'arcs. Ce raffinement consiste tout simplement à sélectionner à chaque étape de l'algorithme de Ford et Fulkerson, le plus court chemin de  $s$  à  $p$ , c'est-à-dire le chemin de  $s$  à  $p$  qui a le plus petit nombre d'arcs. Il existe aussi un algorithme de flot très différent, l'algorithme "preflow-push" de Goldberg et Tarjan (1986), qui a un temps d'exécution de  $O(n^2m)$ . Toutes ces idées se généralisent au cas du problème de flot à coût minimum. Voir [1] pour un état de l'art. •

## 11.4 Programmation dynamique et problèmes de plus court chemin

La programmation dynamique, développée par R. Bellman dans les années 50, est une méthode très générale qui s'applique aux problèmes de décision dans le temps (comme en commande optimale, sauf que la variable de temps est parfois déguisée). Il s'agit pour chaque problème d'identifier une bonne notion d'état. A chaque état on associe une valeur optimale partant de cet état, et l'équation de programmation dynamique relie la valeur d'un état à un instant donné à celles des états auxquels on peut accéder à l'instant suivant. (La programmation dynamique est précisément à l'optimisation ce que le point de vue Markovien est à la théorie des probabilités.)

### 11.4.1 Principe d'optimalité de Bellman

La **programmation dynamique** repose sur le principe d'optimalité de Bellman, qui peut s'énoncer très simplement dans le cas particulier du problème du plus court chemin : si un plus court chemin d'une ville  $A$  à une ville  $B$  passe par une ville  $C$ , alors le sous-chemin allant de  $A$  à  $C$  est encore un plus court chemin de  $A$  à  $C$ . En notant  $d_{XY}$  la distance de  $X$  à  $Y$ , on obtient

$$d_{AB} = \min_C (d_{AC} + d_{CB}) , \quad (11.39)$$



où le min est pris sur l'ensemble des villes  $C$  par lesquelles peut passer un chemin de  $A$  à  $B$ . L'équation (11.39) est un cas particulier d'équation de programmation dynamique, ou **équation de Bellman**, elle va permettre de calculer récursivement  $d$  sachant que  $d_{XY}$  est connue quand  $X$  et  $Y$  sont des villes voisines. La notion d'état apparaît ici naturellement : nous nous sommes posés le problème du calcul de  $d_{AB}$ ,  $A$  et  $B$  étant deux villes fixées, et nous venons de voir qu'il convient de tabuler les distances  $d_{CB}$  (ou de manière duale,  $d_{AC}$ ) pour toutes les villes intermédiaires  $C$ . Les détails du calcul doivent bien sûr être fixés pour fournir un vrai algorithme : c'est ce que nous ferons dans les deux sous-sections qui suivent. Nous traiterons d'abord une version plus simple du problème, où le temps apparaît explicitement, dans la Sous-section 11.4.2 puis reviendrons au problème du plus court chemin, sous une forme plus générale, dans la Sous-section 11.4.3.

### 11.4.2 Problème en horizon fini

Considérons un petit problème de nature économique, qui mérite le nom de problème de contrôle optimal discret en horizon fini. Soit  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  un graphe orienté, muni d'un poids  $c : \mathcal{A} \rightarrow \mathbb{R}$ , que nous interpréterons comme un **coût**. Rappelons qu'un **chemin** est une suite de nœuds reliés consécutivement par des arcs, et qu'un chemin dont le premier et le dernier nœud coïncident est appelé **circuit**. Le coût d'un chemin est par définition la somme des coûts de ses arcs. Fixons un entier  $T$  (l'horizon), et un nœud initial  $i \in \mathcal{N}$ . On veut calculer le coût total partant de  $i$  en horizon  $T$  que l'on note

$$v_i^T = \min_{(\ell_0, \dots, \ell_T) \text{ chemin}, \ell_0 = i} c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} . \quad (11.40)$$

La fonction  $i \mapsto v_i^T$ , que l'on peut représenter par un vecteur  $v^T \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ , est traditionnellement appelée **fonction valeur** (en horizon  $T$ ). Nous convenons en écrivant (11.40) que  $\min \emptyset = +\infty$ , ce qui revient à écrire  $v_i^T = +\infty$  quand il n'y a pas de chemins de longueur  $T$  partant de  $i$  dans le graphe. Un autre cas spécial est  $T = 0$  : les chemins qui apparaissent dans (11.40) sont alors de longueur nulle. On conviendra que pour tout sommet du graphe, il y a un circuit de longueur nulle passant par ce sommet, et que son coût est zéro. En particulier,  $v_i^0 = 0$ .

Une évaluation naïve de (11.40) consisterait à énumérer tous les chemins de longueur  $T$ , dont le nombre croît exponentiellement avec  $T$ . La programmation dynamique va nous permettre de factoriser ce calcul, et ainsi de l'effectuer en temps polynomial (la notion de temps polynomial est définie dans la Remarque 11.6.6).

L'idée pour calculer  $v_i^T$  est de faire varier l'horizon et l'état initial, en calculant  $v_k^t$  pour tout  $0 \leq t \leq T$  et  $k \in \mathcal{N}$ . En effet, on peut écrire pour tout  $i \in \mathcal{N}$ , et  $t \geq 1$ ,

$$v_i^t = \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} (c_{i,k} + v_k^{t-1}) , \quad (11.41)$$

$$v_i^0 = 0 . \quad (11.42)$$

L'équation (11.41) résulte du principe d'optimalité de Bellman : le coût optimal partant de  $i$ , en  $t$  étapes, est obtenu en choisissant le premier mouvement  $i \rightarrow k$ , de manière à minimiser le coût conditionné par ce premier mouvement, c'est-à-dire  $c_{i,k}$  plus le coût optimal partant de  $k$  en horizon  $t - 1$ . La condition initiale (11.42) est triviale : s'il ne reste aucun coup à jouer, on ne paye rien. On vient d'écrire l'équation de programmation dynamique, ou **équation de Bellman**, adaptée à notre problème : elle nous permet de calculer par récurrence la suite de vecteurs  $v^0, v^1, \dots \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ .

Il est facile de résoudre des problèmes plus généraux en modifiant (11.42). Considérons par exemple la nouvelle fonction valeur

$$v_i^T = \min_{(\ell_0, \dots, \ell_T) \text{ chemin}} c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T} \quad , \quad (11.43)$$

où  $\phi \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$  est un vecteur représentant une pénalité associée à l'état final. Lorsque  $\phi = 0$ , on retrouve (11.40). Lorsque  $\phi$  est la fonction indicatrice d'un sommet  $j \in \mathcal{N}$ , i.e.

$$\phi_k = \begin{cases} 0 & \text{si } k = j \\ +\infty & \text{sinon,} \end{cases} \quad (11.44)$$

(11.43) force à terminer dans l'état  $j$ , et  $v_i^T$  donne alors le coût minimal en  $T$  étapes pour aller de  $i$  à  $j$ . La fonction valeur (11.43) vérifie toujours (11.41), avec la nouvelle condition initiale pour l'équation de Bellman

$$v^0 = \phi \quad . \quad (11.45)$$

C'est ici le moment de remarquer que le temps  $T$  qui intervient dans l'équation de Bellman (11.41), (11.45) s'écoule **en sens inverse** du temps physique qui intervient dans les trajectoires (11.43) : dans notre modélisation,  $T$  est le temps qui reste à jouer, aussi une pénalité sur l'état **terminal** de la trajectoire  $(\ell_0, \dots, \ell_T)$  conduit-elle à changer la condition **initiale** de l'équation de Bellman. On dit pour cela que l'équation de Bellman est une équation **rétrograde**. Cette inversion du temps est inhérente aux problèmes de décision : ce n'est souvent qu'à la fin d'une mauvaise partie qu'on comprend comment il ne fallait point jouer au début.

L'intérêt de l'équation de Bellman (11.41) est que  $v^t$  peut se calculer à partir de  $v^{t-1}$  en temps  $O(|\mathcal{A}| + |\mathcal{N}|)$ , où  $|\mathcal{A}|$  et  $|\mathcal{N}|$  désignent respectivement le nombre d'arcs et le nombre de nœuds, voir la Remarque 11.4.1 pour plus de détails. Ainsi,  $v^T$  peut se calculer en temps  $O(T(|\mathcal{A}| + |\mathcal{N}|))$ , à comparer par exemple avec le temps  $O(|\mathcal{N}|p^T)$  d'un algorithme naïf énumérant les chemins de longueur  $T$  dans un graphe où chaque sommet a exactement  $p$  successeurs. On obtient d'autre part très simplement un chemin optimal à partir de l'équation de Bellman (11.41), (11.45) : on pose  $\ell_0 = i$ , puis on choisit  $\ell_1$  réalisant le minimum dans (11.41), i.e.  $v_{\ell_0}^T = c_{\ell_0, \ell_1} + v_{\ell_1}^{T-1}$ , et plus généralement  $\ell_{r+1}$  tel que  $v_{\ell_r}^{T-r} = c_{\ell_r, \ell_{r+1}} + v_{\ell_{r+1}}^{T-r-1}$ . Par construction,  $v_i^T = c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}$ , ce qui montre que  $(\ell_0, \dots, \ell_T)$  est un chemin optimal pour le problème (11.43).

**Exemple 11.4.1** Pour illustrer ce qui précède, considérons le cas d'un chauffeur de taxi maraudant dans la ville imaginaire représentée sur la Figure 11.2. La ville est formée de trois zones, H, un quartier huppé, A un aéroport, et B une banlieue (d'où le chauffeur revient en général à vide). Le chauffeur de taxi souhaite faire  $T$  courses, partant de H, et maximiser le gain total, qui est la somme des gains des courses  $c$ , et d'une prime,  $\phi$ , traduisant sa préférence pour l'état final de la dernière course. On a représenté les gains des courses sur les arcs, et les primes par des arcs sortants. La prime  $-\infty$  en B signifie que le taxi ne veut pas finir sa journée en B. Nous supposons par ailleurs que le taxi a la faculté de choisir ses courses. S'agissant d'un problème de maximisation, l'équation de programmation dynamique s'écrit avec max au lieu de min, et  $-\infty$  au lieu de  $+\infty$  :

$$\begin{aligned} v_H^t &= \max(3 + v_H^{t-1}, 10 + v_A^{t-1}) \\ v_A^t &= \max(12 + v_H^{t-1}, 7 + v_B^{t-1}) \\ v_B^t &= \max(-5 + v_H^{t-1}, -3 + v_A^{t-1}) \end{aligned}$$

$$v_H^0 = 0, \quad v_A^0 = 2, \quad v_B^0 = -\infty.$$

Calculons maintenant la stratégie optimale du taxi en horizon 2 partant de H. Il

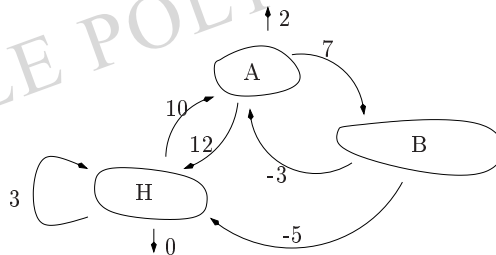


FIGURE 11.2 – Un chauffeur de taxi très déterministe

nous suffit d'évaluer

$$v_H^1 = \max(3 + 0, \underline{10 + 2}) = 12 \quad (11.46)$$

$$v_A^1 = \max(\underline{12 + 0}, 7 + -\infty) = 12 \quad (11.47)$$

$$v_B^1 = \max(-5 + 0, \underline{-3 + 2}) = -1 \quad (11.48)$$

$$v_H^2 = \max(3 + 12, \underline{10 + 12}) = 22, \quad (11.49)$$

où l'on a souligné les termes qui réalisent le maximum. On en déduit que le gain optimal, 22, est obtenu en jouant d'abord le coup qui réalise le max dans (11.49) (on va d'abord de H à A), puis le coup qui réalise le max dans (11.47) (on revient de A à H). •

**Remarque 11.4.1** Une manière classique de coder efficacement en machine un graphe  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  avec  $n = |\mathcal{N}|$  nœuds et  $m = |\mathcal{A}|$  arcs munis d'un coût  $c$ , consiste à définir trois vecteurs : un vecteur  $\mathbf{h} \in \{1, \dots, n\}^m$  (pour "head"), un vecteur  $\mathbf{t} \in \{1, \dots, n\}^m$  (pour "tail"), ainsi qu'un vecteur de réels,  $\mathbf{c} \in \mathbb{R}^m$ . Cette représentation revient à numérotter les arcs de  $\mathcal{A}$  de 1 à  $m$ , en disant que l'arc  $(i, j) \in \mathcal{A}$  qui porte le numéro  $k$  va de  $\mathbf{h}_k = i$  à  $\mathbf{t}_k = j$ , et a pour coût  $c_k = c_{i,j}$ . Le graphe occupe ainsi un espace mémoire  $O(n + m)$ . On voit facilement que le graphe étant codé de la sorte, il est possible de calculer  $v^t$  à partir de  $v^{t-1}$  en utilisant (11.41), en temps  $O(n + m)$ . •

### 11.4.3 Problème du chemin de coût minimum, ou d'arrêt optimal

Considérons maintenant le problème du **chemin de coût minimum** qui, étant donnés deux sommets  $i$  et  $j$ , consiste à trouver un chemin de longueur arbitraire allant de  $i$  à  $j$  et de coût minimum. Il s'agit d'une généralisation du problème du plus court chemin déjà évoqué dans la Sous-section 11.4.1 : contrairement au cas des distances, nous ne supposons pas ici que les coûts sont positifs. En exploitant les notations précédentes, il nous faut maintenant calculer

$$v_i = \inf_{T \in \mathbb{N}} v_i^T = \inf_{\substack{(\ell_0, \dots, \ell_T) \text{ chemin} \\ T \in \mathbb{N}, \ell_0 = i}} c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}, \quad (11.50)$$

où le vecteur  $\phi$  pénalisant un état final autre que  $j$  est donné par (11.44). On peut aussi considérer une pénalité  $\phi \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$  quelconque, ce qui ne change rien aux résultats. On notera que  $v_i$  est maintenant à valeurs dans  $\mathbb{R} = \mathbb{R} \cup \{\pm\infty\}$  (les sommes à droite de (11.50) peuvent valoir  $+\infty$ , par ailleurs, il est possible que ces sommes ne soient pas bornées inférieurement, car l'on considère des chemins de longueur arbitrairement grande). La fonction valeur  $v$  vérifie la nouvelle équation de Bellman

$$v_i = \min(\phi_i, \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} (c_{i,k} + v_k)) \quad , \quad \forall i \in \mathcal{N}. \quad (11.51)$$

Par comparaison avec (11.41), la présence d'un terme supplémentaire dans le min traduit la possibilité de s'arrêter dans n'importe quel état  $k$  avec la pénalité finale  $\phi_k$ , ce qui dans le cas spécial (11.44), interdit de s'arrêter ailleurs qu'en  $j$ .

Il nous faut maintenant montrer que le système (11.51) permet de calculer la fonction valeur. Rappelons qu'on appelle **solution maximale** d'une équation une solution qui majore toutes les autres.

**Théorème 11.4.2** *La fonction valeur  $v$  définie par (11.50) est la solution maximale de l'équation de Bellman (11.51).*

**Démonstration.** On a déjà montré que  $v$  vérifie (11.51). Soit  $v' \in \overline{\mathbb{R}}^{\mathcal{N}}$  une solution quelconque de (11.51). Montrons que  $v' \leq v$ . Soit  $(\ell_0, \dots, \ell_T)$  un chemin quelconque partant de  $i$ . Comme  $v'$  vérifie (11.51), on peut écrire  $v'_{\ell_r} \leq c_{\ell_r, \ell_{r+1}} + v'_{\ell_{r+1}}$ , pour tout  $0 \leq r \leq T-1$ , et  $v'_{\ell_T} \leq \phi_{\ell_T}$ . En enchaînant ces inégalités,

$$v'_i \leq c_{\ell_0, \ell_1} + \dots + c_{\ell_{T-1}, \ell_T} + \phi_{\ell_T}$$

et en prenant l'infimum sur tous les chemins  $(\ell_0, \dots, \ell_T)$  partant de  $i$ ,  $v'_i \leq v_i$ . Ainsi  $v' \leq v$ , ce qui montre que  $v$  est solution maximale de (11.51).  $\square$

**Exercice 11.4.1** Montrer que l'équation de Bellman (11.51) peut avoir plusieurs solutions finies. (Indication : considérer un graphe avec un seul sommet.)

**Exercice 11.4.2** Nous dirons que le graphe est co-accessible pour  $\phi$  si pour tout nœud  $i$  du graphe, il existe un chemin de  $i$  à un nœud  $j$  tel que  $\phi_j \neq +\infty$ . On va montrer que si le graphe n'a pas de circuit de coût négatif et est co-accessible pour  $\phi$ , alors l'équation de Bellman (11.51) possède une unique solution finie, égale à  $v$ . Pour cela, on considère  $v' \in \mathbb{R}^{\mathcal{N}}$  une solution de l'équation de Bellman (11.51). On introduit l'ensemble de continuation

$$C = \{i \in \mathcal{N} \mid \phi_i \geq \min_{k \in \mathcal{N}, (i,k) \in \mathcal{A}} (c_{i,k} + v'_k)\} ,$$

et l'on choisit pour chaque  $i \in C$  un nœud  $\pi(i)$  tel que

$$v'_i = c_{i,\pi(i)} + v'_{\pi(i)} .$$

On définit ainsi une application  $\pi : C \rightarrow \mathcal{N}$ . Montrer que quel que soit  $i \in C$ , il existe un entier  $k$  tel que le  $k$ -ème itéré  $\pi^k(i)$ , n'appartienne pas à  $C$ . Conclure que  $v' \geq v$ .

On peut réécrire (11.51) comme une équation de point fixe  $v = f(v)$ , avec une définition évidente de  $f : \mathbb{R}^{\mathcal{N}} \rightarrow \mathbb{R}^{\mathcal{N}}$ . Aussi le Théorème 11.4.2 suggère-t-il de calculer  $v$  à l'aide d'un algorithme de point fixe. Pour conclure à la convergence de méthodes de point fixe, on recourt souvent à des arguments de contraction. Ici, l'analyse de la convergence fera plutôt appel à la structure d'ordre. On munit en effet  $\mathbb{R}^{\mathcal{N}}$  de l'ordre partiel usuel, défini composante par composante, et pour cet ordre, l'application  $f : x \mapsto f(x)$  est croissante. Puisque  $f(\phi) \leq \phi$ , une récurrence immédiate montre que  $f^{r+1}(\phi) \leq f^r(\phi)$ , pour tout  $r \geq 0$ . La suite  $\{f^r(\phi)\}_{r \geq 0}$  qui décroît converge nécessairement vers un vecteur  $v' \in \mathbb{R}^{\mathcal{N}}$ . (Le lecteur notera ici que l'on autorise la valeur  $-\infty$  pour les coefficients de  $v'$ , sans cela point de convergence.) On a  $f(v') = v'$  par continuité de  $f$ . D'autre part, si  $v''$  est un point fixe quelconque de  $f$ , on a trivialement  $v'' \leq \phi$ , donc puisque  $f$  est croissante,  $v'' = f^r(v'') \leq f^r(\phi)$ , pour tout  $r \geq 0$ , donc en prenant l'infimum sur les  $r \geq 0$ ,  $v'' \leq v'$ , donc  $v'$  est le plus grand point fixe de  $f$ . Puisque d'après le Théorème 11.4.2, la fonction valeur est aussi le plus grand point-fixe de  $f$ , on vient de montrer le résultat suivant.

**Théorème 11.4.3 (Itération sur les valeurs)** La suite  $\{f^r(\phi)\}_{r \geq 0}$  décroît vers la fonction valeur  $v$  définie par (11.50).

Plus généralement, étant donné un opérateur de programmation dynamique  $f$  dont on veut calculer un point fixe, l'algorithme consistant à construire une suite  $f^r(\phi)$ , soit à partir d'une sur-solution  $\phi$  quelconque (on dit que  $\phi$  est une sur-solution de l'équation de point fixe  $x = f(x)$  si  $f(\phi) \leq \phi$ ), soit à partir d'une sous-solution  $\phi$  quelconque (définie en renversant l'inégalité) est appelé **itération sur les valeurs**. Pour des problèmes de programmation

dynamique généraux, notamment en contrôle stochastique, une méthode de type Newton, appelée itération sur les politiques (voir par exemple D. Bertsekas, *Dynamic Programming and Optimal Control*, Vol. I et II, Athena Sci., Belmont, MA, 1995), est souvent plus rapide. L'intérêt de l'itération sur les valeurs est sa simplicité. Dans le cas déterministe, on a même une convergence en temps fini de l'itération sur les valeurs.

**Exercice 11.4.3 (Convergence en temps fini)** Montrer que si  $\mathcal{G}$  n'a pas de circuits de coût strictement négatif, alors  $f^{|\mathcal{N}|-1}(\phi) = v$ . Si au contraire  $\mathcal{G}$  a un circuit de coût strictement négatif, et si  $\mathcal{G}$  est co-accessible pour  $\phi$  (voir l'Exercice 11.4.2 pour cette notion) alors  $f^{|\mathcal{N}|}(\phi) < f^{|\mathcal{N}|-1}(\phi)$ .

L'Exercice 11.4.3 suggère d'implémenter comme suit l'algorithme d'itération sur les valeurs. On calcule par récurrence la suite  $x_0 = \phi$ , et  $x_r = f(x_{r-1})$  pour  $r \geq 0$ , en vérifiant à chaque étape si  $x_r = x_{r-1}$ , auquel cas  $x_r = v$  et l'on s'arrête (l'intérêt de ce test est que le temps de convergence est souvent beaucoup plus petit que  $|\mathcal{N}| - 1$ ). Au pire, on parvient à  $r = |\mathcal{N}|$  et on s'arrête alors : on sait qu'il existe un circuit de coût strictement négatif. En pratique, on programme rarement tel quel cet algorithme d'itération sur les valeurs, mais plutôt la variante suivante de type Gauss-Seidel, appelée algorithme de Ford-Bellman, laquelle met à jour au plus tôt toutes les coordonnées dans l'itération sur les valeurs. Il est plus rapide de programmer cette variante que de la décrire.

**Algorithme 11.4.4 (de Ford et Bellman)** Entrée :  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  et  $c : \mathcal{A} \rightarrow \mathbb{R}$ . Variables :  $v \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$ ,  $b$  Booléen,  $r$  entier,  $i, k \in \mathcal{N}$ .

Initialisation :  $r \leftarrow 0$ ,  $b \leftarrow \text{vrai}$ ; pour tout  $i \in \mathcal{N}$ ,  $v_i \leftarrow \phi_i$ .

Tant que  $r < |\mathcal{N}|$  et  $b$  faire :

$b \leftarrow \text{faux}$ ;

pour tout  $i \in \mathcal{N}$  et pour tout  $(i, k) \in \mathcal{A}$ , si  $c_{i,k} + v_k < v_i$ , faire  $v_i \leftarrow c_{i,k} + v_k$  et  $b \leftarrow \text{vrai}$ .

Si  $r < |\mathcal{N}|$ ,  $v$  est la solution, si  $r = |\mathcal{N}|$ , il existe un circuit de coût strictement négatif.

Afin d'illustrer l'itération sur les valeurs, reprenons le problème du chauffeur de taxi représenté sur la Figure 11.2, et intéressons nous à la valeur des chemins de gain **maximum** et de longueur arbitraire partant de  $H$ . Comme il y a un circuit de gain strictement positif (aller de  $H$  à  $A$ , et revenir, ce qui rapporte 22), notre chauffeur a intérêt à faire un nombre infini de courses, et l'on a en particulier  $v_H = +\infty$ . Afin de rendre le problème moins dégénéré, soustrayons par exemple un impôt de 11 sur le prix de chaque course. L'opérateur  $f$  défini à partir de (11.51) devient

$$\begin{aligned} [f(x)]_H &= \max(0, -8 + x_H, -1 + x_A) \\ [f(x)]_A &= \max(2, 1 + x_H, -4 + x_B) \\ [f(x)]_B &= \max(-16 + x_H, -14 + x_A) . \end{aligned}$$

L'itération sur les valeurs consiste à calculer  $x^0 = \phi$ ,  $x^1 = f(x^0)$ ,  $\dots$ , soit

$$\begin{aligned} x^0 &= \phi = (0, 2, +\infty) \\ x^1_H &= \max(0, -8 + 0, \underline{-1 + 2}) = 1 \\ x^1_A &= \max(\underline{2}, 1 + 0, -4 + -\infty) = 2 \\ x^1_B &= \max(-16 + 0, \underline{-14 + 2}) = -12 \\ x^2 &= x^1 = v, \text{ stop} , \end{aligned}$$

et en considérant les  $\arg \max$ , on voit qu'il est optimal de s'arrêter quand on est en  $A$ , et qu'en  $B$  ainsi qu'en  $H$ , il faut aller en  $A$ .

**Exercice 11.4.4** Comment peut-on compléter l'algorithme de Ford-Bellman pour construire un circuit de coût négatif ?

**Remarque 11.4.5** Il existe bien des variantes de l'algorithme de Ford-Bellman, qui diffèrent par l'ordre dans lequel on parcourt les nœuds  $i$  et les arcs  $(i, k)$  dans la boucle de l'algorithme 11.4.4 : on trouvera le terme générique de "label correcting algorithms" dans la littérature. Les algorithmes de cette famille, et en particulier l'algorithme de Ford-Bellman, sont parmi les plus rapides pour calculer les chemins de coût minimum, dans le cas d'un graphe orienté qui a des circuits (de coût positif ou nul) et dont les coûts peuvent être négatifs. Dans deux cas spéciaux, on sait faire cependant beaucoup mieux que Ford-Bellman. Lorsque le graphe  $\mathcal{G}$  n'a pas de circuit, on effectue un **tri topologique**, c'est-à-dire que l'on munit les sommets d'un ordre total  $\leq$  tel que s'il y a un chemin de  $i$  à  $j$ , alors  $i \leq j$ . Souvent, les nœuds sont déjà naturellement ordonnés (par exemple par temps croissant). Dans le cas contraire, on sait trouver un tel ordre en temps linéaire, voir par exemple [1] pour plus de détails. (On dit qu'un algorithme est en **temps linéaire** si le temps d'exécution est borné par une constante fois la taille du codage machine des données, c'est ce qu'on peut espérer de mieux, voir la Remarque 11.6.6 pour plus de détails sur la notion de temps de calcul.) Une fois les nœuds ordonnés, il suffit d'initialiser  $v$  à  $+\infty$ , et d'effectuer une et une seule fois la substitution  $v_i \leftarrow \min(\phi_i, \min_{k \in \mathcal{N}_i} (c_{i,k} + v_k))$  pour chaque  $i$ , en parcourant les  $i$  par ordre décroissant, pour obtenir la fonction valeur, ce qui prend un temps linéaire. Un autre cas particulier remarquable est celui où les coûts sont positifs ou nuls : dans ce cas, on peut employer un algorithme de type glouton (cf. Section 11.5), appelé algorithme de Moore-Dijkstra, voir par exemple [1].

**Exercice 11.4.5** Nous reprenons l'Exemple 9.1.3, à savoir le problème classique du sac-à-dos. Nous supposons ici que les poids sont des entiers. On prend donc  $n$  objets de poids respectifs  $p_1, \dots, p_n \in \mathbb{N}$ , et d'utilités respectives  $u_1, \dots, u_n \in \mathbb{R}$ , et l'on note  $P \in \mathbb{N}$  le poids maximal que l'on est disposé à porter. On pose  $x_i = 1$  si on met le  $i$ -ème objet dans le sac-à-dos, et  $x_i = 0$  sinon. On veut maximiser l'utilité du sac à dos sous contrainte de poids

$$\max_{\substack{x \in \{0,1\}^n \\ \sum_{1 \leq i \leq n} x_i p_i \leq P}} \sum_{1 \leq i \leq n} x_i u_i . \quad (11.52)$$

Pour cela, on considère, pour tout  $1 \leq t \leq n$  et  $0 \leq Q \leq P$ , le problème

$$\max_{\substack{x \in \{0,1\}^t \\ \sum_{1 \leq i \leq t} x_i p_i \leq Q}} \sum_{1 \leq i \leq t} x_i u_i , \quad (11.53)$$

dont on note  $v_Q^t$  la valeur optimale. On note  $v^t = (v_Q^t)_{0 \leq Q \leq P}$ .

1. Exprimer  $v^t$  en fonction de  $v^{t-1}$  à l'aide d'une équation de programmation dynamique.
2. En déduire un algorithme pour résoudre (11.52). Quel est le temps d'exécution de l'algorithme ?
3. Appliquer l'algorithme à l'exemple suivant :

$$\max_{\substack{x \in \{0,1\}^3 \\ 2x_1 + 3x_2 + 5x_3 \leq 6}} 8x_1 + 2x_2 + 9x_3 . \quad (11.54)$$

**Exercice 11.4.6 (Chemin de coût minimum avec contrainte de temps)** Soit  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$  un graphe orienté, muni de deux valuations,  $c \in \mathbb{R}^{\mathcal{A}}$ , un coût, et  $\tau \in \mathbb{N}^{\mathcal{A}}$ , un temps, et  $\phi \in (\mathbb{R} \cup \{+\infty\})^{\mathcal{N}}$  une pénalité finale. On fixe un nœud source  $s$  et une date limite  $T \in \mathbb{N}$ , et l'on cherche à trouver un chemin  $(\ell_0, \dots, \ell_m)$  de longueur arbitraire, partant de  $s$  (i.e.  $\ell_0 = s$ ), tel que le gain total  $c_{\ell_0, \ell_1} + \dots + c_{\ell_{m-1}, \ell_m} + \phi_{\ell_m}$  soit minimal, sous la contrainte de respecter la date limite  $T$ , i.e. sous la contrainte  $\tau_{\ell_0, \ell_1} + \dots + \tau_{\ell_{m-1}, \ell_m} \leq T$ . On supposera qu'il n'y a pas de circuit dont tous les arcs ont des temps nuls. Formuler un algorithme de programmation dynamique pour résoudre ce problème. Application : trouver par programmation dynamique le chemin de coût minimum du nœud 1 au nœud 6, en temps au plus 10, pour l'exemple de graphe (11.73) qui sera traité plus loin par relaxation Lagrangienne.

**Exercice 11.4.7** Considérons un amateur de théâtre, qui se rend en Juillet pour une journée voir des pièces du festival off d'Avignon. Le festival off comprend plusieurs centaines de pièces. Chaque pièce est caractérisée par un unique lieu de la ville (un théâtre), et une unique plage horaire dans la journée (par exemple, 16h00-17h30). On connaît en outre les temps nécessaires pour aller d'un théâtre à l'autre. En lisant le programme du off avant d'aller au festival, notre spectateur affecte à chaque pièce un plaisir espéré, mesuré sur une échelle de 0 à 5. Son but est de voir dans la journée une suite de pièces du off, de manière à maximiser la somme des plaisirs espérés pour les différentes pièces choisies. Montrer que ce problème peut se ramener à un chemin à coût minimal dans un graphe sans circuits, et qu'il peut se résoudre en un temps quadratique en le nombre de pièces du festival off (on négligera le temps des repas).

**Exercice 11.4.8** On peut imaginer que, sur la planète Mars, des extraterrestres apprennent aux enfants à compter avec l'addition  $(a, b) \mapsto a \oplus b = \min(a, b)$ , et la multiplication  $(a, b) \mapsto a \otimes b = a + b$  (nous empruntons cette plaisanterie à V.P. Maslov, *Méthodes Opératoires*, MIR, 1973). La structure algébrique correspondante,  $(\mathbb{R} \cup \{+\infty\}, \oplus, \otimes)$  est appelée **semi-anneau min-plus**. Elle vérifie les mêmes axiomes que les anneaux, sauf que l'addition, au lieu d'être une loi de groupe, vérifie  $a \oplus a = a$ . Pour nos martiens, l'équation de Bellman associé au problème en horizon fini (11.41) n'est autre qu'un produit de matrice min-plus

$$v_i^T = \bigoplus_{k \in \mathcal{N}} M_{i,k} \otimes v_k^{T-1},$$

avec  $M_{i,k} = c_{i,k}$  si  $(i, k) \in \mathcal{A}$ , et  $M_{i,k} = +\infty$  sinon. Les martiens, qui utilisent les mêmes notations matricielles que nous, mais dans le semi-anneau min-plus, écrivent tout simplement,

$$v^t = M v_{t-1}, \quad \text{et} \quad v^T = M^T \phi.$$

Quant à l'équation de Bellman (11.51), ils l'écrivent

$$v = M v \oplus \phi. \quad (11.55)$$

Montrer que le coût minimum d'un chemin de longueur  $t$  allant de  $i$  à  $k$  est donné par  $(M^t)_{i,k}$ . Montrer que la solution maximale de (11.55) est égale à  $v = M^* \phi$ , où  $M^* = M^0 \oplus M \oplus M^2 \oplus \dots$ . Retrouver ainsi le Théorème 11.4.2. Montrer que si  $\mathcal{G}$  n'a pas de circuit de coût négatif,  $\lim_{T \rightarrow \infty} M^T = +\infty$  (la matrice dont tous les coefficients sont égaux à  $+\infty$ ). Retrouver ainsi le résultat d'unicité de l'Exercice 11.4.2.



**Exercice 11.4.9** Appelons jeu Markovien fini un jeu à deux joueurs, “blanc”, et “noir”, qui déplacent à tour de rôle un jeton sur un graphe fini, à partir d’une position initiale (c’est blanc qui commence). Certains sommets du graphe sont finaux : lorsque le jeton est dans ce sommet, on sait que la partie est gagnante pour blanc, ou gagnante pour noir, ou bien nulle. Peut-on modéliser les échecs ou les dames par un tel jeu ? On suppose que la partie termine toujours. écrire une équation de type programmation dynamique exprimant la valeur d’une position pour blanc. (Indication : cette équation fera intervenir à la fois les lois min et max.) Dédurre que des trois assertions suivantes, une seule est vraie : - les blancs peuvent toujours gagner ; - les noirs peuvent toujours gagner ; - les blancs et les noirs peuvent toujours forcer la partie nulle.

**Remarque 11.4.6** Arrivé à ce point, le lecteur a peut-être l’impression que tout peut se résoudre par programmation dynamique. C’est presque vrai et la programmation dynamique est un outil puissant, sauf que, tout comme les méthodes Markoviennes en probabilités, la programmation dynamique est frappée de ce qu’on appelle la “malédiction de la dimension” : dans le cas de problèmes vraiment difficiles, l’espace d’état nécessaire peut être très gros (penser au jeu des échecs).

## 11.5 Algorithmes gloutons

### 11.5.1 Généralités sur les méthodes gloutonnes

On dit qu’un algorithme pour minimiser un critère est **glouton** (“greedy”, en anglais), s’il construit une solution admissible en se ramenant à une suite de décisions, que l’on prend à chaque fois au mieux en fonction d’un critère local, en ne remettant jamais en question les décisions précédentes. Lorsque la solution admissible ainsi obtenue est sous-optimale, on parle d’**heuristique gloutonne**. Par exemple, si l’on a un certain nombre de colis de taille variées à mettre dans des containers, et si l’on veut minimiser le nombre de containers utilisés (c’est une version du problème dit de “packing”), on peut imaginer une méthode consistant à trier les colis en fonction du volume, et à rentrer les colis dans les containers en commençant par les plus gros : c’est là un exemple typique d’heuristique gloutonne. L’intérêt des heuristiques gloutonnes est d’être souvent très simples à implémenter. Leur défaut est évidemment leur myopie, ainsi que la difficulté à évaluer l’écart à l’optimum. Il est cependant des classes particulières de problèmes pour lesquels une méthode gloutonne fournit une solution optimale (voir à ce sujet la Remarque 11.5.5 ci-dessous). Nous nous contenterons de présenter dans le paragraphe qui suit un exemple fondamental d’algorithme glouton fournissant une solution optimale.

### 11.5.2 Algorithme de Kruskal pour le problème de l’arbre couvrant de coût minimum

On s’intéresse maintenant à un graphe non-orienté. Nous notons  $\mathcal{V}$  l’ensemble des sommets, et  $\mathcal{E}$  l’ensemble des arêtes, qui est un sous-ensemble de l’ensemble des parties à deux éléments de  $\mathcal{V}$ . On a donc  $\{i, j\} \in \mathcal{E}$  s’il y a une arête reliant les sommets  $i$  et  $j$ . Notons que pour bien distinguer le cas non-orienté du cas orienté, nous parlons de

sommets et d'arêtes, alors que nous avons parlé de nœuds et d'arcs pour un graphe orienté (les lettres “ $\mathcal{V}$ ” et “ $\mathcal{E}$ ” renvoient aux mots anglais “vertices” et “edges”).

Un graphe (non-orienté) est dit **connexe** si deux sommets quelconques peuvent être reliés par un chemin. Un graphe (non-orienté) quelconque peut être décomposé en **composantes connexes**, qui sont par définition les classes d'équivalence pour la relation  $R$  telle que  $iRj$  s'il y a un chemin reliant  $i$  et  $j$ . On appelle **forêt** un graphe (non-orienté) sans circuit. Un **arbre** est une forêt connexe. On dit qu'un sous-graphe  $\mathcal{G}'$  couvre un graphe  $\mathcal{G}$  si chaque sommet de  $\mathcal{G}$  est extrémité d'au moins une arête de  $\mathcal{G}'$ . Étant donné un graphe (non-orienté) connexe  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , dont les arêtes sont munies d'une fonction coût  $\mathcal{E} \rightarrow \mathbb{R}$ ,  $\{i, j\} \mapsto c_{ij}$ , le problème de l'**arbre couvrant de coût minimum** consiste à trouver un arbre couvrant  $\mathcal{T}$  dont le coût

$$\sum_{\{i,j\} \in \mathcal{T}} c_{ij} ,$$

est minimum. Ce problème d'optimisation se rencontre par exemple dans les problèmes de câblage, lorsque l'on veut connecter électriquement un ensemble de points tout en minimisant la longueur de fil.

L'algorithme de Kruskal construit une suite de forêts. On part de la forêt comprenant tous les sommets et aucune arête. A chaque étape, on choisit de rajouter à la forêt, parmi toutes les arêtes dont l'adjonction ne crée pas de circuit, celle dont le coût est minimum. L'algorithme termine quand la forêt est un arbre couvrant.

**Théorème 11.5.1** *Si  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  est un graphe (non-orienté) connexe, muni d'une fonction coût arbitraire  $c: \mathcal{E} \rightarrow \mathbb{R}$ , l'algorithme de Kruskal fournit un arbre couvrant de coût minimum.*

Afin de montrer le Théorème 11.5.1, nous énonçons une propriété très élémentaire des arbres couvrants, dont la vérification est laissée au lecteur en guise d'exercice.

**Lemme 11.5.2 (Lemme d'échange)** *Si  $\mathcal{T}$  est un arbre couvrant d'un graphe  $\mathcal{G}$ , et si  $i$  et  $j$  sont deux sommets de  $\mathcal{G}$ , il existe un unique chemin de  $\mathcal{T}$  reliant  $i$  à  $j$ . En outre, si  $\{i, j\}$  est une arête de  $\mathcal{G}$  qui n'appartient pas à  $\mathcal{T}$ , on obtient encore un arbre couvrant si l'on remplace une arête quelconque du chemin de  $\mathcal{T}$  reliant  $i$  à  $j$  par  $\{i, j\}$ .*

Nous pouvons maintenant énoncer la condition d'optimalité.

**Lemme 11.5.3** *Soit  $\mathcal{G}$  et  $c$  comme dans le Théorème 11.5.1, et soit  $\mathcal{T}$  un arbre couvrant  $\mathcal{G}$ . Les assertions suivantes sont équivalentes.*

1.  $\mathcal{T}$  est de coût minimum ;
2. pour chaque arête  $\{i, j\}$  qui n'est pas dans  $\mathcal{T}$ , l'unique chemin de  $\mathcal{T}$  reliant  $i$  et  $j$  est formé d'arêtes dont chacune est de coût inférieur ou égal à  $c_{ij}$  ;
3. pour chaque arête  $\{r, s\}$  de  $\mathcal{T}$ , pour toute composante connexe  $C$  du graphe obtenu en retirant  $\{r, s\}$  de  $\mathcal{T}$ , et pour toute arête  $\{i, j\}$  dont une extrémité et une seule est dans  $C$ ,  $c_{rs} \leq c_{ij}$ .

**Démonstration.** L'implication (non 2)  $\Rightarrow$  (non 1) résulte du Lemme d'échange 11.5.2 : si  $\{i, j\}$  est une arête qui n'est pas dans  $\mathcal{T}$ , et si  $\{r, s\}$  est une arête de l'unique chemin de  $\mathcal{T}$  reliant  $i$  et  $j$ , telle que  $c_{ij} < c_{rs}$ , on obtient en mettant  $\{i, j\}$  à la place de  $\{r, s\}$  dans  $\mathcal{T}$  un nouvel arbre couvrant de coût strictement inférieur à celui de  $\mathcal{T}$ .

Nous montrons maintenant (non 3)  $\Rightarrow$  (non 2). Remarquons tout d'abord que le graphe obtenu en retirant  $\{r, s\}$  de  $\mathcal{T}$  a exactement deux composantes connexes,  $C$  et  $\overline{C} = \mathcal{V} \setminus C$ . Supposons qu'on ait une arête  $\{i, j\}$  avec  $i \in C$  et  $j \in \overline{C}$ , telle que  $c_{ij} < c_{rs}$ . L'unique chemin reliant  $i$  à  $j$  dans  $\mathcal{T}$  contient nécessairement  $\{r, s\}$ , ce qui montre que la condition (2) n'est pas satisfaite.

Nous montrons enfin (3)  $\Rightarrow$  (1). Soit  $\mathcal{T}$  un arbre vérifiant (3), et soit  $\mathcal{T}'$  un arbre de coût optimal, que l'on peut choisir tel que le nombre d'arêtes en commun entre  $\mathcal{T}$  et  $\mathcal{T}'$  soit maximal. On va montrer que  $\mathcal{T} = \mathcal{T}'$ . Dans le cas contraire, on peut trouver une arête  $\{r, s\}$  dans  $\mathcal{T}$  et pas dans  $\mathcal{T}'$ . Le graphe obtenu en rajoutant  $\{r, s\}$  à  $\mathcal{T}'$  contient un circuit,  $\mathcal{C}$ , contenant l'arête  $\{r, s\}$ . Considérons maintenant les deux composantes connexes  $C$  et  $\overline{C}$  du graphe obtenu en enlevant  $\{r, s\}$  à  $\mathcal{T}$ . Comme le circuit  $\mathcal{C}$  contient déjà une arête reliant  $C$  à  $\overline{C}$ , à savoir  $\{r, s\}$ , il doit nécessairement contenir une autre arête,  $\{i, j\}$ , reliant  $C$  et  $\overline{C}$ . Le graphe  $\mathcal{T}''$  obtenu en remplaçant  $\{i, j\}$  par  $\{r, s\}$  dans  $\mathcal{T}'$  est encore un arbre couvrant, d'après le lemme d'échange, et la condition (3) montre que le coût de  $\mathcal{T}''$  est inférieur ou égal à celui de  $\mathcal{T}'$ . Ainsi,  $\mathcal{T}''$  est encore un arbre couvrant de coût minimum, et comme  $\mathcal{T}''$  a en commun avec  $\mathcal{T}$  une arête de plus que  $\mathcal{T}'$ , on contredit l'hypothèse de maximalité dans la définition de  $\mathcal{T}'$ . On a montré  $\mathcal{T}' = \mathcal{T}$ .  $\square$

**Démonstration du Théorème 11.5.1.** Il est immédiat que l'algorithme de Kruskal, appliqué à un graphe connexe, termine avec un arbre couvrant. Cet arbre vérifie l'assertion 2 du Lemme 11.5.3.  $\square$

**Remarque 11.5.4** Il est possible d'implémenter l'algorithme de Kruskal en temps  $O(|\mathcal{E}| \log |\mathcal{E}|)$  où  $|\mathcal{E}|$  est le nombre d'arêtes du graphe, voir [13].  $\bullet$

**Remarque 11.5.5** L'optimalité de méthodes gloutonnes est toujours le signe de fortes propriétés de structure. Par exemple, le lecteur a déjà rencontré un algorithme glouton en algèbre linéaire : on peut voir le problème consistant à fabriquer une base d'un espace vectoriel de dimension finie  $E$  comme un problème d'optimisation, consistant à maximiser le cardinal d'une famille libre de  $E$ . L'algorithme qui consiste à partir d'une famille vide, et à chaque étape, à rajouter dans la famille un vecteur quelconque de  $E$  qui n'est pas combinaison linéaire des vecteurs déjà dans la famille, est un algorithme glouton (qui fournit une solution optimale, c'est-à-dire une base). La preuve de correction de la méthode repose sur un résultat simple d'algèbre linéaire, le lemme d'échange, dont le lecteur aura noté l'analogie avec le Lemme 11.5.2 ci-dessus. Plus généralement, les propriétés qui permettent de montrer que l'algorithme glouton est correct ont été étudiées dans le cadre de la théorie des matroïdes et anti-matroïdes, voir notamment [13].  $\bullet$

## 11.6 Séparation et relaxation de problèmes combinatoires

Dans la résolution d'un problème de RO, une des premières tâches est de reconnaître si une méthode exacte efficace (algorithmes de chemins, flots, algorithme glouton, etc.) s'applique. Que faire, cependant, quand même avec une bonne modélisation, les méthodes polynomiales ne s'appliquent pas? On peut bien sûr recourir à des heuristiques particulières, ou à des méta-heuristiques comme le recuit simulé ou la recherche tabou, voir à ce sujet la Remarque 11.6.7. Dans l'esprit de ce cours, nous nous concentrerons plutôt sur les méthodes exactes, à base de programmation mathématique et d'exploration arborescente, qui permettent de prouver l'optimalité de la solution trouvée, ou tout au moins, de mesurer l'écart à l'optimum.

### 11.6.1 Séparation et évaluation (branch and bound)

Considérons le problème combinatoire très général

$$\min_{x \in X} J(x) , \quad (11.56)$$

avec  $X$  fini (mais gros), et  $J : X \rightarrow \mathbb{R}$ . Dans la méthode de séparation et évaluation (en anglais, "branch and bound"), la **séparation** consiste à représenter l'ensemble  $X$  des solutions admissibles par les feuilles d'un arbre, que l'on va explorer. Les nœuds internes de l'arbre représentent des décisions partielles (correspondant à fixer certaines variables de décision, mais pas toutes), le nœud racine représente une situation initiale, dans laquelle on n'a encore rien décidé. L'**évaluation** s'intéresse, pour chaque nœud interne  $s$  de l'arbre, au **coût conditionnel**

$$v(s) = \min_{s' \text{ est une feuille descendante de } s} J(s') . \quad (11.57)$$

(Le mot "descendant" a le sens généalogique, i.e. nous orientons l'arbre avec la racine en haut, et les feuilles en bas.) Le calcul de ce coût conditionnel en  $s$  est souvent aussi dur que le problème initial (11.56) (en particulier, quand  $s$  est la racine, (11.57) coïncide avec (11.56)), c'est pourquoi nous allons simplifier le problème (11.57), en nous autorisant l'introduction d'un minorant  $b(s)$  du coût conditionnel en  $s$  :

$$b(s) \leq v(s) , \quad (11.58)$$

que l'on devra définir pour tout nœud interne de l'arbre. Nous suivrons l'usage, qui appelle (improprement) **borne inférieure** ou tout simplement **borne** le minorant  $b(s)$ .

#### Énoncé de l'algorithme de séparation et évaluation

L'algorithme de séparation et évaluation consiste à explorer les nœuds de l'arbre, en partant de la racine. En cours d'exploration, on mémorise  $m$ , le coût minimal des

solutions déjà trouvées, ainsi que la solution correspondante. On initialise l'algorithme avec  $m = +\infty$ . Quand on visite pour la première fois un nœud interne  $s$  du graphe, on évalue la borne  $b(s)$ . Si  $b(s) \geq m$ , il ne sert à rien d'explorer les branches filles du nœud  $s$ , car le coût des feuilles qui s'y trouvent n'est pas meilleur que le coût de la meilleure solution rencontrée, et l'on remonte au nœud père de  $s$  afin de poursuivre l'exploration de l'arbre (on peut visualiser cela en disant que l'on **coupe** la branche de l'arbre partant du nœud  $s$ ). Si au contraire  $b(s) < m$ , il est possible que la branche partant de  $s$  contienne une solution améliorant  $m$  : on poursuit dans ce cas l'exploration de l'arbre, en passant aux nœuds fils de  $s$ . Quand on parvient à une feuille de l'arbre, qui représente une solution admissible  $x \in X$  de (11.56), il ne reste qu'à calculer la valeur  $J(x)$  : si  $J(x) < m$ , la meilleure solution trouvée est  $x$ , on pose donc  $m = J(x)$ , et l'on mémorise  $x$  à la place de l'ancienne meilleure solution trouvée. On poursuit alors l'exploration de l'arbre en remontant au nœud père de la feuille courante.

L'algorithme visite au plus une fois chaque feuille. Le pire des cas est celui où la borne  $b$  ne permet jamais de couper de branche : l'algorithme revient dans ce cas à énumérer toutes les solutions admissibles de (11.56).

### Illustration : exemple du voyageur de commerce

Afin de détailler la méthode, considérons le problème du voyageur de commerce, déjà évoqué dans l'Exemple 9.1.4. Nous allons traiter ici la version non-orientée du voyageur de commerce, qui considère un graphe non-orienté **complet**  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  (complet signifie que  $\mathcal{E}$  est formé de toutes les paires de sommets de  $\mathcal{V}$ ). On prendra  $\mathcal{V} = \{1, \dots, n\}$ , avec  $n = |\mathcal{V}|$ . Nous associons à chaque paire de sommets  $\{i, j\}$  un temps qu'on note  $t_{ij}$  (ce qui est une abréviation pour  $t_{\{i,j\}}$ , on a donc  $t_{ij} = t_{ji}$  puisque le graphe est non-orienté). Le but est de trouver un tour, c'est-à-dire une suite de sommets  $\ell_1, \dots, \ell_n$  comprenant chaque sommet de  $\mathcal{G}$  une et une seule fois, et telle que le temps total

$$t_{\ell_1 \ell_2} + t_{\ell_2 \ell_3} + \dots + t_{\ell_n \ell_1} \quad (11.59)$$

soit minimum.

En utilisant l'invariance du critère (11.59) par permutation circulaire, on peut toujours supposer que l'on part du sommet 1, soit  $\ell_1 = 1$ . Le tour est alors spécifié de manière unique par la suite  $\ell_2, \dots, \ell_{n-1}$ . La **séparation** du problème revient à organiser le choix d'un tour en une suite de décisions. Par exemple, on pourra considérer le choix de  $\ell_2 \in \mathcal{V} \setminus \{\ell_1\}$  comme une première décision, le choix de  $\ell_3 \in \mathcal{V} \setminus \{\ell_1, \ell_2\}$  comme une seconde décision, et ainsi de suite jusqu'à  $\ell_{n-1}$ . Un sommet interne de l'arbre de séparation correspondra donc à une sous-suite  $(1 = \ell_1, \ell_2, \dots, \ell_k)$ , avec  $k \leq n - 2$ , que nous appellerons "tour partiel". Il s'agit de borner inférieurement le coût conditionnel (11.57), i.e. de minorer le temps total des tours qui commencent par  $\ell_1, \dots, \ell_k$ . On peut donner par exemple la borne naïve

$$b_1(\ell_1, \dots, \ell_k) = t_{\ell_1 \ell_2} + \dots + t_{\ell_{k-1} \ell_k} + \min_{j \in \mathcal{V} \setminus \{\ell_1, \dots, \ell_{k-1}\}} t_{\ell_k j} +$$

$$\min_{m \in \mathcal{V} \setminus \{\ell_2, \dots, \ell_k\}} t_{m\ell_1} + (n - k - 2) \left( \min_{\substack{j, m \in \mathcal{V} \\ j \neq m}} \min_{\ell_1, \dots, \ell_k} t_{jm} \right). \quad (11.60)$$

En effet, le temps du tour (11.59) est la somme du temps du tour partiel  $(\ell_1, \ell_2, \dots, \ell_k)$ , soit  $t_{\ell_1\ell_2} + \dots + t_{\ell_{k-1}\ell_k}$ , plus du temps de l'arête  $\{\ell_k, \ell_{k+1}\}$ , que l'on minore par le premier min dans (11.60), plus du temps de l'arête  $\{\ell_n, \ell_1\}$ , que l'on minore symétriquement par le second min dans (11.60), et enfin, du temps du chemin  $(\ell_{k+1}, \dots, \ell_n)$ . Ce chemin étant de longueur  $n - k - 2$  et ne comprenant aucun sommet de  $\{\ell_1, \dots, \ell_k\}$ , on peut minorer son temps par le dernier min dans (11.60), ce qui montre que  $b_1(\ell_1, \dots, \ell_k) \leq v(\ell_1, \dots, \ell_k)$ .

Appliquons maintenant l'algorithme de séparation et évaluation, avec la borne  $b_1$ , à un petit exemple de voyageur de commerce. Considérons la ville de type Manhattan représentée sur la Figure 11.3. On prendra l'ensemble  $\mathcal{V} = \{1, \dots, 5\}$  dont les éléments correspondent aux points représentés sur le dessin, de coordonnées  $P_1 = (0, 0)$ ,  $P_2 = (3, 0)$ ,  $P_3 = (1, 1)$ ,  $P_4 = (3, 2)$ , et  $P_5 = (0, 3)$  (dans un repère dirigé vers l'Est et le Sud), et  $t_{ij}$  représentera le temps de marche du point  $P_i$  au point  $P_j$ , c'est-à-dire la norme  $\|P_i - P_j\|_1$ .

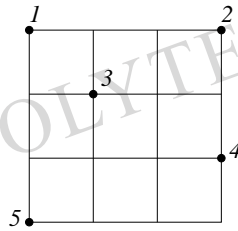


FIGURE 11.3 – Un voyageur de commerce dans Manhattan

Le parcours d'arbre correspondant est représenté sur la Figure 11.4. On part du nœud (1), qui correspond à un tour partiel vide partant du point 1 de la ville. La première étape de l'algorithme consiste à choisir le point suivant de la ville que l'on va visiter, qui peut être 2, 3, 4, ou 5. Choisissons par exemple le point 2, ce qui nous amène au nœud (1, 2) que nous avons représenté à gauche de l'arbre, avec le tour partiel qui lui correspond. Comme nous n'avons pas de solution admissible pour l'instant,  $m = +\infty$ , et comme le test  $b_1(1, 2) < m$  est automatiquement vérifié, nous ne calculons pas  $b_1(1, 2)$ . Poursuivons le parcours en profondeur d'abord : on arrive au nœud (1, 2, 3) (nous avons à nouveau représenté le tour partiel), puis au nœud (1, 2, 3, 4), qui est une feuille, car il détermine de manière unique la solution admissible  $x = (1, 2, 3, 4, 5)$ , de temps  $m = 16$ . Le tour complet ainsi obtenu est représenté à l'endroit de la feuille. Étant arrivée à une feuille, la recherche en profondeur remonte au nœud père, pour redescendre à la feuille suivante, qui représente la solution (1, 2, 3, 5, 4), de temps 18 pire que  $m$ . Le prochain nouveau nœud exploré est (1, 2, 4), et l'on calcule  $b_1(1, 2, 4) = 13 < m$  : on explore donc un premier fils de (1, 2, 4), qui fournit la solution  $x = (1, 2, 4, 5, 3)$ , de temps 14, ce qui est mieux que  $m$  : on pose donc  $m = 14$ . L'autre

fil de  $(1, 2, 4)$  fournit une autre solution de temps 14,  $(1, 2, 4, 3, 5)$ . Le nouveau nœud suivant visité est  $(1, 2, 5)$ , avec  $b_1(1, 2, 5) = 17 \geq m$  : on coupe donc le sous-arbre partant de  $(1, 2, 5)$ , et le nouveau nœud suivant est  $(1, 3)$ . Nous laissons le lecteur finir le calcul, et montrer ainsi que  $x = (1, 2, 4, 5, 3)$  est un tour dont le temps 14 est optimal.

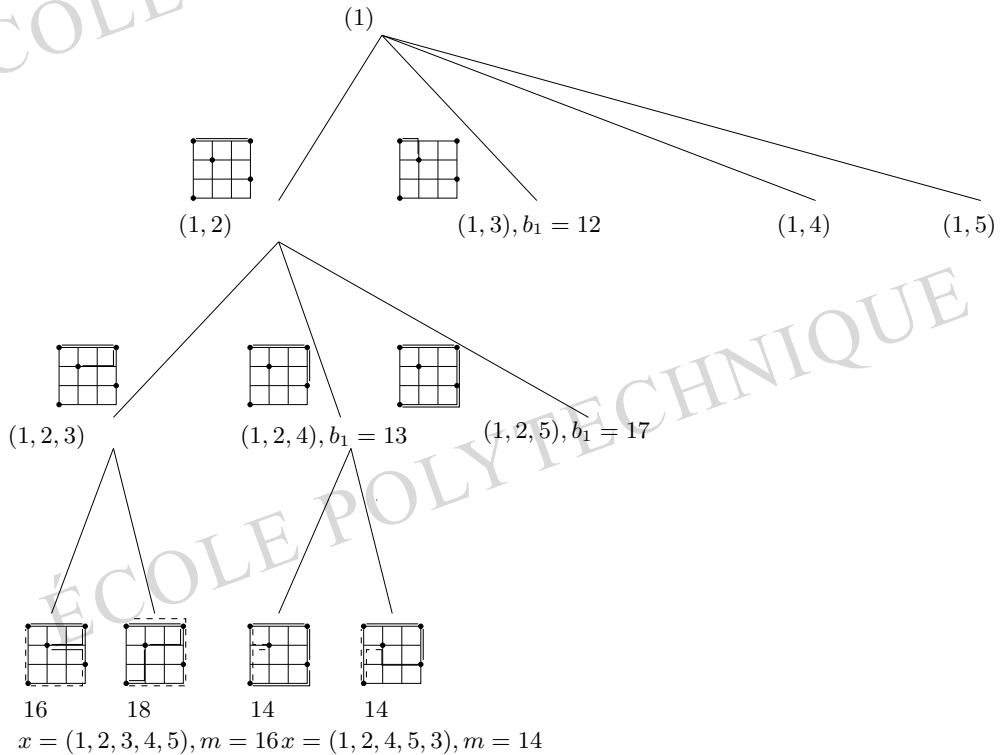


FIGURE 11.4 – Une partie de l'arbre de séparation et évaluation, pour le problème de voyageur de commerce de la Figure 11.3.

**Remarque 11.6.1** Le problème du voyageur de commerce dans Manhattan est un cas particulier du voyageur de commerce métrique, dans lequel  $t_{ij}$  est la distance de  $i$  à  $j$  pour une certaine métrique, ici, celle de la norme  $\|\cdot\|_1$ . Une application classique du voyageur de commerce métrique est le perçage de circuits imprimé, si l'on considère le cas d'une perceuse devant percer une suite de trous (suivant le type de la mécanique de l'outil, on obtient un voyageur de commerce pour la norme euclidienne, pour la norme  $\|\cdot\|_\infty$ , etc.). Le problème du voyageur de commerce (métrique ou général) est un exemple classique de problème NP-difficile (voir la Remarque 11.6.6).

### De l'importance de la qualité de la borne

Le lecteur pourrait croire que la borne  $b_1$  est raisonnable, mais voyons comment l'algorithme se comporte quand on augmente le nombre de sommets du graphe. Nous avons programmé l'algorithme qui précède, et résolu des instances similaires de voyageur de commerce dans Manhattan, mais en faisant varier le nombre  $n$  de sommets. Voici un jeu de résultats typique, donnant le nombre de nœuds de l'arbre de séparation et évaluation, en fonction de  $n$  :

$n$	5	6	7	8	9	10	11	12	13	14	15
arbre	10	20	51	805	2175	10598	58414	199276	499887	1250530	3598585

Le cas  $n = 15$  prend déjà une minute sur un PC usuel. Nous venons de rencontrer ce qu'on appelle l'**explosion combinatoire**.

Rétrospectivement, le caractère grossier de la borne  $b_1$  apparaît : lorsque  $n - k$  est grand, il est mauvais de minorer la longueur du tour partiel complémentaire de  $(\ell_1, \dots, \ell_k)$  par  $(n - k - 2)$  fois le minimum des temps des arêtes entre sommets restants, cela revient à multiplier une erreur par un terme d'ordre  $n$ . On voit ici que dans un algorithme de séparation et d'évaluation, il est indispensable d'avoir une borne qui capture suffisamment la "physique" du problème.

### Borne du 1-arbre pour le voyageur de commerce

Donnons un premier exemple d'une telle borne pour le voyageur de commerce non-orienté. Étant donné un tour partiel  $(\ell_1, \ell_2, \dots, \ell_k)$  de  $\mathcal{G}$ , avec  $k \leq n - 2$ , construisons le graphe  $\mathcal{G}'$  induit par le sous ensemble de sommets  $\mathcal{V}' = \mathcal{V} \setminus \{\ell_1, \dots, \ell_k\}$ , c'est-à-dire le graphe d'ensemble de sommets  $\mathcal{V}'$  et d'ensemble d'arêtes  $\mathcal{E}' = \{\{i, j\} \in \mathcal{E} \mid i, j \in \mathcal{V}'\}$ . Nous notons  $\text{ac}(\mathcal{G}')$  le coût minimum d'un arbre couvrant  $\mathcal{G}'$ . (Rappelons que la notion d'arbre couvrant a été définie dans la Section 11.5, où nous avons aussi vu qu'un arbre couvrant de coût minimum se calcule en temps  $O(|\mathcal{E}| \log |\mathcal{E}|)$  par l'algorithme de Kruskal.) On a la borne inférieure suivante

$$b_2(\ell_1, \dots, \ell_k) = t_{\ell_1 \ell_2} + \dots + t_{\ell_{k-1} \ell_k} + \min_{j \in \mathcal{V} \setminus \{\ell_1, \dots, \ell_{k-1}\}} t_{\ell_k j} + \min_{m \in \mathcal{V} \setminus \{\ell_2, \dots, \ell_k\}} t_{m \ell_1} + \text{ac}(\mathcal{G}') \quad (11.61)$$

Dans le cas spécial où le tour partiel est de longueur nulle, soit  $k = 1$ , on peut raffiner très légèrement la borne  $b_2$  en notant que dans ce cas, les arêtes  $\{\ell_1, \ell_2\}$  et  $\{\ell_k, \ell_1\}$  du tour doivent être distinctes (à condition de supposer que le graphe a au moins trois sommets), ce qui donne la nouvelle borne

$$b'_2(\ell_1) = \min_{j, m \in \mathcal{V} \setminus \{\ell_1\}, j \neq m} (t_{\ell_1 j} + t_{m \ell_1}) + \text{ac}(\mathcal{G}') \quad (11.62)$$

Cette dernière borne, classique, est connue sous le nom de **borne du 1-arbre** (un 1-arbre d'un graphe est un sous-graphe formé d'une part d'un arbre couvrant tous les sommets hormis un sommet distingué noté "1", et d'autre part de deux arêtes



distinctes dont 1 est l'extrémité). Par exemple, pour le graphe de la Figure 11.3, l'algorithme de Kruskal fournit le 1-arbre de coût minimal représenté sur la Figure 11.5. Le coût de ce 1-arbre est  $b'_2(1) = 13$ . En revenant à l'arbre de la Figure 11.4, on voit que remplacer  $b_1$  par  $b_2$  aurait permis de ne pas visiter les descendants du nœud  $(1, 3)$ , car  $b_2(1, 3) = 14$  est supérieur (en fait, égal) à la valeur  $m = 14$  de la meilleure solution rencontrée avant la visite de  $(1, 3)$ .

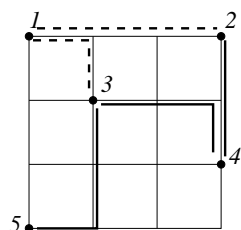


FIGURE 11.5 – Borne du 1-arbre  $b'_2(1)$ , pour le problème du voyageur de commerce de la Figure 11.3. L'arbre couvrant  $\mathcal{G}'$  est en traits gras, les deux arêtes connectant cet arbre au sommet 1 sont en traits pointillés.

### Du choix de l'arbre et de l'ordre d'exploration des branches

Il y a en général plusieurs manières de représenter l'ensemble des solutions admissibles  $X$ , et le choix est souvent suggéré par la technique servant à fabriquer la borne. Ainsi, nous verrons dans la Sous-section 11.6.2 qu'on peut modéliser le voyageur de commerce par un programme linéaire en variables entières, en introduisant pour chaque  $\{i, j\} \in \mathcal{E}$  une variable  $x_{ij}$  valant 1 si l'arête  $\{i, j\}$  appartient au tour considéré et 0 sinon. Avec une telle modélisation, on peut construire un arbre de séparation et d'évaluation binaire, dans lequel une décision élémentaire consiste à fixer la valeur d'une variable  $x_{ij}$  à 0 ou à 1. Un autre paramètre déterminant est l'ordre dans lequel on visite les sommets : il est judicieux d'examiner le plus tôt possible (i.e., près de la racine de l'arbre) les décisions dont on pense qu'elles ont le plus d'influence sur le coût de la solution, le but étant de couper les branches le plus haut possible dans l'arbre. Dans le même esprit, initialiser  $m$  avec la valeur d'une solution obtenue heuristiquement, au lieu de  $+\infty$ , ne fait qu'aider à couper plus tôt les branches.

### 11.6.2 Relaxation de problèmes combinatoires

Une manière systématique d'obtenir des bornes inférieures pour la valeur optimale du coût d'un problème combinatoire consiste à **relâcher** (ou relaxer) le problème, c'est-à-dire à grossir l'ensemble admissible de manière à obtenir un problème plus facile, fournissant une borne inférieure pour le problème initial. Quand l'ensemble admissible est représenté par des contraintes, une manière de relâcher est d'oublier tout simplement certaines contraintes. Nous avons déjà vu un exemple de relaxation avec la borne du 1-arbre pour le voyageur de commerce : les tours sont précisément

les 1-arbres tels que chaque sommet a deux voisins. Nous allons maintenant présenter des techniques générales de relaxation.

### Relaxations continues

L'efficacité des outils de programmation linéaire suggère souvent de modéliser les problèmes combinatoires par des programmes linéaires en nombre entiers : on obtient alors une borne inférieure en relâchant les contraintes d'intégrité.

Présentons pour illustrer cette idée la relaxation du voyageur de commerce proposée par Dantzig, Fulkerson, et Johnson dans un article de 1954, qui résolvait à l'époque un problème à 49 villes. La postérité des idées de cet article permet de résoudre aujourd'hui exactement des instances à plusieurs milliers de villes (pour un historique et un état de l'art récent, on pourra se reporter à "On the solution of traveling salesman problems", D. Applegate, R. Bixby, V. Chvátal, and W. Cook, *Documenta Math.*, Extra volume ICM 1998, III - 645-656, et plus généralement à la toile <http://www.math.princeton.edu/tsp>).

Nous reprenons les notations de la section précédente pour le voyageur de commerce :  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  est un graphe non orienté complet (i.e.  $\mathcal{E}$  contient toutes les arêtes reliant deux éléments de  $\mathcal{V}$ ), avec une fonction temps  $t : \mathcal{E} \rightarrow \mathbb{R}$ ,  $\{i, j\} \mapsto t_{ij}$ . A chaque tour, on associe un vecteur  $x \in \{0, 1\}^{\mathcal{E}}$  tel que  $x_{ij} = 1$  si  $\{i, j\}$  fait partie du tour, et  $x_{ij} = 0$  sinon. Réciproquement, un vecteur  $x \in \{0, 1\}^{\mathcal{E}}$  représente le sous-graphe de  $\mathcal{G}$  ayant pour arêtes les  $\{i, j\}$  tels que  $x_{ij} = 1$ . Il nous faut maintenant exprimer par des contraintes linéaires le fait que  $x \in \{0, 1\}^{\mathcal{E}}$  représente un tour. Comme chaque sommet d'un tour a exactement deux voisins,  $x$  vérifie nécessairement

$$\sum_{k \in \mathcal{V}, \{k, j\} \in \mathcal{E}} x_{kj} = 2, \text{ pour tout } j \in \mathcal{V}. \quad (11.63)$$

Les contraintes (11.63) ne suffisent pas à caractériser un tour, car l'ensemble des arêtes  $\{i, j\}$  telles que  $x_{ij} = 1$  peut ne pas être connexe : en fait, on peut voir que les  $x \in \{0, 1\}^{\mathcal{E}}$  solutions de (11.63) représentent exactement les unions disjointes de circuits. Afin d'éliminer les solutions parasites, on peut rajouter les contraintes suivantes, dites de **sous-tour**,

$$\text{pour tout } \mathcal{S} \subset \mathcal{V}, \text{ tel que } \mathcal{S} \neq \emptyset \text{ et } \mathcal{S} \neq \mathcal{V}, \quad \sum_{k, m \in \mathcal{S}, \{k, m\} \in \mathcal{E}} x_{km} \leq |\mathcal{S}| - 1. \quad (11.64)$$

On voit facilement que  $x \in \{0, 1\}^{\mathcal{E}}$  vérifie (11.63) et (11.64) si, et seulement si, il représente un tour : en effet, tout vecteur associé à un tour vérifie ces contraintes, et réciproquement, comme tout  $x \in \{0, 1\}^{\mathcal{E}}$  vérifiant (11.63) représente une union disjointe de circuits, il suffit de prendre pour  $\mathcal{S}$  l'ensemble des sommets de l'un quelconque de ces circuits pour obtenir  $\sum_{k, m \in \mathcal{S}, \{k, m\} \in \mathcal{E}} x_{km} = |\mathcal{S}|$ , et si  $x$  vérifie les inégalités de sous-tour (11.64), on a nécessairement  $\mathcal{S} = \mathcal{V}$ , ce qui montre que  $x$  représente un tour. Ceci nous permet de formuler le problème de voyageur de commerce comme un

programme linéaire en nombres entiers

$$(VC) : \min \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} \quad \text{sous les contraintes } x \in \{0,1\}^{\mathcal{E}}, \quad (11.63), (11.64),$$

et on obtient aussitôt une borne inférieure en considérant le programme linéaire relaxé en variables continues

$$(VC)_{\text{rel}} : \min \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} \quad \text{sous les contraintes } 0 \leq x_{ij} \leq 1, \quad (11.63), (11.64).$$

Malgré le nombre exponentiel de contraintes dans (11.64), il est possible de résoudre efficacement  $(VC)_{\text{rel}}$  en procédant comme suit. On commence à minimiser  $\sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij}$  sous les contraintes  $0 \leq x_{ij} \leq 1$  et (11.63), en oubliant les contraintes de sous-tour (11.64). On trouve ainsi un premier  $x \in [0,1]^{\mathcal{E}}$ . On va ensuite chercher s'il existe un sous ensemble  $\mathcal{S} \subset \mathcal{V}$ ,  $\mathcal{S} \neq \emptyset$ ,  $\mathcal{S} \neq \mathcal{V}$ , pour lequel l'inégalité de (11.64) est violée, ce qui peut se faire (Exercice 11.6.1) très efficacement. S'il n'en existe pas, on a résolu  $(VC)_{\text{rel}}$ . Si au contraire on a trouvé un  $\mathcal{S}$  tel que (11.64) ne soit pas satisfaite, l'on minimise à nouveau  $\sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij}$  sous les contraintes  $0 \leq t_{ij} \leq 1$  et (11.63), en rajoutant l'inégalité de sous-tour associée à  $\mathcal{S}$ . En poursuivant cette suite de minimisations et rajouts successifs d'inégalités, on aboutit finalement à une solution de  $(VC)_{\text{rel}}$ . Cette méthode peut s'interpréter géométriquement en parlant de **coupes**. Notons en effet  $P$  le polytope des solutions admissible du problème  $(VC)_{\text{rel}}$ , et soit  $t : x \mapsto \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij}$  la forme linéaire que l'on minimise. La méthode que l'on a esquissée revient à définir une suite décroissante de polytopes  $P^1 \supset P^2 \supset \dots \supset P$ . On prend tout d'abord pour  $P^1$  l'ensemble défini par  $0 \leq x_{ij} \leq 1$  et (11.63). On minimise d'abord  $t$  sur  $P^1$ , et le minimum est atteint en un point  $x^1$ . Si  $x^1$  est dans  $P$ , on a résolu  $(VC)_{\text{rel}}$ , sinon détecter une contrainte (11.64) violée par  $x^1$  revient à **séparer**  $x^1$  de  $P$ , c'est-à-dire à trouver un demi-espace  $H^1$  particulier tel que  $x^1 \notin H^1$ , et  $P \subset H^1$ , et l'étape suivante revient à minimiser  $t$  sur le nouveau polytope  $P^2 = P^1 \cap H^1$ , obtenu en "coupant"  $P^1$  par  $H$ . L'on fabrique ainsi une suite décroissante de polytopes dont on peut dire intuitivement qu'ils "approchent"  $P$  au voisinage du point où  $t$  est minimal. De telles techniques peuvent même fournir la solution du problème original  $(VC)$ , si l'on est capable de trouver d'autres coupes, de nature combinatoire, permettant d'approcher suffisamment bien la clôture entière  $P_e$  de  $P$ .

**Exercice 11.6.1** Montrer que dans le problème  $(VC)_{\text{rel}}$ , on peut remplacer les contraintes (11.64) par les contraintes

$$\text{pour tout } \mathcal{S} \subset \mathcal{V}, \text{ tel que } \mathcal{S} \neq \emptyset \text{ et } \mathcal{S} \neq \mathcal{V}, \quad \sum_{k \in \mathcal{V}, m \in \mathcal{S} \setminus \mathcal{V}, \{k,m\} \in \mathcal{E}} x_{km} \geq 2. \quad (11.65)$$

Montrer, en utilisant le théorème de Ford et Fulkerson (Question 4 de l'Exercice 11.3.8) que l'on peut vérifier si un vecteur  $x \in [0,1]^{\mathcal{E}}$  satisfait (11.65) à l'aide d'un algorithme de flot.

**Relaxations Lagrangiennes**

Considérons le problème très général

$$\begin{aligned} & \text{minimiser } J(x) \text{ sous les contraintes :} \\ & x \in X, \\ & F_i(x) \leq 0, \quad i = 1, \dots, m, \\ & F_i(x) = 0, \quad i = m+1, \dots, m+q, \end{aligned} \quad (11.66)$$

où  $J, F_1, \dots, F_{m+q}$  sont des fonctions de  $\mathbb{R}^n$  dans  $\mathbb{R}$ , et  $X$  est un sous-ensemble non-vide de  $\mathbb{R}^n$ . Considérons le Lagrangien  $\mathcal{L} : X \times \Lambda \rightarrow \mathbb{R}$ , avec  $\Lambda = (\mathbb{R}_+)^m \times \mathbb{R}^q$  et

$$\mathcal{L}(x, \lambda) = J(x) + \lambda_1 F_1(x) + \dots + \lambda_{m+q} F_{m+q}(x) .$$

Soit  $J^*$  la valeur optimale de (11.66). Nous avons déjà noté dans la Remarque 10.3.10 qu'on a toujours l'inégalité de dualité faible

$$J^* = \inf_{x \in X} \sup_{\lambda \in \Lambda} \mathcal{L}(x, \lambda) \geq \sup_{\lambda \in \Lambda} \inf_{x \in X} \mathcal{L}(x, \lambda) = \sup_{\lambda \in \Lambda} \mathcal{D}(\lambda) , \quad (11.67)$$

où

$$\mathcal{D} : \Lambda \rightarrow \mathbb{R} \cup \{-\infty\}; \quad \mathcal{D}(\lambda) = \inf_{x \in X} \mathcal{L}(x, \lambda) \quad (11.68)$$

est la fonction duale. La méthode de la **relaxation Lagrangienne** consiste à employer le second membre de (11.67) comme borne inférieure pour la valeur  $J^*$  du problème original (11.66). N'importe quel multiplicateur de Lagrange  $\lambda \in \Lambda$  fournit une borne  $\mathcal{D}(\lambda) \leq J^*$  mais il est naturel de chercher la meilleure borne possible, ce qui revient à maximiser  $\mathcal{D}$ . Or  $\mathcal{D}$ , qui est un infimum de fonctions affines, est concave. Si, comme c'est le cas pour la plupart des problèmes combinatoires,  $X$  est fini,  $\mathcal{D}$  qui est un infimum fini de fonctions affines, est une fonction **non-différentiable** (sauf bien sûr dans des cas dégénérés). Maximiser  $\mathcal{D}$  relève donc de l'**optimisation non-différentiable**, qui traite de la minimisation de fonctions convexes non-différentiables (ou symétriquement, de la maximisation de fonctions concaves non-différentiables).

Nous allons brièvement présenter une méthode très simple, la **méthode de sous-gradient**, qui généralise les méthodes de gradient de la Sous-section 10.5.2. Il nous faut d'abord définir les notions de **sous-gradient** (pour les fonctions convexes) ou de **sur-gradient** (pour les fonctions concaves) qui sont des notions fondamentales en analyse convexe (nous nous contenterons ici de rappels succincts). Il sera utile de considérer des fonctions convexes à valeur dans  $\mathbb{R} \cup \{+\infty\}$ , et symétriquement des fonctions concaves à valeur dans  $\mathbb{R} \cup \{-\infty\}$ , car par exemple la fonction  $\mathcal{D}$  définie par (11.68) est concave et peut prendre la valeur  $-\infty$  (la valeur  $+\infty$  n'est pas possible, car nous avons exclu le cas  $X = \emptyset$ ). Lorsque  $J$  est convexe de  $\mathbb{R}^n$  dans  $\mathbb{R} \cup \{+\infty\}$ , nous notons  $\text{dom } J = \{x \in \mathbb{R}^n \mid J(x) < +\infty\}$  le **domaine** de  $J$ . Symétriquement, si  $J$  est une fonction concave de  $\mathbb{R}^n$  dans  $\mathbb{R} \cup \{-\infty\}$ , on définit  $\text{dom } J = \{x \in \mathbb{R}^n \mid J(x) > -\infty\}$ . Il est d'usage de traiter seulement le cas des fonctions convexes, étant entendu que tous les résultats ont des versions symétriques pour les fonctions concaves. Le lecteur prendra garde dans la suite à symétriser les résultats pour les appliquer à la maximisation de fonctions duales associées à des problèmes de minimisation.

**Définition 11.6.2** Si  $J$  est une fonction convexe de  $\mathbb{R}^n$  dans  $\mathbb{R} \cup \{+\infty\}$ , et si  $x \in \text{dom } J$ , on appelle **sous-différentiel** de  $J$  en  $x$  l'ensemble

$$\partial J(x) = \{p \in \mathbb{R}^n \mid J(y) - J(x) \geq p \cdot (y - x), \quad \forall y \in \mathbb{R}^n\}. \quad (11.69)$$

Les éléments de  $\partial J(x)$  sont appelés **sous-gradients**. Les notions de sur-différentiel et de sur-gradient d'une fonction concave de  $\mathbb{R}^n$  dans  $\mathbb{R} \cup \{-\infty\}$  sont définies symétriquement en renversant l'inégalité dans (11.69).

Il résulte aussitôt de la Définition (11.69) que le sous-différentiel  $\partial J(x)$  est un convexe fermé de  $\mathbb{R}^n$ . D'autre part, si  $0 \in \partial J(x)$ ,  $x$  est évidemment un point de minimum de  $J$ .

**Remarque 11.6.3** Les sous-différentiels peuvent être définis par (11.69) même lorsque  $J$  n'est pas convexe. Ils sont cependant surtout utiles lorsque  $J$  est convexe, car dans ce cas  $\partial J(x) \neq \emptyset$  en tout point  $x$  appartenant à l'intérieur de  $\text{dom } J$  (c'est là une conséquence de la forme géométrique du théorème de Hahn-Banach). La convexité garantit également l'équivalence de la Définition 11.6.2 des sous-gradients avec des définitions locales. On pourrait en effet définir les sous-gradients d'une fonction quelconque  $J$  de  $\mathbb{R}^n$  dans  $\mathbb{R} \cup \{+\infty\}$ , en disant que  $p$  est un sous-gradient de  $J$  en un point  $x \in \text{dom } J$  si  $J$  est localement "au dessus" de la fonction affine  $y \mapsto J(x) + p \cdot (y - x)$ , c'est-à-dire

$$J(y) \geq J(x) + p \cdot (y - x) + o(y - x), \quad \text{lorsque } y \rightarrow x.$$

Lorsque  $J$  est convexe, cette inégalité est équivalente à  $p \in \partial J(x)$ . •

Appliquons maintenant ces notions à la fonction duale  $\mathcal{D}$ . Il sera commode de prolonger  $\mathcal{D}$  à  $\mathbb{R}^{m+q}$  en posant  $\mathcal{D}(\lambda) = -\infty$ , si  $\lambda \notin \Lambda$ , ce qui définit bien une fonction concave de  $\mathbb{R}^{m+q}$  dans  $\mathbb{R} \cup \{-\infty\}$ . Lorsque  $X$  est fini,  $\text{dom } \mathcal{D} = \Lambda$ . Un sur-gradient de la fonction duale  $\mathcal{D}$  se calcule aisément, à l'aide de l'observation suivante.

**Proposition 11.6.4** Supposons  $X$  fini, soit  $\mathcal{D}$  la fonction duale définie par (11.68), et pour tout  $\lambda \in \Lambda$ , posons  $\Gamma(\lambda) = \arg \max_{x \in X} \mathcal{L}(x, \lambda) = \{x \in X \mid \mathcal{L}(x, \lambda) = \mathcal{D}(\lambda)\}$ . Alors, pour tout  $x \in \Gamma(\lambda)$ ,  $F(x) = (F_i(x))_{1 \leq i \leq m+q}$  est un sur-gradient de  $\mathcal{D}$  au point  $\lambda$ .

**Démonstration.** Pour tout  $x \in \Gamma(\lambda)$ , et pour tout  $\mu \in \Lambda$ , on a  $\mathcal{D}(\mu) - \mathcal{D}(\lambda) \geq \mathcal{L}(x, \mu) - \mathcal{L}(x, \lambda) = F(x) \cdot (\mu - \lambda)$ , ce qui montre que  $F(x)$  est un sur-gradient de  $\mathcal{D}$  au point  $\lambda$ . □

**Remarque 11.6.5** La Proposition 11.6.4 est une version faible du résultat suivant sur les sous-différentiels de maxima de fonctions convexes. Si  $J$  de  $\mathbb{R}^n$  dans  $\mathbb{R} \cup \{+\infty\}$  est de la forme  $J(x) = \sup_{i \in I} J_i(x)$ , où  $I$  est un ensemble fini, et si  $x \mapsto J_i(x)$  est convexe, pour tout  $i \in I$ , alors, pour tout  $x \in \text{dom } J$ ,

$$\partial J(x) = \text{co} \left( \bigcup_{\substack{i \in I \\ J_i(x) = J(x)}} \partial J_i(x) \right). \quad (11.70)$$

La preuve de cette identité fait l'objet de l'Exercice 11.6.5. •

Soit  $P_\Lambda$  la projection de  $\mathbb{R}^{m+q}$  dans  $\Lambda$ ,  $\lambda \mapsto (\lambda_1^+, \dots, \lambda_m^+, \lambda_{m+1}, \dots, \lambda_{m+q})$ . L'algorithme de sur-gradient (projeté) pour maximiser la fonction concave  $\mathcal{D}$  consiste à construire la suite

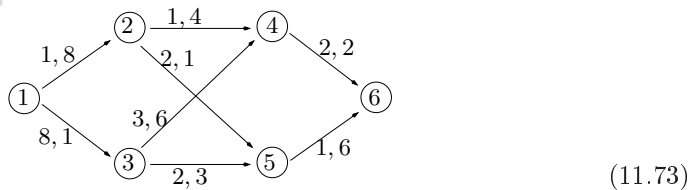
$$\lambda_{k+1} = P_\Lambda \left( \lambda_k + \frac{\rho_k}{\|p_k\|} p_k \right), \quad (11.71)$$

où  $\lambda_0 \in \Lambda$  est choisi arbitrairement, où  $p_k$  est un sur-gradient quelconque de  $\mathcal{D}$  au point  $\lambda_k$ , et où  $\rho_k$  est une suite de réels strictement positifs telle que

$$\rho_k \rightarrow 0, \quad \sum_i \rho_i = +\infty. \quad (11.72)$$

évidemment, la valeur  $\lambda_{k+1}$  n'est bien définie que si  $p_k \neq 0$ . Lorsque  $p_k = 0$ , l'algorithme s'arrête :  $\lambda_k$  est alors le maximum de  $J$  (par définition même des sur-gradients la nullité d'un sur-gradient en un point implique que la fonction est maximale en ce point).

Illustrons maintenant la relaxation Lagrangienne en traitant l'exemple du chemin de coût minimum avec contrainte de temps, déjà mentionné dans l'Exercice 11.4.6 comme application de la programmation dynamique. Nous considérons donc un graphe orienté  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ , l'arc  $(i, j)$  étant muni du coût  $c_{ij}$  et du temps  $\tau_{ij}$ . Pour fixer les idées, on cherchera le chemin de coût minimum et de temps total au plus 10, allant du nœud source  $s = 1$  au nœud puits  $p = 6$ , dans le graphe



On a représenté sur chaque arc les valuations  $c$  et  $\tau$ , dans cet ordre, par exemple, l'arc  $(1, 2)$  coûte  $c_{12} = 1$  et prend  $\tau_{12} = 8$  unités de temps.

Il nous faut tout d'abord formuler ce problème sous la forme (11.66). Pour cela, nous représenterons un chemin par le vecteur Booléen  $x \in \mathbb{R}^{\mathcal{A}}$  tel que  $x_{ij} = 1$  si  $(i, j)$  appartient au chemin, et  $x_{ij} = 0$  sinon. Le problème de plus court chemin d'une source  $s$  à un puits  $p$  en temps au plus  $T$  s'écrit alors

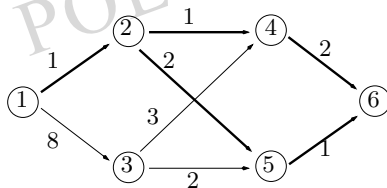
$$\begin{aligned} & \text{minimiser} && \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \\ & \text{sous les contraintes :} && x \in \{0, 1\}^{\mathcal{A}}, \\ & && x \text{ représente un chemin de } s \text{ à } p, \\ & && \sum_{(i,j) \in \mathcal{A}} \tau_{ij} x_{ij} \leq T. \end{aligned} \quad (11.74)$$

Il y a bien sûr plusieurs manières d'écrire un programme (11.66). Ainsi, c'est volontairement que nous avons laissé la contrainte " $x$  représente un chemin de  $s$  à  $p$ " sous

forme littérale. Nous aurions pu expliciter cette contrainte en écrivant la loi des nœuds de Kirchoff (11.31). Mais dualiser la loi des nœuds serait une mauvaise idée : le succès de la relaxation Lagrangienne dépend précisément de la capacité à identifier le plus “petit” ensemble de contraintes dont la relaxation conduit à un problème plus simple, résoluble de préférence par une méthode combinatoire directe. Ici, c’est la contrainte de temps qu’il faut relâcher, car si l’on oublie cette contrainte, on obtient un pur problème de chemin de coût minimum. En dualisant la contrainte de temps, la fonction  $\mathcal{D} : \mathbb{R}_+ \rightarrow \mathbb{R} \cup \{-\infty\}$  s’écrit en effet

$$\begin{aligned} \mathcal{D}(\lambda) &= \inf_{\substack{x \in \{0,1\}^{\mathcal{A}} \\ x \text{ représente un chemin de } s \text{ à } p}} \left( \sum_{(i,j) \in \mathcal{A}} c_{ij} x_{ij} \right) + \lambda \left( \sum_{(i,j) \in \mathcal{A}} \tau_{ij} x_{ij} - T \right) , \\ &= -\lambda T + \inf_{\substack{x \in \{0,1\}^{\mathcal{A}} \\ x \text{ représente un chemin de } s \text{ à } p}} \sum_{(i,j) \in \mathcal{A}} (c_{ij} + \lambda \tau_{ij}) x_{ij} . \end{aligned}$$

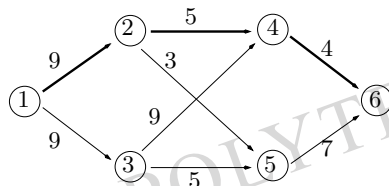
A  $\lambda$  fixé, le calcul de  $\mathcal{D}(\lambda)$  revient à résoudre un problème classique de chemin de coût minimum **sans contraintes de temps** dans le graphe de coûts  $c_{ij} + \lambda \tau_{ij}$ , ce qui peut se faire par programmation dynamique, et qui prend même un temps linéaire dans le cas d’un graphe sans circuits, comme c’est le cas pour l’exemple (11.73). Appliquons maintenant l’algorithme de sur-gradient (11.71) sur cet exemple. Pour  $\lambda_1 = 0$ , il nous faut trouver le chemin de coût minimum  $1 \rightarrow 6$  dans le graphe muni des coûts  $c_{ij} + 0\tau_{ij}$



L’ensemble  $\Gamma(0)$  des chemins optimaux est réduit aux chemins  $(1, 2, 4, 6)$ , et  $(1, 2, 5, 6)$  représentés en traits gras, qui ont pour coût 4. Ainsi,  $\mathcal{D}(0) = 4 - 0T = 4$ . D’après la Proposition 11.6.4, on a le sur-gradient

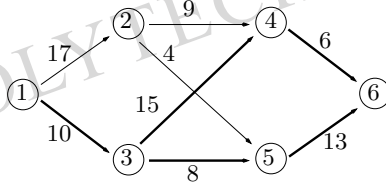
$$p_1 = \tau_{12} + \tau_{24} + \tau_{46} - T = 4 .$$

Avec un pas  $\rho_1 = 1$ , (11.71) donne  $\lambda_2 = 1$ , et le nouveau graphe muni des coûts  $c_{ij} + 1\tau_{ij}$



Comme le chemin optimal n’a pas changé, on a le même sur-gradient  $p_2 = p_1$ . En

prenant par exemple  $\rho_2 = 1$ , on a  $\lambda_3 = 2$ , ce qui donne le graphe



Il y a cette fois ci deux chemins optimaux :  $\Gamma(\lambda_3) = \{(1, 3, 4, 6), (1, 3, 5, 6)\}$ , de coût 21. En particulier, le sur-gradient correspondant au chemin  $(1, 3, 5, 6)$  est  $\tau_{13} + \tau_{35} + \tau_{56} - T = 0$ . L'algorithme de sur-gradient s'arrête donc :  $\max_{\lambda \in \mathbb{R}_+} \mathcal{D}(\lambda) = \mathcal{D}(2) = 31 - 20 = 11$ . Comme le chemin  $(1, 3, 5, 6)$  est de temps  $10 \leq T$  et de coût  $11 = \mathcal{D}(2)$ , nous venons de résoudre non seulement le problème relâché Lagrangien, qui consiste à maximiser  $\mathcal{D}(\lambda)$ , mais également le problème initial (11.74). Le fait que le relâché Lagrangien fournisse une solution du problème initial est cependant exceptionnel. En général, il y a un saut de dualité, mais les  $x$  réalisant le min dans la fonction duale (11.68), évaluée en un point de maximum  $\lambda$  de  $\mathcal{D}$ , peuvent souvent être modifiées sans trop augmenter le coût pour arriver à une solution (sous-optimale) du problème initial. On parle alors d'**heuristiques Lagrangiennes**.

Pour conclure cette illustration de la relaxation Lagrangienne, revisitons le problème du voyageur de commerce dans un graphe non-orienté complet  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ , muni d'une fonction temps  $t : \mathcal{E} \rightarrow \mathbb{R}_+$ ,  $\{i, j\} \mapsto t_{ij}$ . On peut écrire le problème du voyageur de commerce sous forme de programme linéaire en nombre entiers, équivalent à  $(VC)$ ,

$$\begin{aligned} \min \quad & \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} \text{ sous les contraintes} \\ & x \in \{0, 1\}^{\mathcal{E}}, \\ & \sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} = 2, \text{ pour tout } j \in \mathcal{V}, \\ & x \text{ représente un graphe connexe à } |\mathcal{V}| \text{ arêtes couvrant tous les sommets.} \end{aligned}$$

Distinguons un sommet particulier  $1 \in \mathcal{V}$ . Si, pour tout  $j \in \mathcal{V} \setminus \{1\}$ , on relâche les contraintes  $\sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} = 2$ , on obtient la fonction duale  $\mathcal{D} : \mathbb{R}^{\mathcal{V} \setminus \{1\}} \rightarrow \mathbb{R}$

$$\begin{aligned} \mathcal{D}(\lambda) = \quad & \min \sum_{\{i,j\} \in \mathcal{E}} t_{ij} x_{ij} + \sum_{j \in \mathcal{V} \setminus \{1\}} \lambda_j (\sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} - 2) \\ & \text{sous les contraintes} \\ & x \in \{0, 1\}^{\mathcal{E}} \\ & \sum_{k \in \mathcal{V}, \{k,1\} \in \mathcal{E}} x_{k1} = 2 \\ & x \text{ représente un graphe connexe à } |\mathcal{V}| \text{ arêtes couvrant tous les sommets.} \end{aligned}$$

Au terme  $-\sum_{j \in \mathcal{V} \setminus \{1\}} 2\lambda_j$  près, on reconnaît dans  $\mathcal{D}(\lambda)$  le coût minimum d'un 1-arbre pour le graphe  $\mathcal{D} = (\mathcal{V}, \mathcal{E})$ , dont chaque arête  $\{i, j\}$  est munie du coût  $t_{ij} + \lambda_i + \lambda_j$ , pour  $i, j \in \mathcal{V} \setminus \{1\}$ , et dont chaque arête  $\{i, 1\}$ , pour  $i \in \mathcal{V} \setminus \{1\}$ , a pour coût  $t_{i1} + \lambda_i$ . Comme nous l'avons déjà remarqué dans la Sous-section 11.6.1, l'algorithme de Kruskal de la Section 11.5 permet justement de calculer un 1-arbre de coût minimum en temps  $O(|\mathcal{E}| \log |\mathcal{E}|)$ , ce qui nous permet de mettre en œuvre efficacement l'algorithme de



sur-gradient (11.71) pour calculer la borne suivante :  $\sup_{\lambda \in \mathbb{R}^{\mathcal{V} \setminus \{1\}}} \mathcal{D}(\lambda)$ . Cette borne remarquable pour le problème du voyageur de commerce, est due à Held et Karp. La borne du 1-arbre vue dans la Sous-section 11.6.1 coïncide avec la valeur  $\mathcal{D}(0)$  obtenue en ne faisant pas payer la transgression des contraintes  $\sum_{k \in \mathcal{V}, \{k,j\} \in \mathcal{E}} x_{kj} = 2$ , pour tout  $j \in \mathcal{V} \setminus \{1\}$ .

**Exercice 11.6.2** Calculer la fonction duale  $\mathcal{D}(\lambda)$  pour le problème (11.73), et retrouver ainsi la valeur de  $\max_{\lambda \in \mathbb{R}_+} \mathcal{D}(\lambda)$ .

**Exercice 11.6.3** Proposer une relaxation Lagrangienne fournissant une borne supérieure pour le problème du sac-à-dos (11.52), et l'appliquer à l'Exemple (11.54).

**Exercice 11.6.4** La version "orientée" du problème du voyageur de commerce consiste à considérer un graphe orienté  $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ , muni d'une fonction temps  $t : \mathcal{A} \rightarrow \mathbb{R}_+$ , et à chercher un circuit orienté passant une et une seule fois par chaque nœud. On demande de proposer une relaxation Lagrangienne, en faisant en sorte que pour chaque vecteur  $\lambda$  de multiplicateurs de Lagrange, le calcul de la fonction duale  $\mathcal{D}(\lambda)$  revienne à résoudre un problème d'affectation. Expliquer pourquoi cette relaxation est moins intéressante que la borne de Held et Karp, dans le cas particulier du problème du voyageur de commerce non-orienté.

**Exercice 11.6.5** Nous allons prouver la formule (11.70), dans le cas où tous les  $J_i$  sont des fonctions convexes de  $\mathbb{R}^n$  dans  $\mathbb{R}$ . (L'exclusion de la valeur  $+\infty$  n'est qu'une commodité, qui nous permettra d'appliquer directement les résultats du Chapitre 10, prouvés dans le cas de fonctions convexes à valeurs finies). Notons  $C$  le second membre de (11.70). 1. Montrer que  $C \subset \partial J(x)$ . 2. On suppose que  $0 \in \partial J(x)$ . En considérant le programme convexe

$$\min \quad t \quad \text{sous les contraintes} \quad y \in \mathbb{R}^n, \quad t \in \mathbb{R}, \quad J_i(y) \leq t, \quad \forall i \in I,$$

montrer que  $0 \in C$ . 3. Conclure que de manière générale,  $C = \partial J(x)$ .

**Remarque 11.6.6** Bien que ce ne soit pas l'objet de ce cours, disons quelques mots des questions de **complexité**. Un algorithme est dit **polynomial** (resp. **linéaire**) si son temps d'exécution (sur un ordinateur usuel) est borné par une fonction polynôme (resp. affine) de la taille des données d'entrée. On dit qu'un problème est polynomial s'il peut être résolu par un algorithme polynomial. La classe P des problèmes polynomiaux formalise la notion de problème que l'on sait "bien" résoudre en pratique. Par exemple, l'Exercice 11.4.3 montre que le problème consistant à trouver un chemin de coût minimal d'un nœud à un autre, dans un graphe, est polynomial. Un autre exemple de problème polynomial est la programmation linéaire (l'algorithme du simplexe peut prendre un temps exponentiel dans certaines situations dégénérées, mais les algorithmes de points intérieurs esquissés dans la Sous-section 11.2.3 tournent en temps polynomial). Une classe de problèmes combinatoires qui semblent plus difficiles a retenu beaucoup d'attention. Il s'agit de NP. Intuitivement, un problème de décision est dans NP si l'on peut vérifier qu'une solution est correcte en temps

polynomial. Par exemple, le problème appelé Circuit Hamiltonien qui consiste à décider s'il existe un circuit visitant chaque sommet d'un graphe une et une seule fois, est un problème dans NP, car si l'on vous donne un circuit quelconque, vous pouvez vérifier s'il est Hamiltonien en temps polynomial (il suffit de compter le nombre de visites en chaque sommet). Un autre exemple de problème dans NP est le problème Sat, qui consiste à décider si un système d'équations Booléennes admet une solution. Un problème est dit **NP-difficile** s'il est au moins aussi difficile que tous les problèmes de NP, ce qui signifie que si l'on savait résoudre ce problème en temps polynomial, on saurait résoudre tous les problèmes de NP en temps polynomial. Les problèmes **NP-complets** sont des problèmes de décision à la fois NP-difficiles et dans NP : l'existence de tels problèmes est un théorème dû à Cook et Levin. Par exemple, Sat et Circuit Hamiltonien sont des problèmes NP-complets. On dit aussi qu'un problème d'optimisation est NP-difficile quand sa version décision est NP-complète. Par exemple, trouver un circuit Hamiltonien de coût minimum, c'est-à-dire un tour de voyageur de commerce de coût minimum, est un problème NP-difficile ; le problème du sac-à-dos de l'Exercice 11.4.5, ou le problème du plus court chemin avec contraintes de l'Exercice 11.4.6, sont également NP-difficiles. On peut penser que les problèmes NP-difficiles sont véritablement plus difficiles que les problèmes polynômiaux : montrer qu'il en est ainsi, c'est-à-dire montrer que  $P \neq NP$ , est un problème ouvert célèbre. Le lecteur intéressé par ces questions pourra consulter *Computers and intractability*, M. R. Garey et D. S. Johnson, Freeman, 1979, ainsi que *Computational complexity*, C. H. Papadimitriou, Addison Wesley, 1995. •

**Remarque 11.6.7** Pour conclure ce chapitre d'initiation à la RO, mentionnons l'existence de deux approches importantes qui sortent du cadre de ce cours. Pour des problèmes qui ne jouissent pas de bonnes propriétés de structure, ou bien qui sont de trop grande taille pour pouvoir appliquer des méthodes exactes, on recourt souvent à des **méthodes de voisinage** : algorithmes stochastiques (dont le plus classique est le **recuit simulé**, voir par exemple [3]), ou bien **recherche** avec liste **tabou** [26]. Une méthode de voisinage est une heuristique qui consiste à explorer l'espace des solutions par modifications successives d'une solution admissible. Si l'on n'accepte que des modifications qui améliorent le critère, on obtient simplement une heuristique gloutonne, qui peut converger vers un minimum local. Des algorithmes tels que le recuit simulé ou la recherche tabou spécifient comment gérer des modifications qui dégradent provisoirement le critère, et parvenir ainsi plus souvent à un optimum global. Une seconde technique générale, très différente (qui permet souvent de résoudre de manière optimale des problèmes de taille moyenne) est la programmation par contrainte, ou PPC, qui explore intelligemment l'espace des solutions à l'aide de logiciels qui réduisent la combinatoire par des déductions logiques. •

## RAPPELS SUR LES ESPACES DE HILBERT

Nous donnons brièvement quelques propriétés des espaces de Hilbert (pour plus de détails, nous renvoyons aux cours de mathématiques [7], [27]). Pour simplifier la présentation, on ne considère que le cas d'espaces de Hilbert sur  $\mathbb{R}$ .

**Définition 12.1.8** *Un espace de Hilbert réel est un espace vectoriel sur  $\mathbb{R}$ , muni d'un produit scalaire, noté  $\langle x, y \rangle$ , qui est complet pour la norme associée à ce produit scalaire, notée  $\|x\| = \sqrt{\langle x, x \rangle}$ . (On rappelle qu'un espace vectoriel normé est complet si toute suite de Cauchy est une suite convergente dont la limite appartient à cet espace.)*

Dans tout ce qui suit nous noterons  $V$  un espace de Hilbert réel, et  $\langle x, y \rangle$  son produit scalaire associé.

**Définition 12.1.9** *Un ensemble  $K \subset V$  est dit convexe si, pour tout  $x, y \in K$  et tout réel  $\theta \in [0, 1]$ , l'élément  $(\theta x + (1 - \theta)y)$  appartient à  $K$ .*

Un résultat essentiel est le théorème de projection sur un ensemble convexe (voir le théorème 4.2.1 de [7]).

**Théorème 12.1.10 (de projection sur un convexe)** *Soit  $V$  un espace de Hilbert. Soit  $K \subset V$  un convexe fermé non vide. Pour tout  $x \in V$ , il existe un unique  $x_K \in K$  tel que*

$$\|x - x_K\| = \min_{y \in K} \|x - y\|.$$

*De façon équivalente,  $x_K$  est caractérisé par la propriété*

$$x_K \in K, \langle x_K - x, x_K - y \rangle \leq 0 \quad \forall y \in K. \quad (12.1)$$

*On appelle  $x_K$  la projection orthogonale sur  $K$  de  $x$ .*

**Remarque 12.1.11** Le Théorème 12.1.10 permet de définir une application  $P_K$ , appelée opérateur de projection sur l'ensemble convexe  $K$ , en posant  $P_K x = x_K$ . On vérifie sans peine que  $P_K$  est continue et faiblement contractanté, c'est-à-dire que

$$\|P_K x - P_K y\| \leq \|x - y\| \quad \forall x, y \in V. \quad (12.2)$$

•

**Remarque 12.1.12** Un cas particulier de convexe fermé  $K$  est un sous-espace vectoriel fermé  $W$ . Dans ce cas, la caractérisation (12.1) de  $x_W$  devient

$$x_W \in W, \langle x_W - x, z \rangle = 0 \quad \forall z \in W.$$

En effet, dans (12.1) il suffit de prendre  $y = x_K \pm z$  avec  $z$  quelconque dans  $W$ . •

**Démonstration.** Soit  $y^n$  une suite minimisante, c'est-à-dire que  $y^n \in K$  vérifie

$$d_n = \|x - y^n\| \rightarrow d = \inf_{y \in K} \|x - y\| \text{ quand } n \rightarrow +\infty.$$

Montrons que  $y^n$  est une suite de Cauchy. En utilisant la symétrie du produit scalaire, il vient

$$\|x - \frac{1}{2}(y^n + y^p)\|^2 + \|\frac{1}{2}(y^n - y^p)\|^2 = \frac{1}{2}(d_n^2 + d_p^2).$$

Or, par convexité de  $K$ ,  $(y^n + y^p)/2 \in K$ , et  $\|x - \frac{1}{2}(y^n + y^p)\|^2 \geq d^2$ . Par conséquent

$$\|y^n - y^p\|^2 \leq 2(d_n^2 + d_p^2) - 4d^2,$$

ce qui montre que  $y^n$  est une suite de Cauchy. Comme  $V$  est un espace de Hilbert, il est complet, donc la suite  $y^n$  est convergente vers une limite  $x_K$ . Par ailleurs, comme  $K$  est fermé, cette limite  $x_K$  appartient à  $K$ . Par conséquent, on a  $d = \|x - x_K\|$ . Comme toute la suite minimisante est convergente, la limite est forcément unique, et  $x_K$  est le seul point de minimum de  $\min_{y \in K} \|x - y\|$ .

Soit  $x_K \in K$  ce point de minimum. Pour tout  $y \in K$  et  $\theta \in [0, 1]$ , par convexité de  $K$ ,  $x_K + \theta(y - x_K)$  appartient à  $K$  et on a

$$\|x - x_K\|^2 \leq \|x - (x_K + \theta(y - x_K))\|^2.$$

En développant le terme de droite, il vient

$$\|x - x_K\|^2 \leq \|x - x_K\|^2 + \theta^2 \|y - x_K\|^2 - 2\theta \langle x - x_K, y - x_K \rangle,$$

ce qui donne pour  $\theta > 0$

$$0 \geq -2\langle x - x_K, y - x_K \rangle + \theta \|z\|^2.$$

En faisant tendre  $\theta$  vers 0, on obtient la caractérisation (12.1). Réciproquement, soit  $x_K$  qui vérifie cette caractérisation. Pour tout  $y \in K$  on a

$$\|x - y\|^2 = \|x - x_K\|^2 + \|x_K - y\|^2 + 2\langle x - x_K, x_K - y \rangle \geq \|x - x_K\|^2,$$

ce qui prouve que  $x_K$  est bien la projection orthogonale de  $x$  sur  $K$ . □

**Définition 12.1.13** Soit  $V$  un espace de Hilbert pour le produit scalaire  $\langle, \rangle$ . On appelle base hilbertienne (dénombrable) de  $V$  une famille dénombrable  $(e_n)_{n \geq 1}$  d'éléments de  $V$  qui est orthonormale pour le produit scalaire et telle que l'espace vectoriel engendré par cette famille est dense dans  $V$ .

**Proposition 12.1.14** *Soit  $V$  un espace de Hilbert pour le produit scalaire  $\langle \cdot, \cdot \rangle$ . Soit  $(e_n)_{n \geq 1}$  une base hilbertienne de  $V$ . Pour tout élément  $x$  de  $V$ , il existe une unique suite  $(x_n)_{n \geq 1}$  de réels telle que la somme partielle  $\sum_{n=1}^p x_n e_n$  converge vers  $x$  quand  $p$  tend vers l'infini, et cette suite est définie par  $x_n = \langle x, e_n \rangle$ . De plus, on a*

$$\|x\|^2 = \langle x, x \rangle = \sum_{n \geq 1} |\langle x, e_n \rangle|^2. \quad (12.3)$$

On écrit alors

$$x = \sum_{n \geq 1} \langle x, e_n \rangle e_n.$$

**Démonstration.** S'il existe une suite  $(x_n)_{n \geq 1}$  de réels telle que  $\lim_{p \rightarrow +\infty} \sum_{n=1}^p x_n e_n = x$ , alors par projection sur  $e_n$  (et comme cette suite est par définition indépendante de  $p$ ) on a  $x_n = \langle x, e_n \rangle$ , ce qui prouve l'unicité de la suite  $(x_n)_{n \geq 1}$ . Montrons maintenant son existence. Par définition d'une base hilbertienne, pour tout  $x \in V$  et pour tout  $\epsilon > 0$ , il existe  $y$ , combinaison linéaire finie des  $(e_n)_{n \geq 1}$ , tel que  $\|x - y\| < \epsilon$ . Grâce au Théorème 12.1.10 on peut définir une application linéaire  $S_p$  qui, à tout point  $z \in V$ , fait correspondre  $S_p z = z_W$ , où  $z_W$  est la projection orthogonale sur le sous-espace vectoriel  $W$  engendré par les  $p$  premiers vecteurs  $(e_n)_{1 \leq n \leq p}$ . En vertu de (12.1),  $(z - S_p z)$  est orthogonal à tout élément de  $W$ , donc en particulier à  $S_p z$ . On en déduit que

$$\|z\|^2 = \|z - S_p z\|^2 + \|S_p z\|^2, \quad (12.4)$$

ce qui implique

$$\|S_p z\| \leq \|z\| \forall z \in V.$$

Comme  $S_p z$  est engendré par les  $(e_n)_{1 \leq n \leq p}$ , et que  $(z - S_p z)$  est orthogonal à chacun des  $(e_n)_{1 \leq n \leq p}$ , on vérifie facilement que

$$S_p z = \sum_{n=1}^p \langle z, e_n \rangle e_n.$$

Pour  $p$  suffisamment grand, on a  $S_p y = y$  car  $y$  est une combinaison linéaire finie des  $(e_n)_{n \geq 1}$ . Par conséquent

$$\|S_p x - x\| \leq \|S_p(x - y)\| + \|y - x\| \leq 2\|x - y\| \leq 2\epsilon.$$

On en déduit la convergence de  $S_p x$  vers  $x$ . De cette convergence et de l'équation (12.4) on tire

$$\lim_{p \rightarrow +\infty} \|S_p x\|^2 = \|x\|^2,$$

qui n'est rien d'autre que la formule de sommation (12.3), dite de Parseval.  $\square$

L'existence d'une base hilbertienne dénombrable n'est pas garantie pour tous les espaces de Hilbert. La proposition suivante donne une condition nécessaire et suffisante d'existence d'une base hilbertienne dénombrable.

**Proposition 12.1.15** *Soit  $V$  un espace de Hilbert séparable (i.e. il existe une famille dénombrable dense dans  $V$ ). Alors il existe une base hilbertienne dénombrable de  $V$ .*

**Démonstration.** Soit  $(v_n)_{n \geq 1}$  la famille dont le sous-espace vectoriel engendré est dense dans  $V$  (quitte à renuméroter les  $v_n$  et à en supprimer certains, on peut toujours supposer qu'ils sont libres). Par application du procédé de Gram-Schmidt à cette famille, on obtient une famille orthonormale  $(e_n)_{n \geq 1}$ . Comme  $[v_1, \dots, v_n] = [e_1, \dots, e_n]$ , on en déduit que le sous-espace vectoriel engendré par  $(e_n)_{n \geq 1}$  coïncide avec celui engendré par  $(v_n)_{n \geq 1}$  qui est dense dans  $V$ . Donc,  $(e_n)_{n \geq 1}$  est une base hilbertienne.  $\square$

**Définition 12.1.16** Soit  $V$  et  $W$  deux espaces de Hilbert réels. Une application linéaire  $A$  de  $V$  dans  $W$  est dite continue s'il existe une constante  $C$  telle que

$$\|Ax\|_W \leq C\|x\|_V \quad \forall x \in V.$$

La plus petite constante  $C$  qui vérifie cette inégalité est la norme de l'application linéaire  $A$ , autrement dit

$$\|A\| = \sup_{x \in V, x \neq 0} \frac{\|Ax\|_W}{\|x\|_V}.$$

Souvent on utilisera la dénomination équivalente d'opérateur au lieu d'application entre espaces de Hilbert (on parlera ainsi d'opérateur linéaire continu plutôt que d'application linéaire continue). Si  $V$  est de dimension finie, alors toutes les applications linéaires de  $V$  dans  $W$  sont continues, mais ce n'est plus vrai si  $V$  est de dimension infinie.

**Définition 12.1.17** Soit  $V$  un espace de Hilbert réel. Son dual  $V'$  est l'ensemble des formes linéaires continues sur  $V$ , c'est-à-dire l'ensemble des applications linéaires continues de  $V$  dans  $\mathbb{R}$ . Par définition, la norme d'un élément  $L \in V'$  est

$$\|L\|_{V'} = \sup_{x \in V, x \neq 0} \frac{|L(x)|}{\|x\|}.$$

Dans un espace de Hilbert la dualité a une interprétation très simple grâce au théorème de Riesz (voir le théorème 4.3.1 de [7]) qui permet d'identifier un espace de Hilbert à son dual par isomorphisme.

**Théorème 12.1.18 (de représentation de Riesz)** Soit  $V$  un espace de Hilbert réel, et soit  $V'$  son dual. Pour toute forme linéaire continue  $L \in V'$  il existe un unique  $y \in V$  tel que

$$L(x) = \langle y, x \rangle \quad \forall x \in V.$$

De plus, on a  $\|L\|_{V'} = \|y\|$ .

**Démonstration.** Soit  $M = \text{Ker} L$ . Il s'agit d'un sous-espace fermé de  $V$  car  $L$  est continue. Si  $M = V$ , alors  $L$  est identiquement nulle et seul  $y = 0$  convient. Si  $M \neq V$ , alors il existe

$z \in V \setminus M$ . Soit alors  $z_M \in M$  sa projection sur  $M$ . Comme  $z$  n'appartient pas à  $M$ ,  $z - z_M$  est non nul et, par le Théorème 12.1.10, est orthogonal à tout élément de  $M$ . Soit finalement

$$z_0 = \frac{z - z_M}{\|z - z_M\|}.$$

Tout vecteur  $x \in V$  peut s'écrire

$$x = w + \lambda z_0 \text{ avec } \lambda = \frac{L(x)}{L(z_0)}.$$

On vérifie aisément que  $L(w) = 0$ , donc  $w \in M$ . Ceci prouve que  $V = \text{Vect}(z_0) \oplus M$ . Par définition de  $z_M$  et de  $z_0$ , on a  $\langle w, z_0 \rangle = 0$ , ce qui implique

$$L(x) = \langle x, z_0 \rangle L(z_0),$$

d'où le résultat désiré avec  $y = L(z_0)z_0$  (l'unicité est évidente). D'autre part, on a

$$\|y\| = |L(z_0)|,$$

et

$$\|L\|_{V'} = \sup_{x \in V, x \neq 0} \frac{|L(x)|}{\|x\|} = L(z_0) \sup_{x \in V, x \neq 0} \frac{\langle x, z_0 \rangle}{\|x\|}.$$

Le maximum dans le dernier terme de cette égalité est atteint par  $x = z_0$ , ce qui implique que  $\|L\|_{V'} = \|y\|$ .  $\square$

Un résultat essentiel pour pouvoir démontrer le Lemme de Farkas 10.2.17 (utile en optimisation) est la propriété géométrique suivante qui est tout à fait conforme à l'intuition.

**Théorème 12.1.19 (Séparation d'un point et d'un convexe)** *Soit  $K$  une partie convexe non vide et fermée d'un espace de Hilbert  $V$ , et  $x_0 \notin K$ . Alors il existe un hyperplan fermé de  $V$  qui sépare strictement  $x_0$  et  $K$ , c'est-à-dire qu'il existe une forme linéaire  $L \in V'$  et  $\alpha \in \mathbb{R}$  tels que*

$$L(x_0) < \alpha < L(x) \quad \forall x \in K. \quad (12.5)$$

**Démonstration.** Notons  $x_K$  la projection de  $x_0$  sur  $K$ . Puisque  $x_0 \notin K$ , on a  $x_K - x_0 \neq 0$ . Soit  $L$  la forme linéaire définie pour tout  $y \in V$  par  $L(y) = \langle x_K - x_0, y \rangle$ , et soit  $\alpha = (L(x_K) + L(x_0))/2$ . D'après (12.1), on a  $L(x) \geq L(x_K) > \alpha > L(x_0)$  pour tout  $x \in K$ , ce qui achève la démonstration.  $\square$

Nous aurons enfin besoin pour démontrer le Théorème de Minkowski 11.3.1 d'une variante du théorème de séparation, faisant intervenir la notion importante d'hyperplan d'appui. Si  $K$  est un convexe d'un espace de Hilbert  $V$ , on appelle **hyperplan d'appui** de  $K$  en un point  $x$  un hyperplan affine  $H = \{y \in V \mid L(y) = \alpha\}$ , avec  $L \in V'$ ,  $L \neq 0$ , et  $\alpha \in \mathbb{R}$ , tel que  $\alpha = L(x) \leq L(y)$ , pour tout  $y \in C$ .

**Corollaire 12.1.20 (Hyperplan d'appui)** *Il existe un hyperplan d'appui en tout point frontière d'un convexe fermé  $K$  d'un espace de Hilbert de dimension finie.*

**Démonstration.** Soit  $x$  un point frontière de  $K$  : il existe alors une suite  $x_n \in V \setminus K$ , avec  $x_n \rightarrow x$ . Le Théorème de séparation 12.1.19 fournit pour tout  $n$  une forme linéaire  $L_n$  non nulle telle que  $L_n(x_n) \leq L_n(y)$  pour tout  $y \in K$ . On peut choisir  $L_n$  de norme 1. Comme  $V$  est de dimension finie, la sphère unité de  $V'$  est compacte, et quitte à remplacer  $L_n$  par une sous-suite, on peut supposer que  $L_n$  converge vers une forme linéaire  $L$ , qui est non nulle, car de norme 1. Il suffit maintenant de passer à la limite dans  $L_n(x_n) \leq L_n(y)$ , ce que l'on justifie en écrivant  $L_n(x_n) = L_n(x_n - x) + L_n(x)$  et en notant que  $|L_n(x_n - x)| \leq \|L_n\| \|x_n - x\| = \|x_n - x\|$ , pour obtenir  $L(x) \leq L(y)$  quel que soit  $y \in K$ . Ainsi,  $H = \{y \in V \mid L(y) = L(x)\}$  est un hyperplan d'appui de  $K$  en  $x$ .  $\square$

**Remarque 12.1.21** La preuve du Corollaire 12.1.20 ne s'étend pas en dimension infinie : dans ce cas, la suite  $L_n$  a bien une valeur d'adhérence  $L$  pour la topologie faible, mais rien ne dit que  $L \neq 0$ . Comme contre exemple, considérons l'ensemble  $K$  des suites de  $\ell_2$  à termes positifs ou nuls, qui est un convexe fermé de  $\ell_2$  d'intérieur vide. Tout point de  $K$  est donc point frontière, mais si  $x$  est une suite de  $\ell_2$  à termes strictement positifs, il n'existe pas d'hyperplan d'appui en  $x$ .  $\bullet$



## ANALYSE NUMÉRIQUE MATRICIELLE

Cette annexe est consacrée à l'analyse numérique matricielle, et plus précisément aux algorithmes utilisés pour résoudre des systèmes linéaires (notamment ceux issus de la méthode des éléments finis), et pour calculer les valeurs et vecteurs propres d'une matrice auto-adjointe (intervenant dans le calcul des modes propres d'un modèle mécanique). Pour plus de détails nous renvoyons aux ouvrages [2] et [10].

### 13.1 Résolution des systèmes linéaires

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

$$Ax = b, \quad (13.1)$$

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. Bien sûr, on dispose des célèbres formules de Cramer qui, pour une matrice inversible  $A$  de colonnes  $(a^1, \dots, a^n)$ , donnent la solution de (13.1) par ses composantes

$$x_i = \frac{\det(a^1, \dots, a^{i-1}, b, a^{i+1}, \dots, a^n)}{\det A}.$$

On pourrait croire que cette formule explicite suffit pour nos besoins. Mais il n'en est rien, car les formules de Cramer sont **totalelement inadaptées** pour calculer efficacement la solution d'un système linéaire. En effet, leur coût en temps d'exécution sur un ordinateur est prohibitif : il faut calculer  $n + 1$  déterminants et si on utilise la méthode du développement par ligne (ou colonne) chaque déterminant demande plus de  $n!$  multiplications. Au total, la méthode de Cramer nécessite donc plus de  $(n + 1)!$  multiplications, ce qui est inenvisageable : par exemple pour  $n = 50$ , si les calculs sont effectués sur un ordinateur fonctionnant à 1 Gigafllops (un milliard d'opérations par seconde), le temps de calcul est de l'ordre de  $4.8 \cdot 10^{49}$  années ! Même si on utilise une meilleure méthode pour calculer les déterminants la méthode de Cramer n'est pas compétitive par rapport aux algorithmes que nous allons voir (dont le nombre d'opérations sera typiquement de l'ordre de  $n^3$ ).

Nous verrons 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.

### 13.1.1 Rappels sur les normes matricielles

Nous commençons par rappeler la notion de **norme subordonnée** pour les matrices. On note  $\mathcal{M}_n(\mathbb{R})$  (respectivement  $\mathcal{M}_n(\mathbb{C})$ ) l'ensemble des matrices carrées réelles (respectivement complexes) d'ordre  $n$ . Même si l'on considère des matrices réelles, il est nécessaire, pour des raisons techniques qui seront exposées à la Remarque 13.1.4, de les traiter comme des matrices complexes.

**Définition 13.1.1** 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 sur  $\mathcal{M}_n(\mathbb{C})$  ou sur  $\mathcal{M}_n(\mathbb{R})$ .

**Lemme 13.1.2** 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  $\|A\| = \frac{\|Ax_A\|}{\|x_A\|}$ .

3. La matrice identité vérifie  $\|\text{Id}\| = 1$ .

4. Soient  $A$  et  $B$  deux matrices. On a  $\|AB\| \leq \|A\| \|B\|$ .

**Démonstration.** Le premier point est évident. Le second se démontre en remarquant que la fonction continue  $\|Ax\|$  atteint son maximum sur le compact  $\{x \in \mathbb{C}^n, \|x\| = 1\}$ . Le troisième est évident, tandis que le quatrième est une conséquence de l'inégalité  $\|ABx\| \leq \|A\| \|Bx\|$ .  $\square$

**Remarque 13.1.3** Il existe des normes matricielles qui ne sont subordonnées à aucune norme vectorielle. Le meilleur exemple en est la norme euclidienne définie par  $\|A\| = \sqrt{\sum_{i,j=1}^n |a_{ij}|^2}$ . En effet, on a  $\|\text{Id}\| = \sqrt{n}$ , ce qui n'est pas possible pour une norme subordonnée.  $\bullet$

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  $A$  et  $A^t$  la matrice transposée de  $A$ .)

**Exercice 13.1.1** Montrer que

1.  $\|A\|_2 = \|A^*\|_2 = \text{maximum des valeurs singulières de } A$ ,
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 13.1.4** 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 13.1.1, 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 13.1.7 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 13.1.1 de la norme subordonnée. •

**Définition 13.1.5** 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  $A$ .

Le rayon spectral  $\rho(A)$  n'est pas une norme sur  $\mathcal{M}_n(\mathbb{C})$ . En effet, on peut avoir  $\rho(A) = 0$  avec  $A \neq 0$  (prendre, par exemple, une matrice triangulaire avec des zéros sur la diagonale). Cependant, le lemme ci-dessous montre que c'est une norme sur l'ensemble des matrices normales.

**Lemme 13.1.6** 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  $U^*U = \text{Id}$ , 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  $\|A\|_2 = \|\text{diag}(\lambda_i)\|_2 = \rho(A)$ . □

On compare maintenant la norme d'une matrice  $A$  avec son rayon spectral  $\rho(A)$ .

**Proposition 13.1.7** 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  $\epsilon$ ) telle que

$$\|A\| \leq \rho(A) + \epsilon. \quad (13.2)$$

**Démonstration.** Soit  $\lambda \in \mathbb{C}$  une valeur propre de  $A$  telle que  $\rho(A) = |\lambda|$ , et  $x^0 \neq 0$  un vecteur propre associé ( $Ax^0 = \lambda x^0$ ). On a

$$\|\lambda x^0\| = \rho(A)\|x^0\| = \|Ax^0\| \leq \|A\|\|x^0\|,$$

d'où l'on déduit  $\rho(A) \leq \|A\|$ . Comme le vecteur propre  $x^0$  peut être complexe, il est essentiel d'utiliser une norme vectorielle sur  $\mathbb{C}^n$  même pour une matrice réelle (cf. Remarque 13.1.4). Réciproquement, il existe une matrice  $U$  inversible telle que  $T = U^{-1}AU$  soit triangulaire supérieure. Pour tout  $\delta > 0$  on définit une matrice diagonale  $D_\delta = \text{diag}(1, \delta, \delta^2, \dots, \delta^{n-1})$  de telle sorte que la matrice  $T_\delta$  définie par

$$T_\delta = (UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}TD_\delta$$

vérifie

$$T_\delta = \begin{pmatrix} t_{11} & \delta t_{12} & \cdots & \delta^{n-1}t_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \delta t_{n-1n} \\ 0 & \cdots & 0 & t_{nn} \end{pmatrix} \quad \text{avec} \quad T = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & t_{nn} \end{pmatrix}.$$

Étant donné  $\epsilon > 0$ , on peut choisir  $\delta$  suffisamment petit pour que les éléments extra-diagonaux de  $T_\delta$  soient très petits aussi, par exemple pour que, pour tout  $1 \leq i \leq n-1$ ,

$$\sum_{j=i+1}^n \delta^{j-i}|t_{ij}| \leq \epsilon.$$

Alors l'application  $B \rightarrow \|(UD_\delta)^{-1}B(UD_\delta)\|_\infty$  est une norme subordonnée (qui dépend de  $A$  et  $\epsilon$ ) qui vérifie bien (13.2).  $\square$

**Lemme 13.1.8** Soit  $A$  une matrice de  $\mathcal{M}_n(\mathbb{C})$ . Les quatre conditions suivantes sont équivalentes

1.  $\lim_{i \rightarrow +\infty} A^i = 0$ ,
2.  $\lim_{i \rightarrow +\infty} A^i x = 0$  pour tout vecteur  $x \in \mathbb{C}^n$ ,
3.  $\rho(A) < 1$ ,
4. il existe au moins une norme matricielle subordonnée telle que  $\|A\| < 1$ .

**Démonstration.** Montrons tout d'abord que (1) implique (2). L'inégalité

$$\|A^i x\| \leq \|A^i\| \|x\|$$

montre que  $\lim_{i \rightarrow +\infty} A^i x = 0$ . Ensuite, (2) implique (3) car, si  $\rho(A) \geq 1$ , alors il existe  $\lambda$  et  $x \neq 0$  tels que  $Ax = \lambda x$  et  $|\lambda| = \rho(A)$ , et, par conséquent, la suite  $A^i x = \lambda^i x$  ne peut pas converger vers 0. Comme “(3) implique (4)” est une conséquence immédiate de la Proposition 13.1.7, il ne reste plus qu'à montrer que (4) implique (1). Pour cela, on considère la norme matricielle subordonnée telle que  $\|A\| < 1$ , et on a

$$\|A^i\| \leq \|A\|^i \rightarrow 0 \text{ lorsque } i \rightarrow +\infty,$$

ce qui montre que  $A^i$  tend vers 0. □

### 13.1.2 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 13.1.9** Soit une norme matricielle subordonnée que l'on note  $\|A\|$  (voir la Définition 13.1.1). On appelle conditionnement d'une matrice  $A \in \mathcal{M}_n(\mathbb{C})$ , relatif à cette norme, la valeur définie par

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\|.$$

Cette notion de conditionnement va permettre de mesurer l'amplification des erreurs des données (second membre ou matrice) au résultat.

**Proposition 13.1.10** 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

$$Ax = b, \text{ et } A(x + \delta x) = b + \delta b.$$

Alors on a

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \frac{\|\delta b\|}{\|b\|}. \quad (13.3)$$

2. Soit  $x$  et  $x + \delta x$  les solutions respectives des systèmes

$$Ax = b, \text{ et } (A + \delta A)(x + \delta x) = b.$$

Alors on a

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|}. \quad (13.4)$$

De plus, ces inégalités sont optimales.

**Remarque 13.1.11** 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 13.1.10, 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  $\|\delta x\| \leq \|A^{-1}\| \cdot \|\delta b\|$ . Or, on a aussi  $\|b\| \leq \|A\| \|x\|$ , ce qui donne (13.3). Cette inégalité est optimale au sens suivant : pour toute matrice  $A$ , il existe  $\delta b$  et  $x$  (qui dépendent de  $A$ ) tels que (13.3) est en fait une égalité. En effet, d'après une propriété des normes matricielles subordonnées (voir le Lemme 13.1.2) il existe  $x$  tel que  $\|b\| = \|A\| \|x\|$  et il existe  $\delta b$  tel que  $\|\delta x\| = \|A^{-1}\| \|\delta b\|$ .

Pour obtenir (13.4) 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 (13.4). Pour en démontrer l'optimalité, on va montrer que pour toute matrice  $A$  il existe une perturbation  $\delta A$  et un second membre  $b$  pour lesquels il y a égalité. Grâce au Lemme 13.1.2 il existe  $y \neq 0$  tel que  $\|A^{-1}y\| = \|A^{-1}\| \|y\|$ . Soit  $\epsilon$  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 (13.4). □

Les conditionnements les plus utilisés en pratique sont

$$\text{cond}_p(A) = \|A\|_p \|A^{-1}\|_p \text{ pour } p = 1, 2, +\infty,$$

où les normes  $\|A\|_p$  sont explicitement définies dans le Lemme 13.1.1. On vérifie facilement un certain nombre de propriétés du conditionnement.

**Exercice 13.1.2** Soit une matrice  $A \in \mathcal{M}_n(\mathbb{C})$ . Vérifier que

1.  $\text{cond}(A) = \text{cond}(A^{-1}) \geq 1$ ,  $\text{cond}(\alpha A) = \text{cond}(A) \forall \alpha \neq 0$ ,
2. pour une matrice quelconque,  $\text{cond}_2(A) = \frac{\mu_n(A)}{\mu_1(A)}$ , 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,  $\text{cond}_2(A) = \frac{|\lambda_n(A)|}{|\lambda_1(A)|}$ , 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  $U$ ,  $\text{cond}_2(U) = 1$ ,

5. pour toute matrice unitaire  $U$ ,  $\text{cond}_2(AU) = \text{cond}_2(UA) = \text{cond}_2(A)$ .

**Exercice 13.1.3** Montrer que le conditionnement de la matrice de rigidité  $\mathcal{K}_h$ , donnée par (6.12) pour la méthode des éléments finis  $P_1$  appliquée au Laplacien, est

$$\text{cond}_2(\mathcal{K}_h) \approx \frac{4}{\pi^2 h^2}. \quad (13.5)$$

On montrera que les valeurs propres de  $\mathcal{K}_h$  sont

$$\lambda_k = 4h^{-2} \sin^2 \left( \frac{k\pi}{2(n+1)} \right) \quad 1 \leq k \leq n,$$

pour des vecteurs propres  $u^k$  donnés par leurs composantes

$$u_j^k = \sin \left( \frac{jk\pi}{n+1} \right) \quad 1 \leq j, k \leq n.$$

**Remarque 13.1.12** L'estimation (13.5) du conditionnement de la matrice de rigidité  $\mathcal{K}_h$  semble très pessimiste, voire catastrophique. En effet, la méthode des éléments finis converge si  $h = 1/(n+1)$  tend vers zéro. Autrement dit, des résultats précis ne peuvent être obtenus que si la matrice est très grande et très mal conditionnée. Mais dans ce cas, les inévitables erreurs d'arrondi sur le second membre ou sur la matrice risquent d'être énormément amplifiées au point de rendre la solution discrète  $u_h$  très différente de sa limite prédite. Fort heureusement, il n'en est rien en pratique car le second membre  $b_h$  du système linéaire  $\mathcal{K}_h U_h = b_h$  n'est pas quelconque et ne rend pas les inégalités de la Proposition 13.1.10 optimales. Si l'on reprend la démonstration de l'optimalité de ces inégalités, on s'aperçoit qu'elle est obtenue pour un vecteur  $b$  qui est vecteur propre de  $\mathcal{K}_h$  associé à sa plus grande valeur propre  $\lambda_n$ . D'après l'Exercice 13.1.3, un tel vecteur propre oscille fortement sur le maillage (ses composantes changent de signe d'une maille à l'autre). Si le second membre  $b_h$  est plus "régulier" (c'est-à-dire qu'il est combinaison linéaire des  $K$  premiers vecteurs propres  $u^k$  de  $\mathcal{K}_h$ ), on peut améliorer le résultat de la Proposition 13.1.10 en obtenant

$$\frac{\|\delta x\|}{\|x\|} \leq C(K) \frac{\|\delta b\|}{\|b\|},$$

où  $C(K)$  est une constante indépendante de  $n$ . C'est exactement ce qui arrive en pratique, et cette dernière inégalité justifie l'utilisation de la méthode des éléments finis malgré la présence d'erreurs d'arrondi dans les calculs sur ordinateurs. •

### 13.1.3 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 récurrente. En effet, le système

$$\begin{cases} t_{1,1}x_1 + t_{1,2}x_2 + \cdots & \cdots & t_{1,n}x_n = b_1 \\ & t_{2,2}x_2 + \ddots & \cdots & t_{2,n}x_n = b_2 \\ & & \ddots & \vdots \\ & & & t_{n-1,n-1}x_{n-1} + t_{n-1,n}x_n = b_{n-1} \\ & & & & t_{n,n}x_n = b_n \end{cases}$$

se résout en calculant d'abord  $x_n = b_n/t_{n,n}$ , puis  $x_{n-1}$ , et ainsi de suite jusqu'à  $x_1$ . On appelle ce procédé **remontée** (dans le cas d'une matrice triangulaire inférieure, le procédé similaire qui calcule les composantes de la solution de  $x_1$  à  $x_n$  est appelé **descente**). Remarquons que l'on a ainsi résolu 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 13.1.13** *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  $(a_{ij}^k)_{1 \leq i, j \leq n}$  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.

Plus précisément, on effectue les opérations suivantes. On multiplie  $A^k$  par une matrice de permutation  $P^k$  pour obtenir  $\tilde{A}^k = P^k A^k$  telle que son pivot  $\tilde{a}_{kk}^k$  soit non nul. Si  $a_{kk}^k \neq 0$ , alors il suffit de prendre  $P^k = \text{Id}$ . Sinon, s'il existe  $a_{ik}^k \neq 0$  avec  $i \geq k+1$ , on permute la  $k$ -ème ligne avec la  $i$ -ème en prenant  $P^k = (e_1, \dots, e_{k-1}, e_i, e_{k+1}, \dots, e_{i-1}, e_k, e_{i+1}, \dots, e_n)$  (si tous les éléments de la  $k$ -ème colonne sous la diagonale,  $a_{ik}^k$  avec  $i \geq k$ , sont nuls, alors il n'y



a rien à faire!). Puis on multiplie  $\tilde{A}^k$  par une matrice  $E^k$ , définie par

$$E^k = \begin{pmatrix} 1 & & & & & \\ 0 & \ddots & & & & \\ \vdots & & 1 & & & \\ \vdots & & -\frac{\tilde{a}_{k+1,k}^k}{\tilde{a}_{k,k}^k} & 1 & & \\ \vdots & & \vdots & & \ddots & \\ 0 & & -\frac{\tilde{a}_{n,k}^k}{\tilde{a}_{k,k}^k} & & & 1 \end{pmatrix}, \quad (13.6)$$

ce qui élimine tous les coefficients de la  $k$ -ème colonne en dessous de la diagonale. On pose

$$A^{k+1} = E^k \tilde{A}^k = \begin{pmatrix} \tilde{a}_{11}^1 & \cdots & \cdots & \cdots & \cdots & \tilde{a}_{1n}^1 \\ 0 & \ddots & & & & \vdots \\ \vdots & \ddots & \tilde{a}_{k,k}^k & \tilde{a}_{k,k+1}^k & \cdots & \tilde{a}_{k,n}^k \\ \vdots & & 0 & a_{k+1,k+1}^{k+1} & \cdots & a_{k+1,n}^{k+1} \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{n,k+1}^{k+1} & \cdots & a_{n,n}^{k+1} \end{pmatrix}$$

avec  $a_{ij}^{k+1} = \tilde{a}_{ij}^k - \frac{\tilde{a}_{i,k}^k}{\tilde{a}_{k,k}^k} \tilde{a}_{k,j}^k$  pour  $k+1 \leq i, j \leq n$ . La matrice  $A^{k+1}$  a donc bien la forme désirée avec ses  $k$  premières colonnes ayant uniquement des zéros sous la diagonale. Après  $(n-1)$  étapes, la matrice  $A^n$  est triangulaire supérieure et vérifie  $A^n = MA$  avec  $M = E^{n-1}P^{n-1}\dots E^1P^1$ . La matrice  $M$  est inversible car  $\det P^i = \pm 1$  et  $\det E^i = 1$ .

On peut mettre à jour le second membre (c'est-à-dire calculer  $Mb$ ) au fur et à mesure que l'on calcule les matrices  $P^k$  et  $E^k$ . On construit une suite de seconds membres  $(b^k)_{1 \leq k \leq n}$  définis par

$$b^1 = b, \quad b^{k+1} = E^k P^k b^k \text{ pour } 1 \leq k \leq n-1,$$

et on a bien  $b^n = Mb$ . Pour résoudre le système linéaire  $Ax = b$  il ne reste plus qu'à résoudre le système  $A^n x = Mb$  où  $A^n = T$  est une matrice triangulaire supérieure.  $\square$

**Remarque 13.1.14** Insistons sur quelques aspects pratiques de la méthode d'élimination de Gauss.

1. On ne calcule jamais  $M$ ! On n'a pas besoin de multiplier les matrices  $E^k$  et  $P^k$  pour calculer  $Mb$  et  $A^n$ .
2. Si  $A$  n'est pas inversible, on va trouver un coefficient diagonal de  $A^n = T$  nul et on ne va pas pouvoir résoudre  $Tx = Mb$ . Par contre l'élimination est toujours possible.
3. A l'étape  $k$ , on ne modifie que la partie des lignes  $k+1$  à  $n$  comprise entre les colonnes  $k+1$  à  $n$ .
4. Comme sous-produit de l'élimination de Gauss, on peut calculer le déterminant de la matrice  $A$ . En effet, on a  $\det A = \pm \det T$  selon le nombre de permutations effectuées.

5. Pour obtenir une meilleure stabilité numérique dans les calculs sur ordinateurs on peut choisir astucieusement le pivot  $\tilde{a}_{kk}^k$ . Pour éviter la propagation des erreurs d'arrondi il faut choisir le plus grand pivot possible en valeur absolue. Même lorsque le pivot naturel  $a_{kk}^k$  n'est pas nul, on permute pour amener à sa place un plus grand pivot  $\tilde{a}_{kk}^k$ . On dit que l'on fait un pivot partiel si on prend le plus grand pivot possible dans la  $k$ -ème colonne en dessous de la diagonale (comme on a fait dans la démonstration ci-dessus). On dit que l'on fait un pivot total si on prend le plus grand pivot possible dans la sous-matrice inférieure diagonale de taille  $(n - k) \times (n - k)$  (dans ce cas on permute ligne et colonne).

•

### 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 13.1.15** *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*

$$A = LU.$$

**Remarque 13.1.16** L'hypothèse de la Proposition 13.1.15 n'est pas déraisonnable. En effet, elle est vraie si, par exemple,  $A$  est définie positive. En effet, si  $\Delta^k$  n'est pas inversible, alors il existe un vecteur non nul  $x^k \in \text{Ker} \Delta^k$  et en le complétant par des zéros on construit un vecteur non nul  $x = (x^k, 0)$  qui vérifie  $Ax \cdot x = 0$ , ce qui contredit le caractère défini positif de  $A$ .

•

**Démonstration.** Supposons qu'au cours de l'élimination de Gauss il n'y ait pas besoin de faire de permutations pour changer de pivot, c'est-à-dire que tous les pivots naturels  $a_{kk}^k$  sont non nuls. Alors, avec les notations de la Proposition 13.1.13 on a  $A^n = E^{n-1} \dots E^1 A$  avec  $E^k$  définie par (13.6). On pose  $U = A^n$  et  $L = (E^1)^{-1} \dots (E^{n-1})^{-1}$ . Alors on a  $A = LU$  et il reste simplement à vérifier que  $L$  est bien triangulaire inférieure. Un calcul facile montre que  $(E^k)^{-1}$  s'obtient facilement

à partir de  $E^k$  en changeant le signe des éléments sous la diagonale, c'est-à-dire que, posant  $l_{ik} = a_{i,k}^k / a_{k,k}^k$ , pour  $k+1 \leq i \leq n$ , on a

$$E^k = \begin{pmatrix} 1 & & & & \\ 0 & \ddots & & & \\ \vdots & & 1 & & \\ \vdots & & -l_{k+1,k} & \ddots & \\ \vdots & & \vdots & & \ddots \\ 0 & -l_{n,k} & & & 1 \end{pmatrix}, \quad (E^k)^{-1} = \begin{pmatrix} 1 & & & & \\ 0 & \ddots & & & \\ \vdots & & 1 & & \\ \vdots & & +l_{k+1,k} & \ddots & \\ \vdots & & \vdots & & \ddots \\ 0 & +l_{n,k} & & & 1 \end{pmatrix}.$$

Un autre calcul montre que  $L$  est triangulaire inférieure et que sa  $k$ -ème colonne est la même que celle de  $(E^k)^{-1}$

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{pmatrix}.$$

Il faut maintenant vérifier que les pivots ne s'annulent pas sous l'hypothèse faite sur les matrices  $\Delta^k$ . On le vérifie par récurrence. Le premier pivot  $a_{11}$  est non nul car égal à  $\Delta^1$  qui est inversible. On suppose que tous les pivots jusqu'à l'ordre  $k-1$  sont non nuls. Montrons que le nouveau pivot  $a_{kk}^k$  est aussi non nul. Comme les  $k-1$  premiers pivots sont non nuls, on a pu calculer sans encombre la matrice  $A^k$ . On écrit alors l'égalité  $(E^1)^{-1} \dots (E^{k-1})^{-1} A^k = A$  sous la forme d'une égalité entre matrices par blocs

$$\begin{pmatrix} L_{11}^k & 0 \\ L_{21}^k & \text{Id} \end{pmatrix} \begin{pmatrix} U_{11}^k & A_{12}^k \\ A_{21}^k & A_{22}^k \end{pmatrix} = \begin{pmatrix} \Delta^k & A_{12} \\ A_{21} & A_{22} \end{pmatrix},$$

avec  $U_{11}^k$ ,  $L_{11}^k$ , et  $\Delta^k$  des blocs carrés de taille  $k$ , et  $A_{22}^k$ ,  $\text{Id}$ , et  $A_{22}$  des blocs carrés de taille  $n-k$ . En appliquant la règle de multiplication des matrices par blocs, on obtient

$$L_{11}^k U_{11}^k = \Delta^k,$$

où  $U_{11}^k$  est une matrice triangulaire supérieure, et  $L_{11}^k$  une matrice triangulaire inférieure avec des 1 sur la diagonale. On en déduit que la matrice  $U_{11}^k = (L_{11}^k)^{-1} \Delta^k$  est inversible comme produit de matrices inversibles. Son déterminant est donc non nul. Or

$$\det U_{11}^k = \prod_{i=1}^k a_{ii}^k \neq 0,$$

donc le pivot  $a_{kk}^k$  à l'étape  $k$  est non nul.

Il ne reste plus qu'à vérifier l'unicité. Soit deux décompositions LU de la matrice  $A = L_1 U_1 = L_2 U_2$ . On en déduit que  $L_2^{-1} L_1 = U_2 U_1^{-1}$ , où la matrice  $L_2^{-1} L_1$

est triangulaire inférieure et  $U_2 U_1^{-1}$  est triangulaire supérieure en vertu du Lemme 13.1.17. Elles sont donc toutes les deux diagonales, et comme la diagonale de  $L_2^{-1} L_1$  est composée de 1, on a  $L_2^{-1} L_1 = U_2 U_1^{-1} = \text{Id}$ .  $\square$

**Lemme 13.1.17** *Soit  $T$  une matrice triangulaire inférieure. Son inverse (s'il existe) est aussi une matrice triangulaire inférieure et ses éléments diagonaux sont les inverses des éléments diagonaux de  $T$ . Soit  $T'$  une autre matrice triangulaire inférieure. Le produit  $TT'$  est aussi triangulaire inférieure, et ses éléments diagonaux sont le produit des éléments diagonaux de  $T$  et de  $T'$ .*

Nous laissons au lecteur la démonstration élémentaire du Lemme 13.1.17.

**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

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{2,1} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n,1} & \dots & l_{n,n-1} & 1 \end{pmatrix}, \quad U = \begin{pmatrix} u_{1,1} & \dots & \dots & u_{1,n} \\ 0 & u_{2,2} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{n,n} \end{pmatrix},$$

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  $A$

$$\begin{aligned} a_{i,j} &= \sum_{k=1}^i l_{i,k} u_{k,j} \Rightarrow u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k} u_{k,j} & \text{pour } 1 \leq i \leq j, \\ a_{i,j} &= \sum_{k=1}^j l_{i,k} u_{k,j} \Rightarrow l_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} l_{i,k} u_{k,j}}{u_{j,j}} & \text{pour } j+1 \leq i \leq n. \end{aligned}$$

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!

**Algorithme numérique.** On écrit en pseudo-langage informatique l'algorithme correspondant à la méthode de décomposition LU. On vient de voir que l'on parcourt la matrice  $A$  colonne par colonne : à l'étape  $k$  on calcule la  $k$ -ème colonne de  $L$ , puis

on fait apparaître des zéros sous la diagonale de la  $k$ -ème colonne en effectuant des combinaisons linéaires de la  $k$ -ème ligne avec chacune des lignes de  $k+1$  à  $n$ . Comme à l'étape  $k$  les  $k$  premières lignes et les  $k-1$  premières colonnes de la matrice ne sont plus modifiées, on peut stocker dans un même tableau, contenant initialement la matrice  $A$ , les matrices  $A^k$  et  $L^k = (E^1)^{-1} \dots (E^{k-1})^{-1}$  (on ne stocke que les éléments non triviaux de  $L^k$  à la place des zéros de  $A^k$  sous la diagonale de ses  $k-1$  premières colonnes). A la fin ce tableau contiendra les matrices  $L$  et  $U$  ( $L$  dans sa partie inférieure (sans la diagonale de 1) et  $U$  dans sa partie supérieure).

```

Pour  $k = 1, n-1$       ← étape  $k$ 
  Pour  $i = k+1, n$     ← ligne  $i$ 
     $a_{ik} = \frac{a_{ik}}{a_{kk}}$     ← nouvelle colonne de  $L$ 
    Pour  $j = k+1, n$ 
       $a_{ij} = a_{ij} - a_{ik}a_{kj}$  ← combinaison des lignes  $i$  et  $k$ 
    Fin de la boucle en  $j$ 
  Fin de la boucle en  $i$ 
Fin de la boucle en  $k$ 

```

**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.

**Remarque 13.1.18** On utilise aussi la méthode de la factorisation LU pour calculer le déterminant et l'inverse d'une matrice. Pour obtenir  $A^{-1}$ , on décompose  $A$  en

facteurs LU et on résout  $n$  systèmes linéaires avec comme seconds membres les vecteurs de base  $(e_i)_{1 \leq i \leq n}$  (2 substitutions par résolution, mais les vecteurs de base  $e_i$  ont de nombreuses composantes nulles, ce qui rend moins chère l'étape de descente avec  $L$ ). Le nombre d'opérations pour calculer  $A^{-1}$  est

$$N_{op} \approx \frac{n^3}{3} + \sum_{j=1}^n \frac{j^2}{2} + n\left(\frac{n^2}{2}\right) \approx n^3.$$

Pour calculer le déterminant de  $A$ , on décompose  $A$  en facteurs LU et on calcule le déterminant de  $U$  (celui de  $L$  vaut 1), ce qui ne nécessite que de multiplier les éléments diagonaux de  $U$  entre eux ( $n - 1$  multiplications). Le nombre d'opérations pour calculer  $\det A$  est donc  $N_{op} \approx n^3/3$ . •

### 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 13.1.19** *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*

$$A = BB^*.$$

**Démonstration.** Par application de la Proposition 13.1.15, il existe un unique couple de matrices  $(L, U)$  tel que  $A = LU$  avec  $U$  triangulaire supérieure et  $L$  triangulaire inférieure ayant une diagonale de 1. On note  $D$  la matrice diagonale définie par  $D = \text{diag}(\sqrt{u_{ii}})$ . Il est possible de prendre la racine carrée des éléments de la diagonale de  $U$  car un argument de multiplication de matrices par blocs montre que  $\prod_{i=1}^k u_{ii} = \det \Delta^k > 0$ , où  $\Delta^k$  est la sous-matrice diagonale d'ordre  $k$  extraite de  $A$ , donc chaque  $u_{ii}$  est strictement positif. On pose alors  $B = LD$  et  $C = D^{-1}U$  qui vérifient  $A = BC$ . Comme  $A = A^*$ , on en déduit  $C(B^*)^{-1} = B^{-1}(C^*)$ . En vertu du Lemme 13.1.17 la matrice  $C(B^*)^{-1}$  est triangulaire supérieure, tandis que  $B^{-1}C^*$  est triangulaire inférieure. Elles sont donc toutes les deux diagonales. De plus, les éléments diagonaux de  $B$  et  $C$  sont les mêmes, donc la diagonale de  $B^{-1}C^*$  n'est constituée que de 1, c'est-à-dire que  $C(B^*)^{-1} = B^{-1}C^* = \text{Id}$ , donc  $C = B^*$ . Pour montrer l'unicité de la décomposition de Cholesky, on suppose qu'il existe deux factorisations  $A = B_1B_1^* = B_2B_2^*$ , d'où  $B_2^{-1}B_1 = B_2^*(B_1^*)^{-1}$ . Du Lemme 13.1.17, on en déduit que  $B_2^{-1}B_1 = D = \text{diag}(d_1, \dots, d_n)$ , et donc que  $A = B_2B_2^* = B_2(DD^*)B_2^*$ . Comme  $B_2$  est inversible, il vient  $D^2 = \text{Id}$  donc  $d_i = \pm 1$ . Or tous les coefficients diagonaux d'une décomposition de Cholesky sont positifs par hypothèse. Donc  $d_i = 1$ , ce qui implique  $B_1 = B_2$ . □

**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

$$\begin{aligned} a_{jj} &= \sum_{k=1}^j (b_{jk})^2 \Rightarrow b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2} \\ a_{i,j} &= \sum_{k=1}^j b_{jk} b_{i,k} \Rightarrow b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} b_{jk} b_{i,k}}{b_{jj}} \text{ pour } j+1 \leq i \leq n. \end{aligned}$$

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.

**Algorithme numérique.** L'algorithme de Cholesky peut s'écrire de façon très compacte en utilisant un seul tableau qui contient originellement  $A$  et qui se remplit peu à peu du facteur  $B$ . Remarquons qu'il suffit de stocker la moitié inférieure de  $A$  puisque  $A$  est symétrique.

```
Pour j = 1, n
  Pour k = 1, j - 1
    ajj = ajj - (bjk)2
  Fin de la boucle en k
  ajj = √ajj
  Pour i = j + 1, n
    Pour k = 1, j - 1
      aij = aij - bjkbik
    Fin de la boucle en k
    aij = aij / ajj
  Fin de la boucle en i
Fin de la boucle en j
```

**Compte d'opérations.** Pour mesurer l'efficacité de la méthode de Cholesky on compte le nombre d'opérations (uniquement les multiplications) nécessaires à son ac-

complissement. 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 13.1.20** 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 13.1.4** Montrer que les factorisations LU et de Cholesky conservent la structure bande des matrices.

**Remarque 13.1.21** 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 13.1.5** 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  $\mathcal{O}(np^2/6)$ .



Passons au cas des matrices creuses. Expliquons d'abord la manière de stocker ces matrices creuses dans la mémoire des ordinateurs. Comme on ne stocke que les éléments non nuls de la matrice, on obtient un gain de place appréciable. Nous présentons une méthode de stockage, appelée **stockage Morse**, sur un exemple simple, étant entendu qu'on ne l'utilise en pratique que pour des matrices de grande taille. Soit donc la matrice

$$A = \begin{pmatrix} 9 & 0 & -3 & 0 \\ 7 & -1 & 0 & 4 \\ 0 & 5 & 2 & 0 \\ 1 & 0 & -1 & 2 \end{pmatrix}.$$

Les éléments de  $A$  sont stockés, ligne par ligne, dans un tableau à une seule entrée STOCKA. On définit un tableau DEBUTL qui indique le début des lignes de  $A$  dans STOCKA : plus précisément STOCKA(DEBUTL( $i$ )) est le premier élément non nul de la ligne  $i$ . On a aussi besoin d'un tableau INDICC qui indique la colonne de chaque élément stocké dans STOCKA : si  $a_{i,j}$  est stocké dans STOCKA( $k$ ), alors INDICC( $k$ ) =  $j$ . Le nombre d'éléments non nuls de  $A$  est égal à la taille du vecteur INDICC (ou du vecteur STOCKA). Le nombre de lignes de  $A$  est égal à la taille du vecteur DEBUTL. Dans notre exemple, on a

STOCKA	INDICC	DEBUTL
9	1	1
-3	3	3
7	1	6
-1	2	8
4	4	
5	2	
2	3	
1	1	
-1	3	
2	4	

En général, les factorisations LU et de Cholesky d'une matrice creuse produisent des facteurs "pleins" avec peu d'éléments non nuls (il existe des algorithmes qui minimisent ce "remplissage" mais ils sont moins efficaces que dans le cas des matrices bandes). En pratique, les matrices creuses sont souvent associée aux méthodes itératives que nous verrons dans la Sous-section 13.1.4 ci-dessous. En effet, les produits matrice-vecteur sont faciles à évaluer avec ce type de stockage.

Équivalence d'opérations et algorithme de Strassen

On remarque que toutes ces méthodes de résolution de systèmes linéaires ont un compte d'opérations de l'ordre de  $n^3$ , de même que le calcul de l'inverse (voir la Remarque 13.1.18) ou la multiplication de deux matrices d'ordre  $n$ . Il ne s'agit pas d'une coïncidence comme le montre le résultat suivant qui montre, de manière surprenante, que l'inversion d'une matrice

a la même complexité que la multiplication de deux matrices (bien que l'inversion semble, à première vue, une opération bien plus compliquée).

**Lemme 13.1.22** *L'inversion de matrices et la multiplication de matrices ont la même complexité asymptotique, c'est à dire que s'il existe un algorithme pour l'une de ces opérations tel que son nombre d'opérations est borné par  $\mathcal{O}(n^\alpha)$  avec  $\alpha \geq 2$ , alors on peut construire un algorithme pour l'autre opération dont le nombre d'opérations est aussi borné par  $\mathcal{O}(n^\alpha)$ .*

**Démonstration.** Soit  $I(n)$  le nombre d'opérations pour calculer  $A^{-1}$  par un algorithme donné, tel qu'il existe  $C$  et  $\alpha \geq 2$  vérifiant  $I(n) \leq Cn^\alpha$ . Montrons qu'il existe un algorithme pour calculer le produit  $AB$  dont le nombre d'opérations  $P(n)$  est tel qu'il existe  $C'$  pour lequel  $P(n) \leq C'n^\alpha$  avec le même exposant  $\alpha$ . On remarque que

$$\begin{pmatrix} \text{Id} & A & 0 \\ 0 & \text{Id} & B \\ 0 & 0 & \text{Id} \end{pmatrix}^{-1} = \begin{pmatrix} \text{Id} & -A & AB \\ 0 & \text{Id} & -B \\ 0 & 0 & \text{Id} \end{pmatrix}.$$

Par conséquent, le produit  $AB$  est obtenu en inversant une matrice 3 fois plus grande. Donc

$$P(n) \leq I(3n) \leq C3^\alpha n^\alpha.$$

Soit maintenant  $P(n)$  le nombre d'opérations pour calculer  $AB$  par un algorithme donné, tel qu'il existe  $C$  et  $\alpha$  vérifiant  $P(n) \leq Cn^\alpha$ . Montrons qu'il existe un algorithme pour calculer  $A^{-1}$  dont le nombre d'opérations  $I(n)$  est tel qu'il existe  $C'$  pour lequel  $I(n) \leq C'n^\alpha$  avec le même exposant  $\alpha$ . On remarque que

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B\Delta^{-1}CA^{-1} & -A^{-1}B\Delta^{-1} \\ -\Delta^{-1}CA^{-1} & \Delta^{-1} \end{pmatrix}.$$

avec  $\Delta = D - CA^{-1}B$  (appelé parfois complément de Schur). On en déduit

$$I(2n) \leq 2I(n) + 6P(n),$$

si on néglige les additions. En itérant cette formule pour  $n = 2^k$ , on obtient

$$I(2^k) \leq 2^k I(1) + 6 \sum_{i=0}^{k-1} 2^{k-i-1} P(2^i) \leq C \left( 2^k + \sum_{i=0}^{k-1} 2^{k-i-1+\alpha i} \right).$$

Comme  $\alpha \geq 2$ , on en déduit

$$I(2^k) \leq C' 2^{\alpha k}.$$

Si  $n \neq 2^k$  pour tout  $k$ , il existe  $k$  tel que  $2^k < n < 2^{k+1}$ . On inscrit alors la matrice  $A$  dans une plus grande matrice de taille  $2^{k+1}$

$$\begin{pmatrix} A & 0 \\ 0 & \text{Id} \end{pmatrix}$$

où  $\text{Id}$  est l'identité d'ordre  $2^{k+1} - n$ . On obtient

$$I(n) \leq C'(2^{k+1})^\alpha \leq C' 2^{\alpha n},$$

ce qui est le résultat désiré. □

On a longtemps cru que la multiplication de matrices d'ordre  $n$  (et donc leur inversion) ne pouvait pas se faire en moins de  $n^3$  opérations. Mais depuis une découverte de Strassen en 1969, on sait que ce n'est pas vrai. En effet, Strassen a mis au point un algorithme de multiplication de matrices qui demande beaucoup moins d'opérations pour  $n$  grand. Il a obtenu pour son algorithme un compte d'opérations

$$N_{op}(n) = \mathcal{O}(n^{\log_2 7}) \quad \text{avec } \log_2 7 \sim 2,81. \quad (13.7)$$

Ce résultat, tout à fait surprenant, a depuis été amélioré : on a trouvé d'autres algorithmes, de plus en plus compliqués, dont le nombre d'opérations croît moins vite pour  $n$  grand, mais on n'a toujours pas trouvé le meilleur algorithme possible (c'est-à-dire celui qui conduise au plus petit exposant  $\alpha$  tel que  $N_{op}(n) = \mathcal{O}(n^\alpha)$ ). Évidemment, pour de très grandes matrices, le gain de temps procuré par ces algorithmes est très appréciable (l'algorithme de Strassen a effectivement été utilisé sur des super-ordinateurs). Malheureusement, ces algorithmes ont souvent des problèmes de stabilité numérique (ils amplifient les erreurs d'arrondi) qui limitent leur usage pratique.

L'algorithme de Strassen repose sur le lemme suivant d'apparence bien bénigne!

**Lemme 13.1.23 (algorithme de Strassen)** *Le produit de deux matrices d'ordre 2 peut se faire avec 7 multiplications et 18 additions (au lieu de 8 multiplications et 4 additions par la règle habituelle).*

**Démonstration.** Un simple calcul montre que

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix} = \begin{pmatrix} m_1 + m_2 - m_4 + m_6 & m_4 + m_5 \\ m_6 + m_7 & m_2 - m_3 + m_5 - m_7 \end{pmatrix},$$

avec

$$\begin{aligned} m_1 &= (b-d)(\gamma+\delta) & m_5 &= a(\beta-\delta) \\ m_2 &= (a+d)(\alpha+\delta) & m_6 &= d(\gamma-\alpha) \\ m_3 &= (a-c)(\alpha+\beta) & m_7 &= (c+d)\alpha \\ m_4 &= (a+b)\delta \end{aligned}$$

On compte bien 7 multiplications et 18 additions. □

Le point crucial dans le Lemme 13.1.23 est que la règle de multiplication de Strassen est aussi valable si les coefficients des matrices sont dans une algèbre non commutative. En particulier, c'est donc vrai pour des matrices blocs.

Soit alors une matrice de taille  $n = 2^k$ . On découpe cette matrice en 4 blocs de taille  $2^{k-1}$ , et on applique la règle de Strassen. Si on compte non seulement les multiplications mais aussi les additions, le nombre d'opérations  $N_{op}(n)$  pour effectuer le produit de deux matrices vérifie

$$N_{op}(2^k) = 7N_{op}(2^{k-1}) + 18(2^{k-1})^2,$$

car l'addition de deux matrices de taille  $n$  requiert  $n^2$  additions. Une récurrence simple donne

$$N_{op}(2^k) = 7^k N_{op}(1) + 18 \sum_{i=0}^{k-1} 7^i 4^{k-1-i} \leq 7^k (N_{op}(1) + 6).$$

On en déduit facilement que le nombre optimal d'opérations  $N_{op}(n)$  vérifie la borne (13.7).

### 13.1.4 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 13.1.24** 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

$$\begin{cases} x_0 \text{ donné dans } \mathbb{R}^n, \\ Mx_{k+1} = Nx_k + b \quad \forall k \geq 1. \end{cases} \quad (13.8)$$

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 (13.8), 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  $\epsilon$  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 13.1.25** 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 13.1.5 pour la notion de rayon spectral).

**Lemme 13.1.26** La méthode itérative définie par (13.8) 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 13.1.8, on en déduit que  $e_k$  tend vers 0, quel que soit  $e_0$ , si et seulement si  $\rho(M^{-1}N) < 1$ .  $\square$

Dans la pratique le rayon spectral d'une matrice est difficile à calculer (il faut calculer ses valeurs propres). C'est pourquoi on utilise d'autres conditions suffisantes de convergence, comme indiqué ci-dessous.

**Lemme 13.1.27** *Soit  $A$  une matrice hermitienne, définie positive. Soit une décomposition régulière de  $A$  définie par  $A = M - N$  avec  $M$  inversible. Alors la matrice  $(M^* + N)$  est hermitienne. De plus, si  $(M^* + N)$  est aussi définie positive, alors*

$$\rho(M^{-1}N) < 1.$$

**Démonstration.** Tout d'abord, montrons que  $M^* + N$  est bien hermitienne :

$$(M^* + N)^* = M + N^* = (A + N) + N^* = A^* + N^* + N = M^* + N.$$

On définit la norme vectorielle  $|x|_A = \sqrt{\langle Ax, x \rangle}$  sur  $\mathbb{R}^n$  (qui est bien une norme car  $A$  est définie positive). On note  $\|\cdot\|$  la norme matricielle subordonnée à  $|\cdot|_A$ . On va montrer que  $\|M^{-1}N\| < 1$  ce qui implique le résultat désiré grâce à la Proposition 13.1.7. On calcule

$$\|M^{-1}N\|^2 = \max_{|v|_A=1} |M^{-1}Nv|_A^2.$$

Or, d'après le Lemme 13.1.2, il existe  $v$ , dépendant de  $M^{-1}N$ , tel que  $|v|_A = 1$  et

$$|M^{-1}Nv|_A^2 = \|M^{-1}N\|^2.$$

Comme  $N = M - A$ , on obtient, en posant  $w = M^{-1}Av$ ,

$$\begin{aligned} |M^{-1}Nv|_A^2 &= \langle AM^{-1}Nv, M^{-1}Nv \rangle \\ &= \langle AM^{-1}(M - A)v, M^{-1}(M - A)v \rangle \\ &= \langle (Av - AM^{-1}Av), (I - M^{-1}A)v \rangle \\ &= \langle Av, v \rangle - \langle AM^{-1}Av, v \rangle \\ &\quad + \langle AM^{-1}Av, M^{-1}Av \rangle - \langle Av, M^{-1}Av \rangle \\ &= 1 - \langle M^{-1}Av, MM^{-1}Av \rangle \\ &\quad + \langle AM^{-1}Av, M^{-1}Av \rangle - \langle MM^{-1}Av, M^{-1}Av \rangle \\ &= 1 - \langle w, Mw \rangle + \langle Aw, w \rangle - \langle Mw, w \rangle \\ &= 1 - \langle (M^* + N)w, w \rangle. \end{aligned}$$

Or  $\langle (M^* + N)w, w \rangle > 0$ , car  $(M^* + N)$  est définie positive, et  $w \neq 0$  car  $A$  et  $M$  inversibles. Donc

$$\|M^{-1}N\|^2 = 1 - \langle (M^* + N)w, w \rangle < 1,$$

ce qui termine la démonstration.  $\square$

Puisque ces méthodes itératives de résolution de systèmes linéaires sont destinées à être utilisées sur des ordinateurs dont les calculs ne sont pas exacts mais entachés d'erreurs d'arrondi, il est nécessaire de vérifier que ces erreurs ne se propagent pas au point de détruire la convergence des méthodes ou, pire, de les faire converger vers de fausses solutions. Fort heureusement cela n'est pas le cas comme le montre le résultat suivant.

**Lemme 13.1.28** Soit une décomposition régulière de  $A$  définie par  $A = M - N$  avec  $M$  inversible. Soit un second membre  $b \in \mathbb{R}^n$  et la solution  $x \in \mathbb{R}^n$  telle que  $Ax = b$ . On suppose qu'à chaque étape  $k$  la méthode itérative est affectée d'une erreur  $\epsilon_k \in \mathbb{R}^n$  au sens où  $x_{k+1}$  n'est pas exactement donné par (13.8) mais plutôt par

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b + \epsilon_k.$$

On suppose que  $\rho(M^{-1}N) < 1$  et qu'il existe une norme vectorielle et une constante positive  $\epsilon$  telles que pour tout  $k \geq 0$

$$\|\epsilon_k\| \leq \epsilon.$$

Alors il existe une constante  $K$ , qui ne dépend que de  $M^{-1}N$ , telle que

$$\limsup_{k \rightarrow +\infty} \|x_k - x\| \leq K\epsilon.$$

**Démonstration.** On définit encore l'erreur  $e_k = x_k - x$  et on a maintenant  $e_{k+1} = M^{-1}Ne_k + \epsilon_k$ . On en déduit

$$e_k = (M^{-1}N)^k e_0 + \sum_{i=0}^{k-1} (M^{-1}N)^i \epsilon_{k-i-1}. \quad (13.9)$$

Par application de la Proposition 13.1.7 il existe une norme matricielle subordonnée  $\|\cdot\|_s$  telle que  $\|M^{-1}N\|_s < 1$  puisque  $\rho(M^{-1}N) < 1$ . On note de la même façon la norme vectorielle associée. Or toutes les normes vectorielles sur  $\mathbb{R}^n$  sont équivalentes : il existe donc une constante  $C \geq 1$ , qui ne dépend ainsi que de  $M^{-1}N$ , telle que

$$C^{-1}\|y\| \leq \|y\|_s \leq C\|y\| \quad \forall y \in \mathbb{R}^n.$$

En majorant (13.9) il vient

$$\|e_k\|_s \leq \|M^{-1}N\|_s^k \|e_0\|_s + \sum_{i=0}^{k-1} \|M^{-1}N\|_s^i C\epsilon \leq \|M^{-1}N\|_s^k \|e_0\|_s + \frac{C\epsilon}{1 - \|M^{-1}N\|_s}$$

d'où l'on obtient le résultat avec  $K = C^2/(1 - \|M^{-1}N\|_s)$ .  $\square$

Il est temps de donner les exemples les plus classiques de méthodes itératives basées sur une décomposition régulière.

**Définition 13.1.29 (méthode de Jacobi)** Soit  $A = (a_{ij})_{1 \leq i, j \leq n}$ . On note  $D = \text{diag}(a_{ii})$  la diagonale de  $A$ . On appelle méthode de Jacobi la méthode itérative associée à la décomposition

$$M = D, \quad N = D - A.$$

Il s'agit de la plus simple des méthodes itératives. Pour qu'elle soit bien définie, il faut bien sûr que la matrice diagonale  $D$  soit inversible. Par application du Lemme 13.1.27, dans le cas où  $A$  est symétrique réelle, la méthode de Jacobi converge si  $A$  et  $2D - A$  sont définies positives.

**Définition 13.1.30 (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  $D = \text{diag}(a_{ii})$  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

$$M = D - E, \quad N = F.$$

Pour que la méthode de Gauss-Seidel soit bien définie, il faut que la matrice  $D - E$  soit inversible, c'est-à-dire que  $D$  soit inversible (la matrice  $(D - E)$  est facile à inverser car elle est triangulaire). Par application du Lemme 13.1.27, si  $A$  est symétrique réelle définie positive, la méthode de Gauss-Seidel converge.

**Définition 13.1.31 (méthode de relaxation (SOR))** Soit  $\omega \in \mathbb{R}^+$ . On appelle méthode de relaxation (SOR en anglais pour "Successive Over Relaxation"), pour le paramètre  $\omega$ , la méthode itérative associée à la décomposition

$$M = \frac{D}{\omega} - E, \quad N = \frac{1 - \omega}{\omega} D + F$$

Pour que la méthode de relaxation soit bien définie, il faut encore que la matrice  $D$  soit inversible. Pour  $\omega = 1$ , on retrouve la méthode de Gauss-Seidel. Si  $\omega < 1$ , on dit qu'il s'agit d'une méthode de sous-relaxation, tandis que  $\omega > 1$  correspond à une méthode de sur-relaxation. En général, il existe un paramètre optimal  $\omega_{opt}$  qui minimise le rayon spectral de la matrice d'itération  $M^{-1}N$ , et donc qui maximise la vitesse de convergence.

**Exercice 13.1.6** Soit  $A$  une matrice hermitienne définie positive. Montrer que pour tout  $\omega \in ]0, 2[$ , la méthode de relaxation converge.

**Exercice 13.1.7** Montrer que, pour la méthode de relaxation, on a toujours

$$\rho(M^{-1}N) \geq |1 - \omega|, \quad \forall \omega \neq 0,$$

et donc qu'elle ne peut converger que si  $0 < \omega < 2$ .

**Définition 13.1.32 (méthode du gradient)** Soit un paramètre réel  $\alpha \neq 0$ . On appelle méthode du gradient la méthode itérative associée à la décomposition

$$M = \frac{1}{\alpha} \text{Id} \quad \text{et} \quad N = \left( \frac{1}{\alpha} \text{Id} - A \right).$$

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.

**Lemme 13.1.33** Soit  $A$  une matrice diagonalisable de valeurs propres  $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ . Si  $\lambda_1 \leq 0 \leq \lambda_n$ , alors la méthode du gradient ne converge pour aucune valeur de  $\alpha$ . Si  $0 < \lambda_1 \leq \dots \leq \lambda_n$ , alors la méthode du gradient converge si et seulement si  $0 < \alpha < 2/\lambda_n$ , et le paramètre  $\alpha$  optimal, qui minimise  $\rho(M^{-1}N)$ , est

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n} \quad \text{et} \quad \min_{\alpha} \rho(M^{-1}N) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}.$$

**Remarque 13.1.34** Si  $\lambda_1 \leq \dots \leq \lambda_n < 0$ , alors on a un résultat symétrique du cas défini positif en changeant  $\alpha$  en  $-\alpha$ . Pour le paramètre optimal  $\alpha_{opt}$ , le rayon spectral de la matrice d'itération est une fonction croissante du rapport  $\lambda_n/\lambda_1$ . Si  $A$  est auto-adjointe, alors ce rapport n'est autre que le conditionnement  $\text{cond}_2(A)$  de la matrice  $A$ . Par conséquent, plus la matrice  $A$  est bien conditionnée, meilleure est la convergence de la méthode du gradient. •

**Démonstration.** D'après le Lemme 13.1.26, on sait que la méthode du gradient est convergente si et seulement si  $\rho(M^{-1}N) < 1$ . Or  $M^{-1}N = (\text{Id} - \alpha A)$ , donc

$$\rho(M^{-1}N) < 1 \Leftrightarrow |1 - \alpha\lambda_i| < 1 \Leftrightarrow -1 < 1 - \alpha\lambda_i < 1, \quad \forall i.$$

Ceci implique que  $\alpha\lambda_i > 0$  pour tout  $1 \leq i \leq n$ . Par conséquent toutes les valeurs propres de  $A$  doivent être non nulles et du même signe que  $\alpha$ . La méthode du gradient ne converge donc pas si  $\lambda_1 \leq 0 \leq \lambda_n$ , et ceci quelque soit  $\alpha$ . Si, au contraire, on a  $0 < \lambda_1 \leq \dots \leq \lambda_n$ , alors on en déduit qu'il faut  $0 < \alpha < 2/\lambda_n$ . Pour calculer le paramètre optimal  $\alpha_{opt}$ , on remarque que la fonction  $\lambda \rightarrow |1 - \alpha\lambda|$  est décroissante sur  $] -\infty, 1/\alpha]$  puis croissante sur  $[1/\alpha, +\infty[$ , donc

$$\rho(M^{-1}N) = \max\{|1 - \alpha\lambda_1|, |1 - \alpha\lambda_n|\}.$$

Par conséquent, le minimum de la fonction  $\alpha \rightarrow \rho(M^{-1}N)$  est atteint au point d'intersection  $\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n}$  de ces deux droites. □

### 13.1.5 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. Il s'agit d'une amélioration spectaculaire de la méthode du gradient (surtout si elle combinée avec un préconditionnement, voir la Définition 13.1.43). Pour construire la méthode du gradient conjugué, on introduit la notion d'espace de Krylov.

**Définition 13.1.35** Soit  $r_0$  un vecteur dans  $\mathbb{R}^n$ . On appelle espace de Krylov associé au vecteur  $r_0$ , et on note  $K_k$ , le sous-espace vectoriel de  $\mathbb{R}^n$  engendré par les  $k+1$  vecteurs  $\{r_0, Ar_0, \dots, A^k r_0\}$ .



Les espaces de Krylov  $(K_k)_{k \geq 0}$  forment une suite croissante de sous-espaces vectoriels  $K_k \subset K_{k+1} \quad \forall k \geq 0$ . Comme  $K_k \subset \mathbb{R}^n$ , cette suite devient forcément stationnaire à partir d'un certain rang. Plus précisément, nous laissons le lecteur vérifier qu'il existe une dimension critique  $k_0$ , avec  $0 \leq k_0 \leq n-1$ , telle que

$$\begin{cases} \dim K_k = k+1 & \text{si } 0 \leq k \leq k_0, \\ \dim K_k = k_0+1 & \text{si } k_0 \leq k. \end{cases}$$

En prenant  $r_0 = b - Ax_0$ , il est facile de voir que, pour la méthode du gradient, l'itérée  $x_k$  appartient à l'espace affine  $[x_0 + K_{k-1}]$  (défini comme l'ensemble des vecteurs  $x$  tels que  $(x - x_0) \in K_{k-1}$ ), ce qui implique que le résidu  $r_k = b - Ax_k$  appartient à l'espace de Krylov  $K_k$  (associé au résidu initial  $r_0$ ).

Pour améliorer la méthode du gradient, on décide de "mieux choisir"  $x_k$  dans l'espace affine  $[x_0 + K_{k-1}]$ . La méthode du gradient conjugué consiste, à partir d'un vecteur initial  $x_0 \in \mathbb{R}^n$  et de son résidu  $r_0 = b - Ax_0$ , à construire une suite de vecteurs  $x_k \in [x_0 + K_{k-1}]$  tels que  $r_k = b - Ax_k$  est orthogonal au sous-espace  $K_{k-1}$ , pour  $k \geq 1$ .

**Lemme 13.1.36** *Soit  $A$  une matrice symétrique réelle définie positive d'ordre  $n$ . Soit  $x_0 \in \mathbb{R}^n$ ,  $r_0 = b - Ax_0$ , et  $(K_k)_{k \geq 0}$  la suite des espaces de Krylov associés à  $r_0$ . La méthode du gradient conjugué est définie, pour  $k \geq 1$ , par*

$$x_k \in [x_0 + K_{k-1}] \quad \text{et} \quad r_k = b - Ax_k \perp K_{k-1}. \quad (13.10)$$

*Pour tout  $k \geq 1$ , il existe bien un unique vecteur  $x_k$  donné par (13.10). De plus, cette méthode converge vers la solution du système linéaire  $Ax = b$  en, au plus,  $n$  itérations.*

**Remarque 13.1.37** Le Lemme 13.1.36 montre que l'algorithme du gradient conjugué qu'on a conçu comme une méthode itérative est en fait une méthode directe puisqu'il converge en un nombre fini d'itérations (exactement  $k_0 + 1$  où  $k_0$  est la dimension critique de Krylov). Cependant, dans la pratique on l'utilise comme une méthode itérative qui converge "numériquement" en beaucoup moins de  $k_0 + 1$  itérations. Intuitivement il est facile de voir pourquoi le gradient conjugué améliore le gradient simple. En effet, le résidu  $r_k$  est orthogonal à un sous-espace  $K_k$  de plus en plus grand. •

**Démonstration.** Montrons tout d'abord qu'il existe bien un unique  $x_k$  qui satisfait les hypothèses. Comme  $A$  est définie positive, on peut définir le produit scalaire  $\langle x, y \rangle_A = Ax \cdot y$  sur  $\mathbb{R}^n$ . On cherche  $x_k$  sous la forme  $x_k = x_0 + y_k$  avec  $y_k \in K_{k-1}$ , et la condition d'orthogonalité de  $r_k$  devient

$$\langle A^{-1}r_0 - y_k, y \rangle_A = 0 \quad \forall y \in K_{k-1},$$

ce qui n'est rien d'autre que la caractérisation de  $y_k$  comme projection orthogonale de  $A^{-1}r_0$  sur le sous-espace  $K_{k-1}$  (pour le produit scalaire  $\langle \cdot, \cdot \rangle_A$ ). Ceci prouve l'existence et l'unicité de  $x_k$ .

Soit  $k_0$  la dimension critique des espaces de Krylov, c'est-à-dire que, pour tout  $k \geq k_0$ ,  $\dim K_k = k_0 + 1$ . En particulier,  $AK_{k_0} \subset K_{k_0+1} = K_{k_0}$ , donc  $r_{k_0+1} = b - Ax_{k_0+1} = r_0 - Ay_{k_0+1}$  appartient à  $K_{k_0}$  tout en lui étant orthogonal. Par conséquent,  $r_{k_0+1} = 0$  et  $x_{k_0+1}$  est la solution exacte du système linéaire.  $\square$

**Remarque 13.1.38** La méthode du gradient conjugué a été présentée comme un procédé d'orthogonalisation par rapport à l'espace de Krylov. Nous verrons plus loin que, de manière équivalente, on peut l'introduire comme un problème de minimisation. Plus précisément,  $x_k \in [x_0 + K_{k-1}]$  réalise le minimum dans  $[x_0 + K_{k-1}]$  de

$$f(x) = \frac{1}{2}Ax \cdot x - b \cdot x,$$

ou bien encore le résidu  $r_k = b - Ax_k$  minimise dans  $K_k$  la fonction

$$g(r) = \frac{1}{2}A^{-1}r \cdot r$$

avec  $r = b - Ax$ .  $\bullet$

**Exercice 13.1.8** Soit  $A$  une matrice symétrique définie positive. Soit  $(x_k)_{0 \leq k \leq n}$  la suite de solutions approchées obtenues par la méthode du gradient conjugué. On pose  $r_k = b - Ax_k$  et  $d_k = x_{k+1} - x_k$ . Montrer que

(i) l'espace de Krylov  $K_k$  est aussi égal à

$$K_k = [r_0, \dots, r_k] = [d_0, \dots, d_k],$$

(ii) la suite  $(r_k)_{0 \leq k \leq n-1}$  est orthogonale

$$r_k \cdot r_l = 0 \text{ pour tout } 0 \leq l < k \leq n-1,$$

(iii) la suite  $(d_k)_{0 \leq k \leq n-1}$  est conjuguée par rapport à  $A$

$$Ad_k \cdot d_l = 0 \text{ pour tout } 0 \leq l < k \leq n-1.$$

La définition que l'on vient de donner de la méthode du gradient conjugué est purement théorique. En effet, on n'a pas indiqué d'algorithme pour construire  $r_k$  orthogonal à  $K_{k-1}$ , ni dit comment on calcule  $x_k$  dans la pratique. Le théorème suivant donne des formules pratiques particulièrement simples pour calculer ces vecteurs.

**Proposition 13.1.39** 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

$$p_0 = r_0 = b - Ax_0, \text{ et pour } 0 \leq k \begin{cases} x_{k+1} = x_k + \alpha_k p_k \\ r_{k+1} = r_k - \alpha_k A p_k \\ p_{k+1} = r_{k+1} + \beta_k p_k \end{cases} \quad (13.11)$$

avec

$$\alpha_k = \frac{\|r_k\|^2}{Ap_k \cdot p_k} \text{ et } \beta_k = \frac{\|r_{k+1}\|^2}{\|r_k\|^2}.$$

Alors,  $(x_k)_{0 \leq k \leq k_0+1}$  est la suite de solutions approchées de la méthode du gradient conjugué définie par (13.10).

**Démonstration.** Il est facile de montrer par récurrence que les relations

$$r_0 = b - Ax_0 \text{ et } \begin{cases} r_{k+1} = r_k - \alpha_k Ap_k \\ x_{k+1} = x_k + \alpha_k p_k \end{cases}$$

impliquent que la suite  $r_k$  est bien celle du résidu, à savoir  $r_k = b - Ax_k$ . Une autre récurrence facile montre que les relations

$$r_0 = p_0 \text{ et } \begin{cases} r_k = r_{k-1} - \alpha_{k-1} Ap_{k-1} \\ p_k = r_k + \beta_{k-1} p_{k-1} \end{cases}$$

impliquent que  $p_k$  et  $r_k$  appartiennent à l'espace de Krylov  $K_k$ , pour tout  $k \geq 0$ . On en déduit, par la relation de récurrence  $x_{k+1} = x_k + \alpha_k p_k$ , que  $x_{k+1}$  appartient bien à l'espace affine  $[x_0 + K_k]$ . Pour conclure, il nous faut montrer que  $r_{k+1}$  est orthogonal à  $K_k$ . Tout d'abord, nous allons montrer par récurrence que  $r_{k+1}$  est orthogonal à  $r_j$ , pour tout  $0 \leq j \leq k$ , et que  $p_{k+1}$  est conjuguée à  $p_j$ , pour tout  $0 \leq j \leq k$ , c'est-à-dire que  $Ap_{k+1} \cdot p_j = 0$ . A l'ordre 0 on a

$$r_1 \cdot r_0 = \|r_0\|^2 - \alpha_0 Ap_0 \cdot r_0 = 0$$

car  $p_0 = r_0$ , et

$$Ap_1 \cdot p_0 = (r_1 + \beta_0 p_0) \cdot Ap_0 = \alpha_0^{-1} (r_1 + \beta_0 r_0) \cdot (r_0 - r_1) = 0.$$

On suppose que jusqu'à l'ordre  $k$  on a

$$r_k \cdot r_j = 0 \text{ pour } 0 \leq j \leq k-1, \text{ et } Ap_k \cdot p_j = 0 \text{ pour } 0 \leq j \leq k-1.$$

Montrons que c'est encore vrai à l'ordre  $k+1$ . A cause de la formule de récurrence qui donne  $x_{k+1}$  on a

$$r_{k+1} \cdot r_j = r_k \cdot r_j - \alpha_k Ap_k \cdot r_j,$$

et à cause de la relation  $r_j = p_j - \beta_{j-1} p_{j-1}$  on obtient

$$r_{k+1} \cdot r_j = r_k \cdot r_j - \alpha_k Ap_k \cdot p_j + \alpha_k \beta_{j-1} Ap_k \cdot p_{j-1}.$$

A cause de l'hypothèse de récurrence, on en déduit facilement que  $r_{k+1} \cdot r_j = 0$  si  $j \leq k-1$ , tandis que la formule pour  $\alpha_k$  implique que  $r_{k+1} \cdot r_k = 0$ . D'autre part, la formule de récurrence qui donne  $p_{k+1}$  conduit à

$$Ap_{k+1} \cdot p_j = p_{k+1} \cdot Ap_j = r_{k+1} \cdot Ap_j + \beta_k p_k \cdot Ap_j,$$

et comme  $Ap_j = (r_j - r_{j+1})/\alpha_j$  on en déduit

$$Ap_{k+1} \cdot p_j = \alpha_j^{-1} r_{k+1} \cdot (r_j - r_{j+1}) + \beta_k p_k \cdot Ap_j.$$

Pour  $j \leq k-1$ , l'hypothèse de récurrence et l'orthogonalité de  $r_{k+1}$  (que l'on vient d'obtenir) prouve que  $Ap_{k+1} \cdot p_j = 0$ . Pour  $j = k$ , on obtient  $Ap_{k+1} \cdot p_k = 0$  grâce aux formules donnant

$\alpha_k$  et  $\beta_k$ . Ceci termine cette récurrence. Comme la famille  $(r_k)_{0 \leq k \leq k_0}$  est orthogonale, elle est libre tant que  $r_k \neq 0$ . Or  $r_k \in K_k$ , ce qui entraîne  $K_k = [r_0, \dots, r_k]$  car ces deux espaces ont même dimension. Par conséquent,  $r_{k+1}$  est bien orthogonal à  $K_k$ , et la suite  $x_k$  est bien celle du gradient conjugué.  $\square$

**Remarque 13.1.40** On dit que la suite  $(p_k)$  est **conjuguée** par rapport à  $A$  car elle est orthogonale pour le produit scalaire  $\langle x, y \rangle_A = Ax \cdot y$ . C'est cette propriété qui a donné son nom à la méthode.

Le lecteur se demande peut-être comment les formules (13.11) ont pu être "inventées". En fait, il existe une réciproque à la Proposition 13.1.39. Plus précisément, l'Exercice 13.1.8 montre que la suite  $d_k = x_{k+1} - x_k$  est conjuguée par rapport à  $A$  et que le sous-espace engendré par  $(d_0, \dots, d_k)$  coïncide avec  $K_k$ . On en déduit donc un moyen pratique de construction de la suite  $d_k$  : on applique le procédé d'orthonormalisation de Gram-Schmidt à  $(r_0, \dots, A^k r_0)$  pour le produit scalaire  $\langle x, y \rangle_A$ . Le résultat se trouve être la suite  $p_k$  qui est nécessairement colinéaire à  $d_k$  (on trouve que  $d_k = \alpha_k p_k$ ). Grâce à la symétrie de  $A$ , les formules de Gram-Schmidt qui définissent  $p_k$  se simplifient considérablement, et on aboutit (après quelques calculs) aux formules (13.11). •

**Algorithme numérique** Dans la pratique, lorsqu'on met en oeuvre l'algorithme du gradient conjugué, les formules (13.11) de la Proposition 13.1.39 sont programmées de la manière suivante

$$\begin{array}{ll} \text{initialisation} & \left\{ \begin{array}{l} \text{choix initial } x_0 \\ r_0 = p_0 = b - Ax_0 \end{array} \right. \\ \text{itérations } k \geq 1 & \left\{ \begin{array}{l} \alpha_{k-1} = \frac{\|r_{k-1}\|^2}{Ap_{k-1} \cdot p_{k-1}} \\ x_k = x_{k-1} + \alpha_{k-1} p_{k-1} \\ r_k = r_{k-1} - \alpha_{k-1} Ap_{k-1} \\ \beta_{k-1} = \frac{\|r_k\|^2}{\|r_{k-1}\|^2} \\ p_k = r_k + \beta_{k-1} p_{k-1} \end{array} \right. \end{array}$$

Dès que  $r_k = 0$ , l'algorithme a convergé, c'est-à-dire que  $x_k$  est la solution du système  $Ax = b$ . On sait que la convergence est atteinte en  $k_0 + 1$  itérations, où  $k_0 \leq n - 1$  est la dimension critique des espaces de Krylov (que l'on ne connaît pas a priori). Cependant dans la pratique, les calculs sur ordinateurs sont toujours sujets à des erreurs d'arrondi, et on ne trouve pas exactement  $r_{k_0+1} = 0$ . C'est pourquoi, on introduit un "petit" paramètre  $\epsilon$  (typiquement  $10^{-4}$  ou  $10^{-8}$  selon la précision désirée), et on décide que l'algorithme a convergé dès que

$$\frac{\|r_k\|}{\|r_0\|} \leq \epsilon.$$

Par ailleurs, pour des systèmes de grande taille (pour lesquels  $n$  et  $k_0$  sont "grands", de l'ordre de  $10^4$  à  $10^6$ ), la méthode du gradient conjugué est utilisée comme une méthode itérative, c'est-à-dire qu'elle converge, au sens du critère ci-dessus, en un nombre d'itérations bien inférieur à  $k_0 + 1$  (cf. Proposition 13.1.42 ci-dessous).

**Remarque 13.1.41**

1. En général, si on n'a pas d'indications sur la solution, on choisit d'initialiser la méthode du gradient conjugué par  $x_0 = 0$ . Si on résout une suite de problèmes peu différents les uns des autres, on peut initialiser  $x_0$  par la solution précédente.
2. 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  $r_k = b - Ax_k$ .
3. 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  $y$ .
4. 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.

**Exercice 13.1.9** Si on considère la méthode du gradient conjugué comme une méthode directe, montrer que dans le cas le plus défavorable,  $k_0 = n - 1$ , le nombre d'opérations (multiplications seulement) pour résoudre un système linéaire est  $N_{op} = n^3 (1 + o(1))$ .

Nous admettons le résultat de convergence suivant (voir [10]).

**Proposition 13.1.42** Soit  $A$  une matrice symétrique réelle définie positive. Soit  $x$  la solution exacte du système  $Ax = b$ . Soit  $x_k$  la suite de solutions approchées du gradient conjugué. Alors

$$\|x_k - x\|_2 \leq 2\sqrt{\text{cond}_2(A)} \left( \frac{\sqrt{\text{cond}_2(A)} - 1}{\sqrt{\text{cond}_2(A)} + 1} \right)^k \|x_0 - x\|_2.$$

On rappelle que dans le cas d'une matrice symétrique réelle définie positive, le conditionnement est donné par la formule  $\text{cond}_2(A) = \lambda_n / \lambda_1$ , où  $\lambda_1, \lambda_n$  sont respectivement la plus petite et la plus grande valeur propre de  $A$ . La Proposition 13.1.42 améliore le Lemme 13.1.33 ce qui veut dire que la méthode du gradient conjugué converge beaucoup plus vite que la méthode du gradient.

On déduit de ce résultat trois conséquences importantes. Tout d'abord, la méthode du gradient conjugué fonctionne bien comme une méthode itérative. En effet, même si on ne fait pas les  $n$  itérations requises pour converger, on diminue l'erreur commise entre  $x$  et  $x_k$  d'autant plus qu'on itère. D'autre part, la vitesse de convergence dépend de la racine carrée du conditionnement de  $A$ , et non pas du conditionnement lui-même comme pour la méthode du gradient simple. La méthode du gradient conjugué converge donc beaucoup plus vite que celle du gradient simple (on dit que la convergence est quadratique au lieu d'être linéaire). Enfin, la convergence sera d'autant plus rapide que  $\text{cond}_2(A)$  est proche de 1, c'est-à-dire que  $A$  est bien conditionnée.

### Préconditionnement

Comme la vitesse de convergence de la méthode du gradient conjugué dépend du conditionnement de la matrice  $A$ , l'idée du preconditionnement 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 13.1.43** Soit à résoudre le système linéaire  $Ax = b$ . On appelle preconditionnement de  $A$ , une matrice  $C$  (facile à inverser) telle que  $\text{cond}_2(C^{-1}A)$  soit plus petit que  $\text{cond}_2(A)$ . On appelle système preconditionné le système équivalent  $C^{-1}Ax = C^{-1}b$ .

En général, la matrice  $C^{-1}A$  n'est plus symétrique, ce qui pose problème pour appliquer la méthode du gradient conjugué. C'est pourquoi on applique souvent un preconditionnement "symétrique" que nous décrivons maintenant. Supposons (pour simplifier) que  $C$  une matrice symétrique définie positive qui admet une décomposition de Cholesky  $C = BB^*$ . On remplace le système original  $Ax = b$  par le système équivalent

$$\tilde{A}\tilde{x} = \tilde{b}, \text{ où } \tilde{A} = B^{-1}AB^{-*}, \tilde{b} = B^{-1}b \text{ et } \tilde{x} = B^*x. \quad (13.12)$$

La matrice  $\tilde{A}$  étant symétrique définie positive, on peut utiliser l'algorithme du gradient conjugué pour résoudre ce système. Néanmoins, on ne connaît pas toujours la factorisation de Cholesky de  $C$  (ou bien on ne veut pas la faire par souci d'économie). Il existe alors un moyen astucieux de transformer l'algorithme du gradient conjugué pour le système (13.12) en un algorithme où seul  $C$  apparaît (et pas le facteur  $B$ ).

**Exercice 13.1.10** On note avec un tilde  $\tilde{\cdot}$  toutes les quantités associées à l'algorithme du gradient conjugué appliqué au système linéaire (13.12). Soit  $x_k = B^{-*}\tilde{x}_k$ ,  $r_k = B\tilde{r}_k = b - Ax_k$ , et  $p_k = B^{-*}\tilde{p}_k$ . Montrer que l'algorithme du gradient conjugué pour (13.12) peut aussi s'écrire sous la forme

$$\begin{array}{ll} \text{initialisation} & \left\{ \begin{array}{l} \text{choix initial } x_0 \\ r_0 = b - Ax_0 \\ p_0 = z_0 = C^{-1}r_0 \end{array} \right. \\ \\ \text{itérations } k \geq 1 & \left\{ \begin{array}{l} \alpha_{k-1} = \frac{z_{k-1} \cdot r_{k-1}}{Ap_{k-1} \cdot p_{k-1}} \\ x_k = x_{k-1} + \alpha_{k-1}p_{k-1} \\ r_k = r_{k-1} - \alpha_{k-1}Ap_{k-1} \\ z_k = C^{-1}r_k \\ \beta_{k-1} = \frac{z_k \cdot r_k}{z_{k-1} \cdot r_{k-1}} \\ p_k = z_k + \beta_{k-1}p_{k-1} \end{array} \right. \end{array}$$

La technique du preconditionnement 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  $C = \text{diag}(A)$ . Il est malheureusement peu efficace, et on lui préfère souvent le “préconditionnement SSOR” (pour Symmetric SOR). En notant  $D = \text{diag}(A)$  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.

**Exercice 13.1.11** Soit  $A$  la matrice d’ordre  $n$  issue de la discrétisation du Laplacien en dimension  $N = 1$  avec un pas d’espace constant  $h = 1/(n + 1)$

$$A = h^{-1} \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

Montrer que pour la valeur optimale

$$\omega_{opt} = \frac{2}{1 + 2 \sin \frac{\pi}{2n}} \simeq 2 \left( 1 - \frac{\pi}{n} \right)$$

le conditionnement de la matrice  $C_\omega^{-1}A$  est majoré par

$$\text{cond}_2(C_\omega^{-1}A) \leq \frac{1}{2} + \frac{1}{2 \sin \frac{\pi}{2n}},$$

et donc que, pour  $n$  grand, on gagne un ordre en  $n$  dans la vitesse de convergence.

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 13.1.19). 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). Le préconditionnement de Cholesky incomplet est souvent le préconditionnement le plus efficace en pratique.

## 13.2 Calcul de valeurs et vecteurs propres

Dans cette section nous expliquons comment calculer les valeurs propres et les vecteurs propres d'une matrice symétrique réelle. Un exemple typique est la matrice de rigidité issue de l'approximation par éléments finis d'une équation aux dérivées partielles. Dans ce cas, ses valeurs et vecteurs propres sont des approximations des modes propres du modèle physique sous-jacent (voir (7.23)).

Puisque les valeurs propres d'une matrice  $A$  sont les racines de son polynôme caractéristique  $\det(A - \lambda \text{Id})$ , 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  $n$ ,

$$P(\lambda) = (-1)^n (\lambda^n + a_1 \lambda^{n-1} + a_2 \lambda^{n-2} + \cdots + a_{n-1} \lambda + a_n),$$

est le polynôme caractéristique (le développer par rapport à la dernière colonne) de la matrice

$$A = \begin{pmatrix} -a_1 & -a_2 & \cdots & \cdots & -a_n \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & 0 \end{pmatrix}.$$

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 indiquons trois méthodes typiques (il y en a d'autres, éventuellement plus efficaces mais plus compliquées). La méthode de la puissance est la plus simple mais limitée dans son applicabilité. La méthode de Givens-Householder permet de calculer une (ou plusieurs) valeur propre de rang quelconque prédéterminé sans avoir à calculer toutes les valeurs propres. Finalement, la méthode de Lanczos, qui "ressemble" au gradient conjugué, est à la base de nombreux développements récents qui conduisent aux méthodes les plus efficaces pour les grandes matrices creuses.

### 13.2.1 Méthode de la puissance

Une méthode très simple pour calculer la plus grande ou la plus petite valeur propre (en module) d'une matrice (et un vecteur propre associé) est la méthode de



la puissance. 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  $(\lambda_1, \dots, \lambda_n)$  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  $\lambda_n$  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  $\varepsilon$  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 13.2.1** *On suppose que la matrice  $A$  est symétrique réelle, de valeurs propres  $(\lambda_1, \dots, \lambda_n)$ , associées à une base orthonormée de vecteurs propres  $(e_1, \dots, e_n)$ , et que la valeur propre de plus grand module  $\lambda_n$  est simple et positive, c'est-à-dire que  $|\lambda_1|, \dots, |\lambda_{n-1}| < \lambda_n$ . 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*

$$\lim_{k \rightarrow +\infty} \|y_k\| = \lambda_n, \quad \lim_{k \rightarrow +\infty} x_k = x_\infty \text{ avec } x_\infty = \pm e_n.$$

La vitesse de convergence est proportionnelle au rapport  $|\lambda_{n-1}|/|\lambda_n|$

$$\| \|y_k\| - \lambda_n \| \leq C \left| \frac{\lambda_{n-1}}{\lambda_n} \right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left| \frac{\lambda_{n-1}}{\lambda_n} \right|^k.$$

**Remarque 13.2.2** La convergence de la suite de valeurs propres approchées  $\|y_k\|$  est plus rapide que celle des vecteurs propres approchés  $x_k$  (quadratique au lieu de linéaire). La méthode de la puissance fonctionne aussi pour des matrices non symétrique, mais la convergence de  $\|y_k\|$  est seulement linéaire dans ce cas.

**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  $\lambda_n$ . □

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 13.2.3** *On suppose que la matrice  $A$  est symétrique réelle, de valeurs propres  $(\lambda_1, \dots, \lambda_n)$ , associées à une base orthonormée de vecteurs propres  $(e_1, \dots, e_n)$ , et que la valeur propre de plus petit module  $\lambda_1$  est simple et strictement positive, c'est-à-dire que  $0 < \lambda_1 < |\lambda_2|, \dots, |\lambda_n|$ . 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*

$$\lim_{k \rightarrow +\infty} \frac{1}{\|y_k\|} = |\lambda_1|, \quad \lim_{k \rightarrow +\infty} x_k = x_\infty \text{ avec } x_\infty = \pm e_1.$$

La vitesse de convergence est proportionnelle au rapport  $\lambda_1/|\lambda_2|$

$$\left| \frac{1}{\|y_k\|} - \lambda_1 \right| \leq C \left| \frac{\lambda_1}{\lambda_2} \right|^{2k}, \quad \|x_k - x_\infty\| \leq C \left| \frac{\lambda_1}{\lambda_2} \right|^k.$$

La démonstration est similaire à celle de la Proposition 13.2.1 et nous la laissons au lecteur en guise d'exercice.

**Remarque 13.2.4** Les considérations de la Remarque 13.2.2 s'appliquent aussi à la méthode de la puissance inverse. Pour accélérer la convergence, on peut toujours procéder à une translation de la matrice  $A$  qu'on remplace par  $A - \sigma \text{Id}$  avec  $\sigma$  une approximation de  $\lambda_1$ . •

### 13.2.2 Méthode de Givens-Householder

La méthode de Givens-Householder se compose de deux étapes successives : tout d'abord l'algorithme de Householder qui réduit une matrice symétrique  $A$  en une matrice tridiagonale (cette étape s'effectue en un nombre fini d'opérations), ensuite l'algorithme de bisection de Givens qui fournit (de manière itérative) les valeurs propres d'une matrice tridiagonale.

**Lemme 13.2.5 (de Householder)** *Soit  $A$  une matrice symétrique réelle d'ordre  $n$ . Il existe  $(n-2)$  matrices orthogonales  $H_k$  telles que*

$$T = (H_1 H_2 \cdots H_{n-2})^* A (H_1 H_2 \cdots H_{n-2})$$

*soit tridiagonale. Bien sûr,  $A$  et  $T$  ont les mêmes valeurs propres.*

**Démonstration.** A partir de  $A$ , on construit une suite de matrices  $(A_k)_{1 \leq k \leq n-1}$  telle que  $A_1 = A$  et  $A_{k+1} = H_k^* A_k H_k$  avec  $H_k$  une matrice orthogonale choisie de telle manière que  $A_k$  ait la structure par blocs suivante

$$A_k = \begin{pmatrix} T_k & E_k^* \\ E_k & M_k \end{pmatrix}$$

où  $T_k$  est une matrice carrée tridiagonale de taille  $k$ ,  $M_k$  est une matrice carrée de taille  $n-k$ , et  $E_k$  est une matrice rectangulaire à  $(n-k)$  lignes et  $k$  colonnes dont seule la dernière colonne, notée  $a_k \in \mathbb{R}^{n-k}$  est non nulle

$$T_k = \begin{pmatrix} \times & \times & & & \\ & \times & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \times \\ & & & \times & \times \end{pmatrix}, \text{ et } E_k = \begin{pmatrix} 0 & \cdots & 0 & a_{k,1} \\ \vdots & & \vdots & a_{k,2} \\ \vdots & & \vdots & \vdots \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & a_{k,n-k} \end{pmatrix}.$$

Il est clair qu'ainsi  $A_{n-1}$  sera sous forme tridiagonale. On remarque que  $A$  est bien sous cette forme pour  $k=1$ . Soit la matrice  $H_k$  définie par

$$H_k = \begin{pmatrix} \text{Id}_k & 0 \\ 0 & \tilde{H}_k \end{pmatrix},$$

avec  $\text{Id}_k$  la matrice identité d'ordre  $k$  et  $\tilde{H}_k$  la matrice, dite de Householder, d'ordre  $n-k$  définie par

$$\tilde{H}_k = \text{Id}_{n-k} - 2 \frac{v_k (v_k)^*}{\|v_k\|^2}, \quad \text{avec } v_k = a_k + \|a_k\| e_1, \quad (13.13)$$

où  $e_1$  est le premier vecteur de la base canonique de  $\mathbb{R}^{n-k}$ . Remarquons que  $\tilde{H}_k a_k = -\|a_k\| e_1$ , et que  $H_k$  est orthogonale et symétrique. Notons que  $\tilde{H}_k$  n'est bien définie que si  $v_k \neq 0$ , mais si ça n'est pas le cas alors la  $k$ -ème colonne de  $A_k$  est déjà du type désiré, et il suffit de prendre  $H_k = \text{Id}_n$ . Un simple calcul montre que

$$A_{k+1} = H_k^* A_k H_k = \begin{pmatrix} T_k & (\tilde{H}_k E_k)^* \\ \tilde{H}_k E_k & \tilde{H}_k M_k \tilde{H}_k \end{pmatrix} \quad \text{avec} \quad \tilde{H}_k E_k = \begin{pmatrix} 0 & \cdots & 0 & -\|a_k\| \\ \vdots & & \vdots & 0 \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & 0 \end{pmatrix},$$

donc  $A_{k+1}$  se met bien sous la forme désirée.  $\square$

Étudions maintenant l'algorithme de bisection de Givens pour une matrice tridiagonale symétrique réelle

$$A = \begin{pmatrix} b_1 & c_1 & & 0 \\ c_1 & \ddots & \ddots & \\ & \ddots & \ddots & c_{n-1} \\ 0 & & c_{n-1} & b_n \end{pmatrix},$$

où on suppose, sans perte de généralité, que  $c_i \neq 0$  pour  $1 \leq i \leq n-1$ . En effet, s'il existe un indice  $i$  tel que  $c_i = 0$ , alors on voit facilement que

$$\det(A - \lambda \text{Id}) = \det(A_i - \lambda \text{Id}) \det(A_{n-i} - \lambda \text{Id})$$

où  $A_i$  et  $A_{n-i}$  sont deux matrices du même type que  $A$  mais d'ordre  $i$  et  $n-i$  respectivement. Donnons tout d'abord deux lemmes techniques.

**Lemme 13.2.6** *Pour  $1 \leq i \leq n$ , on définit une matrice  $A_i$  de taille  $i$  par*

$$A_i = \begin{pmatrix} b_1 & c_1 & & 0 \\ c_1 & \ddots & \ddots & \\ & \ddots & \ddots & c_{i-1} \\ 0 & & c_{i-1} & b_i \end{pmatrix}.$$

Soit  $p_i(\lambda) = \det(A_i - \lambda I)$  son polynôme caractéristique. La suite  $p_i$  vérifie la formule de récurrence

$$p_i(\lambda) = (b_i - \lambda)p_{i-1}(\lambda) - c_{i-1}^2 p_{i-2}(\lambda) \quad \forall i \geq 2,$$

avec  $p_1(\lambda) = b_1 - \lambda$  et  $p_0(\lambda) = 1$ . De plus, pour tout  $i \geq 1$ , le polynôme  $p_i$  possède les propriétés suivantes

1.  $\lim_{\lambda \rightarrow -\infty} p_i(\lambda) = +\infty$ ,
2. si  $p_i(\lambda_0) = 0$ , alors  $p_{i-1}(\lambda_0)p_{i+1}(\lambda_0) < 0$ ,
3.  $p_i$  admet  $i$  racines réelles distinctes qui séparent strictement les  $(i+1)$  racines de  $p_{i+1}$ .

**Démonstration.** En développant  $\det(A_i - \lambda \text{Id})$  par rapport à la dernière ligne, on obtient la formule de récurrence désirée. La première propriété est évidente par définition du polynôme caractéristique. Pour prouver la deuxième, on remarque dans la formule de récurrence que, si  $p_i(\lambda_0) = 0$ , alors

$$p_{i+1}(\lambda_0) = -c_i^2 p_{i-1}(\lambda_0).$$

Comme  $c_i \neq 0$ , on a  $p_{i-1}(\lambda_0)p_{i+1}(\lambda_0) \leq 0$ . Cette inégalité est en fait stricte, car si  $p_{i-1}(\lambda_0) = p_{i+1}(\lambda_0) = 0$  la relation de récurrence implique que  $p_k(\lambda_0) = 0$  pour tout  $0 \leq k \leq i+1$ , ce qui n'est pas possible puisque  $p_0(\lambda_0) = 1$ .

Pour démontrer la troisième propriété, on commence par remarquer que  $p_i(\lambda)$  admet  $i$  racines réelles, notées  $\lambda_1^i \leq \dots \leq \lambda_i^i$  puisque  $A_i$  est symétrique réelle. Montrons par récurrence que ces  $i$  racines de  $p_i$  sont distinctes et sont séparées par celles de  $p_{i-1}$ . Tout d'abord, cette propriété est vraie pour  $i = 2$ , car

$$p_2(\lambda) = (b_2 - \lambda)(b_1 - \lambda) - c_1^2$$

a deux racines ( $\lambda_1^2, \lambda_2^2$ ) qui encadrent la seule racine  $\lambda_1^1 = b_1$  de  $p_1(\lambda)$ , i.e.,  $\lambda_1^2 < \lambda_1^1 < \lambda_2^2$ . On suppose que  $p_i(\lambda)$  possède  $i$  racines réelles distinctes séparées par celles de  $p_{i-1}$ . Montrons que  $p_{i+1}$  a  $i + 1$  racines réelles distinctes séparées par celles de  $p_i$ . On définit un polynôme  $q_i$  de degré  $2i$  par

$$q_i(\lambda) = p_{i-1}(\lambda)p_{i+1}(\lambda).$$

On connaît déjà  $i - 1$  racines de  $q_i$  (celles de  $p_{i-1}$ ) et il y a  $i$  valeurs de  $\lambda$  (les racines de  $p_i$ ) telles que  $q_i(\lambda) < 0$ , c'est-à-dire que

$$q_i(\lambda_k^{i-1}) = 0 \quad 1 \leq k \leq i-1, \quad q_i(\lambda_k^i) < 0 \quad 1 \leq k \leq i,$$

avec

$$\lambda_1^i < \lambda_1^{i-1} < \lambda_2^i < \dots < \lambda_{i-1}^{i-1} < \lambda_i^i.$$

Entre  $\lambda_k^i$  et  $\lambda_{k+1}^i$ , soit  $q_i$  s'annule en un autre point  $\gamma_k \neq \lambda_k^{i-1}$  et on a donc trouvé une racine supplémentaire de  $q_i$  donc de  $p_{i+1}$ , soit  $q_i$  ne s'annule qu'en  $\lambda_k^{i-1}$ , mais dans ce cas c'est au moins une racine double car sa dérivée  $q_i'$  doit aussi s'annuler en  $\lambda_k^{i-1}$ . Or  $\lambda_k^{i-1}$  est racine simple de  $p_{i-1}$ , donc  $\lambda_k^{i-1}$  est aussi racine de  $p_{i+1}$ . Mais, à cause de la relation de récurrence, cela prouverait que  $\lambda_k^{i-1}$  est une racine pour tous les polynômes  $p_j$  avec  $0 \leq j \leq i+1$ , ce qui n'est pas possible car  $p_0 = 1$ . Par conséquent, on vient de montrer qu'entre chaque couple  $\lambda_k^i, \lambda_{k+1}^i$  il existe une autre racine  $\gamma_k \neq \lambda_k^{i-1}$  du polynôme  $q_i$  donc de  $p_{i+1}$ . Au total, on vient de trouver  $(i-1)$  racines distinctes de  $p_{i+1}$  qui encadrent celles de  $p_i$ . Par ailleurs,  $q_i(\lambda_1^i) < 0$  et  $q_i(\lambda_i^i) < 0$ , tandis que

$$\lim_{\lambda \rightarrow \pm\infty} q_i(\lambda) = +\infty.$$

On en déduit l'existence de deux racines supplémentaires distinctes de  $q_i$ , donc de  $p_{i+1}$ , qui encadrent celles de  $p_i$ .  $\square$

**Lemme 13.2.7** Pour tout  $\mu \in \mathbb{R}$ , on définit

$$\text{sgnp}_i(\mu) = \begin{cases} \text{signe de } p_i(\mu) & \text{si } p_i(\mu) \neq 0, \\ \text{signe de } p_{i-1}(\mu) & \text{si } p_i(\mu) = 0. \end{cases}$$

Soit  $N(i, \mu)$  le nombre de changements de signe entre éléments consécutifs de la famille ordonnée  $E(i, \mu) = \{+1, \text{sgnp}_1(\mu), \text{sgnp}_2(\mu), \dots, \text{sgnp}_i(\mu)\}$ . Alors,  $N(i, \mu)$  est le nombre de racines de  $p_i$  qui sont strictement inférieures à  $\mu$ .

**Démonstration.** On remarque d'abord que  $\text{sgnp}_i(\mu)$  est défini sans ambiguïté puisque, si  $p_i(\mu) = 0$ , alors  $p_{i-1}(\mu) \neq 0$  à cause du point 2 du Lemme 13.2.6. On procède par récurrence sur  $i$ . Pour  $i = 1$ , on vérifie le résultat

$$\begin{aligned} \mu \leq b_1 &\Rightarrow E(1, \mu) = \{+1, +1\} \Rightarrow N(1, \mu) = 0, \\ \mu > b_1 &\Rightarrow E(1, \mu) = \{+1, -1\} \Rightarrow N(1, \mu) = 1. \end{aligned}$$

On suppose le résultat vrai jusqu'à l'ordre  $i$ . Soit  $(\lambda_k^i)_{1 \leq k \leq i}$  les racines de  $p_i$  et  $(\lambda_k^{i+1})_{1 \leq k \leq i+1}$  celles de  $p_{i+1}$ , rangées par ordre croissant. On a

$$\lambda_1^i < \dots < \lambda_{N(i,\mu)}^i < \mu \leq \lambda_{N(i,\mu)+1}^i < \dots < \lambda_i^i,$$

et

$$\lambda_{N(i,\mu)}^i < \lambda_{N(i,\mu)+1}^{i+1} < \lambda_{N(i,\mu)+1}^i,$$

en vertu du point 3 du Lemme 13.2.6. Il y a trois cas possibles.

1. Si  $\lambda_{N(i,\mu)}^i < \mu \leq \lambda_{N(i,\mu)+1}^{i+1}$ , on a  $\operatorname{sgn} p_{i+1}(\mu) = \operatorname{sgn} p_i(\mu)$ , donc  $N(i+1, \mu) = N(i, \mu)$ .
2. Si  $\lambda_{N(i,\mu)+1}^{i+1} < \mu < \lambda_{N(i,\mu)+1}^i$ , on a  $\operatorname{sgn} p_{i+1}(\mu) = -\operatorname{sgn} p_i(\mu)$ , donc  $N(i+1, \mu) = N(i, \mu) + 1$ .
3. Si  $\mu = \lambda_{N(i,\mu)+1}^i$ , on a  $\operatorname{sgn} p_i(\mu) = \operatorname{sgn} p_{i-1}(\mu) = -\operatorname{sgn} p_{i+1}(\mu)$ , donc  $N(i+1, \mu) = N(i, \mu) + 1$ , à cause du deuxième point du Lemme 13.2.6.

Dans tous les cas  $N(i+1, \mu)$  est bien le nombre de racines de  $p_{i+1}$  inférieures strictement à  $\mu$ .  $\square$

**Algorithme pratique de Givens.** On note  $\lambda_1 \leq \dots \leq \lambda_n$  les valeurs propres de  $A$  rangées par ordre croissant. Pour calculer numériquement la  $i$ -ème valeur propre  $\lambda_i$ , on prend un intervalle  $[a_0, b_0]$  dans lequel on est sûr que  $\lambda_i$  se trouve (par exemple  $-a_0 = b_0 = \|A\|_2$ ). On calcule alors le nombre  $N(n, \frac{a_0+b_0}{2})$  défini dans le Lemme 13.2.7 (les valeurs de la suite  $p_j(\frac{a_0+b_0}{2})$ , pour  $1 \leq j \leq n$ , sont calculées par la formule de récurrence du Lemme 13.2.6). S'il se trouve que  $N(n, \frac{a_0+b_0}{2}) \geq i$ , alors on en déduit que  $\lambda_i$  appartient à l'intervalle  $[a_0, \frac{a_0+b_0}{2}]$ . Si au contraire on trouve que  $N(n, \frac{a_0+b_0}{2}) < i$ , alors  $\lambda_i$  appartient à l'autre intervalle  $[\frac{a_0+b_0}{2}, b_0]$ . Dans tous les cas on a divisé par deux l'intervalle initial qui contient  $\lambda_i$ . Par dichotomie, c'est-à-dire en répétant cette procédure de division de l'intervalle contenant  $\lambda_i$ , on approche la valeur exacte de  $\lambda_i$  avec la précision désirée.

### 13.2.3 Méthode de Lanczos

La méthode de Lanczos permet de calculer les valeurs propres d'une matrice symétrique réelle en utilisant la notion d'espace de Krylov, déjà introduite à propos de l'algorithme du gradient conjugué. Cette méthode (et ses nombreuses généralisations) est très efficace pour les matrices de grande taille. Il s'agit plus ici de donner le principe de cette méthode que les détails de son implémentation numérique.

Dans ce qui suit, on note  $A$  une matrice symétrique réelle d'ordre  $n$ ,  $r_0 \neq 0 \in \mathbb{R}^n$  un vecteur donné non nul, et  $K_k$  l'espace de Krylov associé, engendré par les vecteurs  $\{r_0, Ar_0, \dots, A^k r_0\}$ . Rappelons qu'il existe un entier  $k_0 \leq n-1$ , appelé dimension critique de Krylov, tel que, si  $k \leq k_0$ , la famille  $(r_0, Ar_0, \dots, A^k r_0)$  est libre et  $\dim K_k = k+1$ , tandis que si  $k > k_0$  on a  $K_k = K_{k_0}$ .

L'algorithme de Lanczos consiste à construire une suite de vecteurs  $(v_j)_{1 \leq j \leq k_0+1}$  par la formule de récurrence suivante, pour  $2 \leq j \leq k_0+1$ ,

$$\hat{v}_j = Av_{j-1} - (Av_{j-1}, v_{j-1})v_{j-1} - \|\hat{v}_{j-1}\|v_{j-2}, \quad v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}, \quad (13.14)$$

avec  $v_0 = 0$  et  $v_1 = r_0/\|r_0\|$ . Pour tout entier  $k \leq k_0 + 1$ , on définit une matrice  $V_k$  de taille  $n \times k$  dont les colonnes sont les vecteurs  $(v_1, \dots, v_k)$ , ainsi qu'une matrice symétrique **tridiagonale**  $T_k$  de taille  $k \times k$  dont les éléments sont

$$(T_k)_{i,i} = Av_i \cdot v_i, \quad (T_k)_{i,i+1} = (T_k)_{i+1,i} = \|\hat{v}_{i+1}\|, \quad (T_k)_{i,j} = 0 \text{ si } |i - j| \geq 2.$$

Avec ces notations, la récurrence de Lanczos vérifie des propriétés remarquables.

**Lemme 13.2.8** *La suite  $(v_j)_{1 \leq j \leq k_0+1}$  est bien définie par (13.14) car  $\|\hat{v}_j\| \neq 0$  pour tout  $1 \leq j \leq k_0+1$ , tandis que  $\hat{v}_{k_0+2} = 0$ . Pour  $1 \leq k \leq k_0+1$ , la famille  $(v_1, \dots, v_{k+1})$  coïncide avec la base orthonormée de  $K_k$  construite par le procédé de Gram-Schmidt appliqué à la famille  $(r_0, Ar_0, \dots, A^k r_0)$ . De plus, pour  $1 \leq k \leq k_0 + 1$ , on a*

$$AV_k = V_k T_k + \hat{v}_{k+1} e_k^*, \quad (13.15)$$

où  $e_k$  est le  $k$ -ème vecteur de la base canonique de  $\mathbb{R}^k$ ,

$$V_k^* AV_k = T_k \text{ et } V_k^* V_k = \text{Id}_k, \quad (13.16)$$

où  $\text{Id}_k$  est la matrice identité de taille  $k \times k$ .

**Démonstration.** Oublions pour l'instant la définition (13.14) de la suite  $(v_j)_{1 \leq j \leq k_0+1}$  et remplaçons la par la nouvelle définition (dont on montrera qu'elle est équivalente à (13.14))

$$\hat{v}_j = Av_{j-1} - \sum_{i=1}^{j-1} (Av_{j-1} \cdot v_i) v_i, \quad v_j = \frac{\hat{v}_j}{\|\hat{v}_j\|}, \quad j \geq 2, \quad (13.17)$$

avec  $v_1 = r_0/\|r_0\|$ . Bien sûr, (13.17) n'a de sens que si  $\|\hat{v}_j\| \neq 0$ . Si  $\|\hat{v}_j\| = 0$ , on dit que l'algorithme stoppe à l'indice  $j$ . Par définition,  $v_j$  est orthogonal à  $v_i$  pour  $1 \leq i \leq j-1$ . Par récurrence, on vérifie facilement que  $v_j \in K_{j-1}$ . Comme la suite des espaces de Krylov  $K_j$  est strictement croissante pour  $j \leq k_0+1$ , on en déduit que, tant que l'algorithme n'a pas stoppé, les vecteurs  $(v_1, \dots, v_j)$  forment une base orthonormale de  $K_{j-1}$ . Par conséquent,  $v_j$  étant orthogonal à  $(v_1, \dots, v_{j-1})$  est aussi orthogonal à  $K_{j-2}$ . En particulier, la famille  $(v_1, \dots, v_j)$ , définie par (13.17), coïncide avec la base orthonormée de  $K_{j-1}$  construite par le procédé de Gram-Schmidt appliqué à la famille  $(r_0, Ar_0, \dots, A^{j-1} r_0)$ . Ceci prouve que l'algorithme stoppe précisément à la dimension critique de Krylov  $k_0$ , c'est-à-dire que  $\|\hat{v}_j\| \neq 0$  tant que  $j \leq k_0 + 1$  et  $\hat{v}_{k_0+2} = 0$ .

Montrons maintenant que les définitions (13.14) et (13.17) de la suite  $(v_j)$  sont identiques. Comme  $A$  est symétrique, on a

$$Av_{j-1} \cdot v_i = v_{j-1} \cdot Av_i = v_{j-1} \cdot \hat{v}_{i+1} + \sum_{k=1}^i (Av_i \cdot v_k)(v_{j-1} \cdot v_k).$$

Grâce aux propriétés d'orthonormalité des  $(v_k)$ , on en déduit que  $Av_{j-1} \cdot v_i = 0$  si  $1 \leq i \leq j-3$ , et que  $Av_{j-1} \cdot v_{j-2} = \|\hat{v}_{j-1}\|$ . Donc les définitions (13.14) et (13.17) coïncident.

Finalement, la relation (13.15), prise colonne par colonne, n'est rien d'autre qu'une réécriture de (13.14) en éliminant  $\hat{v}_j$ . La propriété  $V_k^* V_k = \text{Id}_k$  découle du caractère orthonormal de la famille  $(v_1, \dots, v_k)$ , tandis que la relation  $V_k^* AV_k = T_k$  s'obtient simplement en multipliant (13.15) à gauche par  $V_k^*$  car  $V_k^* \hat{v}_{k+1} = 0$ .  $\square$

**Remarque 13.2.9** L'algorithme de Lanczos apparaît comme une méthode de réduction sous forme tridiagonale comme l'algorithme de Householder vu précédemment. Néanmoins, l'algorithme de Lanczos n'est pas utilisé en pratique comme une méthode de tridiagonalisation car, pour  $n$  grand, les erreurs d'arrondi détruisent en partie l'orthogonalité des derniers vecteurs  $v_j$  par rapport aux premiers. •

Nous allons maintenant comparer les valeurs propres et vecteurs propres des matrices  $A$  et  $T_{k_0+1}$  (qui ne sont pas de même taille en règle générale). On note  $\lambda_1 < \lambda_2 < \dots < \lambda_m$  les valeurs propres distinctes de  $A$  (avec  $1 \leq m \leq n$ ), et  $P_1, \dots, P_m$  les matrices de projections orthogonales sur les sous-espaces propres correspondants de  $A$ . On rappelle que

$$A = \sum_{i=1}^m \lambda_i P_i, \quad \text{Id} = \sum_{i=1}^m P_i, \quad \text{et } P_i P_j = 0 \text{ si } i \neq j. \quad (13.18)$$

**Lemme 13.2.10** *Les valeurs propres de  $T_{k_0+1}$  sont simples et sont aussi valeurs propres de  $A$ . Réciproquement, si on suppose que  $r_0$  vérifie  $P_i r_0 \neq 0$  pour tout  $1 \leq i \leq m$ , alors toutes les valeurs propres de  $A$  sont aussi valeurs propres de  $T_{k_0+1}$  et  $k_0 + 1 = m$ .*

**Remarque 13.2.11** Dans le cas où  $P_i r_0 \neq 0$  pour tout  $i$ , les matrices  $A$  et  $T_{k_0+1}$  ont exactement les mêmes valeurs propres, mais avec une multiplicité éventuellement différente puisque les valeurs propres de  $T_{k_0+1}$  sont simples. On verra dans la démonstration du Lemme 13.2.10 qu'il existe aussi un lien entre les vecteurs propres de  $A$  et  $T_{k_0+1}$ . La condition demandée sur  $r_0$  pour la réciproque de ce lemme est bien nécessaire. En effet, si  $r_0$  est un vecteur propre de  $A$ , alors  $k_0 = 0$  et la matrice  $T_{k_0+1}$  admet pour unique valeur propre celle qui est associée à  $r_0$ . •

**Démonstration.** Soit  $\lambda$  et  $y \in \mathbb{R}^{k_0+1}$  une valeur propre et un vecteur propre tels que  $T_{k_0+1} y = \lambda y$ . Comme  $\hat{v}_{k_0+2} = 0$ , la relation (13.15) devient, pour  $k = k_0 + 1$ ,  $AV_{k_0+1} = V_{k_0+1}T_{k_0+1}$ , et donc, par application du vecteur  $y$ , on obtient  $A(V_{k_0+1}y) = \lambda(V_{k_0+1}y)$ . Le vecteur  $V_{k_0+1}y$  n'est pas nul car  $y \neq 0$  et les colonnes de  $V_{k_0+1}$  sont libres. Par conséquent,  $V_{k_0+1}y$  est bien un vecteur propre de  $A$  associé à la valeur propre  $\lambda$  qui est donc forcément égale à un des  $\lambda_i$ .

Réciproquement, on introduit le sous-espace vectoriel  $E_m$  de  $\mathbb{R}^n$  engendré par les vecteurs  $(P_1 r_0, \dots, P_m r_0)$ . Si  $P_i r_0 \neq 0$  pour tout  $1 \leq i \leq m$ , ces vecteurs sont libres car les projections  $P_i$  sont deux à deux orthogonales. Par conséquent, la dimension de  $E_m$  est exactement  $m$ . Nous allons montrer que dans ce cas on a  $m = k_0 + 1$ . Par (13.18) on a  $A^k r_0 = \sum_{i=1}^m \lambda_i^k P_i r_0$ , c'est-à-dire que  $A^k r_0 \in E_m$ , donc les espaces de Krylov vérifient  $K_k \subset E_m$  pour tout  $k \geq 0$ . En particulier, ceci implique que  $\dim K_{k_0} = k_0 + 1 \leq m$ . Par ailleurs, dans la base  $(P_1 r_0, \dots, P_m r_0)$  de  $E_m$ , les coordonnées du vecteur  $A^k r_0$  sont  $(\lambda_1^k, \dots, \lambda_m^k)$ . Autrement dit, la famille  $(r_0, Ar_0, \dots, A^{m-1} r_0)$  de  $E_m$  est représentée dans la base  $(P_1 r_0, \dots, P_m r_0)$  par la matrice  $M$  définie par

$$M = \begin{pmatrix} 1 & \lambda_1 & \lambda_1^2 & \dots & \lambda_1^{m-1} \\ \vdots & & & & \vdots \\ 1 & \lambda_m & \lambda_m^2 & \dots & \lambda_m^{m-1} \end{pmatrix}.$$



La matrice  $M$  est une matrice de Van Der Monde d'ordre  $m$  qui est inversible car tous les  $\lambda_i$  sont distincts. Donc la famille  $(r_0, Ar_0, \dots, A^{m-1}r_0)$  est libre puisque  $(P_1 r_0, \dots, P_m r_0)$  l'est. Ceci implique que  $\dim K_{m-1} = m$ , donc  $m-1 \leq k_0$ . D'où l'on déduit finalement que  $m = k_0 + 1$  et  $E_m = K_{k_0}$ .

Grâce à la formule (13.18) on a  $A(P_i r_0) = \lambda_i(P_i r_0)$ . Comme  $P_i r_0$  est non nul, il s'agit bien d'un vecteur propre de  $A$  associé à la valeur propre  $\lambda_i$ . Comme  $E_m = K_{k_0}$  et que les colonnes de  $V_{k_0+1}$  forment une base de  $K_{k_0}$ , on en déduit qu'il existe un vecteur non nul  $y_i \in \mathbb{R}^m$  tel que  $P_i r_0 = V_{k_0+1} y_i$ . On multiplie la première égalité de (13.16) par  $y_i$  pour obtenir

$$T_{k_0+1} y_i = V_{k_0+1}^* A V_{k_0+1} y_i = V_{k_0+1}^* A P_i r_0 = \lambda_i V_{k_0+1}^* P_i r_0 = \lambda_i V_{k_0+1}^* V_{k_0+1} y_i = \lambda_i y_i,$$

autrement dit,  $y_i$  est vecteur propre de  $T_{k_0+1}$  pour la valeur propre  $\lambda_i$ . Ce qui termine la démonstration.  $\square$

Le résultat du Lemme 13.2.10 pourrait laisser croire qu'il faut appliquer la récurrence de Lanczos jusqu'à l'itération maximale  $k_0 + 1$ , puis calculer les valeurs propres de  $T_{k_0+1}$  afin d'en déduire les valeurs propres et les vecteurs propres de  $A$ . Cela rendrait la méthode de Lanczos comparable à celle de Givens-Householder (en général  $k_0$  est de l'ordre de  $n$ , ce qui rend le compte d'opérations similaire dans les deux cas). De plus, appliquée ainsi la méthode de Lanczos serait instable numériquement à cause de la perte d'orthogonalité des vecteurs  $v_j$  causée par les inévitables erreurs d'arrondi (voir la Remarque 13.2.9).

Fort heureusement, le résultat suivant indique qu'il n'est pas nécessaire de faire beaucoup d'itérations dans la récurrence de Lanczos pour obtenir des valeurs propres de  $T_k$  qui soient de bonnes approximations de celles de  $A$  (avec  $k$  beaucoup plus petit que  $k_0$  ou  $n$ ).

**Proposition 13.2.12** *Soit un entier  $1 \leq k \leq k_0 + 1$ . Soit  $\lambda$  une valeur propre de  $T_k$  et  $y \in \mathbb{R}^k$  un vecteur propre non nul associé. Il existe une valeur propre  $\lambda_i$  de  $A$  telle que*

$$|\lambda - \lambda_i| \leq \|\hat{v}_{k+1}\| \frac{|e_k \cdot y|}{\|y\|} \leq \|\hat{v}_{k+1}\|,$$

où  $e_k$  est le  $k$ -ème vecteur de la base canonique de  $\mathbb{R}^k$ .

**Remarque 13.2.13** La première conclusion de la Proposition 13.2.12 est que, si  $\|\hat{v}_{k+1}\|$  est petit, alors les valeurs propres de  $T_k$  sont de bonnes approximations de certaines valeurs propres de  $A$ . La deuxième conclusion est la plus importante en pratique : si la dernière composante d'un vecteur propre de  $T_k$  est petite, alors la valeur propre correspondante est une bonne approximation d'une valeur propre de  $A$ .

•

**Démonstration.** Soit un vecteur propre non nul  $y \in \mathbb{R}^k$  tel que  $T_k y = \lambda y$ . Multipliant (13.15) par  $y$  on obtient

$$A V_k y = V_k T_k y + (e_k \cdot y) \hat{v}_{k+1},$$

d'où l'on déduit

$$A(V_k y) - \lambda(V_k y) = (e_k \cdot y) \hat{v}_{k+1}. \quad (13.19)$$

On décompose alors  $V_k y$  sur la base des vecteurs propres de  $A$

$$V_k y = \sum_{i=1}^m P_i(V_k y).$$

On prend le produit scalaire de (13.19) avec  $V_k y$ , et à l'aide des relations (13.18) on a

$$\sum_{i=1}^m (\lambda_i - \lambda) |P_i(V_k y)|^2 = (e_k \cdot y) (\hat{v}_{k+1} \cdot V_k y). \quad (13.20)$$

En minorant le terme de droite et en majorant celui de gauche par l'inégalité de Cauchy-Schwarz, il vient

$$\min_{1 \leq i \leq m} |\lambda_i - \lambda| \|V_k y\|^2 \leq \|y\| \|\hat{v}_{k+1}\| \|V_k y\|.$$

Comme les colonnes de  $V_k$  sont orthonormées, on a  $\|V_k y\| = \|y\|$ , et par simplification on obtient

$$\min_{1 \leq i \leq m} |\lambda_i - \lambda| \leq \|\hat{v}_{k+1}\|.$$

Cette inégalité peut être améliorée si on n'applique pas Cauchy-Schwarz au terme  $\langle e_k, y \rangle$  dans (13.20). Dans ce cas on trouve

$$\min_{1 \leq i \leq m} |\lambda_i - \lambda| \leq \|\hat{v}_{k+1}\| \frac{|e_k \cdot y|}{\|y\|},$$

ce qui termine la démonstration. □

# Bibliographie

- [1] AHUJA R.K., MAGNANTI T.L, ORLIN J.B, *Network flows*, Prentice Hall, Upper Saddle River, New Jersey (1993).
- [2] ALLAIRE G., KABER S. M., *Algèbre linéaire numérique. Cours et exercices*, Éditions Ellipses, Paris (2002).
- [3] BENAÏM M., EL KAROUI N., *Promenade aléatoire*, Éditions de l'École Polytechnique, Palaiseau (2005).
- [4] BERNARDI Ch., MADAY Y., RAPETTI F., *Discretisation variationnelle de problèmes aux limites elliptiques*, Mathématiques et Applications 45, Springer, Paris (2004).
- [5] BONNANS J., *Optimisation continue*, Mathématiques appliquées pour le Master / SMAI, Dunod, Paris (2006).
- [6] BONNANS J., GILBERT J.-C., LEMARECHAL C., SAGASTIZABAL C., *Optimisation numérique*, Mathématiques et Applications 27, Springer, Paris (1997).
- [7] BONY J.-M., *Cours d'analyse. Théorie des distributions et analyse de Fourier*, Éditions de l'École Polytechnique, Palaiseau (2001).
- [8] BREZIS H., *Analyse fonctionnelle*, Masson, Paris (1983).
- [9] CHVÁTAL V., *Linear programming*, Freeman and Co., New York (1983).
- [10] CIARLET P.G., *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, Paris (1982).
- [11] CIARLET P.G., *The finite element methods for elliptic problems*, North-Holland, Amsterdam (1978).
- [12] CIARLET P.G., LIONS J.-L., *Handbook of numerical analysis*, North-Holland, Amsterdam (1990).
- [13] COOK J., CUNNINGHAM W.H., PULLEYBANK W.R., SCHRIJVER A., *Combinatorial optimization*, Wiley, New York (1998).
- [14] COURANT R., HILBERT R., *Methods of mathematical physics*, John Wiley & Sons, New York (1989).
- [15] CULIOLI J.-C., *Introduction à l'optimisation*, Éditions Ellipses, Paris (1994).

- [16] DANAILA I., HECHT F., PIRONNEAU O., *Simulation numérique en C++*, Dunod, Paris (2003).
- [17] DANAILA I., JOLY P., KABER S. M., POSTEL M., *Introduction au calcul scientifique par la pratique*, Dunod, Paris (2005).
- [18] DAUTRAY R., LIONS J.-L., *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Masson, Paris (1988).
- [19] DUBOIS F., DESPRES B., *Systèmes hyperboliques de lois de conservation ; application à la dynamique des gaz*, Éditions de l'École Polytechnique, Palaiseau (2005).
- [20] DUVAUT G., LIONS J.-L., *Les inéquations en mécanique et en physique*, Dunod, Paris (1972).
- [21] EKELAND I., TEMAM R., *Analyse convexe et problèmes variationnels*, Dunod, Paris (1974).
- [22] ERN A., GUERMOND J.-L., *Éléments finis : théorie, applications, mise en oeuvre*, Mathématiques et Applications 36, Springer, Paris (2002).
- [23] FAURRE P., *Analyse numérique. Notes d'optimisation*, École Polytechnique, Ellipses, Édition Marketing, Paris (1988).
- [24] GEORGE P.L., *Génération automatique de maillages, application aux méthodes d'éléments finis*, Masson, Paris (1991).
- [25] GIRAULT V., RAVIART P.-A., *Finite element methods for Navier-Stokes equations : theory and algorithms*, Springer, Berlin (1986).
- [26] GLOVER F., LAGUNA M., *Tabu search*, Kluwer, Boston (1997).
- [27] GUICHARDET A., *Intégration. Analyse hilbertienne*, École Polytechnique, Ellipses, Édition Marketing, Paris (1989).
- [28] LAUDENBACH F., *Calcul différentiel et intégral*, Éditions de l'École Polytechnique, Palaiseau (2000).
- [29] LEMAITRE J., CHABOCHE J.-L., *Mécanique des matériaux solides*, Dunod, Paris (1985).
- [30] LIONS J.-L., *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod, Paris (1969).
- [31] LIONS J.-L., MAGENES E., *Problèmes aux limites non homogènes et applications*, Dunod, Paris (1970).
- [32] LUCQUIN B., PIRONNEAU O., *Introduction au calcul scientifique*, Masson, Paris (1996).
- [33] MOHAMMADI B., SAÏAC J.-H., *Pratique de la simulation numérique*, Dunod, Paris (2003).
- [34] PADBERG M., *Linear optimization and extensions*, Springer, Berlin (1999).
- [35] PIRONNEAU O., *Méthodes des éléments finis pour les fluides*, Masson, Paris (1988).

- [36] RAVIART P.-A., THOMAS J.-M., *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, Paris (1983).
- [37] SALENÇON J., *Mécanique des milieux continus*, Éditions de l'École Polytechnique, Palaiseau (2002).
- [38] SCHRIJVER A., *Theory of linear and integer programming*, Wiley, New York (1986).
- [39] TEMAM R., *Navier-Stokes equations. Theory and numerical analysis*, AMS Chelsea Publishing, Providence, RI (2001).

# Index

- adjoint 213
- affectation 376
- algorithme
  - de Ford et Bellman 384
  - de Kruskal 388
  - de sur-gradient 400
  - du gradient à pas fixe 341
  - du gradient à pas optimal 340
  - du gradient conjugué 342, 435
  - du gradient projeté 343
  - du simplexe 359
  - d'Uzawa 344
  - glouton 387
- arbre 388
- arcs 374
- arêtes 387
- assemblage de la matrice 182
- auto-adjoint 213
  
- base 357
- base hilbertienne 406
- bien posé 27
  
- calcul des variations 289
- capacité 374
- centré 16
- chemin 379
- circuit 379
- coercive 74
- commande optimale) 332
- compact (sous-espace) 214
- compacte (application) 214
- complet 405
- complexité 403, 428
  
- composantes connexes 388
- condition aux limites
  - de Dirichlet 4
  - de Fourier 4
  - de Neumann 4
  - périodiques 39
  - de transmission 127
- condition CFL 20, 25, 40, 47, 274, 279
- condition de qualification 318
- condition de stabilité de Von Neumann
  - 41, 46, 279
- condition initiale 4
- conditionnement 415
- connexe 388
- consistance 35
- contrainte
  - active 318
  - qualifiée 318, 322
- contrôle 290, 337
- convergence faible 300
- convergence numérique 171
- convexe 296, 405
- convexité
  - $\alpha$ - 297
  - forte 297
  - stricte 296
- coût 291, 374
  
- décentré 17
- décomposition spectrale 215
- définie positive 213
- degré de liberté 179
- dérivée faible 83
- dérivée seconde 308

- différences finies 15
- différentiabilité
  - au sens de Fréchet 304
  - au sens de Gâteaux 306
- diffusif 57, 281
- diffusion numérique 57
- direction admissible 312
- directions alternées 50
- dispersif 57
- distribution 108
- divergence 3
- domaine
  - de dépendance 11, 267
  - d'une fonction convexe 398
- dual 327, 365, 408
- écarts complémentaires 367
- effet régularisant 266
- égalité d'énergie 71, 246
- élasticité 13, 138
- éléments finis
  - de Hermite 172
  - de Lagrange 179
  - rectangulaires 194
  - triangulaires 174
- élimination de Gauss 418
- elliptique 29, 74
- énergie complémentaire 331
- enveloppe
  - convexe 368
  - entière 370
- épigraphe 296
- équation
  - aux dérivées partielles 1
  - de Bellman 379
  - de diffusion 2, 6
  - de la chaleur 2
  - de Schrödinger 12
  - des ondes 10
  - des plaques 14, 173
  - équivalente 56
- équipartition de l'énergie 267
- erreur de troncature 35
- espace
  - d'énergie 86, 240, 246, 250, 255
  - de Banach 83
  - de Hilbert 405
  - de Sobolev 86
- estimation d'énergie 116, 121, 144, 241
- état 378
- état adjoint 334
- explicite 17
- factorisation
  - LU 420
  - de Cholesky 424
- flot 374
- fonction
  - coût ou objectif 291
  - de Green 264
  - propre 218
  - test 71
  - valeur 379
- forêt 388
- forme linéaire 408
- formulation faible 115
- formulation variationnelle 71
- formule de Green 68, 93
- gradient 3, 340, 433
  - conjugué 342, 437
  - projeté 343
  - à pas fixe 341
  - à pas optimal 340
- graphe
  - complet 391
  - non orienté 387
  - orienté 374
- hyperbolique 29
- hyperplan d'appui 369, 409
- implicite 17
- inégalité de Poincaré 78, 90
- inéquation d'Euler 309
- infimum 291
- infini à l'infini 292
- instable 19

- intégration numérique 159
- interpolation 161, 190
- invariance par changement d'échelle 8
- irréversible 8
- Lagrangien 316, 324
- Lamé 13
- Laplacien 3
- lemme
  - de Céa 153
  - de Bramble-Hilbert 192
  - de Farkas 320
- loi des nœuds de Kirchoff 374
- maillage 16, 156, 175, 194
- maillage régulier 188
- maillage uniforme 156
- matrice
  - bande 426
  - creuse 427
  - de masse 230
  - de rigidité 151, 153, 158, 230
  - d'itération 38
- mesure surfacique 68
- méthode
  - de Gauss-Seidel 433
  - de Givens-Householder 445
  - de Jacobi 432
  - de Lanczos 448
  - de la puissance 443
  - de Newton 349
  - de relaxation 433
  - directe 411
  - du gradient 433
  - du gradient conjugué 434
  - itérative 411
- minimum
  - global 291
  - local 291
- mode propre 226
- modélisation 2, 4, 7, 14, 137, 139, 147, 259, 261
- mouvement de corps rigide 142
- multi-indice 98
- multiplicateur de Lagrange 311
- NP-complet 404
- $N$ -rectangle 194
- $N$ -simplexe 174
- Navier-Stokes 147
- nœud 175, 374
- nombre de Péclet 6
- normale extérieure 68
- norme discrète 170
- norme subordonnée 412
- opérateur 408
- ordre
  - d'une e.d.p. 29
  - d'un schéma 35
- ouvert régulier 69
- parabolique 29
- pas d'espace, pas de temps 15
- pénalisation 347, 364
- point
  - entier 368
  - extrémal 368
  - selle 324
- polyèdre 174, 357
- polyèdre entier 371
- polyédrique 174
- polynomial (algorithme) 403
- précision 35
- primal 327, 365
- principe
  - du maximum 9, 21, 38, 129, 169, 259, 266
  - du min-max 220
  - des travaux virtuels 71, 144
- problème
  - aux limites 27, 67
  - bien posé 27
  - de Cauchy 27, 67
  - dual 327, 365
  - primal 325, 363
  - relâché 368
  - spectral 218



- programme linéaire 354
- programmation dynamique 378
- projection orthogonale 405
- propagation à vitesse finie 9, 11, 261, 267
- quadrature 159
- quotient de Rayleigh 220
- raide 275
- recherche opérationnelle 284, 353
- régularité 131
- régulier (maillage) 188
- régulier (ouvert) 69
- relaxation Lagrangienne 398
- réversible 8, 11, 60, 265
- RO 353
- schéma linéaire 37
- schéma multiniveaux 33, 44
- séparable 407
- singularité 135
- solution
  - admissible 356
  - basique 357
  - forte 66
  - faible 115
  - maximale 382
  - optimale 356
  - variationnelle 115
- sommets 156, 387
- sous-différentiel 399
- sous-gradient 399
- splitting 50
- stable 19, 37, 274
- stationnaire 12
- stencil 34
- Stokes 14, 147
- suite minimisante 291
- support borné 68, 89
- support compact 82
- système d'équations 13
- système linéaire 411
- théorème
  - de Kuhn et Tucker 325
  - de Lax-Milgram 74
  - de Minkowski 369
  - de Rellich 97
  - de trace 92
  - transport 285, 375
  - treillis 177, 195
  - triangulation 175
  - unimodulaire 372
  - unisolvant 177, 195
  - valeur propre 213
  - valuation 386
  - variable d'écart 355
  - vecteur propre 213
  - Von Neumann 41, 46, 279

# Index des applications

advection d'une concentration .....	7
cisaillement anti-plan .....	146
commande optimale .....	288, 332
conduction de la chaleur .....	2, 236, 257
contrôle d'une membrane .....	290, 337
convection-diffusion d'une concentration .....	6, 117, 186
diffusion d'une concentration .....	186, 206, 276
corde vibrante .....	10, 59
économie .....	288, 326
élasticité .....	13, 139, 207
élastodynamique .....	226, 237
finance .....	6
fissuration et rupture .....	136
logistique .....	285
mécanique des fluides .....	14, 147, 228, 249
mécanique quantique .....	12, 212, 228
membrane élastique .....	12, 337
milieux poreux .....	236
plaque élastique .....	14, 146, 229
problème d'affectation .....	285, 376
problème de flots .....	374
problème de transport .....	285
problème du sac-à-dos .....	286, 385
propagation d'ondes .....	10, 237, 254, 265
recherche opérationnelle .....	353
théorie cinétique des gaz .....	316
tournée du voyageur de commerce .....	286, 391
vibrations .....	226

# Index des notations

$A^*$ matrice adjointe .....	412
$A^t$ matrice transposée .....	412
$\text{co } X$ enveloppe convexe de $X$ .....	368
$C^k(\Omega), C^k(\overline{\Omega})$ espace des fonctions $k$ fois continûment dérivables .....	66
$C_c^\infty(\Omega)$ espace des fonctions infiniment dérivables à support compact .....	82, 108
$C_c^\infty(\overline{\Omega})$ espace des fonctions infiniment dérivables à support borné .....	89
$\mathcal{D}(\Omega)$ espace des fonctions infiniment dérivables à support compact .....	82
$\mathcal{D}'(\Omega)$ espace des distributions .....	108
$\delta_{ij}$ symbole de Kronecker .....	158
$dx$ mesure volumique (de Lebesgue) .....	68
$ds$ mesure surfacique (de Lebesgue) .....	68
$\partial^\alpha$ dérivée partielle d'ordre $\alpha$ .....	98
$\partial\Omega$ bord d'un ouvert $\Omega$ .....	68
$\text{Epi}(J)$ épigraphe d'une fonction $J$ .....	296
$\text{extr } K$ ensemble des points extrémaux d'un convexe $K$ .....	368
$H^1(\Omega)$ espace de Sobolev .....	86
$H_0^1(\Omega)$ espace de Sobolev à trace nulle .....	90
$H^m(\Omega)$ espace de Sobolev d'ordre $m$ .....	99
$H^{1/2}(\partial\Omega)$ espace des traces .....	94
$H(\text{div})$ espace de champs de vecteurs .....	103
$J'$ différentielle d'une fonction $J$ .....	305
$L^2(\Omega)$ espace de Lebesgue .....	82
$L^\infty(\Omega)$ espace de Lebesgue .....	83
$n$ normale unitaire extérieure .....	68
$ \mathcal{N} $ cardinal d'un ensemble fini $\mathcal{N}$ .....	380
$P_e$ enveloppe entière de $P$ .....	370
$\mathbb{P}_k$ espace des polynômes de degré total $k$ .....	177
$\mathbb{Q}_k$ espace des polynômes de degré en chaque variable $k$ .....	195
$\mathbb{R}_+^N$ demi-espace .....	92
$V'$ dual d'un espace $V$ .....	408

*Dans la même collection*

## **BIOLOGIE**

*Bioinformatique. Génomique et post-génomique* - F. Dardel et F. Képès - 250 pages - ISBN 2-7302-0927-1

## **CHIMIE**

*Chimie moléculaire des éléments de transition* - F. Mathey et A. Sevin - 300 pages - ISBN 2-7302-0714-7

*Les orbitales moléculaires dans les complexes - avec Exercices et Corrigés* - Y. Jean  
350 pages - ISBN 2-7302-1024-5

*Chimie moléculaire, sol-gel et nanomatériaux* - R. Corriu et Nguyen T.-A.  
208 pages - ISBN 978-2-7302-1413-1

*Introduction à la chimie quantique* - P. Hiberty et Nguyen T.-A. - 320 pages - ISBN 978-2-7302-1485-8

## **INFORMATIQUE**

*Introduction à la théorie des langages de programmation* - G. Dowek, J.-J. Lévy  
112 pages - ISBN 978-2-7302-1333-2

*Les principes des langages de programmation* - G. Dowek - 176 pages - ISBN 978-2-7302-1526-8

*Les démonstrations et les algorithmes : introduction à la logique et la calculabilité* - G. Dowek - 200 pages -  
ISBN 978-2-7302-1569-5

## **ÉCONOMIE**

*Introduction à la microéconomie* - N. Curien  
110 pages - ISBN 2-7302-0722-8 - ISBN 978-2-7302-0722-5

*Introduction à l'analyse macroéconomique* - P.-A. Muet - 208 pages - ISBN 2-7302-1140-3

*Économie de l'entreprise* - J.-P. Ponssard, D. Sevy, H. Tanguy (2<sup>e</sup> édition)  
316 pages - ISBN 978-2-7302-1244-1

*Corporate Social Responsibility? From Compliance to Opportunity* - P. Crifo, J.-P. Ponssard - 298 pages -  
ISBN 978-2-7302-1568-8

*Économie du climat. Pistes pour l'après-Kyoto* - J.-P. Ponssard et O. Godard - 314 pages  
ISBN 978-2-7302-1576-3

## **MATHÉMATIQUES**

*Le problème de Kepler. Histoire et théorie* - A. Guichardet - 102 pages - ISBN 978-2-7302-1596-1

*Autour des inégalités isopérimétriques* - Sous la direction de Alessio Figalli - 130 pages - ISBN 978-2-7302-1573-2

*Éléments d'analyse et d'algèbre (et de théorie des nombres)* - Nouvelle édition - P. Colmez - 678 pages -  
ISBN 978-2-7302-1587-9

*Transversalité, Courants et Théorie de Morse. Un cours de topologie différentielle (exercices proposés par François Labourie)* - F. Laudenbach - 200 pages - ISBN 978-2-7302-1585-5

*Groupes et représentations* - D. Renard - 223 pages - ISBN 978-2-7302-1571-8  
*Milieux continus en transformations finies. Hyperélasticité, Rupture, Élastoplasticité* - C. Stolz - 278 pages - ISBN 978-2-7302-1562-6

*Éléments d'analyse et d'algèbre (et de théorie des nombres)* - P. Colmez - 488 pages - ISBN 978-2-7302-1563-3  
*Cours d'analyse* - J.-M. Bony - 272 pages - ISBN 2-7302-0775-1

*Calcul différentiel et intégral* - F. Laudenbach - 220 pages - ISBN 2-7302-0724-4  
*Méthodes mathématiques pour les sciences physiques* - J.-M. Bony - 217 pages - ISBN 2-7302-0723-6

*Les théorèmes de Noether* - Y. Kosmann-Schwarzbach, avec la collaboration de L. Meersseman - 180 pages - ISBN 2-7302-1138-1 (2<sup>e</sup> édition)

*Groupes et symétries. Groupes finis, groupes et algèbres de Lie, représentations* - Y. Kosmann-Schwarzbach - 222 pages - ISBN 978-2-7302-1257-1 (2<sup>e</sup> édition)

*Algèbre corporelle* - A. Chambert-Loir - 200 pages - ISBN 2-7302-1217-5

*Calcul variationnel* - J.-P. Bourguignon - 348 pages - ISBN 978-2-7302-1415-5

*Aspects des systèmes dynamiques* - XUPS1994-1996 - N. Berline et C. Sabbah (Comité éditorial) - 238 pages - ISBN 978-2-7302-1560-2

*Aspects de la théorie du contrôle* - XUPS1999 - N. Berline et C. Sabbah (Comité éditorial) - 208 pages - ISBN 978-2-7302-1530-5

*Groupes finis* - XUPS 2000 - N. Berline et C. Sabbah (Comité éditorial) - 104 pages - ISBN 2-7302-0751-5

*Pavages* - XUPS 2001 - N. Berline et C. Sabbah (Comité éditorial) - 112 pages - ISBN 2-7302-0855-0

*La fonction zêta* - XUPS 2002 - N. Berline et C. Sabbah (Comité éditorial) - 206 pages - ISBN 2-7302-1011-3

*Distributions* - XUPS 2003 - Dans le sillage de Laurent Schwartz - N. Berline et C. Sabbah (Comité éditorial) - 106 pages - ISBN 2-7302-1095-4

*Graphes* - XUPS 2004 - N. Berline et C. Sabbah (Comité éditorial) - 84 pages - ISBN 2-7302-1182-9

*Théorie algorithmique des nombres et équations diophantiennes* - XUPS2005 - N. Berline, A. Plagne et C. Sabbah (Comité éditorial) - 200 pages - ISBN 2-7302-1293-0

*Théorie des jeux. Introduction à la théorie des jeux répétés* - XUPS 2006 - N. Berline, A. Plagne et C. Sabbah (Comité éditorial) - 152 pages - ISBN 978-2-7302-1366-0

*Sur la dynamique des groupes de matrices et applications arithmétiques* - XUPS2007 - N. Berline, A. Plagne et C. Sabbah (Comité éditorial) - 160 pages - ISBN 978-2-7302-1418-6

*Géométrie tropicale* - XUPS2008 - P. Harinck, A. Plagne et C. Sabbah (Comité éditorial)  
128 pages - ISBN 978-2-7302-1529-9

*Les représentations linéaires et le grand théorème de Fermat* - XUPS2009 - P. Harinck, A. Plagne et C. Sabbah (Comité éditorial) - 140 pages - ISBN 978-2-7302-1566-4

*Facettes mathématiques de la mécanique des fluides* - XUPS2010 - P. Harinck, A. Plagne et C. Sabbah (Comité éditorial) - 118 pages - ISBN 978-2-7302-1578-7

*Histoire de mathématiques* - XUPS2011 - P. Harinck, A. Plagne et C. Sabbah (Comité éditorial)  
118 pages - ISBN 978-2-7302-1595-4

Séminaires, équations aux dérivées partielles - Années 1971 à 1992

Séminaires, équations aux dérivées partielles - Années 1993 à 2001

Séminaires, équations aux dérivées partielles - Année 2000-2001 - 400 pages - ISBN 2-7302-0834-8

Séminaires, équations aux dérivées partielles - Année 2001-2002 - 364 pages - ISBN 2-7302-0930-1

Séminaires, équations aux dérivées partielles - Année 2002-2003 - 390 pages - ISBN 2-7302-1041-5

Séminaires, équations aux dérivées partielles - Année 2003-2004 - 404 pages - ISBN 2-7302-1183-7 Sémi-

naires, équations aux dérivées partielles - Année 2004-2005 - 404 pages - ISBN 2-7302-1221-3 Séminaires,

équations aux dérivées partielles - Année 2005-2006 - 366 pages - ISBN 2-7302-1335-X

Séminaires, équations aux dérivées partielles - Année 2006-2007 - 444 pages - ISBN 978-2-7302-1414-8

Séminaires, équations aux dérivées partielles - Année 2008-2009 - 308 pages - ISBN 978-2-7302-1567-1

## **MATHÉMATIQUES APPLIQUÉES**

*Une exploration des signaux en ondelettes* - S. Mallat - 654 pages - ISBN 2-7302-0733-3

*Promenade aléatoire* - M. Benaïm et N. El Karoui - 316 pages - ISBN 2-7302-1168-3

*Analyse numérique et optimisation* - G. Allaire - 480 pages - ISBN 978-2-7302-1255-7

*Systèmes hyperboliques de lois de conservation. Application à la dynamique des gaz* - F. Dubois, B. Després  
208 pages - ISBN 2-7302-1253-1

*Commande et optimisation de systèmes dynamiques* - F. Bonnans et P. Rouchon  
286 pages - ISBN 2-7302-1251-5

*Les outils stochastiques des marchés financiers. Une visite guidée de Einstein à Black-Scholes* - N. El Karoui et E. Gobet - 238 pages - ISBN 978-2-7302-1579-4

*Simulation stochastique et méthodes de Monte-Carlo* - C. Graham et D. Talay - 210 pages - ISBN 978-2-7302-1582-4

*Aléatoire. Introduction à la théorie et au calcul des probabilités* - S. Méléard - 280 pages - ISBN 978-2-7302-1575-6

## **MÉCANIQUE**

*Dynamique de l'atmosphère et de l'océan* - P. Bougeault et R. Sadourny - 312 pages - ISBN 2-7302-0825-9

*Introduction aux écoulements compressibles et aux fluides hétérogènes* - A. Sellier  
175 pages - ISBN 2-7302-0764-5

*Mécanique des milieux continus* - J. Salençon

Tome 1 - *Concepts généraux* - 376 pages - ISBN 978-2-7302-1245-8 (avec CD-Rom)

Tome 2 - *Thermoélasticité* - 344 pages - ISBN 978-2-7302-1419-3 (avec CD-Rom)

Tome 3 - *Milieux curvilignes* - 154 pages - ISBN 2-7302-0962-X

*de l'Élasto-plasticité au Calcul à la rupture* - J. Salençon

(accompagné d'un CD-Rom réalisé par J. Salençon) - 266 pages - ISBN 978-2-7302-0915-1

*Viscoélasticité pour le calcul des structures* - J. Salençon - 160 pages - ISBN 978-2-7302-1557-2

*Fluides et Solides* - E. de Langre - 130 pages - ISBN 978-2-7302-0833-8

*Ondes acoustiques* - A. Chaigne - 224 pages - ISBN 2-7302-0840-2

*Stabilité des matériaux et des structures* - C. Stolz - 206 pages - ISBN 2-7302-1076-8

*Instabilités, Chaos et Turbulence* - P. Manneville - 360 pages - ISBN 2-7302-0913-1

*Vibrations des structures couplées avec le vent* - P. Hémon - 144 pages - ISBN 2-7302-1332-5

*Analyse des solides déformables par la méthode des éléments finis* - M. Bonnet et A. Frangi

316 pages - 2-7302-1349-X

*Dynamique et vibrations* - E. de Langre et A. Chaigne - 152 pages - ISBN 978-2-7302-1521-3

*Modélisation et calcul des milieux continus* - P. Le Tallec - 560 pages - ISBN 978-2-7302-1494-0

*Poutres et arcs élastiques*, P. Ballard et A. Millard - 312 pages - ISBN 978-2-7302-1561-9

*Hydrodynamique de l'environnement*, O. Thual - 328 pages - ISBN 978-2-7302-1564-0

*Microhydrodynamique et fluides complexes*, D. Barthès-Biesel - 292 pages - ISBN 978-2-7302-1572-5

## **PHYSIQUE**

*Physique des Tokamaks* - Jean-Marcel Rax - 436 pages - ISBN 978-2-7302-1580-0

*Semi-conducteurs : les bases de la théorie k.p* - G. Fishman - 742 pages - ISBN 978-2-7302-1497-1

*Énergie nucléaire* - J.-L. Basdevant, J. Rich et M. Spiro - 340 pages - ISBN 2-7302-0901-8

*Mécanique quantique* - J.-L. Basdevant et J. Dalibard

(accompagné d'un CD-Rom de M. Joffre) 520 pages - ISBN 978-2-7302-0914-4

*Problèmes quantiques* - J.-L. Basdevant et J. Dalibard - 214 pages - ISBN 2-7302-1117-9

*Principes de la cosmologie* - J. Rich, adaptation française J.-L. Basdevant - 400 pages - ISBN 2-7302-0925-5

*Introduction à la relativité* - A. Rougé - 188 pages - ISBN 978-2-7302-0940-3

*Relativité restreinte*, La contribution d'Henri Poincaré - A. Rougé - 288 pages - ISBN 978-2-7302-1525-1

*Introduction à la physique subatomique* - A. Rougé - 448 pages - ISBN 2-7302-1231-0

*Physique statistique et illustrations en physique du solide.* - C. Hermann - 292 pages - ISBN 978-2-7302-1022-5

*Bases physiques de la plasticité des solides* - J.-C. Tolédano - 264 pages - ISBN 978-2-7302-1378-3

*Physique des électrons dans les solides. Structure de bandes, Supraconductivité et Magnétisme.* H. Alloul - Tome 1 - 360 pages - ISBN 978-2-7302-1411-7

*Physique des électrons dans les solides. Recueil d'exercices et de problèmes.* H. Alloul  
Tome 2 - 272 pages - ISBN 978-2-7302-1412-4



ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

ÉCOLE POLYTECHNIQUE

Achevé d'imprimer en octobre 2012. Dépôt légal : 3<sup>e</sup> trimestre 2005  
ISBN 978 – 2 – 7302 – 1255 – 7. *Imprimé en France*