



**HAL**  
open science

# Analyse numérique des équations aux dérivées partielles

Raphaèle Herbin

► **To cite this version:**

Raphaèle Herbin. Analyse numérique des équations aux dérivées partielles. Engineering school. Marseille, 2011. cel-00637008v1

**HAL Id: cel-00637008**

**<https://cel.hal.science/cel-00637008v1>**

Submitted on 28 Oct 2011 (v1), last revised 4 Aug 2023 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

**Université Aix Marseille 1**

**Master de mathématiques**

**Analyse numérique  
des équations aux dérivées partielles**

Raphaèle Herbin

26 octobre 2011

# Table des matières

<b>Introduction</b>	<b>3</b>
L'analyse numérique des équations aux dérivées partielles . . . . .	3
Principales méthodes de discrétisation . . . . .	4
Quelques exemples d'équations aux dérivées partielles . . . . .	4
<b>1 Différences finies et volumes finis pour les problèmes de diffusion convection réaction stationnaires</b>	<b>6</b>
1.1 Principe des deux méthodes . . . . .	6
1.1.1 Cas de la dimension 1 . . . . .	6
1.1.2 Cas de la dimension 2 ou 3 . . . . .	10
1.1.3 Questions d'analyse numérique . . . . .	12
1.2 Analyse de la méthode des différences finies . . . . .	12
1.3 Schéma volumes finis pour un problème elliptique unidimensionnel . . . . .	18
1.4 Volumes finis pour la prise en compte de discontinuités . . . . .	24
1.5 Différences finies et volumes finis pour les problèmes de diffusion 2D . . . . .	25
1.5.1 Différences finies . . . . .	25
1.5.2 Volumes finis . . . . .	26
1.6 Exercices . . . . .	30
1.7 Suggestions pour les exercices . . . . .	40
1.8 Corrigés des exercices . . . . .	42
<b>2 Problèmes paraboliques : la discrétisation en temps</b>	<b>59</b>
2.1 Le problème continu, et la discrétisation espace-temps . . . . .	59
2.2 Discrétisation par Euler explicite en temps. . . . .	60
2.2.1 Consistance du schéma . . . . .	60
2.2.2 Stabilité . . . . .	61
2.2.3 Convergence . . . . .	61
2.2.4 Exemple de non convergence . . . . .	62
2.2.5 Stabilité au sens des erreurs d'arrondi . . . . .	63
2.2.6 Stabilité au sens de Von Neumann . . . . .	65
2.3 Schéma implicite et schéma de Crank-Nicolson . . . . .	67
2.3.1 Le $\theta$ -schéma . . . . .	67
2.3.2 Consistance et stabilité . . . . .	68
2.3.3 Convergence du schéma d'Euler implicite. . . . .	69
2.4 Cas de la Dimension 2 . . . . .	70
2.5 Exercices . . . . .	71
2.6 Suggestions pour les exercices . . . . .	80
2.7 Corrigés des exercices . . . . .	80
<b>3 Méthodes variationnelles</b>	<b>97</b>
3.1 Exemples de problèmes variationnels . . . . .	97
3.1.1 Le problème de Dirichlet . . . . .	97
3.1.2 Problème de Dirichlet non homogène . . . . .	101
3.1.3 Problème avec conditions aux limites de Fourier . . . . .	103
3.1.4 Condition de Neumann . . . . .	105
3.1.5 Formulation faible et formulation variationnelle. . . . .	106

3.2	Méthodes de Ritz et Galerkin . . . . .	106
3.2.1	Principe général de la méthode de Ritz . . . . .	106
3.2.2	Méthode de Galerkin . . . . .	110
3.2.3	Méthode de Petrov-Galerkin . . . . .	113
3.3	La méthode des éléments finis . . . . .	113
3.3.1	Principe de la méthode . . . . .	113
3.3.2	Construction du maillage, de l'espace $H_N$ et de sa base $\phi_N$ . . . . .	116
3.4	Exercices . . . . .	118
3.5	Suggestions pour les exercices . . . . .	125
3.6	Corrigés des exercices . . . . .	126
<b>4</b>	<b>Eléments finis de Lagrange</b>	<b>137</b>
4.1	Espace d'approximation . . . . .	137
4.1.1	Cohérence "locale" . . . . .	137
4.1.2	Construction de $H_N$ et conformité . . . . .	141
4.2	Exemples . . . . .	143
4.2.1	Élément fini de Lagrange $P1$ sur triangle ( $d = 2$ ) . . . . .	144
4.2.2	Élément fini triangulaire $P2$ . . . . .	145
4.2.3	Éléments finis sur quadrangles . . . . .	146
4.3	Construction du système linéaire . . . . .	147
4.3.1	Construction de $H_N$ et $\Phi_i$ . . . . .	149
4.3.2	Construction de $\mathcal{K}$ et $\mathcal{G}$ . . . . .	149
4.3.3	Calcul de $a_\Omega$ et $T_\Omega$ , matrices élémentaires. . . . .	151
4.3.4	Calcul de $a_{\Gamma_1}$ et $T_{\Gamma_1}$ (contributions des arêtes de bord "Fourier") . . . . .	154
4.3.5	Prise en compte des noeuds liés dans le second membre . . . . .	155
4.3.6	Stockage de la matrice $\mathcal{K}$ . . . . .	155
4.4	Eléments finis isoparamétriques . . . . .	156
4.5	Analyse d'erreur . . . . .	156
4.5.1	Erreurs de discrétisation et d'interpolation . . . . .	156
4.5.2	Erreur d'interpolation en dimension 1 . . . . .	157
4.5.3	Super convergence . . . . .	160
4.5.4	Traitement des singularités . . . . .	161
4.6	Exercices . . . . .	161
4.7	Corrigés des exercices . . . . .	168
<b>5</b>	<b>Problèmes hyperboliques</b>	<b>179</b>
5.1	Une équation de transport . . . . .	179
5.2	Equation linéaire, cas 1D . . . . .	180
5.3	Schémas numériques pour $u_t + u_x = 0$ . . . . .	183
5.3.1	Schéma explicite différences finies centrées . . . . .	183
5.3.2	Schéma différences finies décentré amont . . . . .	184
5.3.3	Schéma volumes finis décentrés amont . . . . .	186
5.4	Equations hyperboliques non linéaires . . . . .	186
5.5	Schémas pour les équations non linéaires . . . . .	194
5.6	Exercices . . . . .	197
5.7	Suggestions pour les exercices . . . . .	205
5.8	Corrigés des exercices . . . . .	205
Bibliography		

# Introduction

## L'analyse numérique des équations aux dérivées partielles

Pour aborder le calcul numérique (à l'aide d'un outil informatique) des solutions d'un problème "réel", on passe par les étapes suivantes :

1. Description qualitative des phénomènes physiques.  
Cette étape, effectuée par des spécialistes des phénomènes que l'on veut quantifier (ingénieurs, chimistes, biologistes etc.....) consiste à répertorier tous les mécanismes qui entrent en jeu dans le problème qu'on étudie.
2. Modélisation  
Il s'agit, à partir de la description qualitative précédente, d'écrire un modèle mathématique. On supposera ici que ce modèle amène à un système d'équations aux dérivées partielles (EDP). Selon les hypothèses effectuées, la modélisation peut aboutir à plusieurs modèles, plus ou moins complexes. Dans la plupart des cas, on ne saura pas calculer une solution analytique, explicite, du modèle ; on devra faire appel à des techniques de résolution approchée.
3. Analyse mathématique du modèle mathématique.  
Même si l'on ne sait pas trouver une solution explicite du modèle, il est important d'en étudier les propriétés mathématiques, dans la mesure du possible. Il est bon de se poser les questions suivantes :
  - Le problème est-il bien posé ? c'est-à-dire y-a-t'il existence et unicité de la solution ?
  - Les propriétés physiques auxquelles on s'attend sont elles satisfaites par les solutions du modèle mathématique ?  
Si l'inconnue est une concentration, par exemple, peut-on prouver que la solution du modèle sensé la représenter est toujours positive ?
  - Y a-t-il continuité de la solution par rapport aux données ?
4. Discrétisation et résolution numérique  
Un problème posé sur un domaine continu (espace - temps) n'est pas résoluble tel quel par un ordinateur, qui ne peut traiter qu'un nombre fini d'inconnues. Pour se ramener à un problème en dimension finie, on discrétise l'espace et/ou le temps. Si le problème original est linéaire on obtient un système linéaire. Si le problème original est non linéaire (par exemple s'il s'agit de la minimisation d'une fonction) on aura un système non linéaire à résoudre par une méthode *ad hoc* (méthode de Newton...)
5. Analyse numérique  
IL s'agit maintenant de l'analyse mathématique du schéma numérique. En effet, une fois le problème discret obtenu, il est raisonnable de se demander si la solution de ce problème est proche, et en quel sens, du problème continu. De même, si on doit mettre en oeuvre une méthode itérative pour le traitement des non-linéarités, il faut étudier la convergence de la méthode itérative proposée.
6. Mise en oeuvre, programmation et analyse des résultats  
La partie mise en oeuvre est une grosse consommatrice de temps. Actuellement, de nombreux codes commerciaux existent, qui permettent en théorie de résoudre "tous" les problèmes. Il faut cependant procéder à une analyse critique des résultats obtenus par ces codes, qui ne sont pas toujours compatibles avec les propriétés physiques attendues...

## Les principales méthodes de discrétisation

### Les méthodes de différences finies et volumes finis

On considère un domaine physique  $\Omega \subset \mathbb{R}^d$ , où  $d$  est la dimension de l'espace. Le principe des méthodes de différences finies (DF) consiste à se donner un certain nombre de points du domaine, qu'on notera  $(x_1 \dots x_N) \subset (\mathbb{R}^d)^N$ . On approche l'opérateur différentiel en espace en chacun des  $x_i$  par des quotients différentiels. Il faut alors discrétiser la dérivée en temps : on pourra par exemple considérer un schéma d'Euler<sup>1</sup> explicite ou implicite pour la discrétisation en temps.

Les méthodes de volumes finis (VF) sont adaptées aux équations de conservation et utilisées en mécanique des fluides depuis plusieurs décennies. Le principe consiste à découper le domaine  $\Omega$  en des "volumes de contrôle"; on intègre ensuite l'équation de conservation sur les volumes de contrôle; on approche alors les flux sur les bords du volume de contrôle par exemple pas une technique de différences finies.

### Les méthodes variationnelles, méthodes d'éléments finis

On met le problème d'équations aux dérivées partielles sous la forme suivante, dite variationnelle :

$$\begin{cases} a(u, v) = (f, v)_H, & \forall v \in H, \\ u \in H, \end{cases}$$

où  $H$  est un espace de Hilbert<sup>2</sup> bien choisi (par exemple parce qu'il y a existence et unicité de la solution dans cet espace),  $(\cdot, \cdot)_H$  le produit scalaire sur  $H$  et  $a$  une forme bilinéaire sur  $H$ . Dans un tel cadre fonctionnel, la discrétisation consiste à remplacer  $H$  par un sous espace de dimension finie  $H_k$ , construit par exemple à l'aide de fonctions de base éléments finis qu'on introduira plus loin :

$$\begin{cases} a(u_k, v_k) = (f, v_k)_H, & \forall v \in H_k, \\ u_k \in H_k. \end{cases}$$

### Les méthodes spectrales

L'idée de ces méthodes est de chercher une solution approchée sous forme d'un développement sur une certaine famille de fonctions. On peut par exemple écrire la solution approchée sous la forme :  $u = \sum_{i=1}^n \alpha_i(u) p_i$ , où les  $p_i$  sont des fonction polynomiales. On choisit la base  $p_i$  de manière à ce que les dérivées de  $\alpha_i$  et  $p_i$  soient faciles à calculer. Ces dernières méthodes sont réputées coûteuses, mais précises. Elles sont d'ailleurs plus souvent utilisées comme aide à la compréhension des phénomènes physiques sur des problèmes modèles que dans pour des applications industrielles.

## Quelques exemples d'équations aux dérivées partielles

- L'équation de Poisson : en trois dimension d'espace, elle s'écrit :  $-\operatorname{div} \kappa \nabla u = f$ , où  $\operatorname{div}$  est l'opérateur divergence qui s'applique à une fonction vectorielle. Pour  $w : x = (x_1, x_2, x_3) \in \mathbb{R}^3 \mapsto w(x) = w_1(x_1, x_2, x_3) \in \mathbb{R}^3$ ,  $\operatorname{div} w = \partial_1 w_1 + \partial_2 w_2 + \partial_3 w_3$ , la notation  $\partial_i$  désignant la dérivée partielle de par rapport à  $x_i$ , et le symbole nabla  $\nabla$  (parfois aussi appelé del) représente le gradient :

$$\nabla u = \begin{pmatrix} \partial_1 u \\ \partial_2 u \\ \partial_3 u \end{pmatrix}.$$

<sup>1</sup>Leonhard Paul Euler, né le 15 avril 1707 à Bâle et mort le 18 septembre 1783 à Saint-Pétersbourg, est un mathématicien et physicien suisse, qui passa la plus grande partie de sa vie en Russie et en Allemagne. Euler fit d'importantes découvertes dans des domaines aussi variés que le calcul infinitésimal et la théorie des graphes. Il introduisit également une grande partie de la terminologie et de la notation des mathématiques modernes, en particulier pour l'analyse mathématique, comme pour la notion d'une fonction mathématique[2]. Il est également connu pour ses travaux en mécanique, en dynamique des fluides, en optique et en astronomie.

<sup>2</sup>David Hilbert (1862 – 1943) est un très grand mathématicien allemand du XXe siècle. Il est en particulier connu pour les 23 problèmes qu'il a énoncés comme défis aux mathématiciens. Certains de ces problèmes sont à ce jour non résolus.

Un **espace de Hilbert**  $H$  ou espace hilbertien est un espace vectoriel normé complet dont la norme, notée  $\|\cdot\|$ , découle d'un produit scalaire  $(\cdot, \cdot)_H$  : pour tout  $u \in H$ ,  $\|u\|_H^2 = (u, u)_H$ .

En une dimension, elle s'écrit  $(-\kappa u)' = f$ . Le réel  $\kappa$  est un coefficient de diffusion, et le terme  $-\kappa \nabla u$  est un "flux de diffusion". Il peut s'agir de la diffusion d'une espèce chimique, dans ce cas,  $\kappa$  est le coefficient de diffusion (souvent noté  $D$ , en  $m^2 s^{-1}$ ) de la loi de Fick qui donne le flux de diffusion  $J$  (quantité de matière par unité de surface et de temps) en fonction de la concentration  $u$  :

$$J = -\kappa u' \text{ en dimension 1, qui devient } J = -\kappa \nabla u \text{ en dimension supérieure.}$$

Il peut s'agir d'une diffusion thermique. Dans ce cas on parle plus souvent de conduction. Le coefficient de diffusion est dans ce cas souvent noté  $\lambda$ , et on l'appelle la conductivité thermique (en  $W m^{-1} K^{-1}$ ). La loi de Fourier (1807) donne la densité de flux de chaleur  $j$  (en  $W m^{-2}$ ) en fonction de la température  $u$  (en Kelvin) :

$$j = -\lambda u' \text{ en dimension 1, qui devient } J = -\lambda \nabla u \text{ en dimension supérieure.}$$

Dans le cas d'une conduction électrique, le coefficient de diffusion est dans ce cas noté  $\sigma$ , et on l'appelle la conductivité électrique (en  $W m^{-1} K^{-1}$ ). La loi d'Ohm donne la densité de courant électrique  $j$  (en  $W m^{-2}$ ) en fonction du potentiel électrique  $u$  :

$$j = -\sigma u' \text{ en dimension 1, qui devient } j = -\sigma \nabla u \text{ en dimension supérieure.}$$

Avec un coefficient constant  $\kappa = 1$ , l'équation de Poisson s'écrit en une dimension d'espace, :  $-u'' = f$ .

En deux dimensions d'espace, elle s'écrit :  $-\Delta u = f$ . avec  $\Delta u = \partial_{xx}^2 u + \partial_{yy}^2 u$ , où  $\partial_{xx}^2 u$  (resp.  $\partial_{yy}^2 u$ ) désigne la dérivée partielle seconde de  $u$  par rapport à  $x$  (respectivement  $y$ ).

Dans le cas  $f = 0$ , on obtient l'équation de Laplace  $-\Delta u = 0$ .

- L'équation de la chaleur :  $u_t - \Delta u = 0$ , où  $u_t$  désigne la dérivée partielle de  $u$  par rapport au temps  $t$  : la fonction  $u$  est ici une fonction du temps et de l'espace.
- L'équation de transport. En une dimension d'espace, elle s'écrit :  $u_t + cu_x = 0$ , où  $c$  est un réel (la vitesse de transport) et  $u_x$  désigne la dérivée partielle de  $u$  par rapport à la variable d'espace  $x$ . Si on se donne comme condition initiale  $u(x, 0) = u_0(x)$ , la solution de l'équation au temps  $t$  est  $u(x, t) = u_0(x - ct)$  (ceci est facile à vérifier au moins dans le cas régulier).
- L'équation des ondes  $u_{tt} - \Delta u = 0$ .

Considérons maintenant une équation aux dérivées partielles linéaire, de degré 2, de la forme :

$$Au_{xx} + Bu_{xy} + Cu_{yy} = 0$$

L'appellation "elliptique", "parabolique" ou "hyperbolique" d'une équation aux dérivées partielles de cette forme correspond à la nature de la conique décrite par l'équation caractéristique correspondante, c'est-à-dire :

$$Ax^2 + Bxy + Cy^2 = 0.$$

Si  $B^2 - 4AC < 0$ , l'équation est dite elliptique. Si  $B^2 - 4AC = 0$ , elle est parabolique, et si  $B^2 - 4AC > 0$ , elle est hyperbolique. Vérifiez que l'équation de Laplace est elliptique alors que l'équation des ondes est hyperbolique. Notons que tous les exemples que nous avons présentés sont des équations aux dérivées partielles linéaires, c.à.d. des équations qui ne font pas intervenir que des termes linéaires en  $u$  et ses dérivées. Bien sur, de nombreux modèles comportent des équations non linéaires : par exemple l'équation hyperbolique non linéaire de Burgers, qui s'écrit :

$$u_t + (u^2)_x = 0, t \in \mathbb{R}_+, x \in \mathbb{R},$$

avec la condition initiale  $u(x, 0) = u_0(x)$ . Une telle équation est dite "hyperbolique non linéaire". Les équations hyperboliques non linéaires sont discrétisées de manière usuelle par la méthode des volumes finis. Les discrétisations par éléments finis mènent à des schémas instables (c'est-à-dire que les solutions discrètes ne sont pas bornées, indépendamment des paramètres de discrétisation, par des normes "naturelles". En général elles ne vérifient pas non plus certaines propriétés qui semblent naturelles du point de vue de l'intuition physique. Ceci sera précisé dans la suite).

# Chapitre 1

## Différences finies et volumes finis pour les problèmes de diffusion stationnaires

### 1.1 Principe des deux méthodes

#### 1.1.1 Cas de la dimension 1

On considère le problème unidimensionnel

$$-u''(x) = f(x), \quad \forall x \in ]0, 1[, \quad (1.1)$$

$$u(0) = u(1) = 0, \quad (1.2)$$

où  $f \in C([0, 1])$ . Les conditions aux limites (1.2) considérées ici sont dites de type Dirichlet<sup>1</sup> homogène (le terme homogène désigne les conditions nulles). Cette équation modélise par exemple la diffusion de la chaleur dans un barreau conducteur chauffé (terme source  $f$ ) dont les deux extrémités sont plongées dans de la glace.

#### Méthode de différences finies.

On se donne une subdivision de  $[0, 1]$ , c'est à dire une suite de points  $(x_k)_{k=0, \dots, N+1}$  tels que  $x_0 = 0 < x_1 < x_2 < \dots < x_N < x_{N+1} = 1$ .

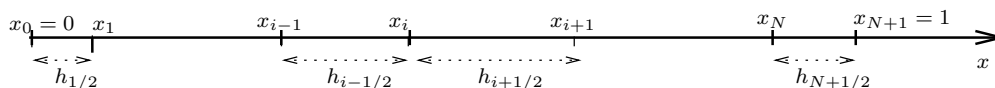


FIG. 1.1 – Maillage différences finies en 1D

Pour  $i = 0, \dots, N$ , on note  $h_{i+1/2} = x_{i+1} - x_i$  et on définit le "pas" du maillage par :

$$h = \max_{i=0, \dots, N} h_{i+1/2}. \quad (1.3)$$

Pour simplifier l'exposé, on se limitera dans un premier temps à un pas constant :

$$h_{i+1/2} = h \quad \forall i \in [0, N].$$

Le principe de la méthode des différences finies consiste à écrire l'équation aux dérivées partielles (1.1) aux points de discrétisation  $x_i$  :

$$-u''(x_i) = f(x_i), \quad \forall i = 1, \dots, N,$$

<sup>1</sup>Johann Peter Gustav Dirichlet, mathématicien allemand, né à Dürren en 1805 et mort à Göttingen en 1859. Il a effectué ses études supérieures à Paris où il a cotoyé les plus grands mathématiciens français de l'époque, dont Legendre, Laplace, Poisson et Fourier. Il retourne ensuite en 1825 en Allemagne où il travaille en particulier avec son ami Jacobi et avec Gauss, dont il reprendra en la chaire à l'Université de Göttingen. Il eut entre autres comme élève Riemann et Kronecker (qui a donné son nom au fameux symbole). Les travaux de Dirichlet ont surtout porté sur les séries de Fourier et l'arithmétique.



puis à approcher l'opérateur différentiel (ici  $-u''$ ) par un quotient différentiel, de manière à en déduire un système d'équations en fonction d'inconnues discrètes sensées représenter des approximations de  $u$  aux points de discrétisation. Voici comment on procède pour l'équation de Poisson unidimensionnelle. Effectuons d'abord un développement de Taylor en  $x_i$ , en supposant que  $u \in C^4([0, 1])$  :

$$u(x_{i+1}) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\zeta_i),$$

$$u(x_{i-1}) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i),$$

avec  $\zeta_i \in [x_i, x_{i+1}]$ ,  $\eta_i \in [x_{i-1}, x_i]$ . En additionnant, on obtient :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2u''(x_i) + O(h^2)$$

Il semble donc raisonnable d'approcher la dérivée seconde  $-u''(x_i)$  par le "quotient différentiel"

$$\frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1}))}{h^2}.$$

Sous des hypothèses de régularité sur  $u$ , on peut montrer (voir lemme 1.13 page 15) que cette approximation est d'ordre 2 au sens

$$R_i = u''(x_i) + \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1}))}{h^2} = O(h^2).$$

On appelle **erreur de consistence** au point  $x_i$  la quantité  $R_i$ . L'approximation de  $u''(x_i)$  par un quotient différentiel suggère de considérer les équations discrètes suivantes :

$$\frac{2u_i - u_{i-1} - u_{i+1}}{h^2} = f(x_i), i = 1, \dots, N.$$

dont les inconnues discrètes sont les  $u_i, i = 1, \dots, N$ . Notons que la première équation fait intervenir  $u_0$  tandis que la dernière fait intervenir  $u_{N+1}$ . Ces valeurs ne sont pas à proprement parler des inconnues, puisqu'elles sont données par les conditions aux limites (1.2). On pose donc  $u_0 = 0$  et  $u_{N+1} = 0$ . Le système complet d'équations s'écrit donc

$$\frac{2u_i - u_{i-1} - u_{i+1}}{h^2} = f(x_i), i = 1, \dots, N. \quad (1.4a)$$

$$u_0 = 0; u_{N+1} = 0. \quad (1.4b)$$

**Remarque 1.1 (Inconnues discrètes et solution exacte) Attention à ne pas confondre  $u_i$  et  $u(x_i)$  :** les équations discrètes (1.4) font intervenir les **inconnues** discrètes  $u_i, i = 1, \dots, N$ , et non pas les **valeurs**  $u(x_i), i = 1, \dots, N$  de la solution exacte. En général, ces valeurs ne sont pas les mêmes. SI la discrétisation a été effectuée correctement (comme c'est le cas ici, et comme nous le démontrerons mathématiquement plus loin), la résolution du système discret nous permettra d'obtenir des valeurs  $u_i, i = 1, \dots, N$  des inconnues discrètes qui seront des bonnes approximations des valeurs  $u(x_i), i = 1, \dots, N$  de la solution exacte que nous ne pouvons pas, dans le cas général, calculer explicitement.

### Méthode des volumes finis.

On ne se donne plus des points mais des volumes de contrôle  $K_i, i = 1, \dots, N$ , avec  $K_i = ]x_{i-1/2}, x_{i+1/2}[$ , et on note  $h_i = x_{i+1/2} - x_{i-1/2}$ . Pour chaque volume de contrôle  $K_i$ , on se donne un point  $x_i \in K_i = ]x_{i-1/2}, x_{i+1/2}[$ . On pourra considérer par exemple (mais ce n'est pas le seul point possible) :  $x_i = 1/2(x_{i+1/2} + x_{i-1/2})$ . On

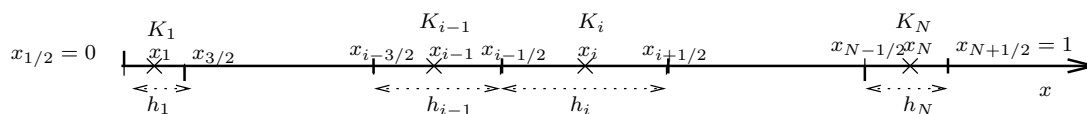


FIG. 1.2 – Maillage volumes finis en 1D

intègre l'équation  $-u'' = f$  sur  $K_i$  :

$$\int_{x_{i-1/2}}^{x_{i+1/2}} -u''(x)dx = \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)dx$$

et on pose  $f_i = \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)dx$ . On obtient :

$$-u'(x_{i+1/2}) + u'(x_{i-1/2}) = h_i f_i, \quad i = 1, \dots, N, \quad (1.5)$$

Cette équation est un bilan de flux. La quantité  $\bar{F}_{i+1/2} = -u'(x_{i+1/2})$  est le flux de diffusion en  $x_{i+1/2}$ . Pour la première maille ( $i = 1$ ), on obtient plus particulièrement :

$$-u'(x_{3/2}) + u'(0) = h_1 f_1, \quad (1.6)$$

et pour la dernière ( $i = N$ ) :

$$-u'(1) + u'(x_{N-1/2}) = h_N f_N. \quad (1.7)$$

On cherche donc à approcher les flux  $-u'(x_{i+1/2})$  aux interfaces  $x_{i+1/2}$  des mailles, et les flux  $u'(0)$  et  $u'(1)$  au bord. Notons que l'opérateur à approcher est ici d'ordre 1, alors qu'il était d'ordre 2 en différences finies pour la même équation.

On se donne une inconnue par maille (ou volume de contrôle  $i$ ), qu'on note  $u_i$ , et on espère approcher ainsi la valeur  $u(x_i)$  (ou  $\frac{1}{h_i} \int_{K_i} u$ ). En supposant  $u$  suffisamment régulière, on peut effectuer deux développements de Taylor à l'ordre 2 de  $u$  entre  $x_{i+1}$  et  $x_{i+1/2}$  et entre  $x_i$  et  $x_{i+1/2}$ ; en soustrayant ces développements de Taylor l'un de l'autre, on se rend compte qu'il est "raisonnable" (voir exercice 13 page 35) d'approcher le terme  $u'(x_{i+1/2})$  dans l'équation (1.5) par le quotient différentiel

$$\frac{u(x_{i+1}) - u(x_i)}{h_{i+1/2}},$$

au sens où l'erreur de consistance sur les flux, définie par :

$$R_{i+1/2} = u'(x_{i+1/2}) - \frac{u(x_{i+1}) - u(x_i)}{h_{i+1/2}}$$

est d'ordre 1 si  $u \in C^2([0, 1], \mathbb{R})$  (ceci est démontré plus loin, voir lemme 1.28 page 22). Le schéma numérique s'écrit donc :

$$-\frac{u_{i+1} - u_i}{h_{i+1/2}} + \frac{u_i - u_{i-1}}{h_{i-1/2}} = h_i f_i, \quad i = 2, \dots, N-1. \quad (1.8)$$

Pour la première et  $N$ -ième équations, on tient compte des conditions aux limites de Dirichlet homogènes (1.2), et on approche  $u'(0)$  dans l'équation (1.6) (resp.  $u'(1)$  dans l'équation (1.7)) par  $\frac{u(x_1) - 0}{h_{1/2}}$  (resp.  $\frac{0 - u(x_N)}{h_{N+1/2}}$ ), ce qui donne comme première et dernière équations du schéma numérique :

$$-\frac{u_2 - u_1}{h_{3/2}} + \frac{u_1}{h_{1/2}} = h_1 f_1, \quad (1.9)$$

$$\frac{u_N}{h_{N+1/2}} + \frac{u_N - u_{N-1}}{h_{N-1/2}} = h_N f_N, \quad (1.10)$$

Là encore comme dans le cas des différences finies, attention : les équations discrètes (1.8)-(1.10) font intervenir les **inconnues** discrètes  $u_i, i = 1, \dots, N$ , et non pas les **valeurs**  $u(x_i), i = 1, \dots, N$  de la solution exacte. En général, ces valeurs ne sont pas les mêmes.

### Autres conditions limites.

**Conditions de Dirichlet non homogènes** Supposons que les conditions aux limites en 0 et en 1 soit maintenant de type Dirichlet non homogènes, c.à.d. :

$$u(0) = a, \quad u(1) = b, \quad (1.11)$$

avec  $a$  et  $b$  pas forcément nuls. Dans ce cas :

1. les équations discrètes du schéma aux différences finies (1.4) restent identiques, mais les valeurs  $u_0$  et  $u_{N+1}$  sont maintenant données par :  $u_0 = a$  et  $u_{N+1} = b$ ;

2. les équations discrètes du *schéma de volumes finis* (1.8) associés aux noeuds internes restent identiques, mais les valeurs  $u'(0)$  et  $u'(1)$  sont maintenant approchés par  $\frac{u(x_1)-a}{h_{1/2}}$  (resp.  $\frac{b-u(x_N)}{h_{N+1/2}}$ ), ce qui donne comme première et dernière équations du schéma numérique :

$$-\frac{u_2 - u_1}{h_{3/2}} + \frac{u_1 - a}{h_{1/2}} = h_1 f_1, \quad (1.12)$$

$$-\frac{b - u_N}{h_{N+1/2}} + \frac{u_N - u_{N-1}}{h_{N-1/2}} = h_N f_N, \quad (1.13)$$

**Conditions de Neumann et Fourier** On appelle *condition de Neumann*<sup>2</sup> une condition qui impose une valeur de la dérivée, par exemple :

$$u'(0) = a. \quad (1.14)$$

On appelle *condition de Fourier*<sup>3</sup> ou *condition de Robin*<sup>4</sup> une condition<sup>5</sup> qui impose une relation entre la valeur de la dérivée et la valeur de la solution, par exemple,

$$u'(1) + \alpha u(1) = b, \quad (1.15)$$

avec  $\alpha > 0$ . Cette condition est donc un mélange des conditions de Dirichlet et de Neumann, qui est souvent utilisée pour exprimer une condition de transfert (thermique par exemple) entre un milieu et l'extérieur.

Enfin, on dit que les conditions aux limites sont *mixtes* si elles sont de type différent sur des portions de frontière du domaine : on a des conditions mixtes dans le cas unidimensionnel si, par exemple, on a une condition de Dirichlet en 0 et une condition de Neumann en 1.

Prenons par exemple le cas de conditions mixtes, en considérant l'équation (1.1) en  $x = 0$  avec les conditions (1.14) et (1.15) en  $x = 1$ . Voyons comment tenir compte de ces nouvelles conditions limites avec les méthodes DF et VF.

1. *Schéma aux différences finies* Pour approcher la condition de Neumann en 0, on effectue un développement de Taylor en 0 à l'ordre 1 :

$$u'(0) = \frac{u(h) - u(0)}{h_{1/2}} + \varepsilon(h) = a.$$

Ceci suggère l'équation discrète suivante pour  $u_0$  en écrivant que :

$$\frac{u_1 - u_0}{h_{1/2}} = a \text{ c.à.d. } u_0 = u_1 - ah_{1/2}.$$

(Rappelons que dans le cas de la condition de Dirichlet homogène  $u(0) = 0$ , la valeur de  $u_0$  était simplement prise comme  $u_0 = 0$ .)

De la même manière, on écrit un développement limité pour la dérivée dans la condition de Fourier (1.15) :

$$\frac{u(1) - u(1-h)}{h_{N+1/2}} + \varepsilon(h) + \alpha u(1) = b,$$

ce qui suggère l'approximation suivante :

$$\frac{u_{N+1} - u_N}{h_{N+1/2}} + \alpha u_{N+1} = b \text{ c.à.d. } u_{N+1} = \frac{u_N + bh_{N+1/2}}{1 + \alpha h_{N+1/2}}.$$

<sup>2</sup>Karl Gottfried Neumann est un mathématicien allemand, né en 1832 à Königsberg et mort 1925 à Leipzig. Il fut l'un des pionniers de la théorie des équations intégrales. Il a laissé son nom aux conditions aux limites que nous mentionnons ici.

<sup>3</sup>Joseph Fourier est un mathématicien et physicien français, né en 1768 à Auxerre et mort en 1830 à Paris. Il est connu pour ses travaux sur la décomposition de fonctions périodiques en séries trigonométriques convergentes appelées séries de Fourier et leur application au problème de la propagation de la chaleur. Il a participé à la révolution, a échappé de peu à la guillotine, et a été nommé préfet de l'Isère par Napoléon. Il a fait construire la route entre Grenoble et Briançon, et fondé en 1810 l'Université Royale de Grenoble, dont il fut le recteur. L'université scientifique de Grenoble et l'un des laboratoires de mathématiques de cette université portent son nom.

<sup>4</sup>Victor Gustave Robin est un mathématicien français né en 1855 et mort en 1897 qui a travaillé en particulier en thermodynamique et sur la théorie du potentiel.

<sup>5</sup>Voir pour plus de précisions à ce sujet l'article *The third boundary condition—was it Robin's ?* par Karl Gustafson and Takehisa Abe, paru en 1998 dans "The Mathematical Intelligencer", Springer.

2. *Schéma de volumes finis* : La condition de Neumann est particulièrement simple à prendre en compte, puisque le schéma de volumes finis fait intervenir l'approximation du flux en 0,  $u'(0)$ , dans l'équation (1.6), que l'on discrétise donc par :

$$a + \frac{u_1 - u_2}{h_{3/2}} = h_1 f_1. \quad (1.16)$$

On tient compte ensuite de la condition de Fourier (1.15) pour approcher le terme  $u'(1)$  dans l'équation (1.6) : on peut par exemple<sup>6</sup> approcher  $u'(1)$  par  $b - \alpha u_N$  ce qui nous donne comme  $N$ -ième équation discrète :

$$F_{N+1/2} - F_{N-1/2} = h_N f_N \text{ avec } F_{N+1/2} = \alpha u_N - b \text{ et } F_{N-1/2} = -\frac{u_N - u_{N-1}}{h_{N-1/2}} \quad (1.17)$$

### 1.1.2 Cas de la dimension 2 ou 3

On considère maintenant le problème de Laplace en dimension 2 ou 3, sur un ouvert borné  $\Omega$  de  $\mathbb{R}^d$ ,  $d = 2$  ou  $3$ , avec conditions aux limites de Dirichlet homogènes :

$$-\Delta u = f, \text{ sur } \Omega, \quad (1.18a)$$

$$u(x) = 0, \text{ sur } \partial\Omega, \quad (1.18b)$$

où  $f$  est une fonction de  $\Omega$  dans  $\mathbb{R}$ .

#### Méthode de différences finies.

Supposons (pour simplifier) que le domaine  $\Omega$  soit un carré (c.à.d.  $d = 2$ , le cas rectangulaire se traite tout aussi facilement). On se donne un pas de maillage constant  $h$  et des points  $x_{i,j} = (ih, jh)$ ,  $i = 1, \dots, N$ ,  $j = 1, \dots, N$ . En effectuant les développements limités de Taylor (comme au paragraphe 1.1.1 page 6) dans les deux directions (voir exercice 18), on approche  $-\partial_i^2 u(x_{i,j})$  (resp.  $-\partial_j^2 u(x_{i,j})$ ) par

$$\frac{2u(x_{i,j}) - u(x_{i+1,j}) - u(x_{i-1,j}))}{h^2} \text{ (resp. par } \frac{2u(x_{i,j}) - u(x_{i,j+1}) - u(x_{i,j-1}))}{h^2} \text{)}.$$

Ce type d'approche est limité à des géométries simples. Pour mailler des géométries compliquées, on utilise souvent des triangles (tétraèdres en dimension 3), auquel cas la méthode des différences finies est plus difficile à généraliser, car on ne peut pas approcher la dérivée seconde comme en maillages cartésiens.

#### Méthode de volumes finis.

On suppose maintenant que  $\Omega$  est un ouvert polygonal de  $\mathbb{R}^2$ , et on se donne un maillage  $\mathcal{T}$  de  $\Omega$ , c.à.d., en gros, un découpage de  $\Omega$  en volumes de contrôle polygonaux  $K$ . En intégrant l'équation (1.18a) sur  $K$ , on obtient :

$$\int_K -\Delta u \, dx = \int_K f \, dx.$$

Par la formule de Stokes, on peut réécrire cette équation :

$$-\int_{\partial K} \nabla u(x) \cdot \mathbf{n}_K(x) d\gamma(x) = \int_K f(x) dx,$$

où  $d\gamma(x)$  désigne l'intégrale par rapport à la mesure uni-dimensionnelle sur le bord de l'ouvert  $\Omega$ , et où  $\mathbf{n}_K$  désigne le vecteur normal unitaire à  $\partial K$  extérieur à  $K$ . Comme  $K$  est polygonal, on peut décomposer  $\partial K$  en arêtes  $\sigma$  qui sont des segments de droite, et en appelant  $\mathcal{E}_K$  l'ensemble des arêtes de  $\partial K$  (trois arêtes dans le cas d'un triangle), on a donc :

$$-\sum_{\sigma \in \mathcal{E}_K} \int_{\sigma} \nabla u \cdot \mathbf{n}_{K,\sigma} d\gamma(x) = \int_K f(x) dx,$$

où  $\mathbf{n}_{K,\sigma}$  désigne le vecteur normal unitaire à  $\sigma$  extérieur à  $K$  (noter que ce vecteur est constant sur  $\sigma$ ). On cherche donc maintenant à approcher la dérivée normale  $\nabla u \cdot \mathbf{n}_{K,\sigma}$  de manière consistante sur chaque arête  $\sigma$ . On se donne

<sup>6</sup>Ce n'est pas la seule possibilité, voir exercice 12.

des inconnues discrètes notées  $(u_K)_{K \in \mathcal{T}}$ , qui, on l'espère vont s'avérer être des approximations de  $u(x_K)$ . Pour une arête  $\sigma = K|L$  séparant les volumes de contrôle  $K$  et  $L$ , il est tentant d'approcher la dérivée normale  $\nabla u \cdot \mathbf{n}_{K,\sigma}$  par le quotient différentiel

$$\frac{u(x_L) - u(x_K)}{d_{K,L}},$$

où  $d_{K,L}$  est la distance entre les points  $x_K$  et  $x_L$ . Cependant, cette approximation ne pourra être justifiée que si la direction du vecteur défini par les deux points  $x_K$  et  $x_L$  est la même que celle de la normale  $\mathbf{n}_{K,\sigma}$ , c.à.d. si le segment de droite  $x_K x_L$  est orthogonal à l'arête  $K|L$ . Pour un maillage triangulaire à angles strictement inférieurs à  $\pi/2$ , ceci est facile à obtenir en choisissant les points  $x_K$  comme intersection des médiatrices du triangle  $K^7$ , voir Figure 1.3.

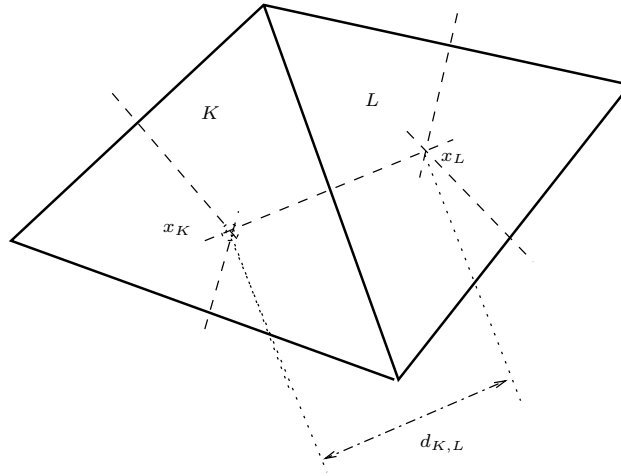


FIG. 1.3 – Exemple de volumes de contrôle pour la méthode des volumes finis en deux dimensions d'espace

Supposons que cette hypothèse, dite d'orthogonalité du maillage, soit satisfaite ; on approche donc  $\nabla u \cdot \mathbf{n}_K|_\sigma$  par  $\frac{u(x_L) - u(x_K)}{d_{K,L}}$  et en notant  $|\sigma|$  la longueur de l'arête  $\sigma$ , on approche :

$$\int_\sigma \nabla u \cdot \mathbf{n}_K d\gamma \text{ par } F_{K,\sigma} = |\sigma| \frac{u_L - u_K}{d_{K,L}}, \text{ pour tout } \sigma \in \mathcal{E}_K \text{ et pour tout } K \in \mathcal{T}.$$

Le schéma volumes finis s'écrit donc

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = |K| f_K, \quad (1.19)$$

où  $|K|$  est la mesure de  $K$ , et  $f_K = \frac{1}{|K|} \int_K f(x) dx$ , et où les flux numériques  $F_{K,\sigma}$  sont définis (en tenant compte des conditions limites pour les arêtes du bord) par :

$$F_{K,\sigma} = \begin{cases} -|\sigma| \frac{u_L - u_K}{d_{K,L}} & \text{si } \sigma = K|L, \\ -|\sigma| \frac{u_K}{d_{K,\sigma}} & \text{si } \sigma \subset \partial\Omega \text{ et } \sigma \in \mathcal{E}_K, \end{cases} \quad (1.20)$$

où  $d_{K,\sigma}$  est la distance entre le point  $x_K$  et l'arête  $\sigma$

### Comparaison des méthodes

Cette introduction aux différences finies et volumes finis nous permet de remarquer que les différences finies sont particulièrement bien adaptées dans le cas de domaines rectangulaires ou parallélépipédiques, pour lesquels on peut facilement définir des maillages structurés (cartésiens dans le cas présent) c.à.d. dont on peut indexer les mailles par un ordre  $(i, j)$  naturel.

<sup>7</sup>On rappelle que les médiatrices d'un triangle se coupent en un point qui est le centre du cercle circonscrit au triangle, alors que les médianes se coupent au barycentre, qui est le centre du cercle inscrit dans le triangle ; ces deux points coïncident dans le cas d'un triangle équilatéral.

Dans le cas de domaines plus complexes, on maille souvent à l'aide de triangles (ou tétraèdres) et dans ce cas la méthode des différences finies ne se généralise pas facilement. On a alors recours soit aux volumes finis, dont on vient de donner le principe, soit aux éléments finis, que nous aborderons ultérieurement.

### 1.1.3 Questions d'analyse numérique

Voici un certain nombre de questions, qui sont typiquement du domaine de l'analyse numérique, auxquelles nous tenterons de répondre dans la suite :

1. Le problème qu'on a obtenu en dimension finie, (avec des inconnues localisées aux noeuds du maillage dans le cas de la méthode des différences finies et dans les mailles dans le cas de la méthode des volumes finis) admet-il une (unique) solution ?
2. La solution du problème discret satisfait-elle les propriétés physiques qui sont vérifiées par la solution du modèle mathématique ?
3. La solution du problème discret converge-t-elle vers la solution du problème continu lorsque le pas du maillage  $h$  tend vers 0 ? Dans le cas des différences finies en une dimension d'espace, le pas du maillage est défini par

$$h = \sup_{i=1, \dots, N} |x_{i+1} - x_i|. \quad (1.21)$$

Dans le cas des volumes finis en une dimension d'espace, il est défini par :

$$h = \sup_{i=1, \dots, N} |x_{i+1/2} - x_{i-1/2}|. \quad (1.22)$$

en deux dimensions d'espace, le pas  $h$  est défini par

$$h = \sup_{K \in \mathcal{T}} \text{diam}(K), \text{ avec } \text{diam}(K) = \sup_{x, y \in K} d(x, y),$$

où  $\mathcal{T}$ , le maillage, est l'ensemble des volumes de contrôle  $K$ . Notons que la réponse à cette question n'est pas évidente *a priori*. La solution discrète peut converger vers la solution continue, elle peut aussi converger mais vers autre chose que la solution du problème continu, et enfin elle peut ne pas converger du tout.

## 1.2 Analyse de la méthode des différences finies

On cherche à discrétiser le problème aux limites suivant :

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & 0 < x < 1, \\ u(0) = u(1) = 0, \end{cases} \quad (1.23)$$

où  $c \in C([0, 1], \mathbb{R}_+)$  et  $f \in C([0, 1], \mathbb{R})$ , qui peut modéliser par exemple un phénomène de diffusion - réaction d'une espèce chimique. On se donne un pas du maillage constant  $h = \frac{1}{N+1}$ , et une subdivision de  $]0, 1[$ , notée  $(x_k)_{k=0, \dots, N+1}$ , avec :  $x_0 = 0 < x_1 < x_2 < \dots < x_N < x_{N+1} = 1$ . Soit  $u_i$  l'inconnue discrète associée au noeud  $i$  ( $i = 1, \dots, N$ ). On pose  $u_0 = u_{N+1} = 0$ . On obtient les équations discrètes en approchant  $u''(x_i)$  par quotient différentiel par développement de Taylor, comme on l'a vu au paragraphe 1.1.1 page 6. On obtient donc le système suivant :

$$\begin{cases} \frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, & i = 1, \dots, N, \\ u_0 = u_{N+1} = 0. \end{cases} \quad (1.24)$$

avec  $c_i = c(x_i)$  et  $f_i = f(x_i)$ . On peut écrire ces équations sous forme matricielle :

$$A_h U_h = b_h, \text{ avec } U_h = \begin{pmatrix} u_1 \\ \vdots \\ u_N \end{pmatrix}, \quad b_h = \begin{pmatrix} f_1 \\ \vdots \\ f_N \end{pmatrix}, \quad (1.25)$$

$$\text{et } A_h = \frac{1}{h^2} \begin{pmatrix} 2 + c_1 h^2 & -1 & 0 & \dots & 0 \\ -1 & 2 + c_2 h^2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 + c_{N-1} h^2 & -1 \\ 0 & \dots & 0 & -1 & 2 + c_N h^2 \end{pmatrix}. \quad (1.26)$$

**Remarque 1.2 (Notations pour les vecteurs et matrices)** Un vecteur  $u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}$  sera aussi noté, par souci de

simplicité,  $u = (u_1, u_2, \dots, u_N)$  (ces deux égalités signifiant que les composantes de  $u$  dans la base canonique de  $\mathbb{R}^N$  sont  $u_1, u_2, \dots, u_N$ ). Attention toutefois à ne pas confondre cette notation avec  $u^t = (u_1 \ u_2 \ \dots \ u_N)$ , qui est une matrice  $1 \times N$ ; c'est la matrice transposée de  $u$  vu comme une matrice  $N \times 1$ . On peut écrire le produit scalaire de deux vecteurs  $u$  et  $v$  de  $\mathbb{R}^N$  avec ces notations :

$$u \cdot v = \sum_{i=1}^N u_i v_i = u^t v = v^t u.$$

Les questions suivantes surgissent alors naturellement :

1. Le système (1.25) admet-il un unique solution ?
2. A-t-on convergence de  $U_h$  vers  $u$  et en quel sens ?

Nous allons répondre par l'affirmative à ces deux questions. Commençons par la première.

**Proposition 1.3** Soit  $c = (c_1, \dots, c_N) \in \mathbb{R}^N$  tel que  $c_i \geq 0$  pour  $i = 1, \dots, N$ ; alors la matrice  $A_h$  définie par (1.26) est symétrique définie positive, et donc inversible.

**Démonstration :** La matrice  $A_h$  est évidemment symétrique. Montrons qu'elle est définie positive. Soit  $v = (v_1 \dots v_N)$ , on pose  $v_0 = v_{N+1} = 0$ . Calculons le produit scalaire  $A_h v \cdot v = v^t A_h v$ . On a :

$$A_h v \cdot v = \frac{1}{h^2} (v_1 \ v_2 \ \dots \ v_N) \begin{pmatrix} 2 + c_1 h^2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 + c_N h^2 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ \vdots \\ v_N \end{pmatrix},$$

c'est-à-dire :

$$A_h v \cdot v = \frac{1}{h^2} \sum_{i=1}^N v_i (-v_{i-1} + (2 + c_i h^2)v_i - v_{i+1}).$$

On a donc, par changement d'indice :

$$A_h v \cdot v = \frac{1}{h^2} \left[ \sum_{i=1}^N (-v_{i-1} v_i) + \sum_{i=1}^N (2 + c_i h^2) v_i^2 - \sum_{j=2}^{N+1} v_{j-1} v_j \right].$$

Et comme on a posé  $v_0 = 0$  et  $v_{N+1} = 0$ , on peut écrire :

$$A_h v \cdot v = \frac{1}{h^2} \sum_{i=1}^N (2 + c_i h^2) v_i^2 + \frac{1}{h^2} \sum_{i=1}^N (-2v_i v_{i-1}),$$

soit encore :

$$A_h v \cdot v = \sum_{i=1}^N c_i v_i^2 + \frac{1}{h^2} \sum_{i=1}^N (-2v_i v_{i-1} + v_i^2 + v_{i-1}^2) + v_N^2.$$

On a donc finalement :

$$A_h v \cdot v = \sum_{i=1}^N c_i v_i^2 + \frac{1}{h^2} \sum_{i=1}^{N+1} (v_i - v_{i-1})^2 \geq 0, \quad \forall v = (v_1, \dots, v_N) \in \mathbb{R}^N.$$

Si on suppose  $A_h v \cdot v = 0$ , on a alors

$$\sum_{i=1}^N c_i h^2 v_i^2 = 0 \text{ et } v_i - v_{i-1} = 0, \quad \forall i = 1, \dots, N+1.$$

On a donc  $v_1 = v_2 = \dots = v_N = v_0 = v_{N+1} = 0$ . Remarquons que ces égalités sont vérifiées même si les  $c_i$  sont nuls. Ceci démontre que la matrice  $A_h$  est bien définie.

**Remarque 1.4 (Existence et unicité de la solution)** *On a montré ci-dessus que  $A_h$  est symétrique définie positive, donc inversible, ce qui entraîne l'existence et l'unicité de la solution de (1.25). On aurait pu aussi démontrer l'existence et l'unicité de la solution de (1.25) directement, en montrant que le noyau de  $(A_h)$  est réduit à  $\{0\}$  (voir exercice 7 page 32). On rappelle qu'en dimension finie, toute application linéaire injective ou surjective est bijective.*

**Remarque 1.5 (Caractère défini et conditions limites)** *Dans la démonstration de la proposition 1.3, si  $c_i > 0$  pour tout  $i = 1, \dots, N$  le terme  $\sum_{i=1}^N c_i h^2 v_i^2 = 0$  permet de conclure que  $v_i = 0$  pour tout  $i = 1, \dots, N$ . Par contre, si on n'a que  $c_i \geq 0$  (ou, bien même  $c_i = 0$  pour tout  $i = 1, \dots, N$ ), on peut encore montrer que  $v_i = 0$ , pour tout  $i = 1, \dots, N$  grâce aux conditions aux limites de Dirichlet homogènes (représentées par le fait qu'on pose  $v_0 = 0$  et  $v_{N+1} = 0$  qui permet d'écrire alors les équations 1 et  $N$  sous la même forme que l'équation  $i$ ); on a en effet  $v_i = v_{i-1}$ , pour tout  $i = 1, \dots, N$ , et  $v_0 = 0$ . En particulier, la matrice de discrétisation de  $-u''$  par différences finies avec conditions aux limites de Neumann homogènes :*

$$\begin{cases} -u'' = f, \\ u'(0) = u'(1) = 0. \end{cases} \quad (1.27)$$

donne une matrice  $A_h$  qui est symétrique et positive, mais non définie (voir exercice 15 page 35). De fait la solution du problème continu (1.27) n'est pas unique, puisque les fonctions constantes sur  $[0, 1]$  sont solutions de (1.27).

Nous allons maintenant nous préoccuper de la question de la convergence.

**Définition 1.6 (Matrices monotones)** *Soit  $A \in \mathcal{M}_N(\mathbb{R})$ , de coefficients  $a_{i,j}$ ,  $i = 1, \dots, N$  et  $j = 1, \dots, N$ . On dit que  $A$  est positive (ou  $A \geq 0$ ) si  $a_{i,j} \geq 0$ ,  $\forall i, j = 1, \dots, N$ . On dit que  $A$  est monotone (ou que  $A$  est une IP-matrice) si  $A$  est inversible et  $A^{-1} \geq 0$ ; voir à ce propos les exercices sur les IP-matrices et les M-matrices du polycopié de L3, à l'adresse <http://www.cmi.univ-mrs.fr/~herbin/PUBLI/anamat.pdf>.*

L'avantage des schémas à matrices monotones est de satisfaire la propriété de conservation de la positivité, qui peut être cruciale dans les applications physiques :

**Définition 1.7 (Conservation de la positivité)** *Soit  $A \in \mathcal{M}_N(\mathbb{R})$ , de coefficients  $a_{i,j}$ ,  $i = 1, \dots, N$  et  $j = 1, \dots, N$ ; on dit que  $A$  conserve la positivité si  $Av \geq 0$  entraîne  $v \geq 0$  (les inégalités s'entendent composante par composante).*

On a en effet la proposition suivante :

**Proposition 1.8 (Monotonie et positivité)** *Soit  $A \in \mathcal{M}_N(\mathbb{R})$ . Alors  $A$  conserve la positivité si et seulement si  $A$  est monotone.*

**Démonstration :** Supposons d'abord que  $A$  conserve la positivité, et montrons que  $A$  inversible et que  $A^{-1}$  a des coefficients  $\geq 0$ . Si  $x$  est tel que  $Ax = 0$ , alors  $Ax \geq 0$  et donc, par hypothèse,  $x \geq 0$ . Mais on a aussi  $Ax \leq 0$ , soit  $A(-x) \geq 0$  et donc par hypothèse,  $x \leq 0$ . On en déduit  $x = 0$ , ce qui prouve que  $A$  est inversible. La conservation de la positivité donne alors que  $y \geq 0 \Rightarrow A^{-1}y \geq 0$ . En prenant  $y = e_1$  on obtient que la première colonne de  $A^{-1}$  est positive, puis en prenant  $y = e_i$  on obtient que la  $i$ -ème colonne de  $A^{-1}$  est positive, pour  $i = 2, \dots, N$ . Donc  $A^{-1}$  a tous ses coefficients positifs.

Réciproquement, supposons maintenant que  $A$  est inversible et que  $A^{-1}$  a des coefficients positifs. Soit  $x \in \mathbb{R}^N$  tel que  $Ax = y \geq 0$ , alors  $x = A^{-1}y \geq 0$ . Donc  $A$  conserve la positivité. ■

**Remarque 1.9 (Principe du maximum)** *On appelle principe du maximum continu le fait que si  $f \geq 0$  alors le minimum de la fonction  $u$  solution du problème 1.23 est atteint sur les bords. Cette propriété mathématique correspond à l'intuition physique qu'on peut avoir du phénomène : si on chauffe un barreau tout en maintenant ses deux extrémités à une température fixe, la température aux points intérieurs du barreau sera supérieure à celle des extrémités. Il est donc souhaitable que la solution approchée satisfasse la même propriété (voir exercice 3 page 30 à ce sujet).*



**Lemme 1.10** Soit  $c = (c_1, \dots, c_N) \in \mathbb{R}^N$ , et  $A_h \in \mathcal{M}_N(\mathbb{R})$  définie par (1.26). Si  $c_i \geq 0$  pour tout  $i = 1, \dots, N$ , alors  $A_h$  est monotone.

**Démonstration :** On va montrer que si  $v \in \mathbb{R}^N$ ,  $A_h v \geq 0$  alors  $v \geq 0$ . On peut alors utiliser la proposition 1.8 pour conclure. Soit  $v = (v_1, \dots, v_N) \in \mathbb{R}^N$ . Posons  $v_0 = v_{N+1} = 0$ . Supposons que  $A_h v \geq 0$ . On a donc

$$-\frac{1}{h^2}v_{i-1} + \left(\frac{2}{h^2} + c_i\right)v_i - \frac{1}{h^2}v_{i+1} \geq 0, \quad i = 1, \dots, N \quad (1.28)$$

Soit

$$p = \min \left\{ i \in \{1, \dots, N\}; v_p = \min_{j=1, \dots, N} v_j \right\}.$$

Supposons que  $\min_{j=1, \dots, N} v_j < 0$ . On a alors  $p \geq 1$  et :

$$\frac{1}{h^2}(v_p - v_{p-1}) + c_p v_p + \frac{1}{h^2}(v_p - v_{p+1}) \geq 0.$$

On en déduit que

$$\frac{2}{h^2}c_p v_p \geq \frac{1}{h^2}(v_{p-1} - v_p) + \frac{1}{h^2}(v_{p+1} - v_p) \geq 0.$$

Si  $c_p > 0$ , on a donc  $v_p \geq 0$ , et donc  $v_i \geq 0$ ,  $\forall i = 1, \dots, N$ . Si  $c_p = 0$ , on doit alors avoir  $v_{p-1} = v_p = v_{p+1}$  ce qui est impossible car  $p$  est le plus petit indice  $j$  tel que  $v_j = \min_{i=1, \dots, N} v_i$ . Donc dans ce cas le minimum ne peut pas être atteint pour  $j = p > 1$ . On a ainsi finalement montré que  $\min_{i \in \{1, \dots, N\}} v_i \geq 0$ , on a donc  $v \geq 0$ . ■

**Définition 1.11 (Erreur de consistance)** On appelle erreur de consistance la quantité obtenue en remplaçant l'inconnue par la solution exacte dans le schéma numérique. Dans le cas du schéma (1.24), l'erreur de consistance au point  $x_i$  est donc définie par :

$$R_i = \frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1})) + c(x_i)u(x_i) - f(x_i). \quad (1.29)$$

L'erreur de consistance  $R_i$  est donc l'erreur qu'on commet en remplaçant l'opérateur  $-u''$  par le quotient différentiel

$$\frac{1}{h^2}(2u(x_i) - u(x_{i-1}) - u(x_{i+1})).$$

Cette erreur peut être évaluée si  $u$  est suffisamment régulière, en effectuant des développements de Taylor.

**Définition 1.12 (Ordre du schéma)** On dit qu'un schéma de discrétisation à  $N$  points de discrétisation est d'ordre  $p$  s'il existe  $C \in \mathbb{R}$ , ne dépendant que de la solution exacte, tel que l'erreur de consistance satisfait :

$$\max_{i=1, \dots, N} (R_i) \leq Ch^p,$$

où  $h$  est le pas du maillage défini par (1.3) (c.à.d. le maximum des écarts  $x_{i+1} - x_i$ ). On dit qu'un schéma de discrétisation est consistant si

$$\max_{i=1, \dots, N} (R_i) \rightarrow 0 \text{ lorsque } h \rightarrow 0,$$

où  $N$  est le nombre de points de discrétisation.

**Lemme 1.13** Si la solution de (1.23) vérifie  $u \in C^4([0, 1])$ , alors le schéma (1.24) est consistant d'ordre 2, et on a plus précisément :

$$|R_i| \leq \frac{h^2}{12} \sup_{[0,1]} |u^{(4)}|, \quad \forall i = 1, \dots, N. \quad (1.30)$$

**Démonstration :** Par développement de Taylor, on a :

$$\begin{aligned} u(x_{i+1}) &= u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\xi_i) \\ u(x_{i-1}) &= u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u'''(x_i) + \frac{h^4}{24}u^{(4)}(\eta_i) \end{aligned}$$

En additionnant ces deux égalités, on obtient que :

$$\frac{1}{h^2}(u(x_{i+1}) + u(x_i) - 2u(x_i)) = u''(x_i) + \frac{h^2}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)),$$

ce qui entraîne que :

$$|R_i| \leq \frac{h^2}{12} \sup_{[0,1]} |u^{(4)}|. \quad (1.31)$$

■

**Remarque 1.14 (Sur l'erreur de consistance)**

1. Si on note  $\bar{U}_h : (u(x_i))_{i=1 \dots N}$  le vecteur dont les composantes sont les valeurs exactes de la solution de (1.23), et  $U_h = (u_1 \dots u_N)$  la solution de (1.24), on a :

$$R = A_h(U_h - \bar{U}_h), \quad (1.32)$$

où  $R \in \mathbb{R}^N$  est le vecteur de composantes  $R_i$ ,  $i = 1, \dots, N$ , erreur de consistance en  $x_i$  définie en (1.29).

2. On peut remarquer que si  $u^{(4)} = 0$ , les développements de Taylor effectués ci-dessus se résument à :

$$-u''(x_i) = \frac{2u(x_i) - u(x_{i-1}) - u(x_{i+1}))}{h^2},$$

et on a donc  $R_i = 0$ , pour tout  $i = 1, \dots, N$ , et donc  $u_i = u(x_i)$ , pour tout  $i = 1 \dots N$ . Dans ce cas (rare !), le schéma de discrétisation donne la valeur exacte de la solution en  $x_i$ , pour tout  $i = 1, \dots, N$ . Cette remarque est bien utile lors de la phase de validation de méthodes numériques et/ou programmes informatiques pour la résolution de l'équation (1.23). En effet, si on choisit  $f$  telle que la solution soit un polynôme de degré inférieur ou égal à 3, alors on doit avoir une erreur entre solution exacte et approchée inférieure à l'erreur machine.

La preuve de convergence du schéma utilise la notion de consistance, ainsi qu'une notion de stabilité, que nous introduisons maintenant :

**Proposition 1.15** On dit que le schéma (1.24) est stable, au sens où la norme infinie de la solution approchée est bornée par un nombre ne dépendant que de  $f$ . Plus précisément, la matrice de discrétisation  $A_h$  satisfait :

$$\|A_h^{-1}\|_\infty \leq \frac{1}{8}. \quad (1.33)$$

inégalité qui peut aussi s'écrire comme une estimation sur les solutions du système (1.25) :

$$\|U_h\|_\infty \leq \frac{1}{8} \|f\|_\infty. \quad (1.34)$$

**Démonstration :** On rappelle que par définition, si  $M \in \mathcal{M}_N(\mathbb{R})$ ,

$$\|M\|_\infty = \sup_{\substack{v \in \mathbb{R}^N \\ v \neq 0}} \frac{\|Mv\|_\infty}{\|v\|_\infty}, \text{ avec } \|v\|_\infty = \sup_{i=1, \dots, N} |v_i|.$$

Pour montrer que  $\|A_h^{-1}\|_\infty \leq \frac{1}{8}$ , on décompose la matrice  $A_h$  sous la forme  $A_h = A_{0h} + \text{diag}(c_i)$  où  $A_{0h}$  est la matrice de discrétisation de l'opérateur  $-u''$  avec conditions aux limites de Dirichlet homogènes, et

$$A_{0h} = \begin{bmatrix} \frac{2}{h^2} & -\frac{1}{h^2} & & & 0 \\ & -\frac{1}{h^2} & & & \\ & & \ddots & & \\ & & & \ddots & -\frac{1}{h^2} \\ 0 & & & -\frac{1}{h^2} & \frac{1}{h^2} \end{bmatrix} \quad (1.35)$$

et  $\text{diag}(c_i)$  désigne la matrice diagonale de coefficients diagonaux  $c_i$ . Les matrices  $A_{0h}$  et  $A_h$  sont inversibles, et on a :

$$A_{0h}^{-1} - A_h^{-1} = A_{0h}^{-1} A_h A_h^{-1} - A_{0h}^{-1} A_{0h} A_h^{-1} = A_{0h}^{-1} (A_h - A_{0h}) A_h^{-1}.$$

Comme  $\text{diag}(c_i) \geq 0$ , on a  $A_h \geq A_{0h}$ , et comme  $A_{0h}$  et  $A_h$  sont monotones, on en déduit que :

$$0 \leq A_h^{-1} \leq A_{0h}^{-1}, \text{ (composante par composante).}$$

On peut maintenant remarquer que si  $B \in \mathcal{M}_N(\mathbb{R})$ , et si  $B \geq 0$  (c.à.d.  $B_{ij} \geq 0$  pour tout  $i$  et  $j$ ), on a

$$\|B\|_\infty = \sup_{\substack{v \in \mathbb{R}^N \\ \|v\|=1}} \sup_{i=1, \dots, N} |(Bv)_i| = \sup_{\substack{v \in \mathbb{R}^N \\ \|v\|=1}} \sup_{i=1, \dots, N} \left| \sum_{j=1}^N B_{ij} v_j \right| \|B\|_\infty = \sup_{i=1, \dots, N} \sum_{j=1}^N B_{ij}.$$

On a donc  $\|A_h^{-1}\| = \sup_{i=1, \dots, N} \sum_{j=1}^N (A_h^{-1})_{ij} \leq \sup_{i=1, \dots, N} \sum_{j=1}^N (A_{0h}^{-1})_{ij}$  car  $A_h^{-1} \leq A_{0h}^{-1}$  ; d'où on déduit que  $\|A_h^{-1}\|_\infty \leq \|A_{0h}^{-1}\|_\infty$ . Il ne reste plus qu'à estimer  $\|A_{0h}^{-1}\|_\infty$ . Comme  $A_{0h}^{-1} \geq 0$ , on a

$$\|A_{0h}^{-1}\|_\infty = \|A_{0h}^{-1} e\|_\infty \text{ avec } e = (1, \dots, 1).$$

Soit  $d = A_{0h}^{-1} e \in \mathbb{R}^N$ . On veut calculer  $\|d\|_\infty$ , où  $d$  vérifie  $A_{0h} d = e$ . Or le système linéaire  $A_{0h} d = e$  n'est autre que la discrétisation par différences finies du problème

$$\begin{cases} -u'' = 1 \\ u(0) = u(1) = 0 \end{cases}, \quad (1.36)$$

dont la solution exacte est :

$$u_0(x) = \frac{x(1-x)}{2},$$

qui vérifie  $u_0^{(4)}(x) = 0$ . On en conclut, par la remarque 1.14, que

$$u_0(x_i) = d_i, \quad \forall i = 1 \dots N.$$

Donc  $\|d\|_\infty = \sup_{i=1, N} \frac{ih(ih-1)}{2}$  où  $h = \frac{1}{N+1}$  est le pas de discrétisation. Ceci entraîne que

$$\|d\|_\infty \leq \sup_{[0,1]} \left| \frac{x(x-1)}{2} \right| = \frac{1}{8}, \text{ et donc que } \|A_h^{-1}\|_\infty \leq \frac{1}{8}.$$

■

**Remarque 1.16 (Sur la stabilité)** *Noter que l'inégalité (1.34) donne une estimation sur les solutions approchées indépendantes du pas de maillage. C'est ce type d'estimation qu'on recherchera par la suite pour la discrétisation d'autres problèmes comme garant de la stabilité d'un schéma numérique.*

**Définition 1.17 (Erreur de discrétisation)** *On appelle erreur de discrétisation en  $x_i$ , la différence entre la solution exacte en  $x_i$  et la  $i$ -ème composante de la solution donnée par le schéma numérique*

$$e_i = u(x_i) - u_i, \quad \forall i = 1, \dots, N. \quad (1.37)$$

**Théorème 1.18** *Soit  $u$  la solution exacte de*

$$\begin{cases} -u'' + cu = f, \\ u(0) = u(1) = 0. \end{cases}$$

*On suppose  $u \in C^4([0, 1])$ . Soit  $u_h$  la solution de (1.24). Alors l'erreur de discrétisation définie par (1.37) satisfait*

$$\max_{i=1, \dots, N} |e_i| \leq \frac{1}{96} \|u^{(4)}\|_\infty h^2.$$

*Le schéma est donc convergent d'ordre 2.*

**Démonstration :** Soit  $U_h = (U_1, \dots, U_n)$  et  $\bar{U}_h = (u(x_1), \dots, u(x_N))$ , on cherche à majorer  $\|\bar{U}_h - U_h\|_\infty$ . On a  $A(\bar{U}_h - U_h) = R$  où  $R$  est l'erreur de consistance (voir remarque 1.14). On a donc

$$\|\bar{U}_h - U_h\|_\infty \leq \|A_h^{-1}\|_\infty \|R\|_\infty \leq \frac{1}{8} \times \frac{1}{12} \|u^{(4)}\|_\infty = \frac{1}{96} \|u^{(4)}\|_\infty$$

■

**Remarque 1.19 (Sur la convergence)** On peut remarquer que la preuve de la convergence s'appuie sur la stabilité (elle-même déduite de la conservation de la positivité) et sur la consistance. Dans certains livres d'analyse numérique, vous trouverez la "formule" : stabilité + consistance  $\implies$  convergence. Il faut toutefois prendre garde au fait que ces notions de stabilité et consistance peuvent être variables d'un type de méthode à un autre (comme nous le verrons en étudiant la méthode des volumes finis, par exemple).

**Remarque 1.20 (Contrôle des erreurs d'arrondi)** On cherche à calculer la solution approchée de  $-u'' = f$ . Le second membre  $f$  est donc une donnée du problème. Supposons que des erreurs soient commises sur cette donnée (par exemple des erreurs d'arrondi, ou des erreurs de mesure). On obtient alors un nouveau système, qui s'écrit  $A_h \tilde{U}_h = b_h + \varepsilon_h$ , où  $\varepsilon_h$  représente la discrétisation des erreurs commises sur le second membre. Si on résout  $A_h \tilde{U}_h = b_h + \varepsilon_h$  au lieu de  $A_h U_h = b_h$ , l'erreur commise sur la solution du système s'écrit

$$E_h = \tilde{U}_h - U_h = A_h^{-1} \varepsilon_h.$$

On en déduit que

$$\|E_h\|_\infty \leq \frac{1}{8} \|\varepsilon_h\|_\infty.$$

On a donc une borne d'erreur sur l'erreur qu'on obtient sur la solution du système par rapport à l'erreur commise sur le second membre. Le problème des erreurs relatives est beaucoup plus subtil, voir à ce propos l'exercice 7 sur le "conditionnement efficace".

### 1.3 Analyse du schéma volumes finis

On va étudier la discrétisation par volumes finis du problème (1.1)–(1.2), qu'on rappelle ici :

$$-u'' = f, \quad x \in ]0, 1[ \tag{1.1}$$

$$u(0) = u(1) = 0. \tag{1.2}$$

**Définition 1.21 (Maillage volumes finis)** On appelle maillage volumes finis de l'intervalle  $[0, 1]$ , un ensemble de  $N$  mailles  $(K_i)_{i=1, \dots, N}$ , telles que  $K_i = ]x_{i-1/2}, x_{i+1/2}[$ , avec  $x_{1/2} = 0 < x_{\frac{3}{2}} < x_{i-1/2} < x_{i+1/2} < \dots < x_{N+1/2} = 1$ , et on note  $K_i = x_{i+1/2} - x_{i-1/2}$ . On se donne également  $N$  points  $(x_i)_{i=1, \dots, N}$  situés dans les mailles  $K_i$ . On a donc :

$$0 = x_{1/2} < x_1 < x_{\frac{3}{2}} < \dots < x_{i-1/2} < x_i < x_{i+1/2} < \dots < x_{N+1/2} = 1.$$

On notera  $h_{i+1/2} = x_{i+1} - x_i$ , et  $h = \max_{i=1, \dots, N} h_i$ , et pour des questions de notations, on posera également  $x_0 = 0$  et  $x_{N+1} = 1$ .

On rappelle que pour obtenir un schéma volumes finis, on part de la forme intégrale (bilan des flux) obtenue en intégrant l'équation (1.1) sur  $K_i = ]x_{i+1/2}, x_{i-1/2}[$  :

$$-u'(x_i + 1/2) + u'(x_i - 1/2) = \int_{K_i} f(x) dx. \tag{1.38}$$

On pose :  $f_i = \frac{1}{h_i} \int_{K_i} f(x) dx$ , et on introduit les inconnues discrètes  $(u_i)_{i=1 \dots N}$  (une par maille) et les équations discrètes du schéma numérique :

$$F_{i+1/2} - F_{i-1/2} = h_i f_i, \quad i = 1, \dots, N, \tag{1.39}$$

où  $F_{i+1/2}$  est le flux numérique en  $x_{i+1/2}$  qui devrait être une approximation raisonnable de  $-u'(x_{i+1/2})$ . On pose alors :

$$F_{i+1/2} = -\frac{u_{i+1} - u_i}{h_{i+1/2}}, \quad i = 1, \dots, N,$$

$$F_{1/2} = -\frac{u_1}{h_{1/2}}, \quad F_{N+1/2} = \frac{u_N}{h_{N+1/2}},$$

pour tenir compte des conditions aux limites de Dirichlet homogènes  $u(0) = u(1) = 0$ . On peut aussi écrire :

$$F_{i+1/2} = -\frac{u_{i+1} - u_i}{h_{i+1/2}}, \quad i = 0, \dots, N, \quad (1.40)$$

$$\text{en posant } u_0 = u_{N+1} = 0. \quad (1.41)$$

On peut écrire le système linéaire obtenu sur  $(u_1, \dots, u_N)$  sous la forme

$$A_h U_h = b_h, \quad (1.42)$$

avec

$$(A_h U_h)_i = \frac{1}{h_i} \left[ \frac{-1}{h_{i+1/2}} (u_{i+1} - u_i) + \frac{1}{h_{i-1/2}} (u_i - u_{i-1}) \right] \text{ et } (b_h)_i = f_i.$$

**Remarque 1.22 (Non consistance au sens des différences finies)**

L'approximation de  $-u''(x_i)$  par

$$\frac{1}{h_i} \left[ \frac{-1}{h_{i+1/2}} (u(x_{i+1}) - u(x_i)) + \frac{1}{h_{i-1/2}} (u(x_i) - u(x_{i-1})) \right]$$

n'est pas consistante dans le cas général : voir exercice 11 page 35.

On peut montrer que les deux schémas différences finies et volumes sont identiques “au bord près” dans le cas d'un maillage uniforme lorsque  $x_i$  est supposé être le centre de la maille : voir exercice 1 page 30.

On va démontrer ici qu'il existe une unique solution  $(u_1, \dots, u_N)$  au schéma (1.39)–(1.41), et que cette solution converge, en un certain sens, vers la solution de problème continu (1.1)–(1.2) lorsque le pas du maillage tend vers 0.

**Proposition 1.23 (Existence de la solution du schéma volumes finis)** Soit  $f \in C([0, 1])$  et  $u \in C^2([0, 1])$  solution du problème (1.1)–(1.2). Soit  $(K_i)_{i=1, \dots, N}$  le maillage par la définition 1.21 page 18. Alors il existe une unique solution  $u_h = (u_1, \dots, u_N)$  de (1.39)–(1.41).

**Démonstration :** Ce résultat se déduit facilement de la proposition suivante, qui donne la stabilité du schéma, c.à.d. une estimation *a priori* sur les solutions approchées. Si  $f_i = 0$  pour tout  $i = 1, \dots, N$ , la proposition 1.25 nous donne que  $\|D_{\mathcal{T}} u_{\mathcal{T}}\| = 0$ , où  $D_{\mathcal{T}} u_{\mathcal{T}}$  est la “dérivée discrète” définie ci-après, et donc  $u_i - u_{i-1} = 0$  pour tout  $i = 1 \dots N$ ; mais comme  $u_0 = 0$ , on en déduit que  $u_i = 0$  pour tout  $i = 1 \dots N$ . Ceci démontre l'unicité de  $(u_i)_{i=1 \dots N}$  solution de (1.39)–(1.41), et donc son existence, puisque le système (1.39)–(1.41) est un système linéaire carré d'ordre  $N$ . (On rappelle qu'une matrice carrée d'ordre  $N$  est inversible si et seulement si son noyau est réduit à  $\{0\}$ ). ■

Nous allons maintenant prouver la stabilité, sous forme d'une estimation dite *a priori*, car on effectue une majoration sur une fonction dont on n'a pas forcément prouvé l'existence : on établit l'estimation *a priori* en premier et on en déduit l'existence.

Pour démontrer cette propriété, on commence par introduire une “dérivée” discrète des fonctions constantes par mailles, qui nous servira dans la suite des démonstrations.

**Définition 1.24 (Dérivée discrète)** On considère le maillage  $(K_i)_{i=1, \dots, N}$  de la définition 1.21 page 18. Soit  $v$  une fonction constante par mailles sur les mailles  $K_i$ , qui représente une approximation d'une fonction définie sur  $[0, 1]$  et nulle en 0 et en 1. En posant  $v_0 = v_{N+1} = 0$ , on peut définir une sorte de “dérivée discrète” de  $v$  par les pentes

$$p_{i+1/2} = \frac{v_{i+1} - v_i}{h_{i+1/2}}, \quad i = 0, \dots, N.$$

On peut alors définir une  $D_{\mathcal{T}} v$ , fonction constante par intervalle et égale à  $p_{i+1/2}$  sur l'intervalle  $K_{i+1/2} = ]x_i, x_{i+1}[$ . La norme  $L^2$  de  $D_{\mathcal{T}} v$  est alors définie par :

$$\|D_{\mathcal{T}} v\|_{L^2([0,1])}^2 = \sum_{i=0}^N h_{i+1/2} p_{i+1/2}^2 = \sum_{i=0}^N \sum_{i=0}^N h_{i+1/2} \frac{(v_{i+1} - v_i)^2}{h_{i+1/2}},$$

$$\text{où } h_{i+1/2} = \frac{h_i + h_{i+1}}{2}$$

**Proposition 1.25 (Stabilité : estimation *a priori* sur les solutions approchées)** Soit  $f \in L^2([0, 1])$ . On considère le maillage  $(K_i)_{i=1, \dots, N}$  de la définition 1.21 page 18 ; pour  $i = 1, \dots, N$ , on note  $f_i$  la valeur moyenne de  $f$  sur la maille  $K_i$ . Si  $u_h$  est la fonction constante par maille dont les valeurs sur les mailles sont des valeurs  $(u_1, \dots, u_N)$  qui vérifient le schéma volumes finis (1.39)–(1.41), alors

$$\|D_{\mathcal{T}}u_{\mathcal{T}}\|_{L^2} \leq \|f\|_{L^2} \quad (1.43)$$

**Démonstration :** La preuve de cette proposition est calquée sur l'estimation *a priori* qu'on peut faire sur les solutions du problème continu : en effet, si  $u$  est une solution qu'on supposera aussi régulière que l'on veut<sup>8</sup> du problème (1.1)–(1.2), alors

$$\|u'\|_{L^2(]0,1[)} \leq \|f\|_{L^2((0,1))} \quad (1.44)$$

Nous allons donc mener les preuves de (1.44) et (1.43) en parallèle. Soit  $u \in C^2(]0, 1[)$  solution de (1.1)–(1.2), et soit  $(u_1, \dots, u_N)$  solution de (1.39)–(1.41).

#### Estimation continue

On multiplie (1.1) par  $u$  et on intègre entre 0 et 1 :

$$-\int_0^1 u''(x)u(x) \, dx = \int_0^1 f(x) \, dx$$

On intègre par parties et on utilise les conditions aux limites (1.2) :

$$\int_0^1 (u'(x))^2 \, dx = \int_0^1 f(x)u(x) \, dx$$

On utilise l'inégalité de Cauchy-Schwarz à droite.

$$\int_0^1 (u'(x))^2 \, dx = \|f\|_{L^2(]0,1[)} \|u\|_{L^2(]0,1[)}$$

L'inégalité de Poincaré (voir un peu plus loin dans la preuve et page 23) s'écrit

$$\|u\|_{L^2(]0,1[)} \leq \|u'\|_{L^2(]0,1[)}$$

Et donc

$$\|u'\|_{L^2(]0,1[)} \leq \|f\|_{L^2((0,1))}$$

Donnons les détails de la sommation par parties dans la preuve de l'estimation discrète. On a :

$$\sum_{i=1}^N -\frac{u_{i+1} - u_i}{h_{i+1/2}} u_i + \sum_{i=1}^N \frac{u_i - u_{i-1}}{h_{i-1/2}} u_i = \sum_{i=1}^N h_i f_i u_i.$$

En effectuant un changement d'indice sur la deuxième somme, on obtient :

$$\sum_{i=1}^N -\frac{u_{i+1} - u_i}{h_{i+1/2}} u_i + \sum_{i=0}^{N-1} \frac{u_{i+1} - u_i}{h_{i+1/2}} u_{i+1} = \sum_{i=1}^N h_i f_i u_i;$$

en regroupant les sommes, on a donc :

$$\sum_{i=0}^N \frac{(u_{i+1} - u_i)^2}{h_{i+1/2}} = \sum_{i=1}^N h_i f_i u_i.$$

<sup>8</sup>En fait la solution unique de (1.1)–(1.2) appartient à l'espace  $H^1(]0, 1[)$ , comme on le verra plus tard.

#### Estimation discrète

On multiplie (1.39) par  $u_i$  et on somme sur  $i$  :

$$\sum_{i=1}^N h_i \left( -\frac{u_{i+1} - u_i}{h_{i+1/2}} + \frac{u_i - u_{i-1}}{h_{i-1/2}} \right) u_i = \sum_{i=1}^N h_i f_i u_i,$$

On somme par parties en utilisant  $u_0 = u_{N+1} = 0$  (les détails sont donnés après la présentation en colonnes) :

$$\sum_{i=0}^N \frac{(u_{i+1} - u_i)^2}{h_{i+1/2}} = \sum_{i=1}^N h_i f_i u_i.$$

On utilise l'inégalité de Cauchy-Schwarz à droite et le fait que  $\int_0^1 f_{\mathcal{T}}(x)^2 \, dx \leq \int_0^1 f(x)^2 \, dx$  (ce qui est encore une conséquence de Cauchy-Schwarz).

$$\|(D_{\mathcal{T}}u_{\mathcal{T}})^2\|_{L^2(]0,1[)} \leq \|f\|_{L^2((0,1))} \|u_{\mathcal{T}}\|_{L^2((0,1))}$$

L'inégalité de Poincaré discrète s'écrit

$$\|u_{\mathcal{T}}\|_{L^2(]0,1[)} \leq \|D_{\mathcal{T}}u_{\mathcal{T}}\|_{L^2(]0,1[)}$$

Et par la définition de  $D_{\mathcal{T}}u$  :

$$\|D_{\mathcal{T}}u_{\mathcal{T}}\|_{L^2} \leq \|f\|_{L^2}$$

Et on se retrouve bien avec la norme  $L^2$  de la dérivée discrète au membre de gauche de l'équation. ■

Il nous faut maintenant donner les démonstrations des inégalités de Poincaré continue et discrète que nous avons utilisé dans la preuve ci-dessus. Là encore, nous allons procéder "en parallèle, car la démonstration en discret est calquée sur la démonstration continue.

**Proposition 1.26 (Poincaré en continu et en discret)**

– Enoncé en continu (pour une fonction régulière) : Soit  $u \in C^1([0, 1])$  telle que  $u(0) = 0$ . Alors

$$\|u\|_{L^2(]0,1])} \leq \|u'\|_{L^2(]0,1])} \quad (1.45)$$

et

$$\|u\|_{L^2(]0,1])} \leq \|u'\|_{L^2(]0,1])} \quad (1.46)$$

– Enoncé en discret On considère le maillage  $(K_i)_{i=1,\dots,N}$  de la définition 1.21 page 18. Soit  $v$  une fonction constante par mailles sur les mailles  $K_i$ , et soit  $D_{\mathcal{T}}v$  sa "dérivée discrète" au sens de la définition 1.24. Alors :

$$\|v\|_{L^2(]0,1[)} \leq \|D_{\mathcal{T}}v\|_{L^2(]0,1])} \quad (1.47)$$

et

$$\|v\|_{\infty} \leq \|D_{\mathcal{T}}v\|_{L^2(]0,1])} \quad (1.48)$$

**Démonstration :** Là encore, on va effectuer les démonstrations en parallèle, vu que la démonstration de l'inégalité "discrète" copie la démonstration de l'inégalité continue.

Inégalité de Poincaré continue

On écrit que  $u$  est l'intégrale de sa dérivée, en utilisant le fait que  $u(0) = 0$  :

$$u(x) = \int_0^x u'(s) ds$$

On majore :

$$|u(x)| \leq \int_0^x |u'(s)| ds \leq \int_0^1 |u'(s)| ds$$

ce qui donne tout de suite (1.46). On utilise l'inégalité de Cauchy-Schwarz à droite.

$$|u(x)|^2 \leq \|u'\|_{L^2(]0,1])}^2$$

On intègre entre 0 et 1 et on aboutit au résultat :

$$\|u'\|_{L^2(]0,1[)} \leq \|u'\|_{L^2(]0,1])}$$

Notons que dans les deux démonstrations, on obtient que  $\|u\|_{\infty} \leq \|u\|_{L^2(]0,1])}$  et que  $\|u\|_{\infty} \leq \|u'\|_{L^2(]0,1])}$  ■

**Définition 1.27 (Erreur de consistance sur le flux)** Soit  $u : [0, 1] \rightarrow \mathbb{R}$  solution du problème (1.1)–(1.2). On se donne une subdivision de  $[0, 1]$ . On appelle  $\bar{F}_{i+1/2} = -u'(x_{i+1/2})$  le flux exact en  $x_{i+1/2}$ , et  $F_{i+1/2}^* = -\frac{u(x_{i+1}) - u(x_i)}{h_{i+1/2}}$  l'approximation du flux exact utilisée pour construire le flux numérique  $F_{i+1/2} = -\frac{u_{i+1} - u_i}{h_{i+1/2}}$ .

Inégalité de Poincaré discrète

En tenant compte du fait que  $v_0 = 0$ , on écrit que :

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

On majore :

$$\begin{aligned} |v_i| &\leq \sum_{k=1}^i |v_k - v_{k-1}| \\ &\leq \sum_{k=1}^{N+1} h_{k+\frac{1}{2}} \frac{|v_k - v_{k-1}|}{h_{k+\frac{1}{2}}} = \int_0^1 |D_{\mathcal{T}}v_{\mathcal{T}}(s)| ds \end{aligned}$$

On utilise l'inégalité de Cauchy-Schwarz à droite.

$$|v_i|^2 \leq \|D_{\mathcal{T}}v_{\mathcal{T}}\|_{L^2(]0,1])}^2,$$

ce qui donne tout de suite (1.48), et en intégrant entre 0 et 1 et on aboutit au résultat :

$$\|v\|_{L^2(]0,1[)} \leq \|D_{\mathcal{T}}v_{\mathcal{T}}\|_{L^2(]0,1])}$$

On dit que le flux numérique est consistant d'ordre  $p$  s'il existe  $C \in \mathbb{R}_+$  ne dépendant que de  $u$  telle que l'erreur de consistance sur le flux, définie par :

$$R_{i+1/2} = \bar{F}_{i+1/2} - F_{i+1/2}^*, \quad (1.49)$$

vérifie

$$|R_{i+1/2}| \leq Ch^p. \quad (1.50)$$

**Lemme 1.28 (Consistance du flux de diffusion)** Soit  $u \in C^2([0,1])$  solution du problème (1.1)–(1.2). Le flux numérique  $F_{i+1/2} = -\frac{u_{i+1}-u_i}{h_{i+1/2}}$  est consistant d'ordre 1. Plus précisément il existe  $C$  ne dépendant que de  $\|u''\|_\infty$  tel que l'erreur de consistance sur les flux définie par (1.49) vérifie :

$$|R_{i+1/2}| = \left| -u'(x_{i+1/2}) + \frac{u(x_{i+1}) - u(x_i)}{h_{i+1/2}} \right| \leq Ch. \quad (1.51)$$

**Démonstration :** La démonstration de ce résultat s'effectue facilement à l'aide de développements de Taylor : voir l'exercice 13 page 35, où l'on montre aussi que si  $x_{i+1/2}$  est au centre de l'intervalle  $[x_i, x_{i+1}]$ , l'erreur de consistance sur les flux est d'ordre 2, i.e. il existe  $C \in \mathbb{R}_+$  ne dépendant que de  $u$  telle que  $R_{i+1/2} \leq Ch^2$ . Notez que cette propriété de consistance est vraie sur les flux, et non pas sur l'opérateur  $-u''$  (voir remarque 1.22) et exercice 11.

**Définition 1.29 (Conservativité)** On dit que le schéma volumes finis (1.39)–(1.41) est conservatif, au sens où, lorsqu'on considère une interface  $x_{i+1/2}$  entre deux mailles  $K_i$  et  $K_{i+1}$ , le flux numérique entrant dans une maille est égal à celui sortant de l'autre.

C'est grâce à la conservativité et à la consistance des flux qu'on va montrer la convergence du schéma volumes finis.

**Théorème 1.30 (Convergence du schéma volumes finis)** On suppose que la solution  $u$  du problème (1.1)–(1.2) vérifie  $u \in C^2([0,1])$ . On pose  $e_i = u(x_i) - u_i$  pour  $i = 1, \dots, N$ , et  $e_0 = e_{N+1} = 0$ . On note  $e_{\mathcal{T}}$  la fonction constante par mailles et égale à  $e_i$  sur la maille  $i$  et  $D_{\mathcal{T}}e_{\mathcal{T}}$  sa dérivée discrète, au sens de la définition 1.24. Il existe  $C \geq 0$  ne dépendant que de  $u$  tel que :

$$\|D_{\mathcal{T}}e_{\mathcal{T}}\|_{L^2([0,1])} = \left( \sum_{i=0}^N \frac{(e_{i+1} - e_i)^2}{h} \right)^{\frac{1}{2}} \leq Ch, \quad (1.52)$$

$$\|e_{\mathcal{T}}\|_{L^2([0,1])} = \left( \sum_{i=1}^N h e_i^2 \right)^{\frac{1}{2}} \leq Ch \quad (1.53)$$

$$\|e_{\mathcal{T}}\|_{\infty} = \max_{i=1, \dots, N} |e_i| \leq Ch. \quad (1.54)$$

(On rappelle que  $h = \sup_{i=1 \dots N} h_i$ .)

**Démonstration :** Ecrivons le schéma volumes finis (1.39) :

$$F_{i+1/2} - F_{i-1/2} = h_i f_i,$$

l'équation exacte intégrée sur la maille  $K_i$  (1.38) :

$$\bar{F}_{i+1/2} - \bar{F}_{i-1/2} = h_i f_i,$$

où  $\bar{F}_{i+1/2}$  est défini dans le lemme 1.28, et soustrayons :

$$\bar{F}_{i+1/2} - F_{i+1/2} - \bar{F}_{i-1/2} + F_{i-1/2} = 0.$$

En introduisant  $R_{i+1/2} = \bar{F}_{i+1/2} - F_{i+1/2}^*$ , on obtient :

$$F_{i+1/2}^* - F_{i+1/2} - F_{i-1/2}^* + F_{i-1/2} = -R_{i+1/2} + R_{i-1/2}$$

ce qui s'écrit encore, au vu de la définition de  $e_i$ ,

$$-\frac{1}{h_{i+1/2}}(e_{i+1} - e_i) + \frac{1}{h_{i-1/2}}(e_i - e_{i-1}) = -R_{i+1/2} + R_{i-1/2}.$$



On multiplie cette dernière égalité par  $e_i$  et on somme de 1 à  $N$  :

$$\sum_{i=1}^N -\frac{1}{h_{i+1/2}}(e_{i+1} - e_i)e_i + \sum_{i=1}^N \frac{1}{h_{i-1/2}}(e_i - e_{i-1})e_i = \sum_{i=1}^N -R_{i+1/2}e_i + \sum_{i=1}^N R_{i-1/2}e_i,$$

ce qui s'écrit encore :

$$\sum_{i=1}^N -\frac{1}{h_{i+1/2}}(e_{i+1} - e_i)e_i + \sum_{i=0}^{N-1} \frac{1}{h_{i+1/2}}(e_{i+1} - e_i)e_{i+1} = \sum_{i=1}^N -R_{i+1/2}e_i + \sum_{i=0}^{N-1} R_{i+1/2}e_{i+1}$$

En réordonnant les termes, on obtient, en remarquant que  $e_0 = 0$  et  $e_{N+1} = 0$  :

$$\sum_{i=0}^N \frac{(e_{i+1} - e_i)^2}{h_{i+1/2}} = \sum_{i=0}^N R_{i+1/2}(e_{i+1} - e_i).$$

Mais

$$\sum_{i=0}^N \frac{(e_{i+1} - e_i)^2}{h_{i+1/2}} = \sum_{i=0}^N h_{i+1/2} \left( \frac{e_{i+1} - e_i}{h_{i+1/2}} \right)^2 = \int_0^1 (D_{\mathcal{T}}e_{\mathcal{T}}(s))^2 ds = \|D_{\mathcal{T}}e_{\mathcal{T}}\|_{L^2(]0,1])}^2.$$

De plus,  $R_{i+1/2} \leq C h$  (par le lemme 1.28). On a donc

$$\|D_{\mathcal{T}}e_{\mathcal{T}}\|_{L^2(]0,1])}^2 \leq C h \sum_{i=0}^N |e_{i+1} - e_i| = C h \sum_{i=0}^N h_{i+1/2} \frac{|e_{i+1} - e_i|}{h_{i+1/2}} = C h \int_0^1 |D_{\mathcal{T}}e_{\mathcal{T}}|,$$

et, par l'inégalité de Cauchy–Schwarz :

$$\|D_{\mathcal{T}}e_{\mathcal{T}}\|_{L^2(]0,1])} \leq C h.$$

On a ainsi démontré (1.52). On obtient (1.53) et (1.54) par l'inégalité de Poincaré discrète (proposition 1.32). ■

**Remarque 1.31 (Espaces fonctionnels et normes discrètes)** On rappelle qu'une fonction  $u$  de l'espace de Lebesgue<sup>9 10</sup>  $L^2(]0, 1])$  admet une dérivée faible dans  $L^2(]0, 1])$  s'il existe  $v \in L^2(]0, 1])$  telle que

$$\int_{]0,1[} u(x)\varphi'(x)dx = - \int_{]0,1[} v(x)\varphi(x)dx, \quad (1.55)$$

pour toute fonction  $\varphi \in C_c^1(]0, 1])$ , où  $C_c^1(]0, 1])$  désigne l'espace des fonctions de classe  $C^1$  à support compact dans  $]0, 1[$ . On peut montrer que  $v$  est unique, voir par exemple [1]. On notera  $v = Du$ . On peut remarquer que si  $u \in C^1(]0, 1])$ , et si on note  $u'$  sa dérivée classique, alors  $Du = u'$  "presque partout". On note  $H^1(]0, 1])$  l'ensemble des fonctions de  $L^2(]0, 1])$  qui admettent une dérivée faible dans  $L^2(]0, 1])$  :

$$H^1(]0, 1]) = \{u \in L^2(]0, 1]) ; Du \in L^2(]0, 1])\}.$$

C'est un espace de Hilbert pour le produit scalaire :

$$(u, v)_{H^1(]0,1])} = \left( \int_{]0,1[} (u(x) v(x) dx + Du(x)Dv(x)) dx. \right.$$

Tout élément de  $H^1(]0, 1])$  (au sens classe d'équivalence de fonctions) admet un représentant continu, et on peut donc définir les valeurs en 0 et en 1 d'une "fonction" de  $H^1(]0, 1])$ .

$$H_0^1(]0, 1]) = \{u \in H^1(]0, 1]) ; u(0) = u(1) = 0\}.$$

<sup>9</sup>Henri-Léon Lebesgue (1875 – 1941) est un mathématicien français. Il est reconnu pour sa théorie d'intégration publiée initialement dans sa dissertation *Intégrale, longueur, aire* à l'université de Nancy en 1902. Il fut l'un des grands mathématiciens français de la première moitié du XXe siècle.

<sup>10</sup>Une fonction  $f$  de  $]0, 1[$  dans  $\mathbb{R}$  est intégrable au sens de Lebesgue si  $f$  est mesurable et  $\int_{]0,1[} |f| dt < +\infty$ . L'espace  $L^2(]0, 1])$  désigne l'ensemble des classes d'équivalence des fonctions de carré intégrable au sens de Lebesgue, pour la relation d'équivalence "égale presque partout", ce qui permet de définir une norme sur  $L^2(]0, 1])$  par  $\|u\|_{L^2(]0,1])} = (\int_{]0,1[} u^2 dt)^{\frac{1}{2}}$ , ce qui en fait un espace complet. Cette norme est associée au produit scalaire  $(u, v)_{L^2(]0,1])} = \int_{]0,1[} u v dt$ , ce qui en fait un espace de Hilbert. Pour plus de détails sur ces questions voir le polycopié *Mesures, intégration, probabilités* sur la page web <http://www-gm3.univ-mrs.fr/~gallouet/>

Pour  $u \in H^1(]0, 1[)$ , on note :

$$\|u\|_{H_0^1} = \left( \int_0^1 (Du(x))^2 dx \right)^{1/2}.$$

C'est une norme sur  $H_0^1$  qui est équivalente à la norme  $\|\cdot\|_{H^1}$  définie par  $\|u\|_{H^1} = \left( \int u^2(x) dx + \int (Du)^2(x) dx \right)^{1/2}$ , ce qui se démontre grâce à l'inégalité de Poincaré<sup>11</sup> qu'on rappelle :

**Proposition 1.32 (Inégalité de Poincaré, cas 1D)** .

$$\|u\|_{L^2(]0,1[)} \leq \|Du\|_{L^2(]0,1[)} \text{ pour tout } u \in H_0^1(]0, 1[). \quad (1.56)$$

La démonstration de cette proposition a été faite dans le cadre de la preuve de la stabilité du schéma en 1D (voir proposition 1.25 pour des fonctions régulières, mais la même preuve s'applique dans le cas d'une fonction  $H^1$ ).

Soit maintenant  $\mathcal{T}$  un maillage volumes finis de  $[0, 1]$  (voir définition 1.21), on note  $X(\mathcal{T})$  l'ensemble des fonctions de  $[0, 1]$  dans  $\mathbb{R}$ , constantes par maille de ce maillage. Pour  $v \in X(\mathcal{T})$ , on note  $v_i$  la valeur de  $v$  sur la maille  $i$ ; on peut écrire les normes  $L^2$  et  $L^\infty$  de  $v$  :

$$\|v\|_{L^2(]0,1[)}^2 = \sum_{i=1}^N h_i v_i^2,$$

et

$$\|v\|_{L^\infty(]0,1[)} = \max_{i=1, \dots, N} |v_i|.$$

Par contre, la fonction  $v$  étant constante par maille, elle n'est pas dérivable au sens classique, ni même au sens faible. On peut toutefois définir une norme  $H^1$  discrète de  $v$  de la manière suivante :

$$|v|_{1, \mathcal{T}} = \left( \sum_{i=0}^N h_{i+1/2} \left( \frac{v_{i+1} - v_i}{h_{i+1/2}} \right)^2 \right)^{1/2},$$

ce qui est la norme  $L^2$  de la dérivée discrète  $D_{\mathcal{T}}v$  (voir définition 1.24). On peut montrer (Exercice 16) que si  $u_{\mathcal{T}} : ]0, 1[ \rightarrow \mathbb{R}$  est définie par  $u_{\mathcal{T}}(x) = u_i \quad \forall x \in K_i$  où  $(u_i)_{i=1, \dots, N}$  solution de (1.39)–(1.41), alors  $|u_{\mathcal{T}}|_{1, \mathcal{T}}$  converge dans  $L^2(]0, 1[)$  lorsque  $h$  tend vers 0, vers  $\|Du\|_{L^2(]0,1[)}$ , où  $u$  est la solution du problème (1.1)–(1.2).

**Remarque 1.33 (Dimensions supérieures)** En une dimension d'espace, on a obtenu une estimation d'erreur en norme " $H_0^1$  discrète" et en norme  $L^\infty$ . En dimension supérieure ou égale à 2, on aura une estimation en  $h$ , en norme  $H_0^1$  discrète, en norme  $L^2$ , mais pas en norme  $L^\infty$ . Ceci tient au fait que l'injection de Sobolev  $H^1(]0, 1[) \subset C(]0, 1[)$  n'est vraie qu'en dimension 1. La démonstration de l'estimation d'erreur en norme  $L^2$  (1.53) se prouve alors directement à partir de l'estimation en norme  $H_0^1$  discrète, grâce à une "inégalité de Poincaré discrète", équivalent discret de la célèbre inégalité de Poincaré continue<sup>12</sup> (voir (1.56) pour la dimension 1).

## 1.4 Volumes finis pour la prise en compte de discontinuités

On considère ici un barreau conducteur constitué de deux matériaux de conductivités  $\lambda_1$  et  $\lambda_2$  différentes, et dont les extrémités sont plongées dans de la glace. On suppose que le barreau est de longueur 1, que le matériau de conductivité  $\lambda_1$  (resp.  $\lambda_2$ ) occupe le domaine  $\Omega_1 = ]0, 1/2[$  (resp.  $\Omega_2 = ]1/2, 1[$ ). Le problème de conduction de la chaleur s'écrit alors :

$$\begin{cases} (-\lambda_1(x)u')' = f(x) & x \in ]0, 1/2[ \\ (-\lambda_2(x)u')' = f(x) & x \in ]1/2, 1[ \\ u(0) = u(1) = 0, \\ -(\lambda_1 u')(1/2) = -(\lambda_2 u')(1/2) \end{cases} \quad (1.57)$$

**Remarque 1.34** La dernière égalité traduit la conservation du flux de chaleur à l'interface  $x = .5$ . On peut noter que comme  $\lambda$  est discontinu en ce point, la dérivée  $u'$  le sera forcément elle aussi.

<sup>11</sup>Henri Poincaré (1854–1912) est un mathématicien, physicien et philosophe français. C'est probablement l'un des plus grands hommes de science de cette époque.

<sup>12</sup>Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^N$ , et  $u \in H_0^1(\Omega)$ , alors  $\|u\|_{L^2(\Omega)} \leq \text{diam}(\Omega) \|Du\|_{L^2(\Omega)}$ .

On choisit de discrétiser le problème par volumes finis. On se donne un maillage volumes finis comme défini par la définition 1.21 page 18, en choisissant les mailles telles que la discontinuité de  $\lambda$  soit située sur un interface de deux mailles qu'on note  $K_k$  et  $K_{k+1}$ . On a donc, avec les notations du paragraphe (1.1.1)  $x_{k+1/2} = 0.5$ . La discrétisation par volumes finis s'écrit alors

$$F_{i+1/2} - F_{i-1/2} = h_i f_i, \quad i = 1, \dots, N,$$

où les flux numériques  $F_{i+1/2}$  sont donnés par

$$F_{i+1/2} = \lambda_* \frac{u_{i+1} - u_i}{h_{i+1/2}}, \quad \text{avec } \lambda_* = \begin{cases} \lambda_1 & \text{si } x_{i+1/2} > 0.5, \\ \lambda_2 & \text{si } x_{i+1/2} < 0.5. \end{cases}$$

Il ne reste donc plus qu'à calculer le flux  $F_{k+1/2}$ , approximation de  $(\lambda u')(x_{k+1/2})$  (avec  $x_{k+1/2} = 0.5$ ). On introduit pour cela une inconnue auxiliaire  $u_{k+1/2}$  que l'on pourra éliminer plus tard, et on écrit une discrétisation du flux de part et d'autre de l'interface.

$$F_{k+1/2} = -\lambda_1 \frac{u_{k+1/2} - u_k}{h_k^+}, \quad \text{avec } h_k^+ = x_{k+1/2} - x_k,$$

$$F_{k+1/2} = -\lambda_2 \frac{u_{k+1} - u_{k+1/2}}{h_{k+1}^-} \quad \text{avec } h_{k+1}^- = x_{k+1} - x_{k+1/2}.$$

L'élimination (et le calcul) de l'inconnue se fait en écrivant la conservation du flux numérique :

$$-\lambda_1 \frac{u_{k+1/2} - u_k}{h_k^+} = -\lambda_2 \frac{u_{k+1} - u_{k+1/2}}{h_{k+1}^-}$$

On en déduit la valeur de  $u_{k+1/2}$

$$u_{k+1/2} = \frac{\frac{\lambda_1}{h_k^+} u_k + \frac{\lambda_2}{h_{k+1}^-} u_{k+1}}{\frac{\lambda_1}{h_k^+} + \frac{\lambda_2}{h_{k+1}^-}}$$

On remplace  $u_{k+1/2}$  par cette valeur dans l'expression du flux  $F_{k+1/2}$ , et on obtient :

$$F_{k+1/2} = \frac{\lambda_1 \lambda_2}{h_k^+ \lambda_2 + h_{k+1}^- \lambda_1} (u_{k+1} - u_k).$$

Si le maillage est uniforme, on obtient

$$F_{k+1/2} = \frac{2\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \left( \frac{u_{i+1} - u_i}{h} \right).$$

Le flux est donc calculé en faisant intervenir la moyenne harmonique des conductivités  $\lambda_1$  et  $\lambda_2$ . Notons que lorsque  $\lambda_1 = \lambda_2$ , on retrouve la formule habituelle du flux.

## 1.5 Différences finies et volumes finis pour les problèmes de diffusion 2D

### 1.5.1 Différences finies

On considère maintenant le problème de diffusion dans un ouvert  $\Omega$  de  $\mathbb{R}^2$  :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega. \end{cases} \quad (1.58)$$

Le problème est bien posé au sens où : Si  $f \in C^1(\bar{\Omega})$ , alors il existe une unique solution  $u \in C^1(\bar{\Omega}) \cap C^2(\Omega)$ , solution de (1.58). Si  $f \in L^2(\Omega)$  et si  $\Omega$  est convexe (ou à bord régulier) alors il existe une unique fonction  $u \in H^2(\Omega)$  au sens faible<sup>13</sup> de (1.58), c.à.d. qui vérifie :

$$\begin{cases} u \in H_0^1(\Omega), \\ \int_{\Omega} \nabla u(x) \nabla v(x) dx = \int_{\Omega} f(x) v(x) dx, \quad \forall v \in H_0^1(\Omega). \end{cases} \quad (1.59)$$

<sup>13</sup>Par définition,  $H^2(\Omega)$  est l'ensemble des fonctions de  $L^2(\Omega)$  qui admet des dérivées faibles jusqu'à l'ordre 2 dans  $L^2(\Omega)$ .

On peut montrer (voir cours Equations aux dérivées partielles) que si  $u \in C^2(\overline{\Omega})$ , alors  $u$  est solution de (1.58) si et seulement si  $u$  est solution faible de (1.58). Pour discrétiser le problème, on se donne un certain nombre de points, alignés dans les directions  $x$  et  $y$ , comme représentés sur la figure 1.4 (on prend un pas de maillage uniforme et égal à  $h$ ). Certains de ces points sont à l'intérieur du domaine  $\Omega$ , d'autres sont situés sur la frontière  $\partial\Omega$ .

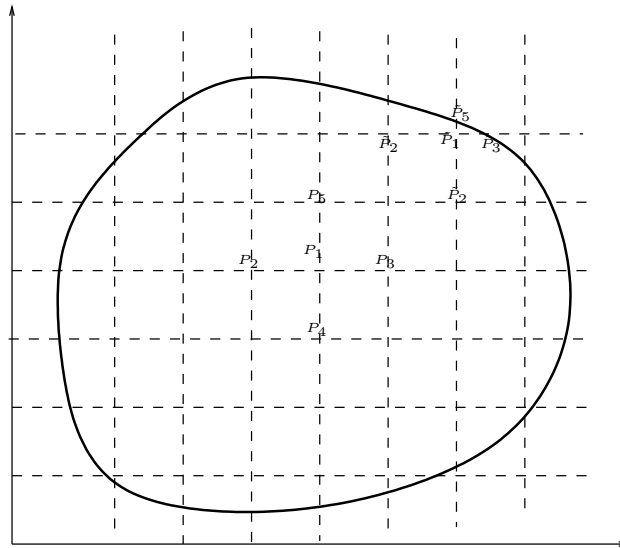


FIG. 1.4 – Discrétisation différences finies bi-dimensionnelle

Comme en une dimension d'espace, les inconnues discrètes sont associées aux noeuds du maillage. On note  $\{P_i, i \in I\}$  les points de discrétisation, et on écrit l'équation aux dérivées partielles en ces points :

$$-\Delta u(P_i) - \frac{\partial^2 u}{\partial x^2}(P_i) - \frac{\partial^2 u}{\partial y^2}(P_i) = f(P_i).$$

### 1er cas :

Dans le cas de points "vraiment intérieurs", tel que le point  $P_1$  sur la figure 1.4, *i.e.* dont tous les points voisins sont situés à l'intérieur de  $\Omega$ , les quotients différentiels

$$\frac{2u(P_1) - u(P_2) - u(P_3)}{h^2} \text{ et } \frac{2u(P_1) - u(P_5) - u(P_4)}{h^2}$$

sont des approximations consistantes à l'ordre 2 de  $-\partial_1^2 u(P_1)$  et  $-\partial_2^2 u(P_1)$ .

Par contre, pour un point "proche" du bord tel que le point  $\tilde{P}_1$ , les mêmes approximations (avec les points  $\tilde{P}_2, \tilde{P}_3, \tilde{P}_4$  et  $\tilde{P}_5$ ) ne seront que d'ordre 1 en raison des différences de distance entre les points (faire les développements de Taylor pour s'en convaincre).

Une telle discrétisation amène à un système linéaire  $A_h U_h = b_h$ , où la structure de  $A_h$  (en particulier sa "largeur de bande", c.à.d. le nombre de diagonales non nulles) dépend de la numérotation des noeuds. On peut montrer que la matrice  $A_h$  est monotone et le schéma est stable. De la consistance et la stabilité, on déduit, comme en une dimension d'espace, la convergence du schéma.

## 1.5.2 Volumes finis

### Le problème modèle

On considère le problème modèle suivant (par exemple de conduction de la chaleur) :

$$-\operatorname{div}(\lambda_i \nabla u(x)) = f(x) \quad x \in \Omega_i, \quad i = 1, 2 \quad (1.60)$$

où  $\lambda_1 > 0$ ,  $\lambda_2 > 0$  sont les conductivités thermiques dans les domaines  $\Omega_1$  et avec  $\Omega_2$ , avec  $\Omega_1 = ]0, 1[ \times ]0, 1[$  et  $\Omega_2 = ]0, 1[ \times ]1, 2[$ . On appelle  $\Gamma_1 = ]0, 1[ \times \{0\}$ ,  $\Gamma_2 = \{1\} \times ]0, 2[$ ,  $\Gamma_3 = ]0, 1[ \times \{2\}$ , et  $\Gamma_4 = \{0\} \times ]0, 2[$  les frontières extérieures de  $\Omega$ , et on note  $I = ]0, 1[ \times \{1\}$  l'interface entre  $\Omega_1$  et  $\Omega_2$  (voir Figure 1.5). Dans la suite, on notera  $\lambda$  la conductivité thermique sur  $\Omega$ , avec  $\lambda|_{\Omega_i} = \lambda_i$ ,  $i = 1, 2$ .

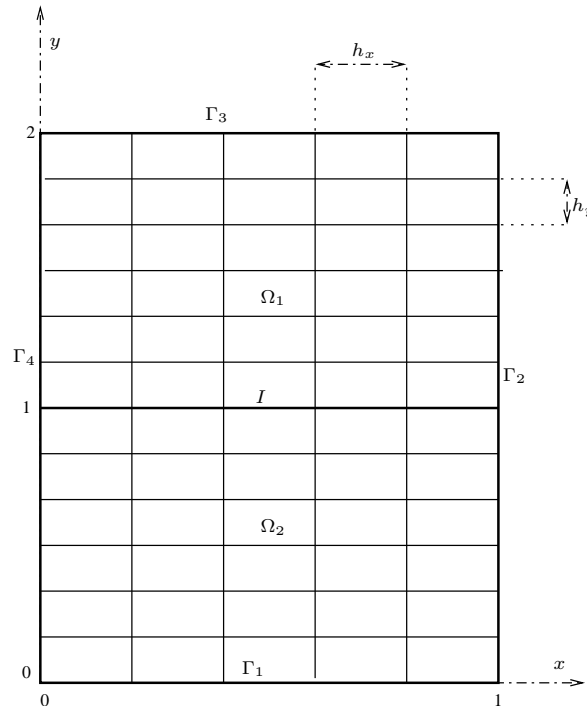


FIG. 1.5 – Domaine d'étude

On va considérer plusieurs types de conditions aux limites, en essayant d'expliquer leur sens physique. On rappelle que le flux de chaleur par diffusion est égal  $\mathbf{q}$  est donné par la loi de Fourier :  $\mathbf{q} = -\lambda \nabla u \cdot \mathbf{n}$ , où  $\mathbf{n}$  est le vecteur normal unitaire à la surface à travers laquelle on calcule le flux.

1. Conditions aux limites de type Fourier (Robin) sur  $\Gamma_1 \cup \Gamma_3$  : On suppose qu'il existe un transfert thermique entre les parois  $\Gamma_1$  et  $\Gamma_3$  et l'extérieur. Ce transfert est décrit par la condition de Fourier (Robin dans la littérature anglo-saxonne), qui exprime que le flux transféré est proportionnel à la différence de température entre l'extérieur et l'intérieur :

$$-\lambda \nabla u \cdot \mathbf{n}(x) = \alpha(u(x) - u_{ext}), \quad \forall x \in \Gamma_1 \cup \Gamma_3. \quad (1.61)$$

où  $\alpha > 0$  est le coefficient de transfert thermique,  $\mathbf{n}$  le vecteur unitaire normal à  $\partial\Omega$  extérieur à  $\Omega$ , et  $u_{ext}$  est la température extérieure (donnée).

2. Conditions aux limites de type Neumann sur  $\Gamma_2$  On suppose que la paroi  $\Gamma_2$  est parfaitement isolée, et que le flux de chaleur à travers cette paroi est donc nul. Ceci se traduit par une condition dite "de Neumann homogène" :

$$-\lambda \nabla u \cdot \mathbf{n} = 0 \quad \forall x \in \Gamma_2. \quad (1.62)$$

3. Conditions aux limites de type Dirichlet sur  $\Gamma_4$  Sur la paroi  $\Gamma_4$ , on suppose que la température est fixée. Ceci est une condition assez difficile à obtenir expérimentalement pour un problème de type chaleur, mais qu'on peut rencontrer dans d'autres problèmes pratiques.

$$u(x) = g(x), \quad \forall x \in \Gamma_4. \quad (1.63)$$

4. Conditions sur l'interface  $I$  : On suppose que l'interface  $I$  est par exemple le siège d'une réaction chimique surfacique  $\theta$  qui provoque un dégagement de chaleur surfacique. On a donc un saut du flux de chaleur à travers de l'interface  $I$ . Ceci se traduit par la condition de saut suivante :

$$-\lambda_1 \nabla u_1(x) \cdot \mathbf{n}_1 - \lambda_2 \nabla u_2(x) \cdot \mathbf{n}_2 = \theta(x), \quad x \in I. \quad (1.64)$$

où  $\mathbf{n}_i$  désigne le vecteur unitaire normal à  $I$  et extérieur à  $\Omega_i$ , et  $\theta$  est une fonction donnée.

### Discrétisation par volumes finis

On se donne un maillage “admissible”  $\mathcal{T}$  de  $\Omega$

$$\bar{\Omega} = \bigcup_{K \in \mathcal{T}} \bar{K}.$$

Par “admissible”, on entend un maillage tel qu’il existe des points  $(x_K)_{K \in \mathcal{T}}$  situés dans les mailles, tels que chaque segment  $x_K x_L$  soit orthogonal à l’arête  $K|L$  séparant la maille  $K$  de la maille  $L$ , comme visible sur la figure 1.6.

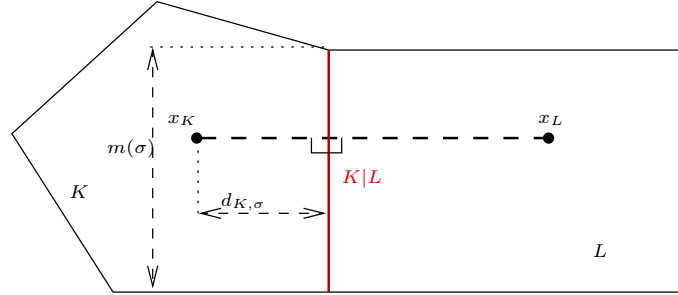


FIG. 1.6 – Condition d’orthogonalité pour un maillage volumes finis

Cette condition permet pour obtenir une approximation consistante du flux de diffusion (c’est-à-dire de la dérivée normale sur l’arête  $K|L$ ) avec deux inconnues discrètes, voir remarque 1.35. Dans le cas présent, le domaine représenté sur la figure 1.5 étant rectangulaire, cette condition est particulièrement facile à vérifier en prenant un maillage rectangulaire. Par souci de simplicité, on prendra ce maillage uniforme, et on notera  $h_x = 1/n$  le pas de discrétisation dans la direction  $x$  et  $h_y = 1/p$  le pas de discrétisation dans la direction  $y$ . Le maillage est donc choisi de telle sorte que l’interface  $I$  coïncide avec un ensemble d’arêtes du maillage qu’on notera  $\mathcal{E}_I$ . On a donc

$$\bar{I} = \bigcup_{\sigma \in \mathcal{E}_I} \bar{\sigma},$$

où le signe  $\bar{\phantom{x}}$  désigne l’adhérence de l’ensemble. On se donne ensuite des inconnues discrètes  $(u_K)_{K \in \mathcal{T}}$  associées aux mailles et  $(u_\sigma)_{\sigma \in \mathcal{E}}$  associées aux arêtes.

Pour obtenir le schéma volumes finis, on commence par établir les bilans par maille en intégrant l’équation sur chaque maille  $K$  (notons que ceci est faisable en raison du fait que l’équation est sous forme conservative, c’est-à-dire sous la forme :  $-\text{div}(\text{flux}) = f$ ). On obtient donc :

$$\int_K -\text{div}(\lambda_i \nabla u(x)) dx = \int_K f(x) dx,$$

soit encore, par la formule de Stokes,

$$\int_{\partial K} -\lambda_i \nabla u(x) \cdot \mathbf{n}(x) d\gamma(x) = m(K) f_K,$$

où  $\mathbf{n}$  est le vecteur unitaire normal à  $\partial\Omega$ , extérieur à  $\Omega$ , et  $\gamma$  désigne le symbole d’intégration sur la frontière. On décompose ensuite le bord de chaque maille  $K$  en arêtes du maillage :  $\partial K = \bigcup_{\sigma \in \mathcal{E}_K} \bar{\sigma}$  où  $\mathcal{E}_K$  représente l’ensemble des arêtes de  $K$ . On obtient alors :

$$\sum_{\sigma \in \mathcal{E}_K} \int_{\sigma} -\lambda_i \nabla u \cdot \mathbf{n}_{K,\sigma} d\gamma(x) = m(K) f_K$$

où  $\mathbf{n}_{K,\sigma}$  est le vecteur unitaire normal à  $\sigma$  extérieur à  $K$ . On écrit alors une “équation approchée” :

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = m(K) f_K,$$

où  $F_{K,\sigma}$  est le flux numérique à travers  $\sigma$ , qui approche le flux exact  $F_{K,\sigma}^* = \int_{\sigma} -\lambda_i \nabla u \cdot \mathbf{n}_{K,\sigma} d\gamma(x)$ . Pour obtenir le schéma numérique, il nous reste à exprimer le flux numérique  $F_{K,\sigma}$  en fonction des inconnues discrètes  $(u_K)_{K \in \mathcal{T}}$  associées aux mailles et  $(u_{\sigma})_{\sigma \in \mathcal{E}}$  associées aux arêtes (ces dernières seront ensuite éliminées) :

$$F_{K,\sigma} = -\lambda_i \frac{u_{\sigma} - u_K}{d_{K,\sigma}} m(\sigma), \quad (1.65)$$

où  $d_{K,\sigma}$  est la distance du point  $x_K$  à l'arête  $\sigma$  et  $m(\sigma)$  est la longueur de l'arête  $\sigma$  (voir Figure 1.6). L'équation associée à l'inconnue  $u_K$  est donc :

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = m(K) f_K.$$

On a ainsi obtenu autant d'équations que de mailles. Il nous reste maintenant à écrire une équation pour chaque arête, afin d'obtenir autant d'équations que d'inconnues.

En ce qui concerne les arêtes intérieures, on écrit la conservativité du flux, ce qui nous permettra d'éliminer les inconnues associées aux arêtes internes. Soit  $\sigma = K|L \subset \Omega_i$ , On a alors :

$$F_{K,\sigma} = -F_{L,\sigma}. \quad (1.66)$$

On vérifiera par le calcul (cf. exercice 21 page 40) que, après élimination de  $u_{\sigma}$ , ceci donne

$$F_{K,\sigma} = -F_{L,\sigma} = \lambda_i \frac{m(\sigma)}{d_{\sigma}} (u_K - u_L), \quad (1.67)$$

où  $d_{\sigma} = d(x_K, x_L)$ .

**Remarque 1.35 (Consistance du flux)** On appelle erreur de consistance associée au flux (1.65) l'expression :

$$R_{K,\sigma} = -\frac{1}{m(\sigma)} \int_{\sigma} \nabla u(x) \cdot \mathbf{n}_{K,\sigma} d\gamma(x) - F_{K,\sigma}^*, \quad \text{où } F_{K,\sigma}^* = -\lambda_i \frac{u(x_{\sigma}) - u(x_K)}{d_{K,\sigma}} m(\sigma),$$

où  $x_{\sigma}$  est l'intersection de  $\sigma$  avec l'arête  $K|L$ ,  $u$  la solution exacte.

On dit que le flux numérique donné par l'expression (1.65) est consistant si

$$\lim_{h(\mathcal{T}) \rightarrow 0} \max_{K \in \mathcal{T}, \sigma \in \mathcal{E}_K} |R_{K,\sigma}| = 0,$$

où  $h(\mathcal{T})$  est le pas du maillage, i.e.  $h(\mathcal{T}) = \max_{K \in \mathcal{T}} \text{diam}(K)$ , avec  $\text{diam}(K) = \sup_{(x,y) \in K^2} d(x,y)$ . On vérifie facilement que si  $u$  est suffisamment régulière et si le segment  $x_K x_L$  est colinéaire au vecteur normal  $\mathbf{n}$ , alors le flux numérique est consistant. Cette propriété, alliée à la propriété de conservativité des flux, permet de démontrer la convergence du schéma, comme on l'a fait dans le cas unidimensionnel.

**Remarque 1.36 (Cas du maillage cartésien de la figure 1.5)** Dans le cas du maillage cartésien considéré pour notre problème, il est naturel de choisir les points  $x_K$  comme les centres de gravité des mailles. Comme le maillage est uniforme, on a donc  $d_{K,\sigma} = \frac{h_x}{2}$  (resp.  $\frac{h_y}{2}$ ) et  $|\sigma| = h_y$  (resp.  $|\sigma| = h_x$ ) pour une arête  $\sigma$  verticale (resp. horizontale).

Ecrivons maintenant la discrétisation des conditions aux limites et interface :

1. **Condition de Neumann sur  $\Gamma_2$**  Sur  $\Gamma_2$ , on a la condition de Neumann (1.62) :  $\lambda_i \nabla u \cdot \mathbf{n} = 0$ , qu'on discrétise par :  $\sigma \in \mathcal{E}_K$  et  $\sigma \subset \Gamma_2$ ,  $F_{K,\sigma} = 0$ .
2. **Condition de Dirichlet sur  $\Gamma_4$**  La discrétisation de la condition de Dirichlet (1.63) peut s'effectuer de la manière suivante :

$$u_{\sigma} = \frac{1}{m(\sigma)} \int_{\sigma} g(y) d\gamma(y).$$

L'expression du flux numérique est alors :

$$F_{K,\sigma} = -\lambda_i \frac{u_{\sigma} - u_K}{d_{K,\sigma}} m(\sigma).$$

3. Condition de Fourier sur  $\Gamma_1 \cup \Gamma_3$  Sur  $\Gamma_1 \cup \Gamma_3$  on a la condition de Fourier (1.61) :

$$-\lambda_i \nabla u \cdot \mathbf{n} = \alpha(u(x) - u_{ext}) \quad \forall x \in \Gamma_1 \cup \Gamma_3$$

qu'on discrétise par

$$F_{K,\sigma} = -m(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} = m(\sigma)\alpha(u_\sigma - u_{ext}) \text{ pour } \sigma \in \Gamma_1 \cup \Gamma_3.$$

Après élimination de  $u_\sigma$  (cf. exercice 21 page 40), on obtient :

$$F_{K,\sigma} = \frac{\alpha\lambda_i m(\sigma)}{\lambda_i + \alpha d_{K,\sigma}} (u_K - u_{ext}). \quad (1.68)$$

4. Condition de saut pour le flux sur  $I$  Si  $\sigma = K|L \in \mathcal{E}_I$ , la discrétisation de la condition de saut (1.64) se discrétise facilement en écrivant :

$$F_{K,\sigma} + F_{L,\sigma} = \theta_\sigma, \text{ avec } \theta_\sigma = \frac{1}{|\sigma|} \int_\sigma \theta(x) d\gamma(x). \quad (1.69)$$

3.

Après élimination de l'inconnue  $u_\sigma$  (voir exercice 21 page 40), on obtient

$$F_{K,\sigma} = \frac{\lambda_1 m(\sigma)}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}} [\lambda_2 (u_K - u_L) + d_{L,\sigma} \theta_\sigma]. \quad (1.70)$$

On a ainsi éliminé toutes les inconnues  $u_\sigma$ , ce qui permet d'obtenir un système linéaire dont les inconnues sont les valeurs  $(u_K)_{K \in \mathcal{T}}$ .

**Remarque 1.37 (Implantation informatique de la méthode)** Lors de l'implantation informatique, la matrice du système linéaire est construite "par arête" (contrairement à une matrice éléments finis, dont nous verrons plus tard la construction "par élément"), c.à.d. que pour chaque arête, on additionne la contribution du flux au coefficient de la matrice correspondant à l'équation et à l'inconnue concernées.

## 1.6 Exercices

**Exercice 1 (Comparaison différences finies- volumes finis, CL Dirichlet non homogènes)** Suggestions en page 40, corrigé en page 42.

On considère le problème :

$$\begin{aligned} -u''(x) + \sin(u(x)) &= f(x), \quad x \in ]0, 1[, \\ u(0) &= a, \quad u(1) = b, \end{aligned} \quad (1.71)$$

1. Ecrire les schémas de différences finies et volumes finis avec pas constant pour le problème (1.71). Pour le schéma volumes finis, on approchera  $\int_{x_{i-1/2}}^{x_{i+1/2}} \sin(u(x)) dx$  par  $(x_{i+1/2} - x_{i-1/2}) \sin(u(x_i))$ .

2. Comparer les schémas ainsi obtenus lorsqu'on suppose que  $u$  reste toujours "petit" et qu'on remplace donc  $\sin u$  par  $u$ .

**Exercice 2 (Comparaison différences finies- volumes finis, conditions mixtes)** Suggestions en page 40.

On considère le problème :

$$\begin{aligned} -u''(x) &= f(x), \quad x \in ]0, 1[, \\ u(0) - u'(0) &= a, \quad u'(1) = b, \end{aligned} \quad (1.72)$$

Ecrire les schémas de différences finies et volumes finis avec pas constant pour le problème (1.72), et comparer les schémas ainsi obtenus.

**Exercice 3 (Principe du maximum)** Suggestions en page 40,



On considère le problème :

$$\begin{cases} -u''(x) + c(x)u(x) = f(x), & 0 < x < 1, \\ u(0) = a, u(1) = b, \end{cases} \quad (1.73)$$

où  $c \in C([0, 1], \mathbb{R}_+)$ , et  $c \in C([0, 1], \mathbb{R})$ , et  $(a, b) \in \mathbb{R}^2$ .

1. Donner la discrétisation par différences finies de ce problème. On appelle  $U_h$  la solution approchée (c.à.d.  $U_h = (u_1, \dots, u_N)$ , où  $u_i$  est l'inconnue discrète en  $x_i$ ).

2. On suppose ici que  $c = 0$ . Montrer que  $u_i \geq \min(a, b)$ , pour tout  $i = 1, \dots, N$ .

#### Exercice 4 (Equation de diffusion réaction)

On s'intéresse au problème elliptique unidimensionnel suivant :

$$\begin{cases} -u''(x) + 2u(x) = x, & x \in ]0, 1[, \\ u(0) = 1, u'(1) + u(1) = 0. \end{cases} \quad (1.74)$$

1. Ecrire une discrétisation de (1.74) par différences finies pour un maillage uniforme. Ecrire le système linéaire obtenu.

2. Ecrire une discrétisation de (1.74) par volumes finis de (1.74) pour un maillage uniforme. Ecrire le système linéaire obtenu.

#### Exercice 5 (Equation de transport-diffusion sous forme non-conservative) *Corrigé en page 43.*

Cet exercice ainsi que le suivant concernent la discrétisation d'une équation de transport-diffusion sous forme non-conservative (exercice 5) puis conservative (exercice 6). On a déjà vu dans le cours que en une dimension d'espace, le terme de diffusion unidimensionnel est de la forme  $-u''$  (tout du moins dans le cas d'un matériau homogène de conductivité constante). On appelle terme de transport (en 1D) un terme de la forme  $v(x)u'(x)$  (forme dite non conservative) ou  $(v(x)u(x))'$  (forme conservative), où  $v$  est la "vitesse de transport" (donnée) et  $u$  l'inconnue, qui est la quantité transportée (une concentration de polluant, par exemple). Remarquez d'abord que si la vitesse  $v$  est constante, les deux formes sont identiques, puisque  $(v(x)u(x))' = v'(x)u(x) + v(x)u'(x) = v(x)u'(x)$ . La deuxième forme est dite conservative car elle est obtenue à partir de l'écriture de la conservation de la masse (par exemple) sur un petit élément  $x + \delta x$ , en passant à la limite lorsque  $\delta x$  tend vers 0. La première forme, non conservative, apparaît dans des modèles de mécanique de fluides (écoulements compressibles polyphasiques, par exemple).

Soient  $v \in C([0, 1], \mathbb{R}_+)$  et  $a_0, a_1 \in \mathbb{R}$ .

1. On considère le problème suivant :

$$\begin{cases} -u''(x) + v(x)u_x(x) = 0, & x \in ]0, 1[, \\ u(0) = a_0, u(1) = a_1. \end{cases} \quad (1.75)$$

On admettra qu'il existe une unique solution  $u \in C([0, 1], \mathbb{R}) \cap C^2(]0, 1[, \mathbb{R})$  à ce problème. On cherche à approcher cette solution par une méthode de différences finies. On se donne un pas de maillage  $h = \frac{1}{N+1}$  uniforme, des inconnues discrètes  $u_1, \dots, u_N$  censées approcher les valeurs  $u(x_1), \dots, u(x_N)$ . On considère le schéma aux différences finies suivant :

$$\begin{cases} \frac{1}{h^2}(2u_i - u_{i+1} - u_{i-1}) + \frac{1}{h}v_i(u_i - u_{i-1}) = 0, & i = 1, \dots, N \\ u(0) = a_0, u(1) = a_1, \end{cases} \quad (1.76)$$

où  $v_i = v(x_i)$ , pour  $i = 1, \dots, N$ .

Noter que le terme de convection  $v(x_i)u'(x_i)$  est approché par  $v(x_i)\frac{u(x_{i+1/2}) - u(x_{i-1/2}))}{h}$ . Comme la vitesse  $v_i$  est positive ou nulle, on choisit d'approcher  $u(x_{i+1/2})$  par la valeur "amont", c.à.d.  $u(x_i)$ ; d'où le schéma.

1. Montrer que le système (1.76) s'écrit sous la forme  $MU = b$  avec  $U = (u_1, \dots, u_N)$ ,  $b \in \mathbb{R}^N$ , et  $M$  est une matrice telle que :

(a)  $MU \geq 0 \Rightarrow U \geq 0$  (les inégalités s'entendent composante par composante),

(b)  $M$  est inversible,

(c) Si  $U$  est solution de  $MU = b$  alors  $\min(a_0, a_1) \leq u_i \leq \max(a_0, a_1)$ .

2. Montrer que  $M$  est une M-matrice, c. à.d. que  $M$  vérifie :

(a)  $m_{i,i} > 0$  pour  $i = 1, \dots, n$ ;

(b)  $m_{i,j} \leq 0$  pour  $i, j = 1, \dots, n, i \neq j$ ;

(c)  $M$  est inversible;

(d)  $M^{-1} \geq 0$ ;

**Exercice 6 (Equation de transport-diffusion sous forme conservative)** *Il est conseillé d'étudier l'exercice 5 avant celui-ci.*

Soit  $v \in C([0, 1], \mathbb{R}_+) \cap C^1(]0, 1[, \mathbb{R})$ , et on considère le problème :

$$\begin{cases} -u''(x) + (vu)'(x) = 0, & x \in ]0, 1[, \\ u(0) = a_0, u(1) = a_1. \end{cases} \quad (1.77)$$

On admettra qu'il existe une unique solution  $u \in C([0, 1], \mathbb{R}) \cap C^2(]0, 1[, \mathbb{R})$  à ce problème. On cherche ici encore à approcher cette solution par une méthode de différences finies. On se donne un pas de maillage  $h = \frac{1}{N+1}$  uniforme, des inconnues discrètes  $u_1, \dots, u_N$  censées approcher les valeurs  $u(x_1), \dots, u(x_N)$ . On considère le schéma aux différences finies suivant :

$$\begin{cases} \frac{2u_i - u_{i+1} - u_{i-1}}{h^2} + \frac{1}{h}(v_{i+\frac{1}{2}}u_i - v_{i-\frac{1}{2}}u_{i-1}) = 0, & i = 1, \dots, N \\ u(0) = a_0, u(1) = a_1, \end{cases} \quad (1.78)$$

où  $v_{i+\frac{1}{2}} = v(\frac{x_i+x_{i+1}}{2})$ , pour  $i = 0, \dots, N$ .

Noter que terme de convection  $(vu)'(x_i)$  peut être approché par  $\frac{1}{h}(v(x_{i+\frac{1}{2}})u(x_{i+\frac{1}{2}}) - v(x_{i-\frac{1}{2}})u(x_{i-\frac{1}{2}}))$ . Comme  $v(x_{i+\frac{1}{2}}) \geq 0$ , on choisit d'approcher  $u(x_{i+\frac{1}{2}})$  par la valeur "amont", c.à.d.  $u(x_i)$ . C'est une valeur amont dans le sens où elle est choisie en amont de l'écoulement, si l'on suppose que  $v$  est la vitesse de l'écoulement. On dit que le schéma est "décentré amont".

1. Montrer que le système (1.78) s'écrit sous la forme  $MU = b$  avec  $U = (u_1, \dots, u_N)$ ,  $b \in \mathbb{R}^N$ ,

2. Pour  $U = (u_1, \dots, u_N)$  et  $W = (w_1, \dots, w_N) \in \mathbb{R}^N$ , calculer  $MU \cdot W$ , et en déduire l'expression de  $(M^t W)_i$ , pour  $i = 1, \dots, N$  (on distinguera les cas  $i = 2, \dots, N-1$ ,  $i = 1$  et  $i = N$ ).

3. Soit  $W \in \mathbb{R}^N$ ;

3. (a) montrer que si  $M^t W \geq 0$  alors  $W \geq 0$ ; en déduire que si  $U \in \mathbb{R}^N$  est tel que  $MU \geq 0$  alors  $U \geq 0$ .

3. (b) en déduire que si  $U \in \mathbb{R}^N$  est tel que  $MU \geq 0$  alors  $U \geq 0$ .

4. Montrer que  $M$  est une M-matrice.

5. Montrer que  $U$  solution de (1.78) peut ne pas vérifier  $\min(a_0, a_1) \leq u_i \leq \max(a_0, a_1)$ .

**Exercice 7 (Conditionnement "efficace".)** *Suggestions en page 40, corrigé en page 43.*

Soit  $f \in C([0, 1])$ . Soit  $N \in \mathbb{N}^*$ ,  $N$  impair. On pose  $h = 1/(N+1)$ . Soit  $A$  la matrice définie par (1.35) page 16, issue d'une discrétisation par différences finies (vue en cours) du problème (1.1)–(1.2) (page 6).

Pour  $u \in \mathbb{R}^N$ , on note  $u_1, \dots, u_N$  les composantes de  $u$ . Pour  $u \in \mathbb{R}^N$ , on dit que  $u \geq 0$  si  $u_i \geq 0$  pour tout  $i \in \{1, \dots, N\}$ . Pour  $u, v \in \mathbb{R}^N$ , on note  $u \cdot v = \sum_{i=1}^N u_i v_i$ .

On munit  $\mathbb{R}^N$  de la norme suivante : pour  $u \in \mathbb{R}^N$ ,  $\|u\| = \max\{|u_i|, i \in \{1, \dots, N\}\}$ . On munit alors  $\mathcal{M}_N(\mathbb{R})$  de la norme induite, également notée  $\|\cdot\|$ , c'est-à-dire  $\|B\| = \max\{\|Bu\|, u \in \mathbb{R}^N \text{ t.q. } \|u\| = 1\}$ , pour tout  $B \in \mathcal{M}_N(\mathbb{R})$ .

**Partie I** Conditionnement de la matrice et borne sur l'erreur relative

1. (Existence et positivité de  $A^{-1}$ ) Soient  $b \in \mathbb{R}^N$  et  $u \in \mathbb{R}^N$  t.q.  $Au = b$ . Remarquer que  $Au = b$  peut s'écrire :

$$\begin{cases} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i, & \forall i \in \{1, \dots, N\}, \\ u_0 = u_{N+1} = 0. \end{cases} \quad (1.79)$$

Montrer que  $b \geq 0 \Rightarrow u \geq 0$ . [On pourra considérer  $p \in \{0, \dots, N+1\}$  t.q.  $u_p = \min\{u_j, j \in \{0, \dots, N+1\}\}$ .]

En déduire que  $A$  est inversible.

- (Préliminaire. . .) On considère la fonction  $\varphi \in C([0, 1], \mathbb{R})$  définie par  $\varphi(x) = (1/2)x(1-x)$  pour tout  $x \in [0, 1]$ . On définit alors  $\phi \in \mathbb{R}^N$  par  $\phi_i = \phi(ih)$  pour tout  $i \in \{1, \dots, N\}$ . Montrer que  $(A\phi)_i = 1$  pour tout  $i \in \{1, \dots, N\}$ .
- (calcul de  $\|A^{-1}\|$ ) Soient  $b \in \mathbb{R}^N$  et  $u \in \mathbb{R}^N$  t.q.  $Au = b$ . Montrer que  $\|u\| \leq (1/8)\|b\|$  [Calculer  $A(u \pm \|b\|\phi)$  avec  $\phi$  défini à la question 2 et utiliser la question 1]. En déduire que  $\|A^{-1}\| \leq 1/8$  puis montrer que  $\|A^{-1}\| = 1/8$ .
- (calcul de  $\|A\|$ ) Montrer que  $\|A\| = \frac{4}{h^2}$ .
- (Conditionnement pour la norme  $\|\cdot\|$ ). Calculer  $\|A^{-1}\|\|A\|$ . Soient  $b, \delta_b \in \mathbb{R}^N$ . Soient  $u, \delta_u \in \mathbb{R}^N$  t.q.  $Au = b$  et  $A(u + \delta_u) = b + \delta_b$ . Montrer que  $\frac{\|\delta_u\|}{\|u\|} \leq \|A^{-1}\|\|A\| \frac{\|\delta_b\|}{\|b\|}$ .  
Montrer qu'un choix convenable de  $b$  et  $\delta_b$  donne l'égalité dans l'inégalité précédente.

### Partie II Borne réaliste sur l'erreur relative : Conditionnement "efficace"

On se donne maintenant  $f \in C([0, 1], \mathbb{R})$  et on suppose (pour simplifier. . .) que  $f(x) > 0$  pour tout  $x \in ]0, 1[$ . On prend alors, dans cette partie,  $b_i = f(ih)$  pour tout  $i \in \{1, \dots, N\}$ . On considère aussi le vecteur  $\varphi$  défini à la question 2 de la partie I.

- Montrer que  $h \sum_{i=1}^N b_i \varphi_i \rightarrow \int_0^1 f(x)\phi(x)dx$  quand  $N \rightarrow \infty$  et que  $\sum_{i=1}^N b_i \varphi_i > 0$  pour tout  $N$ . En déduire qu'il existe  $\alpha > 0$ , ne dépendant que de  $f$ , t.q.  $h \sum_{i=1}^N b_i \varphi_i \geq \alpha$  pour tout  $N \in \mathbb{N}^*$ .
- Soit  $u \in \mathbb{R}^N$  t.q.  $Au = b$ . Montrer que  $N\|u\| \geq \sum_{i=1}^N u_i = u \cdot A\varphi \geq \frac{\alpha}{h}$  (avec  $\alpha$  donné à la question 1).  
Soit  $\delta_b \in \mathbb{R}^N$  et  $\delta_u \in \mathbb{R}^N$  t.q.  $A(u + \delta_u) = b + \delta_b$ . Montrer que  $\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_{L^\infty(]0,1[)}}{8\alpha} \frac{\|\delta_b\|}{\|b\|}$ .
- Comparer  $\|A^{-1}\|\|A\|$  (question I.5) et  $\frac{\|f\|_{L^\infty(]0,1[)}}{8\alpha}$  (question II.2) quand  $N$  est "grand" (ou quand  $N \rightarrow \infty$ ).

### Exercice 8 (Conditionnement, réaction diffusion 1d.)

On s'intéresse au conditionnement pour la norme euclidienne de la matrice issue d'une discrétisation par Différences Finies du problème aux limites suivant :

$$\begin{aligned} -u''(x) + u(x) &= f(x), \quad x \in ]0, 1[, \\ u(0) &= u(1) = 0. \end{aligned} \tag{1.80}$$

Soit  $N \in \mathbb{N}^*$ . On note  $U = (u_j)_{j=1 \dots N}$  une "valeur approchée" de la solution  $u$  du problème (1.80) aux points  $\left(\frac{j}{N+1}\right)_{j=1 \dots N}$ .

1. Montrer que la discrétisation par différences finies de ce problème sur maillage uniforme de pas  $h = \frac{1}{N+1}$  consiste à chercher  $U$  comme solution du système linéaire  $AU = \left(f\left(\frac{j}{N+1}\right)\right)_{j=1 \dots N}$  où la matrice  $A \in M_N(\mathbb{R})$  est définie par  $A = (N+1)^2 B + Id$ ,  $Id$  désigne la matrice identité et

$$B = \begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \dots & 0 & -1 & 2 \end{pmatrix}$$

2. (Valeurs propres de la matrice  $B$ .)

On rappelle que le problème aux valeurs propres

$$\begin{aligned} -u''(x) &= \lambda u(x), \quad x \in ]0, 1[, \\ u(0) &= u(1) = 0. \end{aligned} \tag{1.81}$$

admet la famille  $(\lambda_k, u_k)_{k \in \mathbb{N}^*}$ ,  $\lambda_k = (k\pi)^2$  et  $u_k(x) = \sin(k\pi x)$  comme solution. Montrer que les vecteurs  $U_k = \left(u_k\left(\frac{j}{N+1}\right)\right)_{j=1 \dots N}$  sont des vecteurs propres de la matrice  $B$ . En déduire toutes les valeurs propres de la matrice  $B$ .

3. En déduire les valeurs propres de la matrice  $A$ .

4. En déduire le conditionnement pour la norme euclidienne de la matrice  $A$ .

**Exercice 9 (Erreur de consistance)** Suggestions en page 41 Corrigé en page 45.

On considère la discrétisation à pas constant par le schéma aux différences finies symétrique à trois points (vu en cours) du problème du problème (1.1)–(1.2) (page 6), avec  $f \in C([0, 1])$ . Soit  $N \in \mathbb{N}^*$ ,  $N$  impair. On pose  $h = 1/(N + 1)$ . On note  $u$  la solution exacte,  $x_i = ih$ , pour  $i = 1, \dots, N$  les points de discrétisation, et  $(u_i)_{i=1, \dots, N}$  la solution du système discrétisé.

1. Ecrire le système linéaire obtenu.

2. Montrer que si  $f$  est constante, alors

$$\max_{1 \leq i \leq N} |u_i - u(x_i)| = 0.$$

3. Soit  $N$  fixé, et  $\max_{1 \leq i \leq N} |u_i - u(x_i)| = 0$ . A-t-on forcément que  $f$  est constante sur  $[0, 1]$ ? (justifier la réponse.)

**Exercice 10 (Problème elliptique 1d, discrétisation par différences finies)** <sup>14</sup> Suggestions en page 41.

Soit  $f \in C^2([0, 1])$ . On s'intéresse au problème suivant :

$$\begin{aligned} -u''(x) + \frac{1}{1+x}u'(x) &= f(x), \quad x \in ]0, 1[, \\ u(0) = a \quad u(1) &= b. \end{aligned} \tag{1.82}$$

On admet que ce problème admet une et une seule solution  $u$  et on suppose que  $u \in C^4(]0, 1[)$ . On cherche une solution approchée de (1.82) par la méthode des différences finies. Soit  $n \in \mathbb{N}^*$ , et  $h = \frac{1}{N+1}$ . On note  $u_i$  la valeur approchée recherchée de  $u$  au point  $ih$ , pour  $i = 0, \dots, N + 1$ .

On utilise les approximations centrées les plus simples de  $u'$  et  $u''$  aux points  $ih, i = 1, \dots, n$ . On pose  $u_h = (u_1, \dots, u_n)$ .

1. Montrer que  $u_h$  est solution d'un système linéaire de la forme  $A_h u_h = b_h$ ; donner  $A_h$  et  $b_h$ .

2. Montrer que le schéma numérique obtenu est consistant et donner une majoration de l'erreur de consistance (on rappelle que l'on a supposé  $u \in C^4$ ).

3. Soit  $v \in \mathbb{R}^n$ , montrer que  $A_h v \geq 0 \Rightarrow v \geq 0$  (ceci s'entend composante par composante). Cette propriété s'appelle conservation de la positivité. En déduire que  $A_h$  est monotone.

4. On définit  $\theta$  par

$$\theta(x) = -\frac{1}{2}(1+x)^2 \ln(1+x) + \frac{2}{3}(x^2 + 2x) \ln 2, \quad x \in [0, 1].$$

4.a. Montrer qu'il existe  $C \geq 0$ , indépendante de  $h$ , t.q.

$$\max_{1 \leq i \leq n} \left| \frac{1}{h^2}(-\theta_{i-1} + 2\theta_i - \theta_{i+1}) + \frac{1}{2h(1+ih)}(\theta_{i+1} - \theta_{i-1}) - 1 \right| \leq Ch^2,$$

avec  $\theta_i = \theta(x_i)$ ,  $i = 0, \dots, n + 1$ .

4.b On pose  $\theta_h = (\theta_1, \dots, \theta_N)$ . Montrer que  $(A_h \theta_h)_i \geq 1 - Ch^2$ , pour  $i = 1, \dots, N$ .

4.c Montrer qu'il existe  $M \geq 0$  ne dépendant pas de  $h$  t.q.  $\|A_h^{-1}\|_\infty \leq M$ .

5. Montrer la convergence, en un sens à définir, de  $u_h$  vers  $u$ .

6. Que peut on dire si  $u \notin C^4$ , mais seulement  $u \in C^2$  ou  $C^3$ ?

7. On remplace dans (1.82)  $\frac{1}{1+x}$  par  $\alpha u'(x)$ , avec  $\alpha$  donné (par exemple  $\alpha = 100$ ). On utilise pour approcher (1.82) le même principe que précédemment (approximations centrées de  $u'$  et  $u''$ ). Que peut on dire sur la consistance, la stabilité, la convergence du schéma numérique?

<sup>14</sup>Cet exercice est tiré du livre *Exercices d'analyse numérique matricielle et d'optimisation*, de P.G. Ciarlet et J.M. Thomas, Collection Mathématiques pour la maîtrise, Masson, 1982

**Exercice 11 (Non consistance des volumes finis)** *Suggestions en page 41, corrigé en page 45*

Montrer que la discrétisation de l'opérateur  $-u''$  par le schéma volumes finis n'est pas toujours consistante au sens des différences finies, *i.e.* que l'erreur de consistance définie par (voir remarque 1.22 page 19)

$$R_i = \frac{1}{h_i} \left[ \frac{-1}{h_{i+1/2}} (u(x_{i+1}) - u(x_i)) + \frac{1}{h_{i-1/2}} (u(x_i) - u(x_{i-1})) \right] - u''(x_i)$$

ne tend pas toujours vers 0 lorsque  $h$  tend vers 0.

**Exercice 12 (Condition de Fourier)** *Suggestions en page 41*

On reprend ici la discrétisation du flux sortant en 1 pour le problème (1.1) comme la condition de Fourier (1.15) en 1 et la condition de Neumann (1.14) en 0. On reprend les notations du paragraphe 1.1.1 page 7, et pour simplifier on note  $h = h_N$  et on suppose que  $x_N$  est au milieu de la  $N$ -ème maille, de sorte que  $h_{N+1/2} = \frac{h}{2}$ . On modifie l'approximation de l'équation (1.6) : au lieu d'approcher  $u'(1)$  par  $b - \alpha u_N$ , on introduit une inconnue auxiliaire  $u_{N+1/2}$  sensée approcher la valeur  $u(1)$ , on approche ensuite  $u'(1)$  par  $\frac{u_{N+1/2} - u_{N-1}}{h_{N-1/2}}$ .

1. Montrer que par cette méthode, en éliminant l'inconnue auxiliaire  $u_{N+1/2}$ , on obtient comme  $N$ -ème équation discrète non plus (1.17) mais l'équation suivante :

$$F_{N+1/2} - F_{N-1/2} = h_N f_N \text{ avec } F_{N+1/2} = \frac{1}{1 + \frac{\alpha h}{2}} (\alpha u_N - b) \text{ et } F_{N-1/2} = -\frac{u_N - u_{N-1}}{h_{N-1/2}} \quad (1.83)$$

2. Calculer l'erreur de consistance sur le flux approché  $F_{N+1/2}$  en 1 dans le cas des discrétisations (1.17) et (1.83). Montrer qu'elle est d'ordre 1 dans le premier cas, et d'ordre 2 dans le second.

**Exercice 13 (Consistance des flux)** *Suggestions en page 41, Corrigé en page 46.*

1. Montrer que si  $u \in C^2([0, 1])$ , le flux défini par (1.40) est consistant d'ordre 1 dans le cas d'un maillage général.

2. Montrer que si  $u \in C^3([0, 1])$ , le flux défini par (1.40) est d'ordre 2 si  $x_{i+1/2} = (x_{i+1} + x_i)/2$ .

**Exercice 14 (Conditions aux limites de Neumann)** *Suggestions en page 41*

On considère ici l'équation le problème de diffusion réaction avec conditions aux limites de Neumann homogènes (correspondant à une condition physique de flux nul sur le bord) :

$$\begin{cases} -u''(x) + cu(x) = f(x), & x \in ]0, 1[, \\ u'(0) = u'(1) = 0, \end{cases} \quad (1.84)$$

avec  $c \in \mathbb{R}_+^*$ , et  $f \in C([0, 1])$ . Donner la discrétisation de ce problème par

1. différences finies,
2. volumes finis

Montrer que les matrices obtenues ne sont pas inversibles. Proposer une manière de faire en sorte que le problème soit bien posé, compatible avec ce qu'on connaît du problème continu.

**Exercice 15 (Conditions aux limites de Fourier (ou Robin))** *Suggestions en page 41, corrigé en page 47*

On considère le problème :

$$\begin{cases} -u''(x) + cu(x) = f(x), & x \in ]0, 1[, \\ u'(0) - \alpha(u - \tilde{u}) = 0, \\ u'(1) + \alpha(u - \tilde{u}) = 0, \end{cases} \quad (1.85)$$

avec  $c \in \mathbb{R}_+$ ,  $f \in C([0, 1])$ ,  $\alpha \in \mathbb{R}_+^*$ , et  $\tilde{u} \in \mathbb{R}$ .

Donner la discrétisation de ce problème par

1. différences finies,
2. volumes finis

Dans les deux cas, écrire le schéma sous la forme d'un système linéaire de  $N$  équations à  $N$  inconnues, en explicitant matrice et second membre ( $N$  est le nombre de noeuds internes en différences finies, de mailles en volumes finis).

### Exercice 16 (Convergence de la norme $H^1$ discrète)

Montrer que si  $u_{\mathcal{T}} : ]0, 1[ \rightarrow \mathbb{R}$  est définie par  $u_{\mathcal{T}}(x) = u_i \quad \forall x \in K_i$  où  $(u_i)_{i=1, \dots, N}$  solution de (1.39)–(1.41), alors  $|u_{\mathcal{T}}|_{1, \mathcal{T}}$  converge dans  $L^2(]0, 1[)$  lorsque  $h$  tend vers 0, vers  $\|Du\|_{L^2(]0, 1[)}$ , où  $u$  est la solution du problème (1.1)–(1.2).

### Exercice 17 (Problème elliptique 1d, discrétisation par volumes finis) *Suggestions en page 41, corrigé en page 47*

Soient  $a, b \geq 0, c, d \in \mathbb{R}$  et  $f \in C([0, 1], \mathbb{R})$ ; on cherche à approcher la solution  $u$  du problème suivant :

$$-u''(x) + au'(x) + b(u(x) - f(x)) = 0, \quad x \in [0, 1], \quad (1.86)$$

$$u(0) = c, \quad u(1) = d. \quad (1.87)$$

On suppose (mais il n'est pas interdit d'expliquer pourquoi...) que (1.86)–(1.87) admet une solution unique  $u \in C^2([0, 1], \mathbb{R})$ .

Soient  $N \in \mathbb{N}^*$  et  $h_1, \dots, h_N > 0$  t.q.  $\sum_{i=1}^N h_i = 1$ . On pose  $x_{\frac{1}{2}} = 0, x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + h_i$ , pour  $i = 1, \dots, N$  (de sorte que  $x_{N+\frac{1}{2}} = 1$ ),  $h_{i+\frac{1}{2}} = \frac{h_{i+1} + h_i}{2}$ , pour  $i = 1, \dots, N-1$ , et  $f_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x) dx$ , pour  $i = 1, \dots, N$ .

Pour approcher la solution  $u$  de (1.86)–(1.87), on propose le schéma numérique suivant :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + bh_i u_i = bh_i f_i, \quad i \in \{1, \dots, N\}, \quad (1.88)$$

avec  $(F_{i+\frac{1}{2}})_{i \in \{0, \dots, N\}}$  donné par les expressions suivantes :

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + au_i, \quad i \in \{1, \dots, N-1\}, \quad (1.89)$$

$$F_{\frac{1}{2}} = -\frac{u_1 - c}{\frac{h_1}{2}} + ac, \quad F_{N+\frac{1}{2}} = -\frac{d - u_N}{\frac{h_N}{2}} + au_N. \quad (1.90)$$

En tenant compte des expressions (1.89) et (1.90), le schéma numérique (1.88) donne donc un système de  $N$  équations à  $N$  inconnues (les inconnues sont  $u_1, \dots, u_N$ ).

1. Expliquer comment, à partir de (1.86) et (1.87), on obtient ce schéma numérique.

2. (Existence de la solution approchée.)

(a) On suppose ici que  $c = d = 0$  et  $f_i = 0$  pour tout  $i \in \{1, \dots, N\}$ . Montrer qu'il existe un unique vecteur  $U = (u_1, \dots, u_N) \in \mathbb{R}^N$  solution de (1.88). Ce vecteur est obtenu en prenant  $u_i = 0$ , pour tout  $i \in \{1, \dots, N\}$ . (On rappelle que dans (1.88) les termes  $F_{i+\frac{1}{2}}$  et  $F_{i-\frac{1}{2}}$  sont donnés par (1.89) et (1.90).)

(b) On revient maintenant au cas général (c'est à dire  $c, d \in \mathbb{R}$  et  $f \in C([0, 1], \mathbb{R})$ ). Montrer qu'il existe un unique vecteur  $U = (u_1, \dots, u_N) \in \mathbb{R}^N$  solution de (1.88). (On rappelle, encore une fois, que dans (1.88) les termes  $F_{i+\frac{1}{2}}$  et  $F_{i-\frac{1}{2}}$  sont donnés par (1.89) et (1.90).)

Soient  $\alpha, \beta > 0$ . On suppose, dans tout la suite de l'exercice, qu'il existe  $h > 0$  tel que  $\alpha h \leq h_i \leq \beta h$ , pour tout  $i \in \{1, \dots, N\}$ . On note  $\bar{u}_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u(x) dx$ , pour  $i = 1, \dots, N$ . (On rappelle que  $u$  est la solution exacte de (1.86)–(1.87).)

3. (Non consistance du schéma au sens des différences finies)

(a) Montrer que le système peut se mettre sous la forme  $AU = B$ , où  $B$  est définie par

$$B_1 = bf_1 + \frac{2c}{h_1^2} + \frac{ac}{h_1},$$

$$B_i = bf_i, \quad i = 2, \dots, N-1,$$

$$B_N = bf_n + \frac{2d}{h_N^2}.$$

(b) On pose  $\bar{R} = A\bar{U} - B$  avec  $\bar{U} = (\bar{u}_1, \dots, \bar{u}_N)$ . Vérifier que pour tout  $i \in \{1, \dots, N\}$ ,  $\bar{R}_i$  peut se mettre sous la forme :

$$\bar{R}_i = \bar{R}_i^1 + \bar{R}_i^2$$

où  $\sup_{i=1, \dots, N} |\bar{R}_i^1| \leq C_1$  et  $\sup_{i=1, \dots, N} |\bar{R}_i^2| \leq C_2 h$ .

(c) On se restreint dans cette question au cas où  $a = 0, b > 0, f = 0, c = 1, d = e^{\sqrt{b}}, N = 2q, h_i = h$  si  $i$  est pair et  $h_i = \frac{h}{2}$  si  $i$  est impair, avec  $h = \frac{2}{3N}$ .

Montrer que  $\|\bar{R}\|_\infty$  ne tend pas vers 0 avec  $h$ .

4. (Consistance des flux.) En choisissant convenablement  $(\bar{F}_{i+\frac{1}{2}})_{i \in \{0, \dots, N\}}$ , montrer que :

$$\bar{F}_{i+\frac{1}{2}} - \bar{F}_{i-\frac{1}{2}} + bh_i \bar{u}_i = bh_i f_i, \quad i \in \{1, \dots, N\}, \quad (1.91)$$

et que  $(\bar{F}_{i+\frac{1}{2}})_{i \in \{0, \dots, N\}}$  vérifie les égalités suivantes :

$$\bar{F}_{i+\frac{1}{2}} = -\frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+\frac{1}{2}}} + a\bar{u}_i + R_{i+\frac{1}{2}}, \quad i \in \{1, \dots, N-1\}, \quad (1.92)$$

$$\bar{F}_{\frac{1}{2}} = -\frac{\bar{u}_1 - c}{\frac{h_1}{2}} + ac + R_{\frac{1}{2}}, \quad \bar{F}_{N+\frac{1}{2}} = -\frac{d - \bar{u}_N}{\frac{h_N}{2}} + a\bar{u}_N + R_{N+\frac{1}{2}}, \quad (1.93)$$

avec,

$$|R_{i+\frac{1}{2}}| \leq C_1 h, \quad i \in \{0, \dots, N\}, \quad (1.94)$$

où  $C_1 \in \mathbb{R}$ , et  $C_1$  ne dépend que de  $\alpha, \beta$ , et  $u$ .

5. (Estimation d'erreur.) On pose  $e_i = \bar{u}_i - u_i$ , pour  $i \in \{1, \dots, N\}$  et  $E = (e_1, \dots, e_N)$ .

(a) Montrer que  $E$  est solution du système (de  $N$  équations) suivant :

$$G_{i+\frac{1}{2}} - G_{i-\frac{1}{2}} + bh_i e_i = 0, \quad i \in \{1, \dots, N\}, \quad (1.95)$$

avec  $(G_{i+\frac{1}{2}})_{i \in \{0, \dots, N\}}$  donné par les expressions suivantes :

$$G_{i+\frac{1}{2}} = -\frac{e_{i+1} - e_i}{h_{i+\frac{1}{2}}} + ae_i + R_{i+\frac{1}{2}}, \quad i \in \{1, \dots, N-1\}, \quad (1.96)$$

$$G_{\frac{1}{2}} = -\frac{e_1}{\frac{h_1}{2}} + R_{\frac{1}{2}}, \quad G_{N+\frac{1}{2}} = -\frac{-e_N}{\frac{h_N}{2}} + ae_N + R_{N+\frac{1}{2}}, \quad (1.97)$$

(b) En multipliant (1.95) par  $e_i$  et en sommant sur  $i = 1, \dots, N$ , montrer qu'il existe  $C_2 \in \mathbb{R}$ , ne dépendant que de  $\alpha, \beta$ , et  $u$  tel que :

$$\sum_{i=0}^N (e_{i+1} - e_i)^2 \leq C_2 h^3, \quad (1.98)$$

avec  $e_0 = e_{N+1} = 0$ .

(c) Montrer qu'il existe  $C_3 \in \mathbb{R}$ , ne dépendant que de  $\alpha, \beta$ , et  $u$  tel que :

$$|e_i| \leq C_3 h, \quad \text{pour tout } i \in \{1, \dots, N\}. \quad (1.99)$$

6. (Principe du maximum.) On suppose, dans cette question, que  $f(x) \leq d \leq c$ , pour tout  $x \in [0, 1]$ . Montrer que  $u_i \leq c$ , pour tout  $i \in \{1, \dots, N\}$ . (On peut aussi montrer que  $u(x) \leq c$ , pour tout  $x \in [0, 1]$ .)
7. On remplace, dans cette question, (1.89) et (1.90) par :

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + au_{i+1}, \quad i \in \{1, \dots, N-1\}, \quad (1.100)$$

$$F_{\frac{1}{2}} = -\frac{u_1 - c}{\frac{h_1}{2}} + au_1, \quad F_{N+\frac{1}{2}} = -\frac{d - u_N}{\frac{h_N}{2}} + ad. \quad (1.101)$$

Analyser brièvement le nouveau schéma obtenu (existence de la solution approchée, consistance des flux, estimation d'erreur, principe du maximum).

### Exercice 18 (Discrétisation 2D par différences finies)

Ecrire le système linéaire obtenu lorsqu'on discrétise le problème

$$\begin{cases} -\Delta u = f \text{ dans } \Omega = ]0, 1[ \times ]0, 1[, \\ u = 0 \text{ sur } \partial\Omega. \end{cases} \quad (1.102)$$

par différences finies avec un pas uniforme  $h = 1/N$  dans les deux directions d'espace. Montrer l'existence et l'unicité de la solution du système linéaire obtenu.

### Exercice 19 (Problème elliptique 2d, discrétisation par DF) *Corrigé en page 1.8 page 50*

Soit  $\Omega = ]0, 1[^2 \subset \mathbb{R}^2$ . On se propose d'étudier deux schémas numériques pour le problème suivant :

$$\begin{cases} -\Delta u(x, y) + k \frac{\partial u}{\partial x}(x, y) = f(x, y), & (x, y) \in \Omega, \\ u = 0, & \text{sur } \partial\Omega, \end{cases} \quad (1.103)$$

où  $k > 0$  est un réel donné et  $f \in C(\bar{\Omega})$  est donnée. On note  $u$  la solution exacte de (1.103) et on suppose que  $u \in C^4(\bar{\Omega})$ .

1. (Principe du maximum)

Montrer que pour tout  $\varphi \in C^1(\bar{\Omega})$  t.q.  $\varphi = 0$  sur  $\partial\Omega$ , on a :

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx + \int_{\Omega} k \frac{\partial u}{\partial x}(x) \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx.$$

En déduire que si  $f \leq 0$  sur  $\bar{\Omega}$ , on a alors  $u \leq 0$  sur  $\bar{\Omega}$ .

Soit  $N \in \mathbb{N}$ , on pose  $h = \frac{1}{N+1}$ , et  $u_{i,j}$  est la valeur approchée recherchée de  $u(ih, jh)$ ,  $(i, j) \in \{0, \dots, N+1\}^2$ . On pose  $f_{i,j} = f(ih, jh)$ , pour tout  $(i, j) \in \{1, \dots, N\}^2$ . On s'intéresse à deux schémas de la forme :

$$\begin{cases} a_0 u_{i,j} - a_1 u_{i-1,j} - a_2 u_{i+1,j} - a_3 u_{i,j-1} - a_4 u_{i,j+1} = f_{i,j}, \quad \forall (i, j) \in \{1, \dots, N\}^2, \\ u_{i,j} = 0, \quad (i, j) \in \gamma, \end{cases} \quad (1.104)$$

où  $a_0, a_1, a_2, a_3, a_4$  sont données (ce sont des fonctions données de  $h$ ) et  $\gamma = \{(i, j), (ih, jh) \in \partial\Omega\}$  ( $\gamma$  dépend aussi de  $h$ ). Le premier schéma, schéma [I], correspond au choix suivant des  $a_i$  :

$$a_0 = \frac{4}{h^2}, \quad a_1 = \frac{1}{h^2} + \frac{k}{2h}, \quad a_2 = \frac{1}{h^2} - \frac{k}{2h}, \quad a_3 = a_4 = \frac{1}{h^2}.$$

Le deuxième schéma, schéma [II], correspond au choix suivant des  $a_i$  :

$$a_0 = \frac{4}{h^2} + \frac{k}{h}, \quad a_1 = \frac{1}{h^2} + \frac{k}{h}, \quad a_2 = a_3 = a_4 = \frac{1}{h^2}.$$



## 2. (Consistance)

Donner une majoration de l'erreur de consistance en fonction de  $k$ ,  $h$  et des dérivées de  $u$ , pour les schémas [I] et [II]. Donner l'ordre des schémas [I] et [II].

## 3. (Principe du maximum discret)

Dans le cas du schéma [II] montrer que si  $(w_{i,j})$  vérifie :

$$a_0 w_{i,j} - a_1 w_{i-1,j} - a_2 w_{i+1,j} - a_3 w_{i,j-1} - a_4 w_{i,j+1} \leq 0, \quad \forall (i,j) \in \{1, \dots, N\}^2,$$

on a alors

$$w_{i,j} \leq \max_{(n,m) \in \gamma} (w_{n,m}), \quad \forall (i,j) \in \{1, \dots, N\}^2.$$

Montrer que ceci est aussi vrai dans le cas du schéma [I] si  $h$  vérifie une condition à déterminer.

## 4. (Stabilité)

Montrer que le schéma [II] et le schéma [I] sous la condition trouvée en 3. sont stables (au sens  $\|U\|_\infty \leq C \|f\|_\infty$ , avec une constante  $C$  à déterminer explicitement, où  $U = \{u_{i,j}\}_{(i,j) \in \{0, \dots, N+1\}^2}$  est solution de (1.104). [On pourra utiliser la fonction  $\phi(x, y) = \frac{1}{2}y^2$ ].

En déduire que dans le cas du schéma [II] et du schéma [I] sous la condition trouvée en 3. le problème (1.104) admet, pour tout  $f$ , une et une seule solution.

## 5. (Convergence)

Les schémas [I] et [II] sont-ils convergents ? (au sens  $\max_{(i,j) \in \{0, \dots, N+1\}^2} (|u_{i,j} - u(ih, jh)|) \rightarrow 0$  quand  $h \rightarrow 0$ ). Quel est l'ordre de convergence de chacun des schémas ?

## 6. (Commentaires)

Quels sont, à votre avis, les avantages respectifs des schémas [I] et [II] ?

**Exercice 20 (Implantation de la méthode des volumes finis.)**

On considère le problème de conduction du courant électrique

$$-\operatorname{div}(\mu_i \nabla \phi(x)) = 0 \quad x \in \Omega_i, \quad i = 1, 2 \quad (1.105)$$

où  $\phi$  représente le potentiel électrique,  $j = -\mu \nabla \phi(x)$  est donc le courant électrique,  $\mu_1 > 0$ ,  $\mu_2 > 0$  sont les conductivités thermiques dans les domaines  $\Omega_1$  et avec  $\Omega_2$ , avec  $\Omega_1 = ]0, 1[ \times ]0, 1[$  et  $\Omega_2 = ]0, 1[ \times ]1, 2[$ . On appelle  $\Gamma_1 = ]0, 1[ \times \{0\}$ ,  $\Gamma_2 = \{1\} \times ]0, 2[$ ,  $\Gamma_3 = ]0, 1[ \times \{2\}$ , et  $\Gamma_4 = \{0\} \times ]0, 2[$  les frontières extérieures de  $\Omega$ , et on note  $I = ]0, 1[ \times \{0\}$  l'interface entre  $\Omega_1$  et  $\Omega_2$  (voir Figure 1.5). Dans la suite, on notera  $\mu$  la conductivité électrique sur  $\Omega$ , avec  $\mu|_{\Omega_i} = \mu_i$ ,  $i = 1, 2$ .

On suppose que les frontières  $\Gamma_2$  et  $\Gamma_4$  sont parfaitement isolées. Le potentiel électrique étant défini à une constante près, on impose que sa moyenne soit nulle sur le domaine, pour que le problème soit bien posé.

La conservation du courant électrique impose que

$$\int_{\Gamma_1} j \cdot \mathbf{n} + \int_{\Gamma_3} j \cdot \mathbf{n} = 0,$$

où  $\mathbf{n}$  désigne le vecteur unitaire normal à la frontière  $\partial\Omega$  et extérieure à  $\Omega$ .

Enfin, on suppose que l'interface  $I$  est le siège d'une réaction électrochimique qui induit un saut de potentiel. On a donc pour tout point de l'interface  $I$  :

$$\phi_2(x) - \phi_1(x) = \psi(x), \quad \forall x \in I,$$

où  $\phi_i$  désigne la restriction de  $\phi$  au sous domaine  $i$ . La fonction  $\phi$  est donc discontinue sur l'interface  $I$ . Notons que, par contre, le courant électrique est conservé et on a donc

$$(-\mu \nabla \phi \cdot \mathbf{n})|_2(x) + (-\mu \nabla \phi \cdot \mathbf{n})|_1(x) = 0, \quad \forall x \in I.$$

1. Ecrire le problème complet, avec conditions aux limites.

2. Discrétiser le problème par la méthode des volumes finis, avec un maillage rectangulaire uniforme, (considérer deux inconnues discrètes pour chaque arête de l'interface) et écrire le système linéaire obtenu sur les inconnues discrètes.

**Exercice 21 (Elimination des inconnues d'arêtes.)** Suggestions en page 42, corrigé en page 53

On se place ici dans le cadre des hypothèses et notations du paragraphe 1.5.2 page 26

1. Pour chaque arête interne  $\sigma = K|L$ , calculer la valeur  $u_\sigma$  en fonction de  $u_K$  et  $u_L$  et en déduire que les flux numériques  $F_{K,\sigma}$  et  $F_{L,\sigma}$  vérifient bien (1.67)
2. Pour chaque arête  $\sigma \subset \Gamma_1 \cup \Gamma_3$ , telle que  $\sigma \in \mathcal{E}_K$ , calculer  $u_\sigma$  en fonction de  $u_K$  et montrer que  $F_{K,\sigma}$  vérifie bien (1.68)
3. Pour chaque arête  $\sigma \in \mathcal{E}_I$ , avec  $\sigma = K|L$   $K \in \Omega_1$ , calculer la valeur  $u_\sigma$  en fonction de  $u_K$  et  $u_L$  et en déduire que les flux numériques  $F_{K,\sigma}$  et  $F_{L,\sigma}$  vérifient bien (1.70)
4. Ecrire le système linéaire que satisfont les inconnues  $(u_K)_{K \in \mathcal{T}}$ .

## 1.7 Suggestions pour les exercices

**Exercice 1 page 30 (Comparaison différences finies- volumes finis)**

On rappelle que le schéma différences finies s'obtient en écrivant l'équation en chaque point de discrétisation, et en approchant les dérivées par des quotients différentiels, alors que le schéma volumes finis s'obtient en intégrant l'équation sur chaque maille et en approchant les flux par des quotients différentiels.

1. Ne vous laissez pas impressionner par le sinus, la procédure reste la même que dans le cas linéaire, sauf, bien évidemment que vous allez obtenir un système discret non linéaire.
2. Le système redevient linéaire...

**Exercice 1 page 30 (Comparaison différences finies- volumes finis)**

En différences finies : il faut prendre  $N + 2$  inconnues au lieu des  $N$ . Aux points  $x_1, \dots, x_N$ , on écrit les équations et on remplace les dérivées par des quotients différentiels, comme dans le cours. Pour les conditions limites, on considère en plus les valeurs approchées en 0 et en 1. On écrit alors le quotient différentiel  $\frac{u_1 - u_0}{h}$  pour approcher  $u'(0)$  et de même, on écrit le quotient différentiel  $\frac{u_{N+1} - u_N}{h}$  pour approcher  $u'(1)$ . On obtient deux équations supplémentaires qui permettent d'éliminer  $u_0$  et  $u_{N+1}$ .

En volumes finis, les dérivées en 0 et 1 sont les flux, qui interviennent directement dans la discrétisation des équations. Donc la condition limite en 1 est directement prise en compte dans l'équation de bilan sur la dernière maille. Par contre la condition limite en 0 fait intervenir la valeur  $u(0)$  et il faut donc introduire une inconnue auxiliaire  $u_0$  qu'on élimine par la condition limite approchée.

**Exercice 3 page 30 (Principe du maximum)**

2. Poser  $u_0 = a$ ,  $u_{N+1} = b$ . Considérer  $p = \min\{i = 0, \dots, N + 1 ; u_p = \min_{j=0, \dots, N+1} u_j\}$ . Montrer que  $p = 0$  ou  $N + 1$ .

**Exercice 7 page 32 (Conditionnement efficace)****Partie 1**

1. Pour montrer que  $A$  est inversible, utiliser le théorème du rang.
2. Utiliser le fait que  $\varphi$  est un polynôme de degré 2.
3. Pour montrer que  $\|A^{-1}\| = \frac{1}{8}$ , remarquer que le maximum de  $\varphi$  est atteint en  $x = .5$ , qui correspond à un point de discrétisation car  $N$  est impair.

**Partie 2 Conditionnement efficace**

1. Utiliser la convergence uniforme. 2. Utiliser le fait que  $A\phi = (1 \dots )^t$ .

**Exercice 9 page 34 (Erreur de consistance)**

1. Utiliser l'erreur de consistance.
2. Trouver un contre-exemple.

**Exercice 10 page 34 (Différences finies pour un problème elliptique)**

Questions 1 à 3 : application directe des méthodes de démonstration vues en cours (paragraphe 1.2 page 12).

Question 4 : La fonction  $\theta$  est introduite pour montrer une majoration de  $\|A^{-1}\|$ , puisqu'on a plus  $A\Phi = 1$ , où  $\Phi_i = \varphi(x_i)$  est la et  $\varphi$  est la fonction "miracle" dans le cas  $-u'' = f$ . Une fois qu'on a montré les bonnes propriétés de la fonction  $\theta$  (questions 4.a et 4.b), on raisonne comme dans le cours pour la question 4.c (voir démonstration de la proposition 1.15 page 16).

**Exercice 12 (Traitement de la condition de Fourier)**

2. On rappelle que l'erreur de consistance  $R_{N+1/2}$  sur le flux en  $x_{N+1/2}$  est donné par la formule (1.49) page 22. On a donc :  $R_{N+1/2} = \bar{F}_{N+1/2} - F_{N+1/2}^*$  avec  $\bar{F}_{N+1/2} = -u'(1)$ . Pour déterminer  $F_{N+1/2}^*$ , on remplace l'inconnue discrète par la solution exacte au point associé dans l'expression du flux discret. Après calculs, ceci doit vous amener à :

$$F_{N+1/2}^* \begin{cases} = \alpha u(x_N) - b \text{ pour la discrétisation (1.17)} \\ \frac{2}{\alpha h + 2} (\alpha u(x_N) - b) \text{ pour la discrétisation (1.17)} \end{cases} \quad (1.106)$$

Il faut ensuite effectuer appliquer le théorème des accroissements finis pour montrer qu'elle est d'ordre 1 dans le premier cas, et un développement limité pour montrer quelle est d'ordre 2 dans le second.

**Exercice 11 (Non consistance des volumes finis)**

Prendre  $f$  constante et égale à 1 et prendre  $h_i = h/2$  pour  $i$  pair et  $h_i = h$  pour  $i$  impair.

**Exercice 13 page 35 (Consistance des flux)**

1. Effectuer deux développements de Taylor-Young à l'ordre 2 entre  $x_i$  et  $x_{i+1/2}$  et entre  $x_{i+1}$  et  $x_{i+1/2}$ .
2. Effectuer deux développements de Taylor-Lagrange à l'ordre 2 entre  $x_i$  et  $x_{i+1/2}$  et entre  $x_{i+1}$  et  $x_{i+1/2}$ .

**Exercice 14 page 35 (Conditions aux limites de Neumann)**

1. En différences finies, écrire les équations internes de manière habituelle, et éliminez les inconnues qui apparaissent au bord  $u_0$  et  $u_{N+1}$  en discrétisant convenablement les conditions aux limites. En volumes finis, c'est encore plus simple (flux nul au bord...)
2. Remarquer les constantes sont solutions du problème continu, et chercher alors par exemple une solution à moyenne nulle.

**Exercice 15 page 35 (Conditions aux limites de Fourier (ou Robin) et Neumann)**

Ecrire les équations internes de manière habituelle, et éliminez les inconnues qui apparaissent au bord  $u_0$  et  $u_{N+1}$  en discrétisant convenablement les conditions aux limites.

**Exercice 17 page 36 (Volumes finis 1D)**

1. Pour justifier le schéma : écrire les bilans par maille, et approchez les flux par des quotients différentiels de manière consistante.
- 2 (a) On pourra, par exemple, multiplier (1.88) par  $u_i$  et sommer pour  $i = 1, \dots, N$ , puis conclure en remarquant, en particulier, que  $\sum_{i=1}^N (u_i - u_{i-1})u_i = \frac{1}{2} \sum_{i=1}^{N+1} (u_i - u_{i-1})^2$ , avec  $u_0 = u_{N+1} = 0$ .

2 (b) Pensez au miracle de la dimension finie...

4. Effectuer les développements de Taylor

5. (b) (c) Se débarrasser des termes de convection en remarquant qu'ils ont "le bon signe", et s'inspirer de la démonstration du théorème 1.30 page 22.

### Exercice 18 (Discrétisation 2D par différences finies)

Adapter le cas unidimensionnel, en faisant attention aux conditions limites. Pour montrer l'existence et unicité, calculer le noyau de la matrice.

### Exercice 21 (Elimination des inconnues d'arêtes)

1. Ecrire la conservativité du flux :  $F_{K,\sigma} = -F_{L,\sigma}$  et en déduire la valeur de  $u_\sigma$ .

2. Trouver la valeur de  $u_\sigma$  qui vérifie

$$-m(\sigma)\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} = m(\sigma)\alpha(u_\sigma - u_{ext}).$$

3. Remplacer  $F_{K,\sigma}$  et  $F_{L,\sigma}$  par leurs expressions dans (1.69) et en déduire la valeur de  $u_\sigma$ .

4. Adopter l'ordre lexicographique pour la numérotation des mailles, puis établir l'équation de chaque maille, en commençant par les mailles internes.

## 1.8 Corrigés des exercices

### Exercice 1 page 30

Le schéma différences finies pour l'équation (1.71) s'écrit :

$$\begin{cases} \frac{1}{h^2} (2u_i - u_{i-1} - u_{i+1}) + \sin u_i = f_i, & i = 1, \dots, N, \\ u_0 = a, & u_{N+1} = b. \end{cases}$$

Le schéma volumes finis pour la même équation s'écrit :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + h \sin u_i = h f_i, \quad i = 1, \dots, N \quad (1.107)$$

avec  $F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h}$ ,  $i = 1, \dots, N-1$  et  $F_{\frac{1}{2}} = -\frac{u_1 - a}{\frac{h}{2}}$ .

$$F_{N+\frac{1}{2}} = -\frac{b - u_N}{\frac{h}{2}}$$

En remplaçant les expressions des flux dans l'équation (1.107). On obtient :

$$\frac{1}{h^2} (2u_i - u_{i+1} - u_{i-1}) + \sin u_i = f_i, \quad i = 2, \dots, N-1$$

$$\frac{1}{h^2} (3u_1 - 2u_2 - a) + \sin u_1 = 2f_1$$

$$\frac{1}{h^2} (3u_N - 2u_{N-1} - b) + \sin u_N = 2f_N,$$

2. La différence entre les deux schémas réside uniquement dans la première et la dernière équations.

**Exercice 5 page 31**

1. La matrice  $M$  et le second membre  $b$  sont donnés par :

$$\left\{ \begin{array}{l} (MU)_i = \frac{1}{h^2}(2u_i - u_{i+1} - u_{i-1}) + \frac{1}{h}v_i(u_i - u_{i-1}), \text{ pour } i = 2, \dots, N, \\ (MU)_1 = \frac{1}{h^2}(2u_1 - u_2) + \frac{1}{h}v_1u_1, \\ (MU)_N = \frac{1}{h^2}(2u_N - u_{N-1}) + \frac{1}{h}v_N(u_N - u_{N-1}), \end{array} \right. \quad \text{et } b = \begin{pmatrix} (\frac{1}{h^2} + \frac{v_1}{h})a_0 \\ 0 \\ \vdots \\ 0 \\ \frac{1}{h^2}a_1 \end{pmatrix} \quad (1.108)$$

(a) Supposons  $MU \geq 0$ . Soit  $i_0 = \min\{i; u_i = \min_j u_j\}$ .

(i) Si  $i_0 = 1$ , comme  $\frac{1}{h^2}(2u_1 - u_2) + \frac{1}{h}v_1u_1 \geq 0$ , on a :  $(\frac{1}{h^2} + \frac{1}{h}v_1)u_1 + \frac{1}{h^2}(u_1 - u_2) \geq 0$ , et comme  $u_1 - u_2 \leq 0$ , ceci entraîne  $u_1 \geq 0$ .

(ii) Si  $2 \leq i_0 \leq N - 1$ , on a :

$$\frac{1}{h^2}(u_{i_0} - u_{i_0+1}) + \frac{1}{h^2} - (u_{i_0} - u_{i_0-1}) + \frac{1}{h}v_i(u_{i_0} - u_{i_0-1}) \geq 0.$$

Mais par définition de  $i_0$ , on a  $u_{i_0} - u_{i_0+1} \leq 0$  et, et  $u_{i_0} - u_{i_0-1} < 0$  donc ce cas est impossible.

(iii) Si  $i_0 = N$ , on a :

$$\frac{1}{h^2}u_N + \frac{1}{h^2}(u_N - u_{N-1}) + \frac{1}{h}v_i(u_N - u_{N-1}) \geq 0.$$

Or par définition de  $i_0 = N$ , on a  $u_N < u_{N-1}$ , et donc  $u_N > 0$ .

On a donc montré que si  $MU \geq 0$  alors  $U \geq 0$ .

(b) Comme  $MU \geq 0 \Rightarrow U \geq 0$ , on a donc (en prenant  $U$  puis  $-U$ )  $MU = 0 \Rightarrow U = 0$ , ce qui prouve que  $M$  est inversible,

(c) Soit  $U$  solution de  $MU = b$ . Posons  $\tilde{U} = (u_0, u_1, \dots, u_N, u_{N+1})$ , avec  $u_0 = a_0$  et  $u_{N+1} = a_1$ . Remarquons d'abord que le minimum et le maximum des composantes  $u_i$  de  $\tilde{U}$  ne peuvent être atteints pour  $i = 1, \dots, N$  que si les  $u_i$  sont tous égaux (auquel cas  $u_i = a_0 = a_1$  pour tout  $i = 0, \dots, N + 1$ ). En effet, pour  $i = 1, \dots, N$ , on a :

$$\frac{1}{h^2}(u_i - u_{i+1}) + \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h}v_i(u_i - u_{i-1}) = 0. \quad (1.109)$$

Soit  $i_0 = \min\{i; u_i = \min_j u_j\}$ , et  $i_1 = \min\{i; u_i = \max_j u_j\}$ . Par définition de  $i_0$ , on a  $u_{i_0} - u_{i_0+1} \leq 0$  et, et  $u_{i_0} - u_{i_0-1} < 0$  donc (1.109) est impossible si  $1 < i_0 < N$ . Donc  $i_0 = 1$  ou  $N$ . Soit  $i_1 = \min\{i; u_i = \max_j u_j\}$ , on a  $u_{i_1} - u_{i_1+1} \geq 0$  et, et  $u_{i_1} - u_{i_1-1} > 0$  et (1.109) est encore impossible si  $1 < i_1 < N$ .

On en déduit que  $i_0 = 0$  ou  $N + 1$  et  $i_1 = 0$  ou  $N + 1$ , ce qui prouve que  $\min(a_0, a_1) \leq u_i \leq \max(a_0, a_1)$ .

2.

(a) Par définition,  $m_{i,i} = \frac{1}{h^2} + \frac{1}{h}v_i$ , avec  $v_i \geq 0$ , ce qui prouve le résultat.

(b) Par définition,  $m_{i,j}$  est soit nul, soit égal à  $-\frac{1}{h^2} + \frac{1}{h}v_i$  si  $j = i - 1$ , soit égal à  $-\frac{1}{h^2}$  si  $j = i + 1$ , ce qui prouve le résultat.

(c) On a montré que  $M$  est inversible à la question précédente.

(d) D'après la question 1.2, on sait que si  $MU \geq 0$ , alors  $U \geq 0$ . Soit  $e_i$  le  $i$ -ème vecteur de la base canonique. On a  $e_i = M(M^{-1})e_i \geq 0$ , et donc  $M^{-1}e_i \geq 0$ , ce qui montre que tous les coefficients de  $M^{-1}$  doivent être positifs.

**Exercice 7 page 32 (Conditionnement "efficace")****Partie I**

1. Soit  $u = (u_1, \dots, u_N)$ . On a

$$Au = b \Leftrightarrow \begin{cases} \frac{1}{h^2}(u_i - u_{i-1}) + \frac{1}{h^2}(u_i - u_{i+1}) = b_i, \quad \forall i = 1, \dots, N, \\ u_0 = u_{N+1} = 0. \end{cases}$$

Supposons  $b_i \geq 0, \forall i = 1, \dots, N$ , et soit  $p \in \{0, \dots, N+1\}$  tel que  $u_p = \min(u_i, i = 0, \dots, N+1)$ .

Si  $p = 0$  ou  $N+1$ , alors  $u_i \geq 0 \forall i = 0, N+1$  et donc  $u \geq 0$ .

Si  $p \in \{1, \dots, N\}$ , alors

$$\frac{1}{h^2}(u_p - u_{p-1}) + \frac{1}{h^2}(u_p - u_{p+1}) \geq 0$$

et comme  $u_p - u_{p-1} < 0$  et  $u_p - u_{p+1} \leq 0$ , on aboutit à une contradiction.

Montrons maintenant que  $A$  est inversible. On vient de montrer que si  $Au \geq 0$  alors  $u \geq 0$ . On en déduit par linéarité que si  $Au \leq 0$  alors  $u \leq 0$ , et donc que si  $Au = 0$  alors  $u = 0$ . Ceci démontre que l'application linéaire représentée par la matrice  $A$  est injective donc bijective (car on est en dimension finie).

2. Soit  $\varphi \in C([0, 1], \mathbb{R})$  tel que  $\varphi(x) = 1/2x(1-x)$  et  $\phi_i = \varphi(x_i), i = 1, N$ , où  $x_i = ih$ .

$(A\phi)_i$  est le développement de Taylor à l'ordre 2 de  $\varphi''(x_i)$ , et comme  $\varphi$  est un polynôme de degré 2, ce développement est exact. Donc  $(A\phi)_i = \varphi''(x_i) = 1$ .

3. Soient  $b \in \mathbb{R}^N$  et  $u \in \mathbb{R}^N$  tels que  $Au = b$ . On a :

$$(A(u \pm \|b\|\varphi))_i = (Au)_i \pm \|b\|(A\phi)_i = b_i \pm \|b\|.$$

Prenons d'abord  $\tilde{b}_i = b_i + \|b\| \geq 0$ , alors par la question (1),

$$u_i + \|b\|\phi_i \geq 0 \quad \forall i = 1, \dots, N.$$

Si maintenant on prend  $\bar{b}_i = b_i - \|b\| \leq 0$ , alors

$$u_i - \|b\|\phi_i \leq 0 \quad \forall i = 1, \dots, N.$$

On a donc  $-\|b\|\phi_i \leq \|b\|\phi_i$ .

On en déduit que  $\|u\|_\infty \leq \|b\| \|\phi\|_\infty$ ; or  $\|\phi\|_\infty = \frac{1}{8}$ . D'où  $\|u\|_\infty \leq \frac{1}{8}\|b\|$ .

On peut alors écrire que pour tout  $b \in \mathbb{R}^N$ ,

$$\|A^{-1}b\|_\infty \leq \frac{1}{8}\|b\|, \text{ donc } \frac{\|A^{-1}b\|_\infty}{\|b\|_\infty} \leq \frac{1}{8}, \text{ d'où } \|A^{-1}\| \leq \frac{1}{8}.$$

On montre que  $\|A^{-1}\| = \frac{1}{8}$  en prenant le vecteur  $b$  défini par  $b(x_i) = 1, \forall i = 1, \dots, N$ . On a en effet  $A^{-1}b = \phi$ , et comme  $N$  est impair,  $\exists i \in \{1, \dots, N\}$  tel que  $x_i = \frac{1}{2}$ ; or  $\|\varphi\|_\infty = \varphi(\frac{1}{2}) = \frac{1}{8}$ .

4. Par définition, on a  $\|A\| = \sup_{\|x\|_\infty=1} \|Ax\|$ , et donc  $\|A\| = \max_{i=1, N} \sum_{j=1, N} |a_{i,j}|$ , d'où le résultat.

5. Grâce aux questions 3 et 4, on a, par définition du conditionnement pour la norme  $\|\cdot\|$ ,  $\text{cond}(A) = \|A\| \|A^{-1}\| = \frac{1}{2h^2}$ .

Comme  $A\delta_u = \delta_b$ , on a :

$$\|\delta_u\| \leq \|A^{-1}\| \|\delta_b\| \frac{\|b\|}{\|b\|} \leq \|A^{-1}\| \|\delta_b\| \frac{\|A\| \|u\|}{\|b\|},$$

d'où le résultat.

Pour obtenir l'égalité, il suffit de prendre  $b = Au$  où  $u$  est tel que  $\|u\| = 1$  et  $\|Au\| = \|A\|$ , et  $\delta_b$  tel que  $\|\delta_b\| = 1$  et  $\|A^{-1}\delta_b\| = \|A^{-1}\|$ . On obtient alors

$$\frac{\|\delta_b\|}{\|b\|} = \frac{1}{\|A\|} \text{ et } \frac{\|\delta_u\|}{\|u\|} = \|A^{-1}\|.$$

D'où l'égalité.

## Partie 2 Conditionnement efficace

1. Soient  $\varphi^{(h)}$  et  $f^{(h)}$  les fonctions constantes par morceaux définies par

$$\begin{aligned} \varphi^{(h)}(x) &= \begin{cases} \varphi(ih) = \phi_i \text{ si } x \in ]x_i - \frac{h}{2}, x_i + \frac{h}{2}[ , i = 1, \dots, N, \\ 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in ]1 - \frac{h}{2}, 1]. \end{cases} \text{ et} \\ f^{(h)}(x) &= \begin{cases} f(ih) = b_i \text{ si } x \in ]x_i - \frac{h}{2}, x_i + \frac{h}{2}[ , \\ f(ih) = 0 \text{ si } x \in [0, \frac{h}{2}] \text{ ou } x \in ]1 - \frac{h}{2}, 1]. \end{cases} \end{aligned}$$

Comme  $f \in C([0, 1], \mathbb{R})$  et  $\varphi \in C^2([0, 1], \mathbb{R})$ , la fonction  $f_h$  (resp.  $\varphi_h$ ) converge uniformément vers  $f$  (resp.  $\varphi$ ) lorsque  $h \rightarrow 0$ . On a donc

$$h \sum_{i=1}^N b_i \varphi_i = \int_0^1 f^{(h)}(x) \varphi^{(h)}(x) dx \rightarrow \int_0^1 f(x) \varphi(x) dx \text{ lorsque } h \rightarrow 0.$$

Comme  $b_i > 0$  et  $f_i > 0 \forall i = 1, \dots, N$ , on a évidemment

$$S_N = \sum_{i=1}^N b_i \varphi_i > 0 \text{ et } S_N \rightarrow \int_0^1 f(x) \varphi(x) dx = \beta > 0 \text{ lorsque } h \rightarrow 0.$$

Donc il existe  $N_0 \in \mathbb{N}$  tel que si  $N \geq N_0$ ,  $S_N \geq \frac{\beta}{2}$ , et donc  $S_N \geq \alpha = \min(S_0, S_1 \dots S_{N_0}, \frac{\beta}{2}) > 0$ .

2. On a  $N\|u\| = N \sup_{i=1, \dots, N} |u_i| \geq \sum_{i=1}^N u_i$ . D'autre part,  $A\varphi = (1 \dots 1)^t$  donc  $u \cdot A\varphi = \sum_{i=1}^N u_i$ ; or

$u \cdot A\varphi = A^t u \cdot \varphi = Au \cdot \varphi$  car  $A$  est symétrique. Donc  $u \cdot A\varphi = \sum_{i=1}^N b_i \varphi_i \geq \frac{\alpha}{h}$  d'après la question 1. Comme

$\delta_u = A^{-1} \delta_b$ , on a donc  $\|\delta_u\| \leq \|A^{-1}\| \|\delta_b\|$ ; et comme  $N\|u\| \geq \frac{\alpha}{h}$ , on obtient :  $\frac{\|\delta_u\|}{\|u\|} \leq \frac{1}{8} \frac{hN}{\alpha} \|\delta_b\| \frac{\|f\|_\infty}{\|b\|}$ . Or  $hN = 1$  et on a donc bien :

$$\frac{\|\delta_u\|}{\|u\|} \leq \frac{\|f\|_\infty \|\delta_b\|}{8\alpha \|b\|}.$$

3. Le conditionnement  $\text{cond}(A)$  calculé dans la partie 1 est d'ordre  $1/h^2$ , et donc tend vers l'infini lorsque le pas du maillage tend vers 0, alors qu'on vient de montrer dans la partie 2 que la variation relative  $\frac{\|\delta_u\|}{\|u\|}$  est inférieure à une constante multipliée par la variation relative de  $\frac{\|\delta_b\|}{\|b\|}$ . Cette dernière information est nettement plus utile et réjouissante pour la résolution effective du système linéaire.

### Exercice 9 page 34 (Erreur de consistance)

- Si  $f$  est constante, alors  $-u''$  est constante, et donc les dérivées d'ordre supérieur de  $u$  sont nulles. Donc par l'estimation (1.30) page 15 sur l'erreur de consistance, on a  $R_i = 0$  pour tout  $i = 1, \dots, N$ .  
Si on appelle  $U$  le vecteur de composantes  $u_i$  et  $\bar{U}$  le vecteur de  $\mathbb{R}^N$  de composantes  $u(x_i)$ , on peut remarquer facilement que  $U - \bar{U} = A^{-1}R$ , où  $R$  est le vecteur de composantes  $R_i$ . On a donc  $U - \bar{U} = 0$ , c.q.f.d.
- Il est facile de voir que  $f$  n'est pas forcément constante, en prenant  $f(x) = \sin 2\pi x$ , et  $h = 1/2$ , on n'a donc qu'une seule inconnue  $u_1$  qui vérifie  $u_1 = 0$ , et on a également  $u(1/2) = \sin \pi = 0$ .

### Exercice 11 page 35 (Non consistance des volumes finis)

Par développement de Taylor, pour  $i = 1, \dots, N$ , il existe  $\xi_i \in [x_i, x_{i+1}]$  tel que :

$$u(x_{i+1}) = u(x_i) + h_{i+\frac{1}{2}} u'(x_i) + \frac{1}{2} h_{i+\frac{1}{2}}^2 u''(x_i) + \frac{1}{6} h_{i+\frac{1}{2}}^3 u'''(\xi_i),$$

et donc

$$R_i = -\frac{1}{h_i} \frac{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}}{2} u''(x_i) + u''(x_i) + \rho_i, \quad i = 1, \dots, N, \quad (1.110)$$

où  $|\rho_i| \leq Ch$ ,  $C$  ne dépendant que de la dérivée troisième de  $u$ . Il est facile de voir que, en général,  $R_i$ , ne tend pas vers 0 lorsque  $h$  tend vers 0 (sauf dans des cas particuliers). En effet, prenons par exemple  $f \equiv 1$ ,  $h_i = h$  pour  $i$  pair,  $h_i = h/2$  pour  $i$  impair, and  $x_i = (x_{i+1/2} + x_{i-1/2})/2$ , pour  $i = 1, \dots, N$ . On a dans ce cas  $u'' \equiv -1$ ,  $u''' \equiv 0$ , et donc :

$$R_i = -\frac{1}{4} \text{ si } i \text{ est pair, et } R_i = +\frac{1}{2} \text{ si } i \text{ est impair.}$$

On en conclut que  $\sup\{|R_i|, i = 1, \dots, N\} \not\rightarrow 0$  as  $h \rightarrow 0$ .

### Exercice 13 page 35

1. Par développement de Taylor-Young, pour  $i = 1, \dots, N$  :

$$u(x_{i+1}) = u(x_{i+\frac{1}{2}}) + (x_{i+1} - x_{i+\frac{1}{2}})u'(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_{i+1} - x_{i+\frac{1}{2}})^2u''(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_{i+1} - x_{i+\frac{1}{2}})^2\varepsilon(x_{i+1} - x_{i+\frac{1}{2}}),$$

$$u(x_i) = u(x_{i+\frac{1}{2}}) + (x_i - x_{i+\frac{1}{2}})u'(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_i - x_{i+\frac{1}{2}})^2u''(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_i - x_{i+\frac{1}{2}})^2\varepsilon(x_i - x_{i+\frac{1}{2}}),$$

avec  $\varepsilon(x) \rightarrow 0$  lorsque  $x \rightarrow 0$ . Par soustraction, on en déduit que :

$$u(x_{i+1}) - u(x_i) = (x_{i+1} - x_i)u'(x_{i+\frac{1}{2}}) + \frac{1}{2}(x_{i+1} - x_i)(x_{i+1} + x_i - 2x_{i+\frac{1}{2}})u''(x_{i+\frac{1}{2}}) + \tilde{\rho}_{i+\frac{1}{2}},$$

où

$$\tilde{\rho}_{i+\frac{1}{2}} = \frac{1}{2}((x_{i+1} - x_{i+\frac{1}{2}})^2\varepsilon(x_{i+1} - x_{i+\frac{1}{2}}) - (x_i - x_{i+\frac{1}{2}})^2\varepsilon(x_i - x_{i+\frac{1}{2}})).$$

Donc, en posant  $F_{i+\frac{1}{2}}^* = -\frac{u(x_{i+1}) - u(x_i)}{h_{i+\frac{1}{2}}}$ , on obtient après simplifications :

$$F_{i+\frac{1}{2}}^* = -u'(x_{i+\frac{1}{2}}) - \frac{1}{2} [x_{i+1} + x_i - 2x_{i+\frac{1}{2}}] u''(x_{i+\frac{1}{2}}) + \rho_{i+\frac{1}{2}}$$

avec  $|\rho_{i+\frac{1}{2}}| \leq Ch^2$ , où  $h = \max_{i=1, \dots, N} h_i$  et  $C$  ne dépend que de  $u'''$ .

2. Dans le cas où  $u \in C^3([0, 1])$   $x_{i+\frac{1}{2}} = \frac{x_{i+1} + x_i}{2}$ , on a donc  $|F_{i+\frac{1}{2}}^* + u'(x_{i+\frac{1}{2}})| \leq \rho_{i+\frac{1}{2}}$ , et donc le flux est constant d'ordre 2.

Dans le cas général, on peut seulement majorer  $\frac{1}{2} (x_{i+1} + x_i - 2x_{i+\frac{1}{2}})$  par  $h$ , on a donc un flux constant d'ordre 1.

### Exercice 14 page 35

1. On se donne une discrétisation  $(x_i)_{i=1, \dots, N}$  de l'intervalle  $[0, 1]$ , de pas constant et égal à  $h$ . On écrit l'équation en chaque point  $x_i$ , et on remplace  $-u''(x_i)$  par le quotient différentiel habituel. En appelant  $u_1, \dots, u_N$  les inconnues localisées aux points  $x_1, \dots, x_N$ , et  $u_0, u_{N+1}$  les inconnues auxiliaires localisées en  $x = 0$  et  $x = 1$ , on obtient les équations discrètes associées aux inconnues  $i = 1, \dots, N$ .

$$\frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i,$$

avec  $c_i = c(x_i)$  et  $f_i = f(x_i)$ . Il reste à déterminer  $u_0$  et  $u_{N+1}$ . Ceci se fait en approchant la dérivée  $u'(0)$  ( resp.  $u'(1)$ ) par  $\frac{1}{h}(u(x_1) - u(0))$  ( resp.  $\frac{1}{h}(u(1) - u(x_N))$ ).

Comme  $u'(0) = 0$  et  $u'(1) = 0$ , on obtient donc que  $u_0 = u_1$  et  $u_{N+1} = u_N$ . Le schéma différences finies s'écrit donc :

$$\begin{cases} \frac{1}{h^2}(u_1 - u_2) + c_1 u_1 = f_1 \\ \frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, & i = 2, \dots, N-1, \\ \frac{1}{h^2}(u_N - u_{N-1}) + c_N u_N = f_N. \end{cases}$$



2. On se donne un maillage volumes finis, et on intègre l'équation sur chaque maille, ce qui donne le schéma

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i, i = 1, \dots, N,$$

où  $F_{i+\frac{1}{2}}$  est le flux numérique à l'interface  $x_{i+\frac{1}{2}}$ . Pour  $i = 1, \dots, N - 1$ , ce flux numérique est donné par

$$F_{i+\frac{1}{2}} = \frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}.$$

Pour  $i = 0$  et  $i = N + 1$ , on se sert des conditions de Neumann, qui imposent un flux nul. On écrit donc :

$$F_{\frac{1}{2}} = F_{N+\frac{1}{2}} = 0.$$

### Exercice 15 page 35

1. La discrétisation par différences finies donne comme  $i$ -ème équation (voir par exemple exercice 14 page 35 :

$$\frac{1}{h^2}(2u_i - u_{i-1} - u_{i+1}) + c_i u_i = f_i, i = 1, \dots, N.$$

Il reste donc à déterminer les inconnues  $u_0$  et  $u_{N+1}$  à l'aide de la discrétisation des conditions aux limites, qu'on approche par :

$$\frac{u_1 - u_0}{h} + \alpha(u_0 - \tilde{u}) = 0,$$

$$\frac{u_{N+1} - u_N}{h} + \alpha(u_{N+1} - \tilde{u}) = 0$$

où  $u_0$  et  $u_{N+1}$  sont les valeurs approchées en  $x_0$  et  $x_{N+1}$ , on a donc par élimination :

$$u_0 = \frac{1}{\alpha - \frac{1}{h}} \left( \alpha \tilde{u} - \frac{u_1}{h} \right) \text{ et } u_{N+1} = \frac{1}{\alpha + \frac{1}{h}} \left( \alpha \tilde{u} + \frac{u_N}{h} \right).$$

Ce qui termine la définition du schéma.

2. Par volumes finis, la discrétisation de l'équation s'écrit

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} = h_i f_i, i = 1, \dots, N,$$

et les seuls flux "nouveaux" sont encore  $F_{1/2}$  et  $F_{N+\frac{1}{2}}$ , qu'on obtient à partir de la discrétisation des conditions aux limites. Ceci peut se faire de plusieurs manières.

On peut, par exemple, discrétiser la condition aux limites en 0 par :

$$F_{1/2} + \alpha(u_0 - \tilde{u}) = 0, \text{ avec } F_{1/2} = \frac{u_1 - u_0}{\frac{h_1}{2}}.$$

On a dans ce cas :  $-\alpha(u_0 - \tilde{u}) \times \frac{h_1}{2} = -u_1 + u_0$ , d'où on déduit que  $u_0 = \frac{\alpha \tilde{u} h_1 + 2u_1}{\alpha h_1 + 2}$ , et qui conduit à l'expression suivante pour  $F_{1/2}$  :

$$F_{1/2} = \frac{\alpha}{\alpha h_1 + 2} (2(u_1 - \tilde{u}) - \alpha h_1 \tilde{u}).$$

Le calcul est semblable pour  $F_{N+\frac{1}{2}}$

### Exercice 17 page 36

1. On intègre (1.86) sur une maille  $[x_{i-1/2}, x_{i+1/2}]$  et on obtient :

$$-u'(x_{i+1/2}) + u'(x_{i-1/2}) + a[u(x_{i+1/2}) - u(x_{i-1/2})] + b \int_{x_{i-1/2}}^{x_{i+1/2}} u(x) dx = bh_i f_i. \quad (1.111)$$

Pour justifier le schéma numérique proposé on remarque que :

$$u(x_{i+1}) = u(x_{i+1/2}) + (x_{i+1} - x_{i+1/2})u'(x_{i+1/2}) + \frac{1}{2}(x_{i+1} - x_{i+1/2})^2 u''(\xi_i), \text{ avec } \xi_i \in [x_{i+1/2}, x_{i+1}],$$

et de même

$$u(x_i) = u(x_{i+1/2}) + (x_i - x_{i+1/2})u'(x_{i+1/2}) + \frac{1}{2}(x_i - x_{i+1/2})^2 u''(\gamma_i), \text{ avec } \gamma_i \in [x_{i-1/2}, x_i],$$

dont on déduit :

$$u(x_{i+1}) - u(x_i) = h_{i+1/2}u'(x_{i+1/2}) + \frac{1}{8}(h_{i+1}^2 u''(\xi_i) - h_i^2 u''(\gamma_i)).$$

De plus en utilisant le fait que  $x_i$  est le milieu de  $[x_{i-1/2}, x_{i+1/2}]$  on a (voir démonstration plus loin)

$$\int_{x_{i-1/2}}^{x_{i+1/2}} u dx = u(x_i)h_i + \frac{1}{24}u''(\alpha_i)h_i^3 \quad (1.112)$$

D'où le schéma numérique.

Démontrons la formule (1.112). Pour cela il suffit (par changement de variable) de démontrer que si  $u \in C^2(\mathbb{R})$ , alors pour tout  $\alpha \geq 0$ , on a :

$$\int_{-\alpha}^{\alpha} u dx = 2\alpha u(0) + \frac{1}{3}u''(\alpha_i)\alpha^3. \quad (1.113)$$

Pour cela, on utilise une formule de type Taylor avec reste intégral, qu'on obtient en remarquant que si on pose  $\varphi(t) = u(tx)$ , alors  $\varphi'(t) = xu'(tx)$ , et  $\varphi''(t) = x^2 u''(tx)$ . Or  $\varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)dt$ , et par intégration par parties, on obtient donc :

$$\varphi(1) = \varphi(0) + \varphi'(0) + \int_0^1 \varphi''(t)(1-t)ds.$$

On en déduit alors que

$$u(x) = u(0) + xu'(0) + \int_0^1 x^2 u''(tx)(1-t)dt.$$

En intégrant entre  $-\alpha$  et  $\alpha$ , on obtient alors :

$$\int_{-\alpha}^{\alpha} u(x)dx = 2\alpha u(0) + A, \text{ avec } A = \int_0^1 x^2 u''(tx)(1-t)dt dx.$$

Comme la fonction  $u''$  est continue elle est minorée et majorée sur  $[-\alpha, \alpha]$ . Soient donc  $m = \min_{[-\alpha, \alpha]} u''$  et  $M = \max_{[-\alpha, \alpha]} u''$ . Ces deux valeurs sont atteintes par  $u''$  puisqu'elle est continue. On a donc  $u''([- \alpha, \alpha]) = [m, M]$ . De plus, la fonction  $(x, t) \mapsto x^2(1-t)$  est positive ou nulle sur  $[-\alpha, \alpha] \times [0, 1]$ . On peut donc minorer et majorer  $A$  de la manière suivante

$$m \int_0^1 x^2(1-t)dt dx \leq A \leq M \int_0^1 x^2(1-t)dt dx.$$

Or  $\int_0^1 x^2(1-t)dt dx = \frac{1}{3}\alpha^3$ . On en déduit que  $\frac{1}{3}\alpha^3 m \leq A \leq \frac{1}{3}\alpha^3 M$ , et donc que  $A = \frac{1}{3}\alpha^3 \gamma$ , avec  $\gamma \in [m, M]$ ; mais comme  $u''$  est continue, elle prend toutes les valeurs entre  $m$  et  $M$ , il existe donc  $\beta \in [-\alpha, \alpha]$  tel que  $\gamma = u''(\beta)$ , ce qui termine la preuve de (1.113).

2 (a). On multiplie (1.88) par  $u_i$  et on somme pour  $i = 1, \dots, N$ . On obtient après changement d'indice que

$$\sum_{i=0}^{i=N} \frac{(u_{i+1} - u_i)^2}{h_{i+1/2}} + \frac{a}{2} \sum_{i=0}^{i=N} (u_{i+1} - u_i)^2 + b \sum_{i=0}^{i=N} u_i^2 h_i = 0.$$

Ce qui donne  $u_i = 0$  pour tout  $i = 1 \dots N$ , d'où en mettant le schéma sous la forme matricielle  $AU = B$  on déduit que l'application linéaire représentée par la matrice  $A$  est injective donc bijective (grâce au fait qu'on est en dimension finie) et donc que (1.88) admet une unique solution.

3 (a). Evident.

(b). On pose  $\bar{R} = A\bar{U} - B$ . On a donc  $R_i = R_i^{(1)} + R_i^{(2)}$ , avec

$$R_i^{(1)} = -\frac{1}{h_i} \left[ \left( \frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+1/2}} - u'(x_{i+1/2}) \right) - \left( \frac{\bar{u}_i - \bar{u}_{i-1}}{h_{i-1/2}} - u'(x_{i-1/2}) \right) \right],$$

$$R_i^{(2)} = \frac{a}{h_i} [(\bar{u}_i - u(x_{i+1/2})) - (\bar{u}_{i-1} - u(x_{i+1/2}))].$$

De plus on remarque que

$$\bar{u}_i = \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u dx = u(x_{i+1/2}) - \frac{1}{2} u'(x_{i+1/2}) h_i + \frac{1}{6} u''(x_{i+1/2}) h_i^2 - \frac{1}{24} u^{(3)}(d_i) h_i^3 \text{ avec } d_i \in [x_{i-1/2}, x_{i+1/2}],$$

$$\bar{u}_{i+1} = \frac{1}{h_{i+1}} \int_{x_{i+1/2}}^{x_{i+3/2}} u dx = u(x_{i+1/2}) + \frac{1}{2} u'(x_{i+1/2}) h_{i+1} + \frac{1}{6} u''(x_{i+1/2}) h_{i+1}^2 - \frac{1}{24} u^{(3)}(\delta_i) h_{i+1}^3 \text{ avec } \delta_i \in [x_{i+1/2}, x_{i+3/2}].$$

Ce qui implique que :

$$\frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+1/2}} = u'(x_{i+1/2}) + \frac{1}{3} u''(x_{i+1/2}) (h_{i+1} - h_i) + \frac{1}{24} \frac{1}{h_{i+1/2}} [u^{(3)}(\delta_i) h_{i+1}^3 + u^{(3)}(d_i) h_i^3]$$

et donc

$$\frac{1}{h_i} \left[ \frac{\bar{u}_{i+1} - \bar{u}_i}{h_{i+1/2}} - u'(x_{i+1/2}) \right] = S_i + K_i,$$

avec

$$|S_i| = \left| \frac{1}{3} u''(x_{i+1/2}) \left( \frac{h_{i+1}}{h_i} - 1 \right) + \frac{1}{24} \right| \leq Ch \text{ et } |K_i| = \left| \frac{1}{h_i h_{i+1/2}} [u^{(3)}(\delta_i) h_{i+1}^3 + u^{(3)}(d_i) h_i^3] \right| \leq Ch,$$

où  $C$  ne dépend que de  $u$ . De plus si on pose :  $L_i = \frac{1}{h_i} (\bar{u}_i - u(x_{i+1/2}))$ , par développement de Taylor, il existe  $\tilde{C}$  ne dépendant que de  $u$  telle que  $|L_i| \leq Ch$ . Finalement on conclut que  $|R_i^{(1)}| = |-S_i + S_{i+1}| \leq C_1$  et  $|R_i^{(2)}| = |-K_i + K_{i-1} + a(L_i - L_{i-1})| \leq C_2 h$ .

4. Reprendre les résultats précédents. . . Pour  $|R_{i+1/2}| \leq Ch$  reprendre calcul du 3  $|R_{i+1/2}| = |h_i(-S_i - K_i + L_i)|$ .  
5 (a). On pose  $e_i = \bar{u}_i - u_i$ . Cette définition implique que  $e_i$  est solution du système (1.95)-(1.97). (b). Un calcul similaire à celui de la question 2. donne que

$$b \sum_1^N h_i e_i + \sum_{i=0}^{i=N} \frac{(e_{i+1} - e_i)^2}{h_{i+1/2}} + \frac{a}{2} \sum_{i=0}^{i=N} (e_{i+1} - e_i)^2 = \sum_{i=0}^{i=N} R_{i+1/2} (e_{i+1} - e_i)$$

D'où en utilisant le fait que

$\alpha h \leq h_i \leq \beta h$  et l'inégalité de Cauchy-Schwarz on déduit que

$$\frac{1}{\beta h} \sum_{i=0}^{i=N} (e_{i+1} - e_i)^2 \leq \left( \sum_{i=0}^{i=N} (e_{i+1} - e_i)^2 \right)^{1/2} \left( \sum_{i=0}^{i=N} R_{i+1/2}^2 \right)^{1/2}$$

et en utilisant (1.94), et le fait que  $\sum_{i=0}^{i=N} h_i = 1$  entraîne  $N \leq \frac{1}{\alpha h}$ , on déduit :

$$\sum_{i=0}^{i=N} (e_{i+1} - e_i)^2 \leq C_1 \frac{\beta}{\alpha} h^3.$$

En remarquant que  $e_i = \sum_{j=0}^{j=i-1} (e_{j+1} - e_j)$  on a pour tout  $0 < i \leq N$  que

$$|e_i| \leq \left( \sum_{j=0}^{j=i} (e_{j+1} - e_j)^2 \right)^{1/2} i^{1/2} \leq \left( C_1 \frac{\beta}{\alpha} h^3 \right)^{1/2} N^{1/2}$$

et donc  $|e_i| \leq \frac{\sqrt{C_1 \beta}}{\alpha} h$ , pour tout  $0 < i \leq N$ .

**Exercice 19 page 38**

On note  $(x, y)$  les coordonnées d'un point de  $\mathbb{R}^2$ .

1. En multipliant la première équation de (1.103) par  $\varphi$  et en intégrant par parties, on trouve, pour tout  $\varphi \in C^1(\bar{\Omega})$  t.q.  $\varphi = 0$  sur  $\partial\Omega$  :

$$\int_{\Omega} \nabla u(x, y) \cdot \nabla \varphi(x, y) \, dx dy + \int_{\Omega} k \frac{\partial u(x, y)}{\partial x} \varphi(x, y) \, dx = \int_{\Omega} f(x, y) \varphi(x, y) \, dx dy. \quad (1.114)$$

On suppose maintenant que  $f \leq 0$  sur  $\Omega$ . On se donne une fonction  $\psi \in C^1(\mathbb{R}, \mathbb{R})$  t.q. :

$$\begin{aligned} \psi(s) &= 0, & \text{si } s \leq 0, \\ \psi(s) &> 0, & \text{si } s > 0. \end{aligned}$$

(On peut choisir, par exemple,  $\psi(s) = s^2$  pour  $s > 0$  et  $\psi(s) = 0$  pour  $s \leq 0$ ) et on prend dans (1.114)  $\varphi = \psi \circ u$ . On obtient ainsi :

$$\int_{\Omega} \psi'(u(x, y)) |\nabla u(x, y)|^2 \, dx dy + \int_{\Omega} k \frac{\partial u}{\partial x}(x, y) \psi(u(x, y)) \, dx = \int_{\Omega} f(x, y) \psi(u(x, y)) \, dx dy \leq 0. \quad (1.115)$$

En notant  $G$  la primitive de  $\psi$  s'annulant en 0, on a :  $\frac{\partial}{\partial x} G(u(x, y)) = \psi(u(x, y)) \frac{\partial u}{\partial x}(x, y)$ . Comme  $u = 0$  sur  $\partial\Omega$ , on obtient donc :

$$\int_{\Omega} k \frac{\partial u}{\partial x}(x, y) \psi(u(x, y)) \, dx dy = \int_{\Omega} k \frac{\partial}{\partial x} G(u(x, y)) \, dx dy = \int_{\partial\Omega} k G(u(x, y)) n_x \, d\gamma(x, y) = 0,$$

où  $n_x$  désigne la première composante du vecteur normal  $\mathbf{n}$  à  $\partial\Omega$  extérieur à  $\Omega$ , et  $d\gamma(x, y)$  le symbole d'intégration par rapport à la mesure de Lebesgue unidimensionnelle sur  $\partial\Omega$ . De (1.115) on déduit alors :

$$\int_{\Omega} \psi'(u(x, y)) |\nabla u(x, y)|^2 \, dx dy \leq 0,$$

et donc, comme  $\psi' \geq 0$  et que la fonction  $(x, y) \mapsto \psi'(u(x, y)) |\nabla u(x, y)|^2$  est continue :

$$\psi'(u(x, y)) |\nabla u(x, y)|^2 = 0, \forall (x, y) \in \bar{\Omega}.$$

Ceci donne aussi

$$\nabla \psi(u(x, y)) = 0, \forall (x, y) \in \bar{\Omega}.$$

La fonction  $\psi \circ u$  est donc constante sur  $\bar{\Omega}$ , comme elle est nulle sur  $\partial\Omega$ , elle est nulle sur  $\bar{\Omega}$ , ce qui donne

$$u \leq 0 \text{ sur } \bar{\Omega}$$

2. On s'intéresse ici à la consistance au sens des différences finies. On pose donc

$$\bar{u}_{i,j} = u(ih, jh) \text{ pour } i, j \in \{0, \dots, N+1\}^2.$$

On a bien  $\bar{u}_{i,j} = 0$  pour  $(i, j) \in \gamma$ , et pour  $(i, j) \in \{1, \dots, N\}^2$ , on pose :

$$R_{ij} = a_0 \bar{u}_{i,j} - a_1 \bar{u}_{i-1,j} - a_2 \bar{u}_{i+1,j} - a_3 \bar{u}_{i,j-1} - a_4 \bar{u}_{i,j+1} - f_{i,j}.$$

On rappelle que  $u$  est solution de (2.5.84),  $R_j$  est donc l'erreur de consistance. Dans le cas du schéma [I] on a :

$$R_{ij} = \frac{2\bar{u}_{ij} - \bar{u}_{i+1,j} - \bar{u}_{i-1,j}}{h^2} + \frac{2\bar{u}_{ij} - \bar{u}_{i,j+1} - \bar{u}_{i,j-1}}{h^2} + k \frac{\bar{u}_{i+1,j} - \bar{u}_{i-1,j}}{2h} - f_{ij}.$$

Comme  $u \in C^4(\bar{\Omega})$ , il existe  $\xi_{ij} \in ]0, 1[$  et  $\eta_{ij} \in ]0, 1[$  t.q.

$$\bar{u}_{i+1,j} = \bar{u}_{i,j} + h \frac{\partial u}{\partial x}(ih, jh) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(ih, jh) + \frac{h^3}{6} \frac{\partial^3 u}{\partial x^3}(ih, jh) + \frac{h^4}{24} \frac{\partial^4 u}{\partial x^4}(ih + \xi_{ij}h, jh)$$

$$\bar{u}_{i-1,j} = \bar{u}_{i,j} - h \frac{\partial u}{\partial x}(ih, jh) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(ih, jh) - \frac{h^3}{6} \frac{\partial^3 u}{\partial x^3}(ih, jh) + \frac{h^4}{24} \frac{\partial^4 u}{\partial x^4}(ih + \eta_{ij}h, jh).$$

On obtient des formules analogues pour  $\bar{u}_{i,j+1}$  et  $\bar{u}_{i,j-1}$ , et on en déduit

$$|R_{i,j}| \leq \frac{h^2}{12} \|u_{xxxx}\|_{\infty} + \frac{h^2}{12} \|u_{yyyy}\|_{\infty} + k \frac{h^2}{6} \|u_{xx}\|_{\infty}$$

où  $\|u_{xxxx}\|_{\infty}$  désigne la norme uniforme sur  $\bar{\Omega}$  de la dérivée  $h^2$  de  $u$  par rapport à  $x$  (notations analogues pour  $\|u_{yyyy}\|_{\infty}$  et  $\|u_{xx}\|_{\infty}$ ). On obtient finalement

$$|R_{ij}| \leq C_1 h^2,$$

où  $C_1$  ne dépend que de  $u$  et  $k$ . Comme pour  $h$  petit, on a  $h^2 \leq h$ , on en déduit que schéma [I] est donc d'ordre 2. Pour le schéma [II], on a :

$$R_{ij} = \frac{2\bar{u}_{ij} - \bar{u}_{i+1,j} - \bar{u}_{i-1,j}}{h^2} + \frac{2\bar{u}_{ij} - \bar{u}_{i,j+1} - \bar{u}_{i,j-1}}{h^2} + k \frac{\bar{u}_{ij} - \bar{u}_{i-1,j}}{h} - f_{ij}.$$

D'où l'on déduit

$$|R_{ij}| \leq \frac{h^2}{12} \|u_{xxxx}\|_{\infty} + \frac{h^2}{12} \|u_{yyyy}\|_{\infty} + \frac{kh}{2} \|u_{xx}\|_{\infty},$$

et donc

$$|R_{ij}| \leq C_2 h$$

où  $C_2$  ne dépend que de  $u$  et  $k$ . Le schéma [II] est donc d'ordre 1.

3. Dans le cas du schéma [II], la famille des  $w_{ij}, i, j \in \{0, \dots, N+1\}^2$  vérifie :

$$\frac{1}{h^2} (w_{ij} - w_{i+1,j}) + \left( \frac{1}{h^2} + \frac{k}{h} \right) (w_{ij} - w_{i-1,j}) + \frac{1}{h^2} (w_{ij} - w_{i,j+1}) + \frac{1}{h^2} (w_{ij} - w_{i,j-1}) \leq 0, \forall i, j \in \{1, \dots, N\}$$

On pose  $M = \max\{w_{i,j}, (i, j) \in \{0, \dots, N+1\}^2\}$  et  $m = \max\{w_{i,j}, (i, j) \in \gamma\}$ . Noter que  $\gamma = \{0, \dots, N+1\}^2 \setminus \{1, \dots, N\}^2$ . On a bien sûr  $m \leq M$  et il reste donc à montrer que  $M \leq m$ . Soit  $A = \{(i, j) \in \{0, \dots, N+1\}^2, w_{i,j} = M\}$  et soit  $(\bar{i}, \bar{j}) \in A$  tel que  $\bar{i} = \max\{i, (i, j) \in A\}$  et  $\bar{j} = \max\{j, (i, j) \in A\}$ . On distingue deux cas :

1. Si  $\bar{i} \in \{0, N+1\}$  ou  $\bar{j} \in \{0, \dots, N+1\}$ , on a alors  $(\bar{i}, \bar{j}) \in \gamma$  et donc  $M = w_{\bar{i}, \bar{j}} \leq m$ .
2. Sinon, on a  $\bar{i} \notin \{0, N+1\}$  et  $\bar{j} \notin \{0, N+1\}$ , et donc  $(\bar{i}, \bar{j}) \in \{1, \dots, N\}^2$ . On en déduit que :

$$\begin{aligned} & \frac{1}{h^2} (w_{\bar{i}, \bar{j}} - w_{\bar{i}+1, \bar{j}}) + \left( \frac{1}{h^2} + \frac{k}{h} \right) (w_{\bar{i}, \bar{j}} - w_{\bar{i}-1, \bar{j}}) + \\ & \frac{1}{h^2} (w_{\bar{i}, \bar{j}} - w_{\bar{i}, \bar{j}+1}) + \frac{1}{h^2} (w_{\bar{i}, \bar{j}} - w_{\bar{i}, \bar{j}-1}) \leq 0, \end{aligned}$$

ce qui est impossible car  $w_{\bar{i}, \bar{j}} = M$  et donc

$$\begin{aligned} w_{\bar{i}, \bar{j}} - w_{\bar{i}+1, \bar{j}} & \geq 0, \\ w_{\bar{i}, \bar{j}} - w_{\bar{i}-1, \bar{j}} & \geq 0, \\ w_{\bar{i}, \bar{j}} - w_{\bar{i}, \bar{j}+1} & \geq 0, \\ w_{\bar{i}, \bar{j}} - w_{\bar{i}, \bar{j}-1} & > 0, \end{aligned}$$

noter que la dernière inégalité est bien stricte car  $(\bar{i}+1, \bar{j}) \notin A$  (c'est l'intérêt du choix de  $\bar{i}$ ). On a donc bien montré que  $M \leq m$ .

Dans le cas du schéma [III], si on a  $\bar{i} \notin \{0, N+1\}$  et  $\bar{j} \notin \{0, N+1\}$ , et donc  $(\bar{i}, \bar{j}) \in \{1, \dots, N\}^2$  le même raisonnement que celui du schéma 1 donne :

$$\begin{aligned} & \left( \frac{1}{h^2} - \frac{k}{2h} \right) (u_{\bar{i}\bar{j}} - u_{\bar{i}+1, \bar{j}}) + \left( \frac{1}{h^2} + \frac{h}{2h} \right) (u_{\bar{i}, \bar{j}} - u_{\bar{i}-1, \bar{j}}), \\ & + \frac{1}{h^2} (u_{\bar{i}\bar{j}} - u_{\bar{i}, \bar{j}+1}) + \frac{1}{h^2} (u_{\bar{i}\bar{j}} - u_{\bar{i}, \bar{j}-1}) \leq 0. \end{aligned}$$

On ne peut conclure à une contradiction que si  $\frac{1}{h^2} - \frac{k}{2h} \geq 0$ . Le schéma [III] vérifie

$$w_{i,j} \leq \max_{(k,\ell) \in \gamma} (w_{k,\ell}) \quad \forall i, j \in \{1, \dots, N\}^2$$

lorsque  $h$  vérifie la condition (dite Condition de stabilité) :

$$h \leq \frac{2}{k} \quad (1.116)$$

4. La fonction  $\phi$  vérifie

$$\begin{aligned} \phi_x &= 0 \\ \phi_y &= y, \quad \phi_{yy} = 1, \end{aligned}$$

et donc  $-\Delta\phi + k\frac{\partial\phi}{\partial x} = -1$ . On pose maintenant  $\phi_{i,j} = \phi(ih, jh)$  pour  $i, j \in \{0, \dots, N+1\}^2$  (Noter que  $\phi$  ne vérifie pas la condition  $\phi_{i,j} = 0$  si  $(i, j) \in \gamma$ ) Comme  $\phi_{xx} = \phi_{xxx} = \phi_{xxxx} = \phi_{yyyy} = 0$ , les calculs de la question 2 montrent que pour les schémas [I] et [III],

$$a_0\phi_{i,j} - a_1\phi_{i-1,j} - a_2\phi_{i+1,j} - a_3\phi_{i,j-1} - a_4\phi_{i,j+1} = -1$$

pour  $i, j \in \{1, \dots, N\}^2$ .

En posant  $w_{i,j} = u_{i,j} + C\phi_{i,j}$  pour  $(i, j) \in \{0, \dots, N+1\}^2$  (et  $U$  solution de (1.104)) on a donc

$$a_0w_{i,j} - a_1w_{i-1,j} - a_2w_{i+1,j} - a_3w_{i,j-1} - a_4w_{i,j+1} = f_{i,j} - C \quad \forall i, j \in \{1, \dots, N\}$$

On prend  $C = \|f\|_\infty$ , de sorte que  $f_{i,j} - C \leq 0$  pour tout  $(i, j)$  pour le schéma [II] et pour le schéma [I] avec  $h \leq 2/k$ , la question 3 donne alors pour  $(i, j) \in \{1, \dots, N\}^2$ ,

$$w_{ij} \leq \max\{w_{k\ell}, (k\ell) \in \gamma\} \leq \frac{C}{2},$$

car  $u_{i,j} = 0$  si  $(i, j) \in \gamma$  et  $-\max_{\Omega} \phi = \frac{1}{2}$ . On en déduit pour  $(i, j) \in \{1, \dots, N\}^2$ ,

$$w_{ij} \leq \frac{C}{2} = \frac{1}{2}\|f\|_\infty.$$

Pour montrer que  $-w_{ij} \leq \frac{1}{2}\|f\|_\infty$ , on prend maintenant  $w_{ij} = C\phi_{ij} - u_{i,j}$  pour  $(i, j) \in \{0, \dots, N+1\}^2$ , avec  $C = \|f\|_\infty$ . On a donc

$$a_0w_{i,j} - a_1w_{i-1,j} - a_2w_{i+1,j} - a_3w_{i,j-1} - a_4w_{i,j+1} = -C - f_{i,j} \leq 0, \forall i, j \in \{1, \dots, N\}.$$

Ici encore, pour le schéma [II] ou le schéma [I] avec la condition  $h \leq \frac{2}{k}$ , la question 3 donne

$$w_{ij} \leq \max\{w_{k\ell}, (k\ell) \in \gamma\} = \frac{C}{2}$$

donc  $u_{ij} \geq -\frac{C}{2} = -\frac{\|f\|_\infty}{2}$  pour tout  $(i, j) \in \{1, \dots, N\}^2$ . Pour le schéma [II] ou le schéma [I] avec la condition  $h \leq \frac{2}{k}$ , on a donc :

$$\|U\|_\infty \leq \frac{1}{2}\|f\|_\infty. \quad (1.117)$$

Le système (1.104) peut être vu comme un système linéaire de  $N^2$  équation, à  $N^2$  inconnues (qui sont les  $u_{i,j}$  pour  $(i, j) \in \{1, \dots, N\}^2$ ). Si le second membre de ce système linéaire est nul, l'inégalité (1.117)(I) prouve que la solution est nulle. Le système (1.104) admet donc, pour tout  $f$ , au plus une solution. Ceci est suffisant pour affirmer qu'il admet, pour tout  $f$ , une et une seule solution.

5. pour  $(i, j) \in \{0, \dots, N+1\}^2$  on pose

$$e_{ij} = u(ih, jh) - u_{i,j}.$$

On a donc, pour les schémas [I] et [II], avec les notations de la question 2 :

$$a_0 e_{ij} - a_1 e_{i-1,j} - a_2 e_{i+1,j} - a_3 e_{i,j-1} - a_4 e_{i,j+1} = R_{ij}, \quad \forall i, j \in \{1, \dots, N\}^2.$$

avec les questions 2 et 4, on a donc, pour le schéma [I], si  $h \leq \frac{2}{k}$  :

$$\max\{|e_{ij}|, (i, j) \in \{1, \dots, N\}^2\} \leq \frac{1}{2} C_1 h^2,$$

où  $C_1$  et  $C_2$  ne dépendent que de  $u$  et  $k$  (et sont données à la question 2). Les 2 schémas sont convergents. Le schéma [I] converge en “ $h^2$ ” et le schéma [II] en “ $h$ ”.

6. Le schéma [1] converge plus vite mais a une condition de stabilité  $k \leq \frac{2}{h}$ . Le schéma [II] est inconditionnellement stable.

### Exercice 21 page 40

1. On a vu au paragraphe 1.5.2 page 26 que si  $\sigma$  est une arête du volume de contrôle  $K$ , alors le flux numérique  $F_{K,\sigma}$  s’écrit :

$$F_{K,\sigma} = \lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} m(\sigma).$$

On cherche à éliminer les inconnues auxiliaires  $u_\sigma$ . Pour cela, si  $\sigma$  est une arête interne,  $\sigma = K|L$ , on écrit la conservativité du flux numérique :

$$F_{K,\sigma} = -F_{L,\sigma},$$

Ce qui entraîne, si  $\sigma$  n’est pas une arête de l’interface  $I$ , que :

$$-\lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} m(\sigma) = \lambda_i \frac{u_\sigma - u_L}{d_{L,\sigma}} m(\sigma)$$

On en déduit que

$$u_\sigma \left( \frac{1}{d_{K,\sigma}} + \frac{1}{d_{L,\sigma}} \right) = \frac{u_K}{d_{K,\sigma}} + \frac{u_L}{d_{L,\sigma}},$$

soit encore que

$$u_\sigma = \frac{d_{K,\sigma} d_{L,\sigma}}{d_\sigma} \left( \frac{u_K}{d_{K,\sigma}} + \frac{u_L}{d_{L,\sigma}} \right).$$

Remplaçons alors dans (1.8). On obtient :

$$\begin{aligned} F_{K,\sigma} &= \lambda_i \left( \frac{d_{L,\sigma}}{d_\sigma} \left( \frac{u_K}{d_{K,\sigma}} + \frac{u_L}{d_{L,\sigma}} \right) - \frac{u_K}{d_{K,\sigma}} \right) \\ &= -\frac{\lambda_i}{d_\sigma} \left( \frac{d_{L,\sigma}}{d_{K,\sigma}} u_K + u_L - u_K - \frac{d_{L,\sigma}}{d_{K,\sigma}} u_K \right) \end{aligned}$$

On obtient donc finalement bien la formule (1.67).

2. Considérons maintenant le cas d’une arête  $\sigma \subset \Gamma_1 \cup \Gamma_3$ , où l’on a une condition de Fourier, qu’on a discrétisée par :

$$F_{K,\sigma} = -m(\sigma) \lambda_i \frac{u_\sigma - u_K}{d_{K,\sigma}} = m(\sigma) \alpha (u_\sigma - u_{ext}).$$

On a donc

$$u_\sigma = \frac{1}{\alpha + \frac{\lambda_i}{d_{K,\sigma}}} \left( \frac{\lambda_i u_K}{d_{K,\sigma}} + \alpha u_{ext} \right)$$

On remplace cette expression dans l’égalité précédente. Il vient :

$$F_{K,\sigma} = \frac{m(\sigma) \alpha}{\alpha + \frac{\lambda_i}{d_{K,\sigma}}} \left( \frac{\lambda_i}{d_{K,\sigma}} u_K + \alpha u_{ext} - \alpha u_{ext} - \frac{\lambda_i}{d_{K,\sigma}} u_{ext} \right),$$

Ce qui, après simplifications, donne exactement (1.68).

3. Considérons maintenant une arête  $\sigma = K|L$  appartenant à l'interface  $I$ . La discrétisation de la condition de saut de flux sur  $I$ . S'écrit :

$$F_{K,\sigma} + F_{L,\sigma} = \int_{\sigma} \theta(x) d\gamma(x) = m(\sigma)\theta_{\sigma}$$

Supposons que  $K$  (resp.  $L$ ) soit situé dans le milieu de conductivité (resp.  $\lambda_2$ ). En remplaçant  $F_{K,\sigma}$  et  $F_{L,\sigma}$  par leurs expressions, on obtient :

$$-\lambda_1 m(\sigma) \frac{u_{\sigma} - u_K}{d_{K,\sigma}} - \lambda_2 m(\sigma) \frac{u_{\sigma} - u_L}{d_{L,\sigma}} = m(\sigma)\theta_{\sigma}$$

On en déduit que

$$u_{\sigma} \left( \frac{\lambda_1}{d_{K,\sigma}} + \frac{\lambda_2}{d_{L,\sigma}} \right) = \left( \frac{\lambda_1 u_K}{d_{K,\sigma}} + \frac{\lambda_2 u_L}{d_{L,\sigma}} - \theta_{\sigma} \right)$$

En remplaçant  $u_{\sigma}$  dans l'expression de  $F_{K,\sigma}$ , on obtient :

$$F_{K,\sigma} = -\frac{m(\sigma)}{d_{K,\sigma}} \lambda_1 \frac{1}{\frac{\lambda_1}{d_{K,\sigma}} + \frac{\lambda_2}{d_{L,\sigma}}} \left( \frac{\lambda_1 u_K}{d_{K,\sigma}} + \frac{\lambda_2 u_L}{d_{L,\sigma}} - \theta_{\sigma} - \frac{\lambda_1 u_K}{d_{K,\sigma}} - \frac{\lambda_2 u_K}{d_{L,\sigma}} \right)$$

En simplifiant, on obtient :

$$F_{K,\sigma} = -\frac{m(\sigma)\lambda_1}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}} (\lambda_2 u_L - \lambda_2 u_K - d_{L,\sigma}\theta_{\sigma}),$$

ce qui est exactement (1.70). On obtient alors l'expression de  $F_{L,\sigma}$  :

$$F_{L,\sigma} = m(\sigma)\theta_{\sigma} - F_{K,\sigma},$$

ce qui donne, après simplifications :

$$F_{L,\sigma} = \frac{\lambda_2 m(\sigma)}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}} [\lambda_1 (u_L - u_K) + d_{K,\sigma}\theta_{\sigma}].$$

On vérifie bien que  $F_{K,\sigma} + F_{L,\sigma} = m(\sigma)\theta_{\sigma}$ .

4. Le système linéaire que satisfont les inconnues  $(u_K)_{K \in \mathcal{M}}$  s'écrit

$$AU = b$$

avec  $U = (u_K)_{K \in \mathcal{T}}$ . Pour construire les matrices  $A$  et  $b$ , il faut se donner une numérotation des mailles. On suppose qu'on a  $n \times 2p$  mailles ; on considère un maillage uniforme du type de celui décrit sur la figure 1.5 page 27 et on note  $h_x = \frac{1}{n}$  (resp.  $h_y = \frac{1}{p}$ ) la longueur de la maille dans la direction  $x$  (resp.  $y$ ). Comme le maillage est cartésien, il est facile de numéroter les mailles dans l'ordre "lexicographique" ; c'est-à-dire que la  $k$ -ième maille a comme centre le point  $x_{i,j} = ((i - \frac{1}{2})h_x, (j - \frac{1}{2})h_y)$ , avec  $k = n(j - 1) + i$ . On peut donc déterminer le numéro de la maille (et de l'inconnue associée)  $k$  à partir de la numérotation cartésienne  $(i, j)$  de la maille.

$$k = n(j - 1) + i$$

Remarquons que, comme on a choisi un maillage uniforme, on a pour tout  $K \in \mathcal{T}$  :  $m(K) = h_x h_y$ , pour toute arête intérieure verticale  $\sigma$  :  $d_{\sigma} = h_x m(\sigma) = h_y$  et pour toute arête intérieure horizontale,  $d_{\sigma} = h_y$  et  $m(\sigma) = h_x$ . Pour chaque numéro de maille, nous allons maintenant construire l'équation correspondante.

Mailles internes  $i = 2, \dots, n - 1$  ;  $j = 2, \dots, p - 1, p + 1, \dots, 2p - 1$ .

L'équation associée à une maille interne  $K$  s'écrit

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = m(K) f_K.$$

Avec l'expression de  $F_{K,\sigma}$  donnée par (1.67), ceci amène à :

$$2\lambda_m \left( \frac{h_x}{h_y} + \frac{h_y}{h_x} \right) u_k - \lambda_m \frac{h_x}{h_y} (u_{k-n} + u_{k+n}) - \lambda_m \frac{h_y}{h_x} (u_{k+1} + u_{k-1}) = h_x h_y f_k,$$



avec  $m = 1$  si  $j \leq p - 1$  et  $m = 2$  si  $j \geq p + 1$ .

Mailles du bord  $\Gamma_2$  Les mailles du bord  $\Gamma_2$  sont repérées par les indices  $(n, j)$ ,  $j = 2$  à  $p - 1$ ,  $j = p + 1$  à  $2p - 1$ , (on exclut pour l'instant les coins).

L'équation des flux est la même que pour les mailles internes, mais le flux sur la frontière  $\Gamma_2$  est nul. Ceci donne :

$$\lambda_m \left( 2 \frac{h_x}{h_y} + \frac{h_y}{h_x} \right) u_k - \lambda_m \frac{h_x}{h_y} (u_{k-n} + u_{k+n}) - \lambda_m \frac{h_y}{h_x} u_{k-1} = h_x h_y f_k,$$

avec  $k = n(j - 1) + n$ ,  $j = 2$  à  $p - 1$ ,  $j = p + 1$  à  $2p - 1$  et  $m = 1$  si  $j \leq p - 1$ ,  $m = 2$  si  $j \geq p + 1$ .

Mailles de bord  $\Gamma_4$  Les mailles du bord  $\Gamma_4$  sont repérées par les indices  $(1, j)$ ,  $j = 2$  à  $p - 1$ ,  $j = p + 1$  à  $2p - 1$ . Pour ces mailles, il faut tenir compte du fait que sur une arête de  $\Gamma_4$ , le flux  $F_{K,\sigma}$  est donné par :

$$F_{K,\sigma} = -\lambda_m \frac{g_\sigma - u_K}{d_{K,\sigma}} m(\sigma)$$

$$\text{avec } g_\sigma = \frac{1}{m(\sigma)} \int g(y) d\gamma(y).$$

D'où on tire l'équation relative à la maille  $k = n(j - 1) + 1$ ,  $j = 2, \dots, p - 1, p + 1, \dots, 2p - 1$  :

$$\lambda_m \left( 2 \frac{h_x}{h_y} + 3 \frac{h_y}{h_x} \right) u_k - \lambda_m \frac{h_x}{h_y} (u_{k-n} + u_{k+n}) - \lambda_m \frac{h_y}{h_x} u_{k+1} = h_x h_y f_k + 2 \frac{h_y}{h_x} \lambda_m g_j,$$

avec  $g_j = g_{\sigma_j}$  et  $m = 1$  si  $j \leq p - 1$ ,  $m = 2$  si  $j \geq p + 1$ .

Mailles du bord  $\Gamma_1 \cup \Gamma_3$  Pour  $j = 1$ , où  $j = 2p$ ,  $i = 2 \dots n - 1$ . On tient compte ici de la condition de Fourier sur la maille qui appartient au bord, pour laquelle l'expression du flux est :

$$F_{K,\sigma} = \frac{\alpha \lambda_m m(\sigma)}{\lambda_m + \alpha d_{K,\sigma}} (u_K - u_{ext}).$$

Pour une arête  $\sigma$  horizontale, on note :  $C_{F,\sigma} = \frac{\alpha m(\sigma)}{\lambda_m + \alpha d_{K,\sigma}}$ . Notons que  $C_{F,\sigma}$  est égal à

$$C_F = \frac{2\alpha h_x}{2\lambda_m + \alpha h_y}.$$

Notons que ce coefficient ne dépend pas de  $\sigma$ .

Les équations s'écrivent donc :

$$\lambda_1 \left( 2 \frac{h_y}{h_x} + \frac{h_x}{h_y} + C_F \right) u_k - \lambda_1 \frac{h_x}{h_y} u_{k+n} - \lambda_1 \frac{h_y}{h_x} (u_{k+1} + u_{k-1}) = h_x h_y f_k + \lambda_1 C_F u_{ext},$$

$k = 2, \dots, n - 1,$

$$\lambda_2 \left( \frac{2h_y}{h_x} + \frac{h_x}{h_y} + C_F \right) u_k - \lambda_2 \frac{h_x}{h_y} u_{k-n} - \lambda_2 \frac{h_y}{h_x} (u_{k+1} + u_{k-1}) = h_x h_y f_k + \lambda_2 C_F u_{ext},$$

$k = 2n(p - 1) + 2, \dots, 2np - 1,$

Mailles des coins extérieurs : Il suffit de synthétiser les calculs déjà faits :

• coin sud-est :  $i = 1$ ,  $j = 1$ ,  $k = 1$  ; un bord Dirichlet, un bord Fourier :

$$\lambda_1 \left( \frac{3h_y}{h_x} + \frac{h_x}{h_y} + C_F \right) u_1 - \lambda_1 \frac{h_y}{h_x} u_2 - \lambda_1 \frac{h_x}{h_y} u_{n+1} = h_x h_y f_1 + \lambda_1 C_F u_{ext} + \frac{2h_y}{h_x} \lambda_1 g_1$$

• coin sud-ouest :  $i = 1n$ ,  $j = 1$ ,  $k = n$  ; un bord Fourier, un bord Neumann :

$$\lambda_1 \left( \frac{h_y}{h_x} + \frac{h_x}{h_y} + C_F \right) u_1 - \lambda_1 \frac{h_y}{h_x} u_{n-1} - \lambda_1 \frac{h_x}{h_y} u_{2n} = h_x h_y f_n + \lambda_1 C_F u_{ext}$$

• coin nord-ouest :  $i = 2n$ ,  $j = 2p$ ,  $k = 2np$ .

On a encore un bord Fourier, un bord Neumann, et l'équation s'écrit :

$$\lambda_2 \left( \frac{h_y}{h_x} + \frac{h_x}{h_y} + C_F \right) u_{2np} - \lambda_2 \frac{h_y}{h_x} u_{2np-1} - \lambda_2 \frac{h_x}{h_y} u_{2n(p-1)} = h_x h_y f_{2np} + \lambda_2 C_F u_{ext}$$

- coin nord-est :  $i = 1, j = 2p \quad k = n(2p - 1) + 1$  un bord Dirichlet, un bord Fourier :

$$\lambda_2 \left( \frac{3h_y}{h_x} + \frac{h_x}{h_y} + C_F \right) u_k - \lambda_2 \frac{h_y}{h_x} u_{k+1} - \lambda_2 \frac{h_x}{h_y} (u_{k-n} = h_x h_y f_k + \lambda_2 C_F u_{ext}, + \frac{2h_y}{h_x} \lambda_2 g_k.$$

Interface L'expression du flux sur une arête de l'interface est donnée par (1.70). On pose, pour chaque arête  $\sigma$  de l'interface,

$$s_{I,\sigma} = \frac{m(\sigma)}{\lambda_1 d_{L,\sigma} + \lambda_2 d_{K,\sigma}}.$$

Notons que dans le cas du maillage uniforme considéré, ce coefficient est égal à :

$$s_I = \frac{2h_x}{(\lambda_1 + \lambda_2)h_y},$$

et qu'il est indépendant de l'arête  $\sigma$ . Tenant compte de ce flux, on obtient, pour  $k = n(p-1) + i, i = 2, \dots, N-1$

$$\lambda_1 \left( \frac{2h_y}{h_x} + \frac{h_x}{h_y} + \lambda_2 s_I \right) u_k - \lambda_1 \frac{h_y}{h_x} u_{k+1} - \lambda_1 \frac{h_y}{h_x} u_{k-1} - \lambda_1 \frac{h_x}{h_y} u_{k-n} - \lambda_1 s_I u_{k+n} = h_x h_y f_k + \lambda_1 s_I \frac{h_y}{2} \theta_i,$$

avec

$$\theta_i = \int_{\sigma_i} \theta(x) d\gamma(x).$$

Et de même, pour  $k = np + i, i = 2, \dots, N-1$ ,

$$\lambda_1 \left( \frac{2h_y}{h_x} + \frac{h_x}{h_y} + \lambda_1 s_I \right) u_k - \lambda_2 \frac{h_y}{h_x} u_{k+1} - \lambda_2 \frac{h_y}{h_x} u_{k-1} - \lambda_2 \frac{h_x}{h_y} u_{k+n} - \lambda_2 s_I u_{k-n} = h_x h_y f_k + \lambda_2 s_I \frac{h_y}{2} \theta_i.$$

Il ne reste plus qu'à traiter les coins des interfaces.

- $i = 1, j = p \quad k = n(p-1) + 1$ . Dirichlet sous l'interface

$$\lambda_1 \left( \frac{3h_y}{h_x} + \frac{h_x}{h_y} + \lambda_2 s_I \right) u_k - \lambda_1 \frac{h_y}{h_x} u_{k+1} - \lambda_1 \frac{h_x}{h_y} u_{k+n} - \lambda_1 s_I u_{k+n} = h_x h_y f_k + \lambda_1 s_I \frac{h_y}{2} \theta_i + \frac{2h_y}{h_x} \lambda_1 g_j$$

- $i = 1, j = p+1 \quad k = np + 1$ , Dirichlet, dessus de l'interface

$$\lambda_2 \left( \frac{3h_y}{h_x} + \frac{h_x}{h_y} + \lambda_1 s_I \right) u_k - \lambda_2 \frac{h_y}{h_x} u_{k+1} - \lambda_2 \frac{h_x}{h_y} u_{k+n} - \lambda_2 s_I u_{k-n} = h_x h_y f_k + \lambda_2 s_I \frac{h_y}{2} \theta_i + \frac{2h_y}{h_x} \lambda_2 g_j$$

- $i = n, j = p, k = n(p-1) + n$ . Neumann, sous l'interface.

$$\lambda_1 \left( \frac{h_y}{h_x} + \frac{h_x}{h_y} + \lambda_2 s_I \right) u_k - \lambda_1 \frac{h_y}{h_x} u_{k-1} - \lambda_1 \frac{h_x}{h_y} u_{k-n} - \lambda_1 s_I u_{k+n} = h_x h_y f_k + \lambda_1 s_I \frac{h_y}{2} \theta_i$$

- $i = n, j = p+1, k = np + n$ , Neuman, dessus de l'interface

$$\lambda_2 \left( \frac{h_y}{h_x} + \frac{h_x}{h_y} + \lambda_1 s_I \right) u_k - \lambda_2 \frac{h_y}{h_x} u_{k-1} - \lambda_2 \frac{h_x}{h_y} u_{k+n} - \lambda_2 s_I u_{k-n} = h_x h_y f_k + \lambda_2 s_I \frac{h_y}{2} \theta_i.$$

On a ainsi obtenu  $2np$  équations à  $2np$  inconnues. Notons que chaque équation fait intervenir au plus 5 inconnues.

## Exercice 20 page 39

1. Le problème complet s'écrit :

$$\begin{cases} -\operatorname{div}(\mu_i \nabla \phi)(x) = 0, & x \in \Omega_i, \quad i = 1, 2 \\ \nabla \phi(x) \cdot \mathbf{n}(x) = 0, & x \in \Gamma_2 \cup \Gamma_4, \\ \int_{\Gamma_1} \mu_1 \nabla \phi(x) \cdot \mathbf{n}(x) d\gamma(x) + \int_{\Gamma_3} \mu_2 \nabla \phi(x) \cdot \mathbf{n}(x) d\gamma(x) = 0, \\ \phi_2(x) - \phi_1(x) = 0, & \forall x \in I, \\ -(\mu \nabla \phi \cdot \mathbf{n})|_2(x) - (\mu \nabla \phi \cdot \mathbf{n})|_1(x) = 0, & \forall x \in I. \end{cases}$$

2. On se donne le même maillage rectangulaire uniforme qu'à l'exercice précédent. On note  $\phi_K$  l'inconnue associée à la maille  $K$  (ou  $\phi_k$  si on la référence la maille  $K$  par son numéro  $k = n(j-1) + i$ , où  $i \in \{1, \dots, n\}$  et  $j \in \{1, \dots, 2p\}$ ). Pour une maille intérieure, l'équation obtenue est la même que (1.8) en remplaçant  $\lambda_m$  par  $\mu_m$ . Etudions maintenant le cas d'une maille proche de l'interface. Comme indiqué, on va considérer deux inconnues discrètes par arête de l'interface. Soient  $K$  et  $L$  ayant en commun l'arête  $\tilde{\sigma} \subset I$ ,  $K$  est située au dessous de  $L$ . Les équations associées à  $K$  et  $L$  s'écrivent alors

$$\sum_{\sigma \in \xi_K} F_{K,\sigma} = 0 \text{ et } \sum_{\sigma \in \xi_L} F_{L,\sigma} = 0.$$

Pour les arêtes  $\sigma \in \xi_K$  autres que  $\tilde{\sigma}$ , le flux s'écrit de manière habituelle

$$F_{K,\sigma} = \mu_1 \frac{\phi_K - \phi_M}{d_\sigma}, \text{ avec } \sigma = K|M.$$

Pour l'arête  $\sigma = \tilde{\sigma}$ , on a  $F_{K,\sigma} = \mu_1 \frac{\phi_K - \phi_\sigma^-}{d_\sigma} m(\sigma)$  et  $F_{L,\sigma} = \mu_2 \frac{\phi_L - \phi_\sigma^+}{d_\sigma} m(\sigma)$ , où les deux inconnues discrètes  $\phi_\sigma^+$  et  $\phi_\sigma^-$  sont reliées par les relations :

$$\phi_\sigma^+ - \phi_\sigma^- = \psi_\sigma \left( = \frac{1}{m(\sigma)} \int_\sigma \psi(x) d\gamma(x) \right)$$

$$F_{K,\sigma} + F_{L,\sigma} = 0.$$

On peut alors éliminer  $\phi_\sigma^+$  et  $\phi_\sigma^-$  ; en utilisant par exemple  $\phi_\sigma^+ = \psi_\sigma + \phi_\sigma^-$  et en remplaçant dans la deuxième équation, on obtient :

$$-\mu_1 \frac{\phi_\sigma^- - \phi_K}{d_{K,\sigma}} + \mu_2 \frac{\phi_\sigma^- + \psi_\sigma - \phi_L}{d_{L,\sigma}} = 0,$$

ce qui donne :

$$\phi_\sigma^- = \frac{1}{\frac{\mu_1}{d_{K,\sigma}} + \frac{\mu_2}{d_{L,\sigma}}} \left( \frac{\mu_1}{d_{K,\sigma}} \phi_K + \frac{\mu_2}{d_{L,\sigma}} \phi_L - \frac{\mu_2}{d_{L,\sigma}} \psi_\sigma \right).$$

En remplaçant cette expression dans les flux, on obtient :

$$F_{K,\sigma} = -F_{L,\sigma} = m(\sigma) \frac{\mu_1 \mu_2}{\mu_1 d_{L,\sigma} + \mu_2 d_{K,\sigma}} (\phi_K - \phi_L + \psi_\sigma)$$

On peut alors écrire l'équation discrète associée à une maille de numéro  $k$  située sous l'interface (avec  $k = n(p-1) + i$ ,  $i = 2, \dots, n-1$ ). On pose :

$$\frac{\mu_1 \mu_2}{\mu_1 d_{L,\sigma} + \mu_2 d_{K,\sigma}} = \frac{\mu_I}{d_\sigma}$$

( $\mu_I$  est donc la moyenne harmonique pondérée entre  $\mu$  et  $\mu_2$ ). Notons que pour une arête de  $I$ ,  $d_\sigma = h_y$ , et  $m(\sigma) = h_x$ . L'équation associée à la maille  $k$  s'écrit donc :

$$\left( 2\mu_1 \frac{h_y}{h_x} + \mu_1 \frac{h_x}{h_y} + \frac{\mu_I h_x}{h_y} \right) u_k - \mu_1 \frac{h_y}{h_x} (u_{k-1} + u_{k+1}) - \mu_1 \frac{h_x}{h_y} u_{k-n} - \mu_I \frac{h_x}{h_y} u_{k+n} = -\mu_I \frac{h_x}{h_y} \psi_i,$$

où  $\psi_i$  est le saut de potentiel à travers l'arête  $\sigma_i$  de l'interface considérée ici. De même, l'équation associée à une maille  $k$  avec  $k = np + i$ ,  $i = 2, \dots, n-1$ , située au dessus de l'interface s'écrit :

$$\left( 2\mu_1 \frac{h_y}{h_x} + \mu_1 \frac{h_x}{h_y} + \mu_I \frac{h_x}{h_y} \right) u_k - \mu_1 \frac{h_y}{h_x} (u_{k-1} + u_{k+1}) - \mu_1 \frac{h_x}{h_y} u_{k+n} - \mu_I \frac{h_x}{h_y} u_{k-n} = +\mu_I \frac{h_x}{h_y} \psi_i.$$

La discrétisation des conditions aux limites de Neumann sur  $\Gamma_2$  et  $\Gamma_4$  est effectuée de la même manière que pour le cas du problème thermique, voir exercice 21.

Il ne reste plus qu'à discrétiser la troisième équation du problème (1.8), qui relie les flux sur la frontière  $\Gamma_1$  avec les flux sur la frontière  $\Gamma_3$ . En écrivant la même condition avec les flux discrets, on obtient :

$$\mu_1 \sum_{i=1}^n \frac{2h_x}{h_y} (u_i - u_{B,i}) + \mu_2 \sum_{i=1}^n \frac{2h_x}{h_y} (u_{H,i} - u_{k(i)}) = 0,$$

où :  $\mu_{B,i}$  représente l'inconnue discrète sur la  $i$ -ème arête de  $\Gamma_1$  et  $\mu_{H,i}$  l'inconnue discrète sur la  $i$ -ème arête de  $\Gamma_3$ , et  $k(i) = n(p-1) + i$  est le numéro de la maille jouxtant la  $i$ -ème arête de  $\Gamma_3$ .

Remarquons que tel qu'il est posé, le système n'est pas inversible : on n'a pas assez d'équations pour éliminer les inconnues  $u_{B,i}$  et  $u_{H,i}$ ,  $i = 1 \dots N$ . On peut par exemple pour les obtenir considérer une différence de potentiel fixée entre  $\Gamma_1$  et  $\Gamma_3$ , et se donner un potentiel fixé sur  $\Gamma_1$ .

## Chapitre 2

# Problèmes paraboliques : la discrétisation en temps

On a vu au paragraphe ?? comme exemple type de problème parabolique l'équation de la chaleur instationnaire :

$$u_t - \Delta u = f$$

qui fait intervenir la dérivée en temps d'ordre 1,  $u_t$ , ainsi qu'un opérateur différentiel d'ordre 2 en espace. Pour que ce problème soit bien posé, il faut spécifier des conditions aux limites sur la frontière de  $\Omega$ , et une condition initiale en  $t = 0$ .

### 2.1 Le problème continu, et la discrétisation espace-temps

On considère maintenant le même problème en une dimension d'espace. Au temps  $t = 0$ , on se donne une condition initiale  $u_0$ , et on considère des conditions aux limites de type Dirichlet homogène. Le problème unidimensionnel s'écrit :

$$\begin{cases} u_t - u_{xx} = 0, \forall x \in ]0, 1[, \forall t \in ]0, T[ \\ u(x, 0) = u_0(x), \forall x \in ]0, 1[, \\ u(0, t) = u(1, t) = 0, \forall t \in ]0, T[, \end{cases} \quad (2.1)$$

où  $u(x, t)$  représente la température au point  $x$  et au temps  $t$ ,  $u_t$  désigne la dérivée partielle première de  $u$  par rapport à  $t$  et  $u_{xx}$  la dérivée partielle seconde de  $u$  par rapport à  $x$ . On admettra le théorème d'existence et unicité suivant :

**Théorème 2.1 (Résultat d'existence et unicité)** *Si  $u_0 \in C(]0, 1[, \mathbb{R})$  alors il existe une unique fonction  $u \in C^2(]0, 1[ \times ]0, T[, \mathbb{R}) \cap C([0, 1] \times [0, T], \mathbb{R})$  qui vérifie (2.1).*

On a même  $u \in C^\infty(]0, 1[ \times ]0, T[, \mathbb{R})$ . Ceci est appelé, effet "régularisant" de l'équation de la chaleur.

**Proposition 2.2 (Principe du maximum)** *Sous les hypothèses du théorème 2.1, soit  $u$  la solution du problème (2.1) :*

1. *si  $u^0(x) \geq 0$  pour tout  $x \in [0, 1]$ , alors  $u(x, t) \geq 0$ , pour tout  $t \geq 0$  pour tout  $x \in ]0, 1[$ .*
2.  $\|u\|_{L^\infty([0, 1] \times [0, T])} \leq \|u\|_{L^\infty([0, 1])}$ .

Ces dernières propriétés peuvent être importantes dans le modèle physique ; supposons par exemple que  $u$  représente une fraction massique. Par définition de la fraction massique, celle-ci est toujours comprise entre 0 et 1. La proposition précédente nous dit que la quantité  $u$  donné par le modèle mathématique supposé représenter la fraction massique d'une espèce qui diffuse dans un milieu, par exemple, est aussi comprise entre 0 et 1, dès que la fraction massique initiale  $u_0$  est dans l'intervalle  $[0, 1]$  ce qui est plutôt une bonne nouvelle : le modèle mathématique respecte les bornes de la physique. Mais on ne peut pas en général calculer la solution de (2.1) de manière analytique. On a recours à la discrétisation en temps et en espace pour se ramener à un système d'équations de dimension finie. Il est souhaitable pour la validité du calcul que la solution approchée obtenue par la résolution de ce système, qui est supposée approcher une fraction massique soit aussi comprise à tout instant entre 0 et 1. On dit souvent d'une

méthode de discrétisation (ou d'un schéma de discrétisation) qu'elle (ou il) est "robuste" ou "stable" s'il préserve les bornes imposées par la physique (0 et 1 dans le cas de la fraction massique évoquée ci-dessus).

Pour calculer une solution approchée, on se donne une discrétisation en temps et en espace, qu'on notera  $\mathcal{D}$ . On choisit pour l'instant de discrétiser par différences finies en temps et en espace. La discrétisation consiste donc à se donner un ensemble de points  $t_n, n = 1, \dots, M$  de l'intervalle  $]0, T[$ , et un ensemble de points  $x_i, i = 1, \dots, N$ . Pour simplifier, on considère un pas constant en temps et en espace. Soit  $h = \frac{1}{N+1} = \Delta x$  le pas de discrétisation en espace, et  $k = \Delta t = \frac{T}{M}$ , le pas de discrétisation en temps. On pose alors  $t_n = nk$  pour  $n = 0, \dots, M$  et  $x_i = ih$  pour  $i = 0, \dots, N+1$ . On cherche à calculer une solution approchée  $u_{\mathcal{D}}$  du problème (2.1); plus précisément, on cherche à déterminer  $u_{\mathcal{D}}(x_i, t_n)$  pour  $i = 1, \dots, N$ , et  $n = 1, \dots, M$ . Les inconnues discrètes sont notées  $u_i^{(n)}$ ,  $i = 1, \dots, N$  et  $n = 1, \dots, M$ .

## 2.2 Discrétisation par Euler explicite en temps.

L'approximation en temps par la méthode d'Euler explicite consiste à écrire la première équation de (2.1) en chaque point  $x_i$  et temps  $t_n$ , à approcher la dérivée en temps  $u_t(x_i, t_n)$  par le quotient différentiel :

$$\frac{u(x_i, t_{n+1}) - u(x_i, t_n)}{k},$$

et la dérivée en espace  $-u_{xx}(x_i, t_n)$  par le quotient différentiel :

$$\frac{1}{h^2}(2u(x_i, t_n) - u(x_{i-1}, t_n) - u(x_{i+1}, t_n)).$$

**Remarque 2.3** On a choisi une discrétisation en espace de type différences finies, mais on aurait aussi bien pu prendre un schéma de volumes finis ou d'éléments finis.

On obtient le schéma suivant :

$$\begin{cases} \frac{u_i^{(n+1)} - u_i^{(n)}}{k} + \frac{1}{h^2}(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}) = 0, & i = 1, \dots, N, \quad n = 1, \dots, M, \\ u_i^0 = u_0(x_i), & i = 1, \dots, N, \\ u_0^{(n)} = u_{N+1}^{(n)} = 0, & \forall n = 1, \dots, M. \end{cases} \quad (2.2)$$

le schéma est dit explicite, car la formule ci-dessus donne  $u_i^{(n+1)}$  de manière explicite en fonction des  $(u_i^{(n)})_{i=1, \dots, N}$ . En effet on a :

$$u_i^{(n+1)} = u_i^{(n)} - \lambda(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}),$$

avec  $\lambda = \frac{k}{h^2}$ .

### 2.2.1 Consistance du schéma

Soit  $\bar{u}_i^{(n)} = u(x_i, t_n)$  la valeur exacte de la solution en  $x_i$  et  $t_n$  : L'erreur de consistance  $R_i$  en  $(x_i, t_n)$  peut s'écrire comme la somme des erreurs de consistance en temps et en espace :  $R_i^{(n)} = \tilde{R}_i^{(n)} + \hat{R}_i^{(n)}$  avec :

$$\tilde{R}_i^{(n)} = \frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} - u_t(x_i, t_n) \text{ et } \hat{R}_i^{(n)} = \frac{1}{h^2} \left( 2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)} \right) - u_{xx}(x_i, t_n).$$

**Proposition 2.4** Le schéma (2.2) est consistant d'ordre 1 en temps et d'ordre 2 en espace, c'est à dire qu'il existe  $C \in \mathbb{R}_+$  ne dépendant que de  $u$  tel que :

$$|R_i^{(n)}| \leq C(k + h^2). \quad (2.3)$$

**Démonstration :** On a vu lors de l'étude des problèmes elliptiques que l'erreur de consistance en espace  $\tilde{R}_i^{(n)}$  est d'ordre 2 (voir formule (1.30) page 15). Un développement de Taylor en temps donne facilement que  $\hat{R}_i^{(n)}$  est d'ordre 1 en temps. ■

### 2.2.2 Stabilité

On a vu à la proposition 2.2 page 59 que la solution exacte vérifie :

$$\|u\|_{L^\infty(]0,1[ \times ]0,T[)} \leq \|u_0\|_{L^\infty(]0,1[)}$$

Si on choisit correctement les pas de temps et d'espace, nous allons voir qu'on peut avoir l'équivalent discret sur la solution approchée.

**Définition 2.5** On dit qu'un schéma est  $L^\infty$ -stable si la solution approchée est bornée dans  $L^\infty$  indépendamment du pas du maillage.

**Proposition 2.6** Si la condition de stabilité

$$\lambda = \frac{k}{h^2} \leq \frac{1}{2} \quad (2.4)$$

est vérifiée, alors le schéma (2.2) est  $L^\infty$ -stable au sens où :

$$\sup_{\substack{i=1,\dots,N \\ n=1,\dots,M}} |u_i^{(n)}| \leq \|u_0\|_\infty$$

**Démonstration :** On peut écrire le schéma sous la forme

$$u_i^{(n+1)} = u_i^{(n)} - \lambda(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}),$$

soit encore :

$$u_i^{(n+1)} = (1 - 2\lambda)u_i^{(n)} + \lambda u_{i-1}^{(n)} + \lambda u_{i+1}^{(n)}.$$

Si  $0 \leq \lambda \leq \frac{1}{2}$ , on a  $\lambda \geq 0$  et  $1 - 2\lambda \geq 0$ , et la quantité  $u_i^{(n+1)}$  est donc combinaison convexe de  $u_i^{(n)}$ ,  $u_{i-1}^{(n)}$  et  $u_{i+1}^{(n)}$ . Soit  $M^{(n)} = \max_{i=1,\dots,N} u_i^{(n)}$ , on a alors :

$$u_i^{(n+1)} \leq (1 - 2\lambda)M^{(n)} + \lambda M^{(n)} + \lambda M^{(n)}, \quad \forall i = 1, \dots, N,$$

et donc  $u_i^{(n+1)} \leq M^{(n)}$ . On en déduit en passant au maximum que :

$$M^{(n+1)} \leq M^{(n)}.$$

On montre de la même manière que

$$\min_{i=1,\dots,N} u_i^{(n+1)} \geq \min_{i=1,\dots,N} u_i^{(n)}.$$

On en déduit  $\max_{i=1,\dots,N} (u_i^{(n+1)}) \leq \max u_i^0$  et  $\min_{i=1,\dots,N} (u_i^{(n+1)}) \geq \min u_i^0$  d'où le résultat. ■

### 2.2.3 Convergence

**Définition 2.7** Soit  $u$  la solution du problème (2.1) et  $(u_i^{(n)})_{\substack{i=1,\dots,N \\ n=1,\dots,M}}$  la solution de (2.2). On appelle erreur de discrétisation au point  $(x_i, t_n)$  la quantité  $e_i^n = u(x_i, t_n) - u_i^n$ .

**Théorème 2.8** Sous les hypothèses du théorème 2.1, et sous la condition de stabilité (2.4), il existe  $C \in \mathbb{R}_+$  ne dépendant que de  $u$  tel que

$$\|e_i^{(n+1)}\|_\infty \leq \|e_i^{(0)}\|_\infty + TC(k + h^2), \text{ pour tout } i = 1, \dots, N \text{ et } n = 0, \dots, M - 1.$$

Ainsi, si  $\|e_i^{(0)}\|_\infty = 0$ , alors  $\max_{i=1,\dots,N} \|e_i^{(n)}\|$  tend vers 0 lorsque  $k$  et  $h$  tendent vers 0, pour tout  $n = 1, \dots, M$ . Le schéma (2.2) est donc convergent.

**Démonstration :** On note  $\bar{u}_i^{(n)} = u(x_i, t_n)$ . On a donc, par définition de l'erreur de consistance,

$$\frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} - \frac{1}{h^2}(2\bar{u}_i^{(n)} - \bar{u}_{i-1}^{(n)} - \bar{u}_{i+1}^{(n)}) = R_i^{(n)}. \quad (2.5)$$

D'autre part, le schéma numérique s'écrit :

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} - \frac{1}{h^2}(2u_i^{(n)} - u_{i-1}^{(n)} - u_{i+1}^{(n)}) = 0. \quad (2.6)$$

Retranchons (2.6) à (2.5), on obtient :

$$\frac{e_i^{(n+1)} - e_i^{(n)}}{k} - \frac{1}{h^2}(2e_i^{(n)} - e_{i+1}^{(n)} - e_{i-1}^{(n)}) = R_i^{(n)},$$

soit encore :

$$e_i^{(n+1)} = (1 - 2\lambda)e_i^{(n)} + \lambda e_{i-1}^{(n)} + \lambda e_{i+1}^{(n)} + kR_i^{(n)}$$

Or  $(1 - 2\lambda)e_i^{(n)} + \lambda e_{i-1}^{(n)} + \lambda e_{i+1}^{(n)} \leq \|e^{(n)}\|_\infty$ , car  $\lambda \leq \frac{1}{2}$ , et donc comme le schéma est consistant, l'inégalité (2.3) entraîne que :

$$|e_i^{(n+1)}| \leq \|e^{(n)}\|_\infty + kC(k + h^2).$$

On a donc par récurrence :

$$\|e_i^{(n+1)}\|_\infty \leq \|e^{(0)}\|_\infty + MkC(k + h^2)$$

ce qui démontre le théorème. ■

Donnons maintenant un exemple où lorsque la condition (2.3) n'est pas vérifiée, le schéma est instable.

## 2.2.4 Exemple de non convergence

Montrons que si la condition  $\lambda \leq \frac{1}{2}$  n'est pas respectée, on peut construire une condition initiale pour lequel le schéma n'est pas stable. Soit  $u_0 \in C([-1, 1], \mathbb{R})$  qui vérifie (voir Figure (2.1)) :

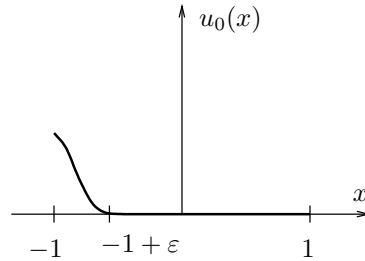


FIG. 2.1 – Condition initiale pour le contre exemple

$$\begin{cases} u_0(x) \geq 0 \\ u_0(x) \neq 0 \text{ si } x \in ]-1; -1 + \epsilon[ \\ u_0(x) = 0 \text{ si } x > -1 + \epsilon \end{cases}$$

On considère le problème :

$$\begin{cases} u_t - u_{xx} = 0, \forall x \in ]-1, 1[; \forall t > 0. \\ u(x, 0) = u_0(x), \forall x \in ]-1, 1[ \\ u(1, t) = u(-1, t) = 0, \forall t > 0. \end{cases} \quad (2.7)$$

On peut montrer que la solution exacte  $u$  de (2.7) vérifie  $u(x, t) > 0, \forall x \in ]-1, 1[, \forall t > 0$ . En particulier, pour un temps  $T > 0$  donné, on a  $u(0, T) > 0$ . Soit  $M \in \mathbb{N}$  et  $k = T/M$ . Soit  $u_i^{(n)}$  la solution approchée par (2.2), sensée



approcher  $u(x_i, t_n)$  ( $i \in \{-N, \dots, N\}, n \in N$ ). On va montrer que  $u_0^M = 0$  pour  $k$  et  $h$  choisis de manière non admissible ; ceci montre que le schéma ne peut pas converger. Calculons  $u_0^M$  :

$$u_0^M = (1 - 2\lambda)u_0^{M-1} + \lambda u_{-1}^{M-1} + \lambda u_1^{M-1}.$$

Donc  $u_0^M$  dépend de

$$\begin{aligned} & u^{(M-1)} \text{ sur } [-h, h] \\ & u^{(M-2)} \text{ sur } [-2h, 2h] \\ & \vdots \\ & u^{(0)} \text{ sur } [-Mh, Mh] = \left[-\frac{T}{k}h, \frac{T}{k}h\right] \end{aligned}$$

Par exemple, si on prend  $\frac{h}{k} = \frac{1}{2T}$  on obtient :  $[-\frac{T}{k}h, \frac{T}{k}h] = [-\frac{1}{2}, \frac{1}{2}]$ , et donc, si  $\varepsilon < \frac{1}{2}$ , on a  $u_0^M = 0$ . On peut donc remarquer que si  $\frac{h}{k} = \frac{1}{2T}$ , même si  $h \rightarrow 0$  et  $k \rightarrow 0$ ,

$$u_0^M \not\rightarrow u(0, T).$$

Le schéma ne converge pas ; notons que ceci n'est pas en contradiction avec le résultat de convergence 2.8 page 61, puisqu'ici, on n'a pas satisfait à la condition  $\frac{k}{h^2} \leq \frac{1}{2}$ .

## 2.2.5 Stabilité au sens des erreurs d'arrondi

On considère le schéma d'Euler explicite pour l'équation (2.1). On appelle  $u$  la solution exacte de (2.1),  $u_{\mathcal{D}}$  la solution exacte de (2.2),  $u_{num}$  la solution effectivement calculée par l'ordinateur. On peut écrire :

$$u - u_{num} = u - u_{\mathcal{D}} + u_{\mathcal{D}} - u_{num}.$$

On sait que l'erreur de discrétisation  $u - u_{\mathcal{D}}$  tend vers 0 lorsque  $h$  et  $k$  tendent vers 0, sous condition de stabilité (2.4), c.à.d.

$$\lambda \leq \frac{1}{2}.$$

Pour contrôler l'erreur entre la solution  $u_{\mathcal{D}}$  du schéma (2.2) et la solution numérique obtenue  $u_{num}$ , on cherche à estimer l'amplification de l'erreur commise sur la donnée initiale. Rappelons que le schéma s'écrit :

$$u_i^{(n+1)} = (1 - 2\lambda)u_i^{(n)} + \lambda u_{i-1}^{(n)} + \lambda u_{i+1}^{(n)},$$

avec  $\lambda = \frac{k}{h^2}$ . Ce schéma se met sous la forme  $u^{(n+1)} = AU^{(n)}$ , avec

$$A = \begin{pmatrix} 1 - 2\lambda & \lambda & 0 & \dots & 0 \\ \lambda & 1 - 2\lambda & \lambda & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \lambda & 1 - 2\lambda & \lambda \\ 0 & \dots & 0 & \lambda & 1 - 2\lambda \end{pmatrix}$$

**Définition 2.9 (Stabilité au sens des erreurs d'arrondi)** *Supposons que l'on commette une erreur  $\varepsilon^0$  sur la condition initiale. La nouvelle condition initiale  $\tilde{u}^0$ , s'écrit donc  $\tilde{u}^0 = u^0 + \varepsilon^0$ . A cette nouvelle condition initiale correspond une nouvelle solution calculée  $\tilde{u}^{(n)} = u^{(n)} + \varepsilon^{(n)}$ . On dit que le schéma est stable au sens des erreurs d'arrondi s'il existe  $C > 0$  indépendant de  $n$  tel que  $\varepsilon^{(n)} \leq C\varepsilon^{(0)}$ .*

On peut trouver une condition suffisante pour que le schéma 2.2 soit stable au sens des erreurs d'arrondi. En effet, on va démontrer le résultat suivant :

**Proposition 2.10** *On suppose que  $\lambda = \frac{k}{h^2} < \frac{1}{2}$ . Alors le schéma 2.2 est stable au sens des erreurs d'arrondi.*

**Démonstration :** Soit donc une condition initiale perturbée  $\tilde{u}^0 = u^0 + \varepsilon^0$  à laquelle on associe une nouvelle solution calculée  $\tilde{u}^{(n)} = u^{(n)} + \varepsilon^{(n)}$ . On a  $\varepsilon^{(n)} = A^n \varepsilon^0$ . Comme  $A$  est symétrique,  $A$  est diagonalisable dans  $\mathbb{R}$ . Soient  $\mu_1, \dots, \mu_N$  les valeurs propres de  $A$ , et  $e_1, \dots, e_N$  les vecteurs propres associés, c'est-à-dire tels que  $Ae_i = \mu_i e_i, \forall i = 1, \dots, N$ . On décompose la perturbation  $\varepsilon^0$  sur la base des vecteurs propres :

$$\varepsilon^0 = \sum_{i=1}^N a_i e_i. \text{ On a donc } A^n \varepsilon^0 = \sum_{i=1}^N a_i \mu_i^n e_i = \varepsilon^{(n)}.$$

Si on prend par exemple :  $\varepsilon^0 = a_i e_i$ , on obtient  $\varepsilon^{(n)} = a_i \mu_i^n e_i$ . Il y a diminution de l'erreur d'arrondi sur  $\varepsilon^0$  si

$$\sup_{i=1 \dots N} |\mu_i| \leq 1$$

c'est-à-dire si  $\rho(A) \leq 1$ , où  $\rho(A)$  désigne le rayon spectral de  $A$ . Calculons  $\rho(A)$ . On écrit :  $A = I + \lambda B$  où  $B$  est la matrice symétrique définie négative, définie par :

$$B = \begin{pmatrix} -2 & 1 & 0 & \dots & 0 \\ 1 & -2 & 1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & 1 & -2 & 1 \\ 0 & \dots & 0 & 1 & -2 \end{pmatrix} \quad (2.8)$$

Soit  $\mathcal{VP}(A)$  l'ensemble de valeurs propres de  $A$ . Alors  $\mathcal{VP}(A) = \{1 + \lambda \mu, \mu \in \mathcal{VP}(B)\}$ . Or  $\mathcal{VP}(B) = \{-4 \sin^2 \frac{j\pi}{2(N+1)}, j = 1, \dots, N\}$  (voir Lemme 2.11 plus loin). Pour que  $\varepsilon^{(n)} < \varepsilon^0$ , il faut donc que :

$$\sup_{j=1, \dots, N} \left| 1 - 4\lambda \sin^2 \frac{j\pi}{2(N+1)} \right| < 1,$$

c.à.d.

$$\lambda \sin^2 \frac{j\pi}{2(N+1)} < \frac{1}{2}.$$

Une condition suffisante pour avoir une diminution de l'erreur est donc que  $\lambda < \frac{1}{2}$ . ■

**Lemme 2.11 (Valeurs propres de  $B$ )** L'ensemble  $\mathcal{VP}(B)$  des valeurs propres de la matrice  $B$  définie par (2.8) est donné par :

$$\mathcal{VP}(B) = \left\{ -4 \sin^2 \frac{j\pi}{2(N+1)}, j = 1, \dots, N \right\}.$$

**Démonstration :** Les valeurs propres  $B$  peuvent se calculer à partir des valeurs propres de l'opérateur continu ; on commence donc par chercher  $u$  solution de :

$$\begin{cases} -u'' + \alpha u = 0, \\ u(0) = u(1) = 0. \end{cases}$$

Cherchons  $u(x)$  sous la forme :

$$u(x) = a \cos \sqrt{\alpha} x + b \sin \sqrt{\alpha} x$$

Comme  $u(0) = 0$ , on a :  $a = 0$ . De même,  $u(1) = B \sin \sqrt{\alpha} = 0$ , et donc  $\sqrt{\alpha} = k\pi$ . Les valeurs propres et vecteurs propres associés de l'opérateur continu sont donc :  $(k^2 \pi^2, \sin k\pi x) \quad k \in \mathbb{N}^*$ . Pour  $k = 1, \dots, N$ , soit  $v^{(k)} \in \mathbb{R}^N$  tel que  $v_i^{(k)} = \sin k\pi i h$ . Calculons  $Bv^{(k)}$  :

$$(Bv^{(k)})_i = v_{i-1}^{(k)} - 2v_i^{(k)} + v_{i+1}^{(k)}$$

et donc

$$(Bv^{(k)})_i = \sin k\pi(i-1)h - 2 \sin k\pi i h + \sin k\pi(i+1)h$$

En développant, on obtient :

$$(Bv^{(k)})_i = \sin k\pi i h \cos(-k\pi h) + \cos k\pi i h \sin(-k\pi h) - 2 \sin k\pi i h + \sin k\pi i h \cos k\pi h + \cos k\pi i h \sin k\pi h.$$

Après simplifications, il vient :

$$(Bv^{(k)})_i = 2 \sin k\pi i h (-1 + \cos k\pi h).$$

Or,  $\cos k\pi h = 1 - 2 \sin^2 \frac{k\pi h}{2}$ . On a donc :

$$\begin{aligned} (Bv^{(k)})_i &= 2 \sin k\pi i h \times \left(-2 \sin^2 \frac{k\pi h}{2}\right) \\ &= -4 \sin^2 \frac{k\pi h}{2} (v^{(k)})_i, \quad \forall k = 1 \dots N. \end{aligned}$$

On a  $h = \frac{1}{N+1}$ , et donc les valeurs propres de  $B$  s'écrivent  $\mu_k = -4 \sin^2 \frac{k\pi}{2(N+1)}$ ,  $k = 1, \dots, N$ . ■

### 2.2.6 Stabilité au sens de Von Neumann

L'analyse de stabilité au sens de Von Neumann<sup>1</sup> consiste à étudier l'impact du schéma sur un mode de Fourier isolé. Pour que le mode de Fourier en question soit solution du problème continu, on remplace les conditions de Dirichlet homogènes du problème (2.1) par des conditions périodiques, et pour alléger les notations, on considère l'intervalle  $]0, 2\pi[$  comme intervalle d'étude en espace plutôt que l'intervalle  $]0, 1[$ .

**Problème continu avec conditions aux limites périodiques** On considère le problème avec conditions aux limites périodiques

$$\begin{cases} u_t - u_{(xx)} = 0, & t \in ]0, T[, x \in ]0, 2\pi[, \\ u(0, t) = u(2\pi, t), & \forall t \in ]0, T[, \\ u(x, 0) = u_0(x). \end{cases} \quad (2.9)$$

Le problème (2.9) est bien posé, au sens où  $\forall u_0 \in C([0, 2\pi])$ , il existe une unique  $u \in C^2(]0, 2\pi[ \times ]0, T[, \mathbb{R})$  solution de (2.9). On suppose que  $u_0 \in L^2(]0, 2\pi[)$ . On rappelle que  $L^2$  est un espace de Hilbert, et que  $\{e^{inx}, n \in \mathbb{Z}\}$  est une base hilbertienne<sup>2</sup> de  $L^2(]0, 2\pi[)$ . On décompose donc la condition initiale dans cette base hilbertienne :  $u_0(x) = \sum_{n \in \mathbb{Z}} c_n(0) e^{inx}$  (au sens de la convergence dans  $L^2$ ). Dans un premier temps, calculons formellement les solutions de (2.9) sous la forme d'un développement dans la base hilbertienne :

$$u(x, t) = \sum_{n \in \mathbb{Z}} c_n(t) e^{inx}.$$

En supposant qu'on ait le droit de dériver terme à terme, on a donc :

$$u_t(x, t) = \sum_{n \in \mathbb{Z}} c'_n(t) e^{inx} \text{ et } u_{xx}(x, t) = \sum_{n \in \mathbb{Z}} -c_n(t) n^2 e^{inx}.$$

On obtient, en remplaçant dans l'équation

$$c'_n(t) = -n^2 c_n(t)$$

c'est-à-dire  $c_n(t) = c_n(0) e^{-n^2 t}$  en tenant compte de la condition initiale. On a donc finalement :

$$u(x, t) = \sum_{n \in \mathbb{Z}} c_n(0) e^{-n^2 t} e^{inx}. \quad (2.10)$$

Justifions maintenant ce calcul formel. On a :

$$\sum_{n \in \mathbb{Z}} |c_n(0)|^2 = \|u^0\|_{L^2}^2 < +\infty$$

De plus, en dérivant (2.10) terme à terme, on obtient :

$$u_t - u_{xx} = 0.$$

La condition de périodicité est bien vérifiée par  $u$  donnée par (2.10). Enfin on a bien :  $u(x, t) \rightarrow u_0(t)$  lorsque  $t \rightarrow 0$ , donc la condition initiale est vérifiée. On peut remarquer qu'il y a "amortissement" des coefficients de Fourier  $c_n(0)$  lorsque  $t$  augmente, c.à.d. qu'on a :  $c_n(t) \leq c_n(0)$ ,  $\forall t > 0$ .

<sup>1</sup>John von Neumann (1903-1957), mathématicien et physicien américain d'origine hongroise, a apporté d'importantes contributions tant en mécanique quantique, qu'en analyse fonctionnelle, en théorie des ensembles, en informatique, en sciences économiques ainsi que dans beaucoup d'autres domaines des mathématiques et de la physique. Il a de plus participé aux programmes militaires américains.

<sup>2</sup>Soit  $H$  un espace de Hilbert,  $(e_i)_{i \in \mathbb{Z}}$  est une base hilbertienne de  $H$  si :  $(e_i)_{i \in \mathbb{Z}}$  est une famille orthonormée telle que  $\forall x \in H, \exists (x_i)_{i \in \mathbb{Z}} \subset \mathbb{R}; x = \sum_{i \in \mathbb{Z}} x_i e_i$  au sens de la convergence dans  $H$ , avec  $x_i = (x, e_i)$ , où  $(\cdot, \cdot)$  désigne le produit scalaire sur  $H$ .

**Discrétisation du problème (2.9)** Si on utilise le schéma (2.2), pour la discrétisation de (2.9) on a :

$$u_j^{(n+1)} = (1 - 2\lambda)u_j^{(n)} + \lambda u_{j-1}^{(n)} + \lambda u_{j+1}^{(n)}. \quad (2.11)$$

On prend comme condition initiale  $u^0(x) = a_p e^{ipx}$ , pour  $p \in \mathbb{Z}$  fixé. En discrétisant, on obtient :  $u_j^0(x) = a_p e^{ipjh}$ , pour  $j = 1, \dots, N$ , avec  $h = \frac{2\pi}{N+1}$ . On a bien  $u_0^0 = u_{N+1}^0 = 0$ . Calculons :

$$u_j^{(1)} = (1 - 2\lambda)a_p e^{ipjh} + \lambda a_p e^{ip(j-1)h} + \lambda a_p e^{ip(j+1)h}$$

donc :  $u_j^{(1)} = a_p e^{ipjh} \xi_p$ . On appelle  $\xi_p$  le facteur d'amplification associé à la fonction  $e^{ipx}$  (appelé aussi “ $p$ -ième mode”). On a donc :

$$\begin{cases} u_j^{(1)} = \xi_p u_j^{(0)} \\ \vdots \\ u_j^{(n)} = (\xi_p)^n u_j^{(0)} \end{cases}$$

On dit que le schéma est “stable au sens de Von Neumann” : si :

$$|\xi_p| < 1, \quad \forall p.$$

Calculons  $\xi_p$  :

$$\begin{aligned} \xi_p &= 1 - 2\lambda + 2\lambda \cos ph \\ &= 1 - 2\lambda + 2\lambda(1 - 2\sin^2 \frac{ph}{2}) \\ &= 1 - 4\lambda \sin^2 \left( \frac{2\pi}{N+1}, \frac{p}{2} \right). \end{aligned}$$

Pour avoir  $|\xi_p| < 1$ , il faut  $\lambda \sin^2 \left( \frac{2\pi}{N+1}, \frac{p}{2} \right) < \frac{1}{4}$ . Une condition suffisante pour que le schéma soit stable au sens de Von Neumann est que :  $\lambda < \frac{1}{2}$ . Remarquons que c'est la même condition que pour la stabilité des erreurs d'arrondis.

**Convergence du schéma avec la technique de Von Neumann** Soit  $u \in C^2([0, 2\pi[ \times ]0, T[, \mathbb{R})$  la solution exacte de (2.9) on a  $u(jh, nk) = \sum_{p \in \mathbb{Z}} c_p(0) e^{-p^2 nk} e^{ipjh}$  où  $h = \frac{2\pi}{N+1}$  est le pas de discrétisation en espace et

$k = \frac{T}{M}$  le pas de discrétisation en temps. Soit  $u_{\mathcal{D}}$  la solution de (2.2), et :

$$u_{\mathcal{D}}(jh, nk) = \sum_{p \in \mathbb{Z}} c_p(0) \xi_p^{(n)} e^{ipjh}.$$

On cherche à montrer la convergence de  $u_{\mathcal{D}}$  vers  $u$  au sens suivant :

**Proposition 2.12** Soit  $u_0 = \sum_{n \in \mathbb{Z}} c_n(0) e^{inx}$  et  $u$  la solution du problème (2.9). On note  $u_{\mathcal{D}}$  la solution approchée obtenue par le schéma d'Euler explicite (2.11). Alors  $\forall \varepsilon > 0, \exists \eta \geq 0$  tel que si  $k \leq \eta$  et  $\frac{k}{h^2} \leq \frac{1}{2}$ , alors

$$|u(jh, nk) - u_{\mathcal{D}}(jh, nk)| \leq \varepsilon, \forall j = 1 \dots N, n = \frac{T}{k}.$$

**Démonstration :** On note  $(u - u_{\mathcal{D}})_j^{(n)}$  la quantité  $u(jh, nk) - u_{\mathcal{D}}(jh, nk)$ . On fera l'hypothèse supplémentaire :

$$\sum_{p \in \mathbb{Z}} |c_p(0)| < +\infty.$$

Donc pour tout  $\varepsilon \in \mathbb{R}_+$ , il existe  $A \in \mathbb{R}$  tel que  $2 \sum_{|p| \geq A} |c_p(0)| \leq \varepsilon$ . On écrit alors :

$$(u - u_{\mathcal{D}})_j^{(n)} \leq \sum_{|p| \leq A} c_p(0) (e^{-p^2 nk} - \xi_p^n) e^{ipjh} + \sum_{|p| \geq A} c_p(0) (e^{-p^2 nk} - \xi_p^n) e^{ipjh}$$

On a donc :

$$(u - u_{\mathcal{D}})_j^{(n)} \leq X + 2 \sum_{|p| \geq A} |c_p(0)|, \text{ avec } X = \sum_{|p| \leq A} |c_p(0)| (e^{-p^2 nk} - \xi_p^n)$$

et  $2 \sum_{|p| \geq A} |c_p(0)| \leq 2\varepsilon$ . Montrons maintenant que  $X \rightarrow 0$  lorsque  $h \rightarrow 0$ . Remarquons que

$$e^{-p^2 nk} - \xi_p^n = e^{-p^2 T} - \xi_p^n, \text{ et } \xi_p = 1 - 4\lambda \sin^2 \frac{ph}{2}.$$

Or,  $\sin^2 \frac{ph}{2} = \frac{p^2 h^2}{4} + O(h^4)$ , et  $\lambda = \frac{k}{h^2}$ . Donc :  $4\lambda \sin^2 \frac{ph}{2} = p^2 k + O(kh^2)$ . On en déduit :

$$(\xi_p)^n = \left(1 - 4\lambda \sin^2 \frac{ph}{2}\right)^{T/k} \text{ et donc } \ln \xi_p^n = \frac{T}{k} \ln \left(1 - 4\lambda \sin^2 \frac{ph}{2}\right) = -Tp^2 + O(h^2).$$

On en déduit que  $\xi_p^n \rightarrow e^{-p^2 T}$  lorsque  $h \rightarrow 0$ . Tous les termes de  $X$  tendent vers 0, et  $X$  est une somme finie ; on a donc montré que  $(u - u_{\mathcal{D}})_j^{(n)}$  tend vers 0 lorsque  $h$  tend vers 0. ■

**Remarque 2.13** On peut adapter la technique de Von Neumann au cas Dirichlet homogène sur  $[0, 1]$ , en effectuant le développement de  $u$  par rapport aux fonctions propres de l'opérateur  $u''$  avec conditions aux limites de Dirichlet :

$$u(x, t) = \sum c_n(t) \sin(n\pi x).$$

L'avantage du développement en série de Fourier est qu'il marche pour  $n$  importe quel opérateur linéaire à condition d'avoir pris des conditions aux limites périodiques.

## 2.3 Schéma implicite et schéma de Crank-Nicolson

### 2.3.1 Le $\theta$ -schéma

Commençons par un petit rappel sur les équations différentielles (voir aussi polycopié d'analyse numérique de licence, sur le site web <http://www.cmi.univ-mrs.fr/herbin>) On considère le problème de Cauchy :

$$\begin{cases} y'(t) = f(y(t)), t > 0. \\ y(0) = y_0 \end{cases} \quad (2.12)$$

Soit  $k$  un pas (constant) de discrétisation, on rappelle que les schémas d'Euler explicite et implicite pour la discrétisation de ce problème s'écrivent respectivement :

$$\text{Euler explicite : } \frac{y^{(n+1)} - y^{(n)}}{k} = f(y^{(n)}), n \geq 0 \quad (2.13)$$

$$\text{Euler implicite : } \frac{y^{(n+1)} - y^{(n)}}{k} = f(y^{(n+1)}), n \geq 0, \quad (2.14)$$

avec  $y^{(n)} = y_0$ . On rappelle également que le  $\theta$ -schéma, où  $\theta$  est un paramètre de l'intervalle  $[0, 1]$  s'écrit :

$$y^{(n+1)} = y^{(n)} + k\theta f(y^{(n+1)}) + k(1 - \theta)f(y^{(n)}). \quad (2.15)$$

Remarquons que pour  $\theta = 0$  on retrouve le schéma (2.13) et pour  $\theta = 1$  le schéma (2.14). On peut facilement adapter le  $\theta$  schéma à la résolution des équations paraboliques. Par exemple, le  $\theta$ -schéma pour la discrétisation en temps du problème (2.1), avec une discrétisation par différences finies en espace s'écrit :

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} = \frac{\theta}{h^2} (-2u_i^{(n+1)} + u_{i-1}^{(n+1)} + u_{i+1}^{(n+1)}) + \frac{1-\theta}{h^2} (-2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)}), ; n \geq 0, i = 1 \quad (2.16)$$

$$u_i^{(0)} = u_0(x_i), i = 1, \dots, N.$$

Si  $\theta = 0$ , on retrouve le schéma d'Euler explicite ; si  $\theta = 1$ , celui d'Euler implicite. Dans ce cas où  $\theta = \frac{1}{2}$  ce schéma s'appelle schéma de Crank<sup>3</sup>-Nicolson<sup>4</sup>. Notons que dès que  $\theta > 0$ , le schéma est implicite, au sens où on n'a pas d'expression explicite de  $u_i^{(n+1)}$  en fonction des  $u_j^{(n)}$ .

<sup>3</sup>John Crank (6 February 1916 –3 October 2006) mathématicien britannique.

<sup>4</sup>Phyllis Nicolson (21 Septembre 1917 – 6 Octobre 1968) mathématicienne britannique.

### 2.3.2 Consistance et stabilité

**Proposition 2.14 (Consistance du  $\theta$ -schéma)** *Le  $\theta$  schéma (2.16) pour la discrétisation du problème (2.1) est d'ordre 2 en espace. Il est d'ordre 2 en temps si  $\theta = \frac{1}{2}$ , et d'ordre 1 sinon.*

**Démonstration :** On pose  $\bar{u}_j^n = u(x_j, t_n)$ ,  $h = \frac{1}{N+1}$ ,

$$R_j^{(n)} = \frac{\bar{u}_j^{(n+1)} - \bar{u}_j^{(n)}}{k} + \frac{\theta}{h^2} \left( -2\bar{u}_i^{(n+1)} + \bar{u}_{i-1}^{(n+1)} + \bar{u}_{i+1}^{(n+1)} \right) + \frac{1-\theta}{h^2} \left( -2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)} \right)$$

On va montrer, en effectuant des développements limités, que :  $|R_j^{(n)}| \leq C(k + h^2)$  si  $\theta \neq \frac{1}{2}$  et que  $|R_j^{(n)}| \leq C(k^2 + h^2)$  si  $\theta = \frac{1}{2}$ . En effet, on décompose

$$R_j^{(n)} = T_j^{(n,1)} + \theta T_j^{(n,2)} + (1-\theta)T_j^{(n,3)}$$

avec :

$$\begin{aligned} T_j^{(n,1)} &= \frac{\bar{u}_j^{(n+1)} - \bar{u}_j^{(n)}}{k}, \quad T_j^{(n,2)} = \frac{\theta}{h^2} \left( -2\bar{u}_i^{(n+1)} + \bar{u}_{i-1}^{(n+1)} + \bar{u}_{i+1}^{(n+1)} \right) \\ T_j^{(n,3)} &= \frac{1-\theta}{h^2} \left( -2u_i^{(n)} + u_{i-1}^{(n)} + u_{i+1}^{(n)} \right) \end{aligned}$$

Effectuons un développement limité pour calculer  $T_j^{(n,1)}$  :

$$T_j^{(n,1)} = (\bar{u}_t)(x_j, t_n) + \frac{k}{2}(u_{tt})(x_j, t_n) + R_1 \quad \text{avec } |R_1| \leq Ck^2.$$

Faisons de même pour  $T_j^{(n,2)}$  :

$$T_j^{(n,2)} = \theta(\bar{u}_{xx}(x_j, t_{n+1}) + R_2) \quad \text{avec } |R_2| \leq Ch^2.$$

Or  $\bar{u}_{xx}(x_j, t_{n+1}) = \bar{u}_{xx}(x_j, t_n) + k\bar{u}_{xxt}(x_j, t_n) + R_3$  avec  $|R_3| \leq Ck^2$ , donc :

$$T_j^{(n,2)} = \theta(u_{xx}(x_j, t_n) + k u_{xxt}(x_j, t_n) + R_4) \quad \text{avec } |R_4| \leq C(h^2 + k^2).$$

De même pour  $T_j^{(n,3)}$ , on a :

$$T_j^{(n,3)} = (1-\theta)u_{xx}(x_j, t_n) + R_5, \quad \text{avec } |R_5| \leq Ck^2.$$

En regroupant, on obtient que

$$R_j^{(n)} = u_t(x_j, t_n) - u_{xx}(x_j, t_n) \frac{k}{2} \frac{\partial}{\partial t} u_t(x_j, t_n) + \theta k(u_{xx})(x_j, t_n) + R$$

avec  $R = R_1 + R_4 + R_5$

• Si  $\theta = \frac{1}{2}$ , on a un schéma d'ordre 2 en temps et en espace.

En effet,  $\frac{k}{2}(\bar{u}_{tt})(x_j, t_n) - \theta k(\bar{u}_{xxt})(x_j, t_n) = \frac{\partial}{\partial t} \left( k \left[ \frac{1}{2}(\bar{u}_t)(x_j, t_n) - \theta(\bar{u}_{xx})(x_j, t_n) \right] \right)$  et  $u_t - u_{xx} = 0$ .

• Si  $\theta \neq \frac{1}{2}$  : on a un schéma d'ordre 2 en espace et d'ordre 1 en temps. ■

**Proposition 2.15 (Stabilité au sens de Von Neumann)** *Si  $\theta \geq \frac{1}{2}$  le  $\theta$ -schéma est inconditionnellement stable.*

*En particulier, les schémas d'Euler implicite et de Crank-Nicolson sont inconditionnellement stables. Si  $\theta < \frac{1}{2}$  le schéma est stable sous condition.*

$$\lambda \leq \frac{1}{2(1-2\theta)}.$$

(On retrouve en particulier que le schéma d'Euler explicite n'est que si  $\lambda \leq \frac{1}{2}$ ).

**Démonstration :** On remplace les conditions aux limites de Dirichlet sur  $[0, 1]$  par des conditions périodiques sur  $[0, 2\pi]$ . La solution exacte s'écrit alors :

$$u = \sum_{p \in \mathbb{Z}} c_p(0) e^{-p^2 t} e^{ipx}.$$

Prenons comme condition initiale  $u_0(x) = e^{ipx}$ . On a :

$$u_j^{(n+1)} - u_j^{(n)} = \frac{k}{h^2} \left[ -\theta(2u_j^{(n+1)} - u_{j-1}^{(n+1)} - u_{j+1}^{(n+1)}) - (1-\theta)(2u_j^{(n)} - u_{j-1}^{(n)} - u_{j+1}^{(n)}) \right],$$

ce qui s'écrit encore, avec :  $\lambda = \frac{k}{h^2}$  :

$$(1 + 2\lambda)u_j^{(n+1)} - \lambda\theta u_{j-1}^{(n+1)} - \lambda\theta u_{j+1}^{(n+1)} = (1 - 2\lambda(1-\theta))u_j^{(n)} + \lambda(1-\theta)u_{j-1}^{(n)} + \lambda(1-\theta)u_{j+1}^{(n)}. \quad (2.17)$$

En discrétisant la condition initiale (mode de Fourier) on obtient  $u_j^{(0)} = e^{ipjh}$  et on cherche le facteur d'amplification  $\xi_p$  tel que  $u_j^1 = \xi_p u_j^0 = \xi_p e^{ipjh}$  ; en appliquant le schéma ci-dessus pour  $n = 0$ , on obtient :

$$(1 + 2\lambda\theta)\xi_p - \lambda\theta\xi_p[e^{-iph} + e^{iph}] = [1 - 2\lambda(1-\theta)] + \lambda(1-\theta)[e^{iph} + e^{-iph}]$$

et donc :

$$\xi_p = \frac{1 - 2\lambda(1-\theta) + 2\lambda(1-\theta)\cos ph}{(1 + 2\lambda\theta) - 2\lambda\cos ph} = \frac{1 - 4\lambda(1-\theta)\sin^2 ph/2}{1 + 4\lambda\theta\sin^2 ph/2}$$

Pour que le schéma soit stable au sens de Von Neumann, il faut que :  $|\xi_p| < 1$  pour tout  $p$ , soit encore :

$$1 - 4\lambda(1-\theta)\sin^2 \frac{ph}{2} < 1 + 4\lambda\theta\sin^2 \frac{ph}{2} \quad (2.18)$$

et

$$4\lambda(1-\theta)\sin^2 \frac{ph}{2} - 1 < 1 + 4\lambda\theta\sin^2 \frac{ph}{2} \quad (2.19)$$

L'inégalité (2.18) est toujours vérifiée. En ce qui concerne l'inégalité (2.19), on distingue deux cas :

1. Si  $\theta \leq \frac{1}{2}$  alors  $0 \leq 1 - \theta \leq \theta$  et dans ce cas (2.19) est toujours vraie.
2. Si  $\theta < \frac{1}{2}$ , on veut :

$$4\lambda \left[ (1-\theta)\sin^2 \frac{ph}{2} - \theta\sin^2 \frac{ph}{2} \right] < 2$$

Il faut donc que

$$\lambda < \frac{1}{2} \left\{ (1-2\theta)\sin^2 \frac{ph}{2} \right\}^{-1}$$

Une condition suffisante est donc :

$$\lambda \leq \frac{1}{2(1-2\theta)} \text{ si } \theta < \frac{1}{2}. \quad \blacksquare$$

### 2.3.3 Convergence du schéma d'Euler implicite.

Prenons  $\theta = 1$  dans le  $\theta$ -schéma : on obtient le schéma d'Euler implicite :

$$(1 + 2\lambda)u_j^{(n+1)} - \lambda u_{j-1}^{(n+1)} - \lambda u_{j+1}^{(n+1)} = u_j^{(n)} \quad (2.20)$$

On rappelle que ce schéma est inconditionnellement stable au sens de Von Neumann. On va montrer de plus qu'il est  $L^\infty$ -stable :

**Proposition 2.16 (Stabilité  $L^\infty$  pour Euler implicite)** Si  $(u_j^{(n)})_{j=1,\dots,N}$  est solution du schéma (2.20), alors :

$$\max_{j=1,\dots,N} u_j^{(n+1)} \leq \max_{j=1,\dots,N} u_j^{(n)} \leq \max_{j=1,\dots,N} u_j^{(0)} \quad (2.21)$$

de même :

$$\min_{j=1,\dots,N} u_j^{(n+1)} \geq \min_{j=1,\dots,N} u_j^{(n)} \geq \min_{j=1,\dots,N} u_j^{(0)} \quad (2.22)$$

Le schéma (2.20) est donc  $L^\infty$  stable.

**Démonstration :** Prouvons l'estimation (2.21), la preuve de (2.22) est similaire. Soit  $j_0$  tel que  $u_{j_0}^{(n+1)} = \max_{j=1, \dots, N} u_j^{(n+1)}$ . Par définition du schéma d'Euler implicite (2.20), On a :

$$u_{j_0}^{(n)} = (1 + 2\lambda)u_{j_0}^{(n+1)} - \lambda u_{j_0-1}^{(n+1)} - \lambda u_{j_0+1}^{(n+1)}.$$

On en déduit :  $u_{j_0}^{(n+1)} \leq \max_{j=1, \dots, N} u_j^{(n)}$ , ce qui prouve que

$$\max_{j=1, \dots, N} u_j^{(n+1)} \leq \max_{j=1, \dots, N} u_j^{(n)}.$$

Donc le schéma (2.20) est  $L^\infty$  stable. ■

**Théorème 2.17** Soit  $e^{(n)}$  l'erreur de discrétisation, définie par

$$e_j^{(n)} = u(x_j, t_n) - u_j^{(n)} \text{ pour } j = 1, \dots, N.$$

Alors  $\|e^{(n+1)}\|_\infty \leq \|e^{(0)}\|_\infty + TC(k + h^2)$ . Si  $\|e^{(0)}\|_\infty = 0$ , le schéma est donc convergent (d'ordre 1 en temps et 2 en espace).

**Démonstration :** En utilisant la définition de l'erreur de consistance, on obtient :

$$(1 + 2\lambda)e_j^{(n+1)} - \lambda e_{j-1}^{(n)} - \lambda e_{j+1}^{(n)} = e_j^{(n)} + R_j^{(n)}$$

et donc :

$$\|e^{(n+1)}\|_\infty \leq \|e^h\|_\infty + kC(k + h^2)$$

On en déduit, par récurrence sur  $n$ , que :

$$\|e^{(n+1)}\|_\infty \leq \|e^0\|_\infty + TC(k + h^2)$$

d'où la convergence du schéma. ■

On peut montrer que le schéma saute-mouton (ou "Leap-frog")

$$\frac{u_j^{(n+1)} - u_j^{(n-1)}}{2k} = \frac{1}{h^2}(u_{j-1}^{(n)} - 2u_j^{(n)} + u_{j+1}^{(n)})$$

est d'ordre 2 en espace et en temps (voir exercice 30 page 74. Malheureusement il est aussi inconditionnellement instable. On peut le modifier pour le rendre stable, en introduisant le schéma de Dufort-Frankel, qui s'écrit :

$$\frac{u_j^{(n+1)} - u_j^{(n-1)}}{2k} = \frac{1}{h^2}(u_{j-1}^{(n)} - (u_j^{(n+1)} + u_j^{(n-1)}) + u_{j+1}^{(n)}).$$

Ce schéma est consistant et inconditionnellement stable (voir exercice 30 page 74).

## 2.4 Cas de la Dimension 2

Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^2$ , on considère le problème suivant :

$$\begin{cases} u_t - \Delta u = 0 & x \in \Omega, t \in ]0, T[ \\ u(x, 0) = u_0(x) & x \in \Omega \\ u(x, t) = g(t) & x \in \partial\Omega \quad \forall t \in ]0, T[ \end{cases}$$

Si le domaine est rectangulaire, ce problème se discrétise facilement à l'aide de  $\theta$  schéma en temps et de différences finies en espace, en prenant un maillage rectangulaire. On peut montrer, comme dans le cas 1D, la consistance, la stabilité, la  $L^\infty$  stabilité, la stabilité au sens de Von Neumann



## 2.5 Exercices

**Exercice 22 (Existence de solutions “presque classiques”)** *Corrigé en page 80*

Soit  $u_0 \in L^2(]0, 1[)$ . On s'intéresse au problème :

$$\begin{aligned} u_t(x, t) - u_{xx}(x, t) &= 0, \quad x \in ]0, 1[, \quad t \in \mathbb{R}_+^*, \\ u(0, t) = u(1, t) &= 0, \quad t \in \mathbb{R}_+^*, \\ u(x, 0) &= u_0(x), \quad x \in ]0, 1[. \end{aligned} \quad (2.23)$$

1. On définit  $u : [0, 1] \times \mathbb{R}_+^* \rightarrow \mathbb{R}$  par :

$$u(x, t) = \sum_{n \in \mathbb{N}^*} e^{-n^2 \pi^2 t} a_n \sin(n\pi x), \quad x \in [0, 1], \quad t \in \mathbb{R}_+^*, \quad (2.24)$$

avec  $a_n = (\int_0^1 u_0(x) \sin(n\pi x) dx) / (\int_0^1 \sin^2(n\pi x) dx)$ .

Montrer que  $u$  est bien définie de  $[0, 1] \times \mathbb{R}_+^*$  dans  $\mathbb{R}$  et est solution de (2.23) au sens suivant :

$$\begin{aligned} u &\in C^\infty([0, 1] \times \mathbb{R}_+^*, \mathbb{R}), \\ u_t(x, t) - u_{xx}(x, t) &= 0, \quad \forall x \in [0, 1], \quad \forall t \in \mathbb{R}_+^*, \\ u(0, t) = u(1, t) &= 0, \quad \forall t \in \mathbb{R}_+^*, \\ \|u(\cdot, t) - u_0\|_{L^2(]0, 1[)} &\rightarrow 0, \quad \text{quand } t \rightarrow 0. \end{aligned} \quad (2.25)$$

2. Montrer qu'il existe une unique fonction  $u$  solution de (2.25).

**Exercice 23 (Discrétisation par DF)**

Soit  $u_0 \in C(]0, 1[)$ . On s'intéresse au problème :

$$\begin{aligned} u_t(x, t) + u_x(x, t) - u_{xx}(x, t) &= 0, \quad x \in ]0, 1[, \quad t \in ]0, T[, \\ u(0, t) &= a, \quad t \in \mathbb{R}_+^*, \\ u'(1, t) &= b \\ u(x, 0) &= u_0(x), \quad x \in ]0, 1[, \end{aligned} \quad (2.26)$$

avec  $T > 0$ ,  $a$  et  $b \in \mathbb{R}$  donnés.

Ecrire une discrétisation espace-temps du problème (2.26) avec le schéma d'Euler explicite en temps et par différences finies avec un maillage uniforme en espace, en utilisant un schéma décentré amont pour le terme d'ordre 1  $u_x(x, t)$ .

**Exercice 24 (Exemple de schéma non convergent)** *Suggestions en page 80, corrigé en page 82*

Soit  $u_0 \in L^2(]-4, 4[)$ . On note  $u$  l'unique solution (au sens vu en cours ou en un sens inspiré de l'exercice précédent) du problème suivant :

$$\begin{aligned} u_t(x, t) - u_{xx}(x, t) &= 0, \quad x \in ]-4, 4[, \quad t \in ]0, 1[, \\ u(-4, t) = u(4, t) &= 0, \quad t \in ]0, 1[, \\ u(x, 0) &= u_0(x), \quad x \in ]-4, 4[. \end{aligned} \quad (2.27)$$

On sait que la solution de (2.27) est de classe  $C^\infty$  sur  $[-4, 4] \times ]0, 1[$  (voir l'exercice précédent). On admettra que si  $u_0 \geq 0$  p.p. sur  $]-4, 4[$  et  $u_0 \neq 0$  (dans  $L^2(]-4, 4[)$ ) alors  $u(x, t) > 0$  pour tout  $x \in ]-4, 4[$  et tout  $t \in ]0, 1[$ .

On suppose maintenant que  $u_0 \in C([-4, 4], \mathbb{R})$ ,  $u_0(-4) = u_0(4) = 0$ ,  $u_0 \geq 0$  sur  $]-4, 4[$ ,  $u_0$  nulle sur  $[-3, 4]$  et qu'il existe  $a \in ]-4, -3[$  t.q.  $u_0(a) > 0$ . On a donc  $u(x, t) > 0$  pour tout  $x \in ]-4, 4[$ .

Avec les notations du cours, on considère la solution de (2.27) donnée par le schéma d'Euler explicite (2.2) avec le pas de temps  $k = 1/(M + 1)$  et le pas d'espace  $h = 8/(N + 1)$  ( $M, N \in \mathbb{N}^*$ ,  $N$  impair). La solution approchée est définie par les valeurs  $u_i^n$  pour  $i \in \{-(N + 1)/2, \dots, (N + 1)/2\}$  et  $n \in \{0, \dots, M + 1\}$ . La valeur  $u_i^n$  est censée être une valeur approchée de  $\bar{u}_i^n = u(ih, nk)$ .

1. Donner les équations permettant de calculer  $u_i^n$  pour  $i \in \{-(N + 1)/2, \dots, (N + 1)/2\}$  et  $n \in \{0, \dots, M + 1\}$ .

2. On suppose maintenant que  $k = h$ . Montrer que  $u_i^n = 0$  pour  $i \geq 0$  et  $n \in \{0, \dots, M+1\}$ . En déduire que  $\max\{|u_i^{M+1} - \bar{u}_i^{M+1}|, i \in \{-(N+1)/2, \dots, (N+1)/2\}\}$  ne tends pas vers 0 quand  $h \rightarrow 0$  (c'est-à-dire quand  $N \rightarrow \infty$ ).

### Exercice 25 (Schéma implicite et principe du maximum)

1. Soient  $T > 0, v \in C^1([0, 1], \mathbb{R}_+), a_0, a_1 \in \mathbb{R}$  et  $u_0 \in C([0, 1])$ . On considère le problème d'évolution suivant :

$$\begin{cases} u_t(x, t) - u_{xx}(x, t) + v(x)u_x(x, t) = 0, & x \in ]0, 1[, t \in ]0, T[, \\ u(0) = a_0, u(1) = a_1, \\ u(x, 0) = u_0(x). \end{cases} \quad (2.28)$$

dont on cherche à approcher la solution par différences finies. On choisit pour cela le schéma de la question 1 de l'exercice 5 pour la discrétisation en espace, et on discrétise par le schéma d'Euler implicite en temps avec un pas de temps uniforme  $k = \frac{T}{P}$  où  $P \geq 1$ .

- 1.1 Ecrire le schéma ainsi obtenu et montrer qu'il admet une solution qu'on notera  $U = (u_i^{(p)})_{\substack{i=1, \dots, N, \\ p=1, \dots, P}}$ , où  $u_i^{(p)}$  est censé être une approximation de  $u(x_i, t_p)$ , où  $t_p = pk, p = 0, \dots, P$ .

- 1.2 Montrer que

$$\min_{[0,1]}(\min u_0, \min(a_0, a_1)) \leq u_i^{(p)} \leq \max_{[0,1]}(\max u_0, \max(a_0, a_1)), \quad \text{pour tout } i = 1, \dots, N \text{ et } p = 1, \dots, P.$$

2. Soit  $T > 0$ , et  $u_0 \in C([0, 1])$ . On considère maintenant le problème d'évolution suivant :

$$\begin{cases} u_t(x, t) - u_{xx}(x, t) + (vu)_x(x, t) = 0, & x \in ]0, 1[, t \in ]0, T[, \\ u(0) = a_0, u(1) = a_1, \\ u(x, 0) = u_0(x). \end{cases} \quad (2.29)$$

dont on cherche à approcher la solution par différences finies. On choisit pour cela le schéma de la question 2 de l'exercice 6 pour la discrétisation en espace, et on discrétise par le schéma d'Euler implicite en temps avec un pas de temps uniforme  $k = \frac{T}{P}$  où  $P \geq 1$ .

- 2.1 Ecrire le schéma ainsi obtenu et montrer qu'il admet une solution qu'on notera  $U = (u_i^{(p)})_{\substack{i=1, \dots, N, \\ p=1, \dots, P}}$ , où  $u_i^{(p)}$  est censé être une approximation de  $u(x_i, t_p)$ , où  $t_p = pk, p = 0, \dots, P$ .

- 2.2 Montrer que si  $a_0 \geq 0, a_1 \geq 0$  et  $u_0 \geq 0$ , alors on a :  $u_i^{(p)} \geq 0$ , pour tout  $i = 1, \dots, N$  et  $p = 1, \dots, P$ .

3. On considère maintenant  $\Omega = ]0, 1]^2$ ; soient  $v \in C^\infty(\Omega, (\mathbb{R}_+)^2)$  ( $v(x)$  est donc un vecteur de  $\mathbb{R}^2$ ),  $a \in C(\partial\Omega, \mathbb{R})$  et  $u_0 \in C(\Omega, \mathbb{R}_+)$ . En s'inspirant des schémas étudiés aux questions 3 et 4, donner une discrétisation en espace et en temps des deux problèmes suivants (avec pas uniforme) :

$$\begin{cases} u_t - \Delta u + v \cdot \nabla u = 0, \\ u|_{\partial\Omega} = a, \\ u(\cdot, 0) = u_0. \end{cases} \quad (2.30)$$

$$\begin{cases} u_t - \Delta u + \operatorname{div}(vu) = 0, \\ u|_{\partial\Omega} = a, \\ u(\cdot, 0) = u_0. \end{cases} \quad (2.31)$$

### Exercice 26 (Schémas explicites centré et décentré) Corrigé en page 83

Soient  $\alpha > 0, \mu > 0, T > 0$  et  $u_0 : \mathbb{R} \rightarrow \mathbb{R}$ . On s'intéresse au problème suivant :

$$\begin{cases} u_t(x, t) + \alpha u_x(x, t) - \mu u_{xx}(x, t) = 0, & x \in ]0, 1[, t \in ]0, T[, \\ u(0, t) = u(1, t) = 0, & t \in ]0, T[, \\ u(x, 0) = u_0(x), & x \in ]0, 1[. \end{cases} \quad (2.32)$$

On rappelle que  $u_t = \frac{\partial u}{\partial t}$ ,  $u_x = \frac{\partial u}{\partial x}$  et  $u_{xx} = \frac{\partial^2 u}{\partial x^2}$ . On suppose qu'il existe  $u \in C^4([0, 1] \times [0, T])$  solution (classique) de (2.32) (noter que ceci implique  $u_0(0) = u_0(1) = 0$ ). On pose  $A = \min\{u_0(x), x \in [0, 1]\}$  et  $B = \max\{u_0(x), x \in [0, 1]\}$  (noter que  $A \leq 0 \leq B$ ).

On discrétise le problème (2.32). On reprend les notations du cours. Soient  $h = 1/(N+1)$  et  $k = T/M$  ( $N, M \in \mathbb{N}^*$ ).

1. Schéma explicite décentré. Pour approcher la solution  $u$  de (2.32), on considère le schéma suivant :

$$\begin{aligned} \frac{1}{k}(u_i^{n+1} - u_i^n) + \frac{\alpha}{h}(u_i^n - u_{i-1}^n) - \frac{\mu}{h^2}(u_{i+1}^n - 2u_i^n + u_{i-1}^n) &= 0, \\ i \in \{1, \dots, N\}, n \in \{0, \dots, M-1\}, \\ u_0^n &= u_{N+1}^n = 0, n \in \{1, \dots, M\}, \\ u_i^0 &= u_0(ih), i \in \{0, \dots, N+1\}. \end{aligned} \quad (2.33)$$

On pose  $\bar{u}_i^n = u(ih, nk)$  pour  $i \in \{0, \dots, N+1\}$  et  $n \in \{0, \dots, M\}$ .

- (Consistance) Montrer que l'erreur de consistance du schéma (2.33) est majorée par  $C_1(k+h)$ , où  $C_1$  ne dépend que de  $u, T, \alpha$  et  $\mu$ .
- (Stabilité) Sous quelle condition sur  $k$  et  $h$  (cette condition peut dépendre de  $\alpha$  et  $\mu$ ) a-t-on  $A \leq u_i^n \leq B$  pour tout  $i \in \{0, \dots, N+1\}$  et tout  $n \in \{0, \dots, M\}$ ? Sous cette condition, en déduire  $\|u^n\|_\infty \leq \|u_0\|_{L^\infty(]0,1])}$  pour tout  $n \in \{0, \dots, M\}$  (avec  $\|u^n\|_\infty = \max\{|u_i^n|, i \in \{0, \dots, N+1\}\}$ )
- (Estimation d'erreur) On pose  $e_i^n = \bar{u}_i^n - u_i^n$ .  
Sous la condition sur  $k$  et  $h$  trouvée précédemment, montrer que  $|e_i^n| \leq C_2(k+h)$  pour tout  $i \in \{0, \dots, N+1\}$  et tout  $n \in \{0, \dots, M\}$  avec  $C_2$  ne dépendant que de  $u, T, \alpha$  et  $\mu$ .

2. Schéma explicite centré.

On change dans le schéma (2.33) la quantité  $(\alpha/h)(u_i^n - u_{i-1}^n)$  par  $(\alpha/2h)(u_{i+1}^n - u_{i-1}^n)$ .

- (Consistance) Montrer que l'erreur de consistance est maintenant majorée par  $C_3(k+h^2)$ , où  $C_3$  ne dépend que de  $u, T, \alpha$  et  $\mu$ .
- Reprendre les questions de stabilité et d'estimation d'erreur du schéma (2.33).

### Exercice 27 (Discrétisation d'un problème parabolique.) Suggestions en page 80.

Dans cet exercice on s'intéresse à des schémas numériques pour le problème :

$$\begin{cases} u_t + u_x - \varepsilon u_{xx} = 0 & (x, t) \in \mathbb{R}^+ \times ]0, T[ \\ u(1, t) = u(0, t) = 0 & t \in ]0, T[ \\ u(x, 0) = u_0(x) & x \in ]0, 1[ \end{cases} \quad (2.34)$$

où  $u_0$  et  $\varepsilon$  sont donnés ( $\varepsilon > 0$ ). On reprend dans la suite de l'exercice les notations du cours.

1. Donner un schéma d'approximation de (2.34) différences finies à pas constant et centré en espace et Euler explicite à pas constant en temps. Montrer que l'erreur de consistance est majorée par  $C(k+h^2)$ ; avec  $C$  dépendant de la solution exacte de (2.34). Sous quelle(s) condition(s) sur  $k$  et  $h$  a-t'on  $\|u^n\|_\infty \leq \|u^0\|_\infty, \forall n \leq M$ ?

Donner un résultat de convergence pour ce schéma.

2. Même question que 1. en remplaçant Euler explicite par Euler implicite.

3. En s'inspirant du schéma de Crank-Nicolson (vu en cours) construire un schéma d'ordre 2 (espace et temps). Sous quelle(s) condition(s) sur  $k$  et  $h$  a-t'on  $\|u^n\|_2 \leq \|u^0\|_2, \forall n \leq M$ ? Donner un résultat de convergence pour ce schéma.

4. Dans les schémas trouvés aux questions 1., 2. et 3. on remplace l'approximation de  $u_x$  par une approximation décentrée à gauche. Quel est l'ordre des schémas obtenus et sous quelle(s) condition(s) sur  $k$  et  $h$  a-t'on  $\|u^n\|_\infty \leq \|u^0\|_\infty$  où  $\|u\|_2$ ,  $\forall n \leq M$ ? Donner un résultat de convergence pour ces schémas.

### Exercice 28 (Equation de diffusion réaction) Suggestions en page 80, corrigé 87

Soit  $u_0$  une fonction donnée de  $[0, 1]$  dans  $\mathbb{R}$ . On s'intéresse ici à la discrétisation du problème suivant :

$$u_t(t, x) - u_{xx}(t, x) - u(t, x) = 0, t \in \mathbb{R}^+, x \in [0, 1], \quad (2.35)$$

$$u(t, 0) = u(t, 1) = 0, t \in \mathbb{R}_+^*; u(0, x) = u_0(x), x \in [0, 1]. \quad (2.36)$$

On note  $u$  la solution de (2.35), (2.36), et on suppose que  $u$  est la restriction à  $\mathbb{R}_+ \times [0, 1]$  d'une fonction de classe  $C^\infty$  de  $\mathbb{R}^2$  dans  $\mathbb{R}$ .

Pour  $h = \frac{1}{N+1}$  ( $N \in \mathbb{N}^*$ ) et  $k > 0$ , on pose  $x_i = ih, i \in \{0, \dots, N+1\}, t_n = nk, n \in \mathbb{N}, \bar{u}_i^n = u(x_i, t_n)$ , et on note  $u_i^n$  la valeur approchée recherchée de  $\bar{u}_i^n$ .

On considère deux schémas numériques, (2.37)–(2.39) et (2.38)–(2.39) définis par les équations suivantes :

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{(u_{i+1}^{n+1} + u_{i-1}^{n+1} - 2u_i^{n+1})}{h^2} - u_i^{n+1} = 0, \quad n \in \mathbb{N}, i \in \{1, \dots, N\}, \quad (2.37)$$

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{(u_{i+1}^{n+1} + u_{i-1}^{n+1} - 2u_i^{n+1})}{h^2} - u_i^n = 0, \quad n \in \mathbb{N}, i \in \{1, \dots, N\}, \quad (2.38)$$

$$u_0^{n+1} = u_{N+1}^{n+1} = 0, \quad n \in \mathbb{N}; \quad u_i^0 = u_0(x_i), \quad i \in \{0, \dots, N+1\}. \quad (2.39)$$

Pour  $n \in \mathbb{N}$ , on note  $u^n = (u_1^n, \dots, u_N^n) \in \mathbb{R}^N$ .

- (Consistance) Soit  $T > 0$ . Pour  $n \in \mathbb{N}$ , et  $i \in \{1, \dots, N\}$ , on note  $R_i^n$  l'erreur de consistance (définie en cours) du schéma numérique (2.37), (2.39) [resp. du schéma numérique (2.38), (2.39)]. Montrer qu'il existe  $C \in \mathbb{R}$ , ne dépendant que de  $u$  et  $T$ , t. q.  $|R_i^n| \leq C(k + h^2)$ , pour tout  $i \in \{1, \dots, N\}$  et tout  $n \in \mathbb{N}$ , t. q.  $kn \leq T$ .
- Montrer que le schéma (2.37), (2.39) [resp. (2.38), (2.39)] demande, à chaque pas de temps, la résolution du système linéaire  $Au^{n+1} = a$  [resp.  $Bu^{n+1} = b$ ] avec  $A, B \in \mathbb{R}^{N,N}$  et  $a, b \in \mathbb{R}^N$  à déterminer.  
Montrer que  $B$  est inversible (et même s.d.p.) pour tout  $h > 0$  et  $k > 0$ . Montrer que  $A$  est inversible (et même s.d.p.) pour tout  $h > 0$  et  $k \in ]0, 1[$ .
- (Stabilité) Pour  $n \in \mathbb{N}$ , on pose  $\|u^n\|_\infty = \sup_{i \in \{1, \dots, N\}} |u_i^n|$ . Soit  $T > 0$ . On considère le schéma (2.38), (2.39). Montrer qu'il existe  $C_1(T) \in \mathbb{R}$ , ne dépendant que de  $T$ , t. q.  $\|u^n\|_\infty \leq C_1(T)\|u_0\|_\infty$ , pour tout  $h > 0, k > 0$ , et  $n \in \mathbb{N}$  tel que  $kn \leq T$ .  
Soit  $\alpha \in [0, 1]$ . On considère le schéma (2.37), (2.39). Montrer qu'il existe  $C_2(T, \alpha) \in \mathbb{R}$ , ne dépendant que de  $T$  et de  $\alpha$ , t. q.  $\|u^n\|_\infty \leq C_2(T, \alpha)\|u_0\|_\infty$ , pour tout  $h > 0, k \in ]0, \alpha[$ , et  $n \in \mathbb{N}$  tel que  $kn \leq T$ .
- (Estimation d'erreur) Pour  $n \in \mathbb{N}$  et  $i \in \{1, \dots, N\}$ , on pose  $e_i^n = \bar{u}_i^n - u_i^n$ . Soit  $T > 0$ . Donner, pour  $kn \leq T$ , des majorations de  $\|e^n\|_\infty$  en fonction de  $T, C, C_1(T), C_2(T, \alpha)$  (définis dans les questions précédentes),  $k$  et  $h$  pour les deux schémas étudiés.

### Exercice 29 (Discrétisation par VF)

Ecrire une discrétisation espace-temps du problème (2.26) de l'exercice 23 avec le schéma de Crank-Nicolson en temps, et par volumes finis avec un maillage uniforme en espace, en utilisant un schéma décentré amont pour le terme d'ordre 1  $u_x(x, t)$ .

### Exercice 30 (Schémas "saute-mouton" et Dufort-Frankel) Corrigé en page 89

On considère le problème suivant :

$$\begin{cases} u_t(x, t) - u_{xx}(x, t) &= 0, \quad x \in ]0, 1[, \quad t \in ]0, T[, \\ u(0, t) = u(1, t) &= 0, \quad t \in ]0, T[, \\ u(x, 0) &= u_0(x), \quad x \in ]0, 1[. \end{cases} \quad (2.40)$$

Pour trouver une solution approchée de ((2.40)), on considère le schéma "saute-mouton" :

$$\begin{cases} \frac{u_j^{n+1} - u_j^{(n-1)}}{2k} = \frac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{h^2}, \quad j = 1, \dots, N-1, \quad n = 1, \dots, M-1, \\ u_0^{n+1} = u_{N+1}^{n+1} = 0, \quad n = 1, \dots, M-1, \end{cases} \quad (2.41)$$

où  $(u_j^0)_{j=1, \dots, N}$  et  $(u_j^1)_{j=1, \dots, N}$  sont supposés connus,  $h = 1/N, k = T/M$ .

- Montrer que le schéma (2.41) est consistant. Quel est son ordre ?
- Montrer que le schéma (2.41) est inconditionnellement instable au sens de Von Neumann.

On modifie "légèrement" le schéma (2.41) en prenant

$$\begin{cases} \frac{u_j^{n+1} - u_j^{(n-1)}}{2k} = \frac{u_{j-1}^n - (u_j^{n+1} + u_j^{(n-1)}) + u_{j+1}^n}{h^2}, \quad j = 1, \dots, N, \quad n = 1, \dots, M-1, \\ u_0^{n+1} = u_{N+1}^{n+1} = 0, \quad n = 1, \dots, M-1, \end{cases} \quad (2.42)$$

(schéma de Dufort-Frankel).

3. Montrer que le schéma (2.42) est consistant avec (2.40) quand  $h, k \rightarrow 0$  sous la condition  $\frac{k}{h} \rightarrow 0$ .
4. Montrer que (2.42) est inconditionnellement stable.

**Exercice 31 (Schéma de Gear)**

On considère le problème suivant :

$$\begin{cases} u_t - u_{xx} = 0, \quad \forall x \in ]0, 1[, \quad \forall t \in ]0, T[ \\ u(x, 0) = u_0(x), \quad \forall x \in ]0, 1[, \\ u(0, t) = u(1, t) = 0, \quad \forall t \in ]0, T[ \end{cases} \quad (2.43)$$

On suppose que  $u_0 \in C(]0, 1[, \mathbb{R})$ . On rappelle que dans ce cas, il existe une unique fonction  $u \in C^2(]0, 1[ \times ]0, T[, \mathbb{R}) \cap C([0, 1] \times [0, T], \mathbb{R})$  qui vérifie (2.43). On cherche une approximation de la solution de ce problème, par une discrétisation par différences finies en espace et en temps. On se donne un ensemble de points  $\{t_n, n = 1, \dots, M\}$  de l'intervalle  $]0, T[$ , et un ensemble de points  $\{x_i, i = 1, \dots, N\}$ . Pour simplifier, on considère un pas constant en temps et en espace. Soit :  $h = \frac{1}{N+1}$  le pas de discrétisation en espace, et  $k = \frac{T}{M}$ , le pas de discrétisation en temps. On pose alors  $t_n = nk$  pour  $n = 0, \dots, M$  et  $x_i = ih$  pour  $i = 0, \dots, N+1$ . On cherche à calculer une solution approchée  $u_{\text{app}}$  du problème (2.43); plus précisément, on cherche à déterminer  $u_{\text{app}}(x_i, t_n)$  pour  $i = 1, \dots, N$ , et  $n = 1, \dots, M$ . Les inconnues discrètes sont notées  $u_i^{(n)}$ ,  $i = 1, \dots, N$  et  $n = 1, \dots, M$ .

On considère le schéma suivant :

$$\begin{cases} \frac{1}{2k} (3u_i^{(n+1)} - 4u_i^{(n)} + u_i^{(n-1)}) + \frac{1}{h^2} (2u_i^{(n+1)} - u_{i-1}^{(n+1)} - u_{i+1}^{(n+1)}) = 0, & \begin{cases} i = 1, \dots, N, \\ n = 1, \dots, M, \end{cases} \\ u_i^0 = u_0(x_i), \quad i = 1, \dots, N, \\ u_i^1 = u_1(x_i), \quad i = 1, \dots, N, \\ u_0^{(n)} = u_{N+1}^{(n)} = 0, \quad \forall n = 1, \dots, M, \end{cases} \quad (2.44)$$

où  $u_1(x_i) = u(x_i, k)$  est supposée connue.

1. Montrer que ce schéma est consistant d'ordre 2 en temps et en espace.
2. Montrer que le schéma s'écrit sous forme matricielle :

$$U^{n+1} = BW^n \quad (2.45)$$

$$\text{où } U^{n+1} = \begin{pmatrix} u_1^{n+1} \\ \vdots \\ u_N^{n+1} \end{pmatrix}, B = (3Id + \frac{2k}{h^2}A)^{-1}, A = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix}, \quad (2.46)$$

$$\text{et } W^n \text{ ne dépend que de } U^{n-1} = \begin{pmatrix} u_1^{n-1} \\ \vdots \\ u_N^{n-1} \end{pmatrix} \text{ et } U^n = \begin{pmatrix} u_1^n \\ \vdots \\ u_N^n \end{pmatrix}. \quad (2.47)$$

Donner l'expression de  $W^n$  en fonction de  $U^{n-1}$  et  $U^n$ .

3. En posant

$$V^n = \begin{pmatrix} U^n \\ U^{n-1} \end{pmatrix} \in \mathbb{R}^{2M},$$

mettre le schéma sous la forme

$$V^{n+1} = MV^n$$

(donner la matrice  $M$  en fonction de  $A$ ).

4. Montrer que  $\mu$  est valeur propre de  $M$  si et seulement si  $\mu^2 - 4\beta\mu + \beta = 0$  où  $\beta$  est une valeur propre de la matrice  $B$ .
5. Montrer que les valeurs propres de la matrice  $M$  sont toutes de module strictement inférieur à 1.
6. Montrer qu'il existe  $C \in \mathbb{R}$ , qui ne dépend pas de  $n$ , tel que  $|U^n|_2 \leq C$ , où  $|\cdot|_2$  désigne la norme euclidienne dans  $\mathbb{R}^N$ .

**Exercice 32 (Problème parabolique non linéaire)** *Corrigé en page 92*

On se propose, dans cet exercice, de montrer l'existence d'une solution faible au problème (2.48)-(2.50), à partir de l'existence de la solution approchée donnée par un schéma numérique. L'inconnue de ce problème est la fonction  $u$  de  $]0, 1[ \times ]0, T[$  dans  $\mathbb{R}$ , elle doit être solution des équations suivantes :

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 \varphi(u)}{\partial x^2}(x, t) = v(x, t), \quad x \in ]0, 1[, t \in ]0, T[, \quad (2.48)$$

$$\frac{\partial \varphi(u)}{\partial x}(0, t) = \frac{\partial \varphi(u)}{\partial x}(1, t) = 0, \quad t \in ]0, T[, \quad (2.49)$$

$$u(x, 0) = u_0(x), \quad x \in ]0, 1[, \quad (2.50)$$

où  $\varphi, v, T, u_0$  sont donnés et sont t.q.

1.  $T > 0, v \in L^\infty(]0, 1[ \times ]0, T[)$ ,
2.  $\varphi$  croissante, lipschitzienne de  $\mathbb{R}$  dans  $\mathbb{R}$ ,
3.  $u_0 \in L^\infty(]0, 1[)$  et  $\varphi(u_0)$  lipschitzienne de  $[0, 1]$  dans  $\mathbb{R}$ .

Un exemple important est donné par  $\varphi(s) = \alpha_1 s$  si  $s \leq 0$ ,  $\varphi(s) = 0$  si  $0 \leq s \leq L$  et  $\varphi(s) = \alpha_2(s - L)$  si  $s \geq L$ , avec  $\alpha_1, \alpha_2$  et  $L$  donnés dans  $\mathbb{R}_+^*$ . Noter pour cet exemple que  $\varphi' = 0$  sur  $]0, L[$ .

Les ensembles  $]0, 1[$  et  $D = ]0, 1[ \times ]0, T[$  sont munis de leur tribu borélienne et de la mesure de Lebesgue sur cette tribu.

On appelle "solution faible" de (2.48)-(2.50) une solution de :

$$u \in L^\infty(]0, 1[ \times ]0, T[, \quad (2.51)$$

$$\int_D (u(x, t) \frac{\partial \psi}{\partial t}(x, t) + \varphi(u(x, t)) \frac{\partial^2 \psi}{\partial x^2}(x, t) + v(x, t) \psi(x, t)) dx dt + \int_{]0, 1[} u_0(x) \psi(x, 0) dx = 0, \quad (2.52)$$

$$\forall \psi \in C_T^{2,1}(\mathbb{R}^2),$$

où  $\psi \in C_T^{2,1}(\mathbb{R}^2)$  signifie que  $\psi$  est une fonction de  $\mathbb{R}^2$  dans  $\mathbb{R}$  deux fois continûment dérivable par rapport à  $x$ , une fois continûment dérivable par rapport à  $t$  et t.q.  $\frac{\partial \psi}{\partial x}(0, t) = \frac{\partial \psi}{\partial x}(1, t) = 0$ , pour tout  $t \in [0, T]$  et  $\psi(x, T) = 0$  pour tout  $x \in [0, 1]$ .

**Question 1** (Solution classique versus solution faible)

On suppose, dans cette question seulement, que  $\varphi$  est de classe  $C^2$ ,  $v$  est continue sur  $[0, 1] \times [0, T]$  et  $u_0$  est continue sur  $[0, 1]$ . Soit  $u \in C^2(\mathbb{R}^2, \mathbb{R})$ . On note encore  $u$  la restriction de  $u$  à  $]0, 1[ \times ]0, T[$ . Montrer que  $u$  est solution de (2.51)-(2.52) si et seulement si  $u$  vérifie (2.48)-(2.50) au sens classique (c'est-à-dire pour tout  $(x, t) \in [0, 1] \times [0, T]$ ).

On cherche maintenant une solution approchée de (2.48)-(2.50).

Soient  $N, M \in \mathbb{N}^*$ . On pose  $h = \frac{1}{N}$  et  $k = \frac{T}{M}$ . On va construire une solution approchée de (2.48)-(2.50) à partir de la famille  $\{u_i^n, i = 1, \dots, N, n = 0, \dots, M\}$  (dont on va prouver l'existence et l'unicité) vérifiant les équations suivantes :

$$u_i^0 = \frac{1}{h} \int_{(i-1)h}^{ih} u_0(x) dx, \quad i = 1, \dots, N, \quad (2.53)$$

$$\frac{u_i^{n+1} - u_i^n}{k} - \frac{\varphi(u_{i-1}^{n+1}) - 2\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})}{h^2} = v_i^n, \quad i = 1, \dots, N, n = 0, \dots, M-1, \quad (2.54)$$

avec  $u_0^{n+1} = u_1^{n+1}, u_{N+1}^{n+1} = u_N^{n+1}$ , pour tout  $n = 0, \dots, M-1$  et  $v_i^n = \frac{1}{kh} \int_{nk}^{(n+1)k} \int_{(i-1)h}^{ih} v(x, t) dx dt$ , pour tout  $i = 1, \dots, N$ , pour tout  $n = 0, \dots, M$ .

**Question 2** (Existence et unicité de la solution approchée)

Soit  $n \in \{0, \dots, M-1\}$ . On suppose connu  $\{u_i^n, i = 1, \dots, N\}$ . On va prouver dans cette question l'existence et l'unicité de  $\{u_i^{n+1}, i = 1, \dots, N\}$  vérifiant (2.54) (avec  $u_0^{n+1} = u_1^{n+1}$ ,  $u_{N+1}^{n+1} = u_N^{n+1}$ ).

1. Soit  $a > 0$ , Pour  $s \in \mathbb{R}$ , on pose  $g_a(s) = s + a\varphi(s)$ . Montrer que  $g_a$  est une application strictement croissante bijective de  $\mathbb{R}$  dans  $\mathbb{R}$ .
2. Soit  $\bar{w} = (\bar{w}_i)_{i=1, \dots, N} \in \mathbb{R}^N$ . On pose  $\bar{w}_0 = \bar{w}_1$  et  $\bar{w}_{N+1} = \bar{w}_N$ . Montrer qu'il existe un et un seul couple  $(u, w) \in \mathbb{R}^N \times \mathbb{R}^N$ ,  $u = (u_i)_{i=1, \dots, N}$ ,  $w = (w_i)_{i=1, \dots, N}$ , t.q. :

$$\varphi(u_i) = w_i, \text{ pour tout } i \in \{1, \dots, N\}, \quad (2.55)$$

$$u_i + \frac{2k}{h^2}w_i = \frac{k}{h^2}(\bar{w}_{i-1} + \bar{w}_{i+1}) + u_i^n + kv_i^n, \text{ pour tout } i = 1, \dots, N. \quad (2.56)$$

On peut donc définir une application  $F$  de  $\mathbb{R}^N$  dans  $\mathbb{R}^N$  par  $\bar{w} \mapsto F(\bar{w}) = w$  où  $w$  est solution de (2.55)–(2.56).

3. On munit  $\mathbb{R}^N$  de la norme usuelle  $\|\cdot\|_\infty$ . Montrer que l'application  $F$  est strictement contractante. [On pourra utiliser la monotonie de  $\varphi$  et remarquer que, si  $a = \varphi(\alpha)$  et  $b = \varphi(\beta)$ , on a  $|\alpha - \beta| \geq (1/L)|a - b|$ , où  $L$  ne dépend que de  $\varphi$ .]
4. Soit  $\{u_i^{n+1}, i = 1, \dots, N\}$  solution de (2.54). On pose  $w = (w_i)_{i=1, \dots, N}$ , avec  $w_i = \varphi(u_i^{n+1})$  pour  $i \in \{1, \dots, N\}$ . Montrer que  $w = F(w)$ .
5. Soit  $w = (w_i)_{i=1, \dots, N}$  t.q.  $w = F(w)$ . Montrer que pour tout  $i \in \{1, \dots, N\}$  il existe  $u_i^{n+1} \in \mathbb{R}$  t.q.  $w_i = \varphi(u_i^{n+1})$ . Montrer que  $\{u_i^{n+1}, i = 1, \dots, N\}$  est solution de (2.54).
6. Montrer qu'il existe une unique famille  $\{u_i^{n+1}, i = 1, \dots, N\}$  solution de (2.54).

**Question 3** (Estimation  $L^\infty([0, 1] \times [0, T])$  sur  $u$ )

On pose  $A = \|u_0\|_{L^\infty([0, 1])}$  et  $B = \|v\|_{L^\infty([0, 1] \times [0, T])}$ . Montrer, par récurrence sur  $n$ , que  $u_i^n \in [-A - nkB, A + nkB]$  pour tout  $i = 1, \dots, N$  et tout  $n = 0, \dots, M$ . [On pourra, par exemple, considérer (2.54) avec  $i$  t.q.  $u_i^{n+1} = \min\{u_j^{n+1}, j = 1, \dots, N\}$ .]

En déduire qu'il existe  $c_{u_0, v, T} \in \mathbb{R}_+$  t.q.  $\|u^n\|_{L^\infty([0, 1])} \leq c_{u_0, v, T}$ .

**Question 4** (Estimation de la dérivée p.r. à  $x$  de  $\varphi(u)$ )

Montrer qu'il existe  $C_1$  (ne dépendant que de  $T, \varphi, v$  et  $u_0$ ) t.q., pour tout  $n = 0, \dots, M-1$ ,

$$\sum_{n=0}^{M-1} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq C_1 \frac{h}{k}. \quad (2.57)$$

[Multiplier (2.54) par  $u_i^{n+1}$  et sommer sur  $i$  et sur  $n$  et utiliser l'inégalité  $a^2 - ab \geq \frac{a^2}{2} - \frac{b^2}{2}$ .]

**Question 5** (Estimation de la dérivée p.r. à  $t$  de  $\varphi(u)$ )

Montrer qu'il existe  $C_2$  (ne dépendant que de  $T, \varphi, v$  et  $u_0$ ) t.q.

$$\sum_{n=0}^{M-1} h \sum_{i=0}^{N+1} (\varphi(u_{i+1}^n) - \varphi(u_i^{n+1}))^2 \leq C_2 k. \quad (2.58)$$

et

$$\sum_{i=0}^{N+1} (\varphi(u_i^{n+1}) - \varphi(u_{i+1}^n))^2 \leq C_2 h, \text{ pour tout } n \in \{0, \dots, M\}. \quad (2.59)$$

[indication : multiplier (2.54) par  $\varphi(u_i^{n+1}) - \varphi(u_i^n)$  et sommer sur  $i$  et  $n$ ]

Dans la suite de l'exercice, il s'agit de passer à la limite (quand  $N, M \rightarrow \infty$ ) pour trouver une solution de (2.48)–(2.50).



Pour  $M \in \mathbb{N}^*$  donné, on prend  $N = M^2$  (et donc  $h$  et  $k$  sont donnés et  $k = T\sqrt{h}$ ), on définit (avec les  $u_i^n$  trouvés dans les questions précédentes) une fonction,  $u_h$ , sur  $[0, 1] \times [0, T]$  en posant

$$u_h(x, t) = \frac{t - nk}{k} u_h^{(n+1)}(x) + \frac{(n+1)k - t}{k} u_h^{(n)}(x), \text{ si } t \in [nk, (n+1)k]$$

et

$$u_h^{(n)}(x) = u_i^n, \text{ si } x \in ](i-1)h, ih[, \quad i = 1, \dots, N, \quad n = 0, \dots, M.$$

Enfin, on définit  $\varphi(u_h)$  par  $\varphi(u_h)(x, t) = \varphi(u_h(x, t))$ .

**Question 6** Montrer que les suites  $(u_h)_{M \in \mathbb{N}^*}$  et  $(\varphi(u_h))_{M \in \mathbb{N}^*}$  sont bornées dans  $L^\infty(]0, 1[ \times ]0, T[)$  (on rappelle que  $h$  est donné par  $M$ ).

**Question 7**

Montrer qu'il existe  $C$  (ne dépendant que de  $T, \varphi, v$  et  $u_0$ ) t.q. l'on ait, pour tout  $M \in \mathbb{N}^*$  :

1. Pour tout  $t \in [0, T]$ ,

$$\int_{\mathbb{R}} |\varphi(u_h)(x + \eta, t) - \varphi(u_h)(x, t)|^2 dx \leq C\eta,$$

pour tout  $\eta \in \mathbb{R}_+^*$ , avec  $\varphi(u_h)(\cdot, t)$  prolongée par 0 hors de  $[0, 1]$ .

2.  $\|\varphi(u_h)(\cdot, t) - \varphi(u_h)(\cdot, s)\|_{L^2(]0, 1])} \leq C|t - s|$ , pour tout  $t, s \in [0, T]$ .

Une conséquence des questions 6 et 7 ( que l'on admet ici est que l'on peut trouver une suite  $(h_n)_{n \in \mathbb{N}}$  et  $u \in L^\infty(]0, 1[ \times ]0, T[)$  telle que, en posant  $u_n = u_{h_n}$  (on rappelle que  $k_n = T\sqrt{h_n}$ ), l'on ait, quand  $n \rightarrow \infty$ ,

1.  $h_n \rightarrow 0$  et  $k_n \rightarrow 0$ ,
2.  $u_n \rightarrow u$  dans  $L^\infty(]0, 1[ \times ]0, T[)$  pour la topologie faible-\*,
3.  $\varphi(u_n) \rightarrow \varphi(u)$  dans  $L^p(]0, 1[ \times ]0, T[)$ , pour tout  $p \in [1, \infty[$ .

**Question 8** Montrer que la fonction  $u$  ainsi trouvée est solution de (2.51),(2.52).

**Remarque.** On peut aussi montrer l'unicité de la solution de (2.51),(2.52).

## 2.6 Suggestions pour les exercices

### Exercice 24 (Exemple de schéma non convergent)

1. Ecrire le schéma d'Euler explicite.
2. Démontrer par récurrence que

$$\text{Si } n \in \{0, \dots, M+1\}, i \in \left\{-\frac{N+1}{2}, \dots, \frac{N+1}{2}\right\} \text{ et } i \geq -\frac{N+1}{4} + n \text{ alors } u_i^n = 0.$$

En déduire que  $u_i^n = 0$  pour  $n \in \{0, \dots, M+1\}$  et  $i \in \{0, \dots, \frac{N+1}{2}\}$  et conclure.

### Exercice 27 (Discretisation d'un problème parabolique)

1. Calculer l'erreur de consistance et la majorer par des développements de Taylor. Chercher ensuite les conditions pour que :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Pour étudier la convergence du schéma, majorer l'erreur de discrétisation :  $e_j^n = \bar{u}_j^n - u_j^n$  où  $u_j^n$  est calculé par (2.70), et  $\bar{u}_j^n$  est la solution du problème (2.34) en  $x_j = jh$  et  $t_n = nk$ .

Même chose pour les questions suivantes...

### Exercice 28 (Problème de diffusion réaction)

1. Effectuer des développements de Taylor...
3. Montrer par récurrence que  $\max_{j=1, \dots, N} u_j^n \leq (1+k)^n \max_{j=1, \dots, N} u_j^0$  et que  $\min_{j=1, \dots, N} u_j^{(n)} \geq (1+k)^n \min_{j=1, \dots, N} u_j^{(0)}$ .
4. Utiliser l'équation, le schéma, et l'erreur de consistance.

### Exercice 30 (Schémas de Saute-Mouton et Dufort-Frankel)

1. Effectuer des développements de Taylor pour majorer l'erreur de consistance.
2. Montrer que le facteur d'amplification  $\xi_n$  obtenu par l'analyse de stabilité de Von Neumann satisfait :

$$\xi_{n+1} - \alpha \xi_n - \xi_{n-1} = 0, \quad n \geq 2.$$

Etudier ensuite les racines de l'équation  $r^2 - \alpha r - 1 = 0$  et montrer que l'une de ses racines est, en module, supérieure à 1.

4. Reprendre la méthode développée à la question 2, en montrant que l'équation caractéristique pour  $\xi$  est maintenant :

$$p(r) = ar^2 + br + c = 0,$$

avec

$$a = \frac{1}{2k} + \frac{1}{h^2}, \quad b = -\frac{2 \cos(ph)}{h^2} \text{ et } c = \frac{1}{h^2} - \frac{1}{2h}.$$

Etudier ensuite les racines de cette équation.

## 2.7 Corrigés des exercices

### Exercice 22 page 71

On note  $\|\cdot\|_2 = \|\cdot\|_{L^2(]0,1[)}$ .

- 1) Pour  $n \in \mathbb{N}^*$ , on a

$$\int_0^1 \sin^2(n\pi x) dx = \int_0^1 \frac{1 - \cos(2n\pi x)}{2} dx = \frac{1}{2},$$

et

$$\int_0^1 |u_0(x) \sin(n\pi x)| dx \leq \|u_0\|_2 \left( \int_0^1 \sin^2(n\pi x) dx \right)^{1/2} = \frac{r_2}{2} \|u_0\|_2.$$

La quantité  $a_n$  est donc bien définie et

$$|a_n| \leq r_2 \|u_0\|_2$$

Pour tout  $t > 0$  et  $x \in [0, 1]$ , on a

$$|e^{-n^2\pi^2 t^2} a_n \sin(n\pi x)| \leq r_2 \|u_0\|_2 e^{-n^2\pi^2 t^2} \quad \forall n \in \mathbb{N}^*.$$

Ceci montre que la série  $\sum_{n>0} e^{-n^2\pi^2 t^2} a_n \sin(n\pi x)$  est absolument convergente et donc que  $u$  est bien définie pour

tout  $t > 0$  et tout  $x \in [0, 1]$  et même pour tout  $x \in \mathbb{R}$ .

On remarque ensuite que  $u$  est de classe  $C^\infty$  sur  $\mathbb{R} \times \mathbb{R}_+^*$ , en appliquant les théorèmes classiques de dérivation terme à terme d'une série. En effet, soit  $\varepsilon > 0$ , pour tout  $x \in \mathbb{R}$  et  $t > \varepsilon$  on a

$$\left| e^{-n^2\pi^2 t^2} a_n \sin(n\pi x) \right| \leq r_2 \|u_0\|_2 e^{-n^2\pi^2 \varepsilon^2}, \quad \forall n \in \mathbb{N}^*$$

Comme  $(x, t) \rightarrow e^{-n^2\pi^2 t^2} a_n \sin(n\pi t)$  est continue (pour tout  $n \in \mathbb{N}^*$ ), on en déduit que  $u$  est continue sur  $\mathbb{R} \times ]\varepsilon, \infty[$ , et finalement sur  $\mathbb{R} \times ]0, \infty[$  car  $\varepsilon > 0$  est arbitraire.

Pour dériver terme à terme la série définissant  $u$ , il suffit également d'obtenir sur  $]\varepsilon, \infty[ \times \mathbb{R}$  (pour tout  $\varepsilon > 0$ ) une majoration du terme général de la série des dérivées par le terme général d'une série convergente (indépendant de  $(x, t) \in \mathbb{R} \times ]\varepsilon, \infty[$ ). On obtient cette majoration en remarquant que, pour  $(x, t) \in \mathbb{R} \times ]\varepsilon, \infty[$ ,

$$| -n^2\pi^2 e^{-n^2\pi^2 t^2} a_n \sin(n\pi x) | \leq n^2\pi^2 e^{-n^2\pi^2 \varepsilon^2} r_2 \|u_0\|_2$$

On montre ainsi finalement que  $u$  est de classe  $C^1$  par rapport à  $t$  et que

$$u_t(x, t) = \sum_{n>0} -n^2\pi^2 e^{-n^2\pi^2 t^2} a_n \sin(n\pi x), \quad x \in \mathbb{R}, t > 0.$$

En itérant ce raisonnement on montre que  $u$  est de classe  $C^\infty$  par rapport à  $t$  sur  $\mathbb{R} \times \mathbb{R}_+^*$ .

Un raisonnement similaire montre que  $u$  est de classe  $C^\infty$  par rapport à  $x$  sur  $\mathbb{R} \times \mathbb{R}_+^*$  et que l'on peut dériver terme à terme la série définissant  $u$ . On obtient donc aussi

$$u_{xx}(x, t) = \sum_{n>0} -n^2\pi^2 e^{-n^2\pi^2 t^2} a_n \sin(n\pi x), \quad x \in \mathbb{R}, t > 0,$$

et ceci donne  $u_t = u_{xx}$  sur  $\mathbb{R} \times \mathbb{R}_+^*$  et donc aussi un  $[0, 1] \times \mathbb{R}_+^*$ . Le fait que  $u(0, t) = u(1, t)$  pour tout  $t > 0$  est immédiat car  $\sin n\pi t = \sin 0 = 0$ , pour tout  $n \in \mathbb{N}^*$ .

Il reste à montrer que  $u(\cdot, t) \rightarrow u_0$  dans  $L^2(]0, 1[)$  quand  $t \rightarrow 0$ .

On définit  $e_n \in L^2(]0, 1[)$  par  $e_n(x) = \sqrt{2} \sin(n\pi x)$ . La famille  $\{e_n, n \in \mathbb{N}^*\}$  est une base hilbertienne de  $L^2(]0, 1[)$ .

On a donc :

$$\sum_{n=1}^N a_n \sin n\pi x \rightarrow u_0, \quad \text{dans } L^2(]0, 1[), \quad \text{quand } n \rightarrow \infty,$$

et

$$\sum_{n=1}^{\infty} a_n^2 = 2 \|u_0\|_2^2.$$

On remarque maintenant que

$$u(x, t) - u_0(x) = u(x, t) - u^{(N)}(x, t) + u^{(N)}(x, t) - u_0^{(N)}(x) - u_0^{(N)}(x) - u_0(x),$$

avec

$$u^{(N)}(x, t) = \sum_{n=1}^N a_n e^{-n^2\pi^2 t^2} \sin(n\pi x)$$

$$u_0^{(N)}(x) = \sum_{n=1}^N a_n \sin(n\pi x).$$

Il est clair que, pour tout  $N \in \mathbb{N}^*$ , on a  $u^{(N)}(\cdot, t) \rightarrow u_0^{(N)}$  uniformément sur  $\mathbb{R}$ , quand  $N \rightarrow \infty$ , et donc  $u^{(N)}(\cdot, t) \rightarrow u_0^{(N)}$  dans  $L^2(]0, 1[)$ .

Comme

$$\|u(\cdot, t) - u^{(N)}(\cdot, t)\|_2^2 = \sum_{n=N+1}^{\infty} a_n^2 \frac{1}{2} e^{-2n^2\pi^2 t^2} \leq \sum_{n=N+1}^{\infty} a_n^2 \frac{1}{2} = \|u_0^{(N)} - u_0\|_2^2 \rightarrow 0$$

quand  $N \rightarrow \infty$ , on en déduit que  $u(\cdot, t) \rightarrow u_0$ ,  $qdt \rightarrow 0$ , dans  $L^2(]0, 1[)$ .

2) On note  $w$  la différence de 2 solutions de (2.25). On a donc

$$\begin{aligned} w &\in C^\infty([0, 1] \times \mathbb{R}_+^*, \mathbb{R}) \\ w_t - w_{xx} &= 0 \text{ sur } [0, 1] \times \mathbb{R}_+^* \\ w(0, t) = w(1, t) &= 0 \text{ pour } t > 0 \\ w(\cdot, t) &\rightarrow 0, \text{ dans } L^2(]0, 1[), \text{ quand } t \rightarrow 0 \end{aligned}$$

Soit  $0 < \varepsilon < T < \infty$ . On intègre l'équation  $ww_t - ww_{xx} = 0$  sur  $]0, 1[ \times ]\varepsilon, T[$ . En utilisant une intégration par parties (noter que  $w \in C^\infty([0, 1] \times [\varepsilon, T])$ ), on obtient :

$$\frac{1}{2} \int_0^1 w^2(x, T) dx - \frac{1}{2} \int_0^1 w^2(x, \varepsilon) dx + \int_0^1 \int_\varepsilon^T w_x^2(x, t) dx dt = 0.$$

D'où l'on déduit  $\|w(\cdot, T)\|_2 \leq \|w(\cdot, \varepsilon)\|_2$ . Quand  $\varepsilon \rightarrow 0$ , on a  $\|w(\cdot, \varepsilon)\|_2 \rightarrow 0$ , on a donc  $\|w(\cdot, T)\|_2 = 0$  et donc, comme  $w(\cdot, t)$  est contenue sur  $[0, 1]$ ,  $wX \in [0, 1]$ . Comme  $T > 0$  est arbitraire, on a finalement

$$w(x, t) = 0 \quad \forall t \in [0, 1], \forall t > 0$$

Ce qui montre bien l'unicité de la solution de (2.25).

## Exercice 24 page 71

1) La formule pour calculer  $u_i^0$  est :

$$u_1^0 = u_0(ih, 0), \quad i = -\frac{N+1}{2}, \dots, \frac{N+1}{2}$$

Soit maintenant  $n \in \{0, \dots, M\}$ . On a :

$$\begin{aligned} u_i^{n+1} &= 0 \quad \text{pour} \quad i = -\frac{N+1}{2} \quad \text{et} \quad i = \frac{N+1}{2} \\ u_i^{n+1} &= u_i^n + \frac{k}{h^2} (u_{i+1}^n + u_{i-1}^n - 2u_i^n), \quad i = -\frac{N+1}{2} + 1, \dots, \frac{N+1}{2} - 1. \end{aligned}$$

2) On va montrer, par récurrence (finie) sur  $n$ , que :

$$\text{Si } n \in \{0, \dots, M+1\}, i \in \left\{ -\frac{N+1}{2}, \dots, \frac{N+1}{2} \right\} \text{ et } i \geq -\frac{N+1}{4} + n \text{ alors } u_i^n = 0. \quad (2.60)$$

Pour initialiser la récurrence, on suppose que  $n = 0$  et  $i \geq -\frac{N+1}{4}$ . On a alors

$$ih \geq -\frac{N+1}{4} \quad \frac{8}{N+1} = -2 > -3$$

et donc  $u_i^0 = 0$ .

Soit maintenant  $n \in \{0, \dots, M\}$ . On suppose que l'hypothèse de récurrence est vérifiée jusqu'au rang  $n$ , et on démontre la propriété au rang  $n+1$ . Soit donc  $i \in \left\{ -\frac{N+1}{2}, \dots, \frac{N+1}{2} \right\}$  tel que  $i \geq -\frac{N+1}{4} + (n+1)$ . Alors :

– Si  $i = \frac{N+1}{2}$  on a bien  $u_i^{N+1} = 0$ .

– Si  $i < \frac{N+1}{2}$ , les indices  $i-1$ ,  $i$  et  $i+1$  sont tous supérieurs ou égaux à  $-\frac{N+1}{4} + n$ , et donc par hypothèse de récurrence,

$$u_i^{n+1} = u_i^n \left(1 - \frac{2k}{h^2}\right) + \frac{k}{h^2} u_{i+1}^n + \frac{k}{h^2} u_{i-1}^n = 0.$$

On a donc bien démontré (2.60). On utilise maintenant l'hypothèse  $k = h$ , c'est-à-dire  $\frac{1}{M+1} = \frac{8}{N+1}$ . On a alors

$$-\frac{N+1}{4} + M + 1 = -2(M+1) + M + 1 = -(M+1) < 0.$$

On en déduit que si  $n \in \{0, \dots, M+1\}$  et  $i \geq 0$ , alors  $i \geq -\frac{N+1}{4} + n$ . On en déduit que  $u_i^n = 0$  pour  $n \in \{0, \dots, M+1\}$  et  $i \in \{0, \dots, \frac{N+1}{2}\}$ . On remarque alors que

$$\begin{aligned} \max \left\{ |u_i^{M+1} - \bar{u}_i^{M+1}|, i \in \left\{ -\frac{N+1}{2}, \dots, \frac{N+1}{2} \right\} \right\} &\geq \max \left\{ |\bar{u}_i^{M+1}|, i \in \left\{ 0, \dots, \frac{N+1}{2} \right\} \right\} \\ &\geq \inf_{[0,4]} u(x, 1) > 0, \end{aligned}$$

et donc ne tend pas vers 0 quand  $h \rightarrow 0$ .

## Exercice 26 page 72 : schéma implicite et principe du maximum

### 1. Schéma explicite décentré

(a) Par définition, l'erreur de consistance en  $(x_i, t_n)$  s'écrit : On s'intéresse ici à l'ordre du schéma au sens des différences finies. On suppose que  $u \in C^4([0, 1] \times [0, T])$  est solution de (2.32) et on pose

$$\bar{u}_i^n = u(ih, nk), \quad i = 0, \dots, N, \quad k = 0, \dots, M.$$

Pour  $i = 1, \dots, N-1$  et  $k = 1, \dots, M-1$ , l'erreur de consistance en  $(x_i, t_k)$  est définie par :

$$R_i^n = \frac{1}{k}(\bar{u}_i^{n+1} - \bar{u}_i^n) - \frac{\alpha}{h}(\bar{u}_i^n - \bar{u}_{i-1}^n) - \frac{\mu}{h^2}(\bar{u}_{i-1}^n - 2\bar{u}_i^n + \bar{u}_{i+1}^n). \quad (2.61)$$

Soit  $i \in \{1, \dots, N-1\}$ ,  $k \in \{1, \dots, M-1\}$ . On cherche une majoration de  $R_i^n$  en utilisant des développements de Taylor. En utilisant ces développements, on obtient qu'il existe  $(\xi_\ell, t_\ell) \in [0, 1] \times [0, T]$ ,  $\ell = 1, \dots, 4$ , t.q. :

$$\bar{u}_i^{n+1} = \bar{u}_i^n + k u_t(ih, nk) + \frac{k^2}{2} u_{tt}(\xi_1, t_1), \quad (2.62)$$

$$\bar{u}_{i-1}^n = \bar{u}_i^n - h u_x(ih, nk) + \frac{h^2}{2} u_{xx}(\xi_2, t_2), \quad (2.63)$$

$$\bar{u}_{i-1}^n = \bar{u}_i^n - h u_x(ih, nk) + \frac{h^2}{2} u_{xx}(ih, nk) - \frac{h^3}{6} u_{xxx}(ih, nk) - \frac{h^4}{24} u_{xxxx}(\xi_3, t_3), \quad (2.64)$$

$$\bar{u}_{i+1}^n = \bar{u}_i^n + h u_x(ih, nk) + \frac{h^2}{2} u_{xx}(ih, nk) + \frac{h^3}{6} u_{xxx}(ih, nk) + \frac{h^4}{24} u_{xxxx}(\xi_4, t_4). \quad (2.65)$$

On en déduit :

$$\begin{aligned} R_i^n &= u_t(ih, nk) + \frac{k}{2} u_{tt}(\xi_1, t_1) + \alpha u_x(ih, nk) + \alpha \frac{h}{2} u_{xx}(\xi_2, t_2) \\ &\quad - \mu u_{xx}(ih, nk) - \mu \frac{h^2}{24} (u_{xxxx}(\xi_3, t_3) + \mu u_{xxxx}(\xi_4, t_4)), \end{aligned}$$

et donc, comme  $u$  est solution de (2.32), pour  $h$  assez petit, on a :

$$|R_i^n| \leq C_1(h+k),$$

où  $C_1$  ne dépend que de  $u$ . Le schéma (2.33) est donc consistant d'ordre 1 en temps et en espace.

- (b) Cherchons les conditions pour que  $u_i^{n+1}$  s'écrive comme combinaison convexe de  $u_i^n$ ,  $u_{i-1}^n$  et  $u_{i+1}^n$ . On peut réécrire le schéma (2.33) :

$$u_i^{n+1} = au_i^n + bu_{i+1}^n + cu_{i-1}^n, \text{ avec } a = 1 - \frac{\alpha k}{h} - \frac{2\mu k}{h^2}, b = \frac{\mu k}{h^2} \text{ et } c = \frac{\alpha k}{h} + \frac{\mu k}{h^2}.$$

Il est facile de voir que  $a + b + c = 1$ , et que  $b \geq 0$ ,  $c \geq 0$ . Il reste à vérifier que  $a \geq 0$ ; pour cela, il faut et il suffit que  $\frac{\alpha k}{h} + \frac{2\mu k}{h^2} \leq 1$ . Cette condition s'écrit encore :

$$k \leq \frac{h^2}{\alpha h + 2\mu}. \quad (2.66)$$

Si  $h$  et  $k$  vérifient la condition (2.66), on pose :  $M^n = \max_{i=1 \dots N} u_i^n$  (resp.  $m^n = \min_{i=1 \dots N} u_i^n$ ). Comme  $u_i^{n+1}$  est une combinaison convexe de  $u_i^n$ ,  $u_{i-1}^n$  et  $u_{i+1}^n$ , on a alors :  $u_i^{n+1} \leq M^n \quad \forall i = 1, \dots, N$  (resp.  $u_i^{n+1} \geq m^n \quad \forall i = 1, \dots, N$ ) et donc :  $M^{n+1} \leq M^n$  (resp.  $m^{n+1} \geq m^n$ ). On a ainsi montré que :

$$\|u^{n+1}\|_\infty \leq \|u^n\|_\infty.$$

On a de même :

$$\|u^n\|_\infty \leq \|u^{n-1}\|_\infty.$$

⋮

$$\|u^1\|_\infty \leq \|u^0\|_\infty.$$

En sommant ces inégalités, on obtient :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Donc, sous la condition (2.66), on a  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$  et donc  $\|u^n\|_\infty \leq \|u^0\|_\infty$ , pour tout  $n = 1, \dots, N$ .

- (c) En retranchant l'égalité (2.61) au schéma (2.33), on obtient l'équation suivante sur  $e_i^n$  :

$$\frac{1}{k}(e_i^{n+1} - e_i^n) + \frac{\alpha}{h}(e_i^n - e_{i-1}^n) - \frac{\mu}{h^2}(e_{i-1}^n - 2e_i^n + e_{i+1}^n) = R_i^n.$$

ce qu'on peut encore écrire :

$$e_i^{n+1} = \left(1 - \frac{k\alpha}{h} - 2\frac{k\mu}{h^2}\right)e_i^n + e_{i-1}^n \frac{k\mu}{h^2} + kR_i^n. \quad (2.67)$$

Sous la condition de stabilité (2.66), on obtient donc :

$$\begin{aligned} |e_i^{n+1}| &\leq \|e^{n+1}\|_\infty + C_1(k+h)k, \\ |e_i^n| &\leq \|e^{n-1}\|_\infty + C_1(k+h)k, \\ &\vdots \leq \vdots + \vdots \\ |e_i^1| &\leq \|e^0\|_\infty + C_1(k+h)k, \end{aligned}$$

Si à  $t = 0$ , on a  $\|e^0\| = 0$ , alors on éduit des inégalités précédentes que  $|e_i^n| \leq C_1 T(k+h)$  pour tout  $n \in \mathbb{N}$ . Le schéma est donc convergent d'ordre 1.

## 2. Schéma explicite centré.

- (a) (Consistance) En utilisant les développements de Taylor (2.62) (2.64) et (2.65), et les développements suivants :

$$\bar{u}_{i-1}^n = \bar{u}_i^n - hu_x(ih, nk) + \frac{h^2}{2}u_{xx}(ih, nk) - \frac{h^3}{6}u_{xxx}(\xi_5, t_5),$$

$$\bar{u}_{i+1}^n = \bar{u}_i^n + hu_x(ih, nk) + \frac{h^2}{2}u_{xx}(ih, nk) + \frac{h^3}{6}u_{xxx}(\xi_6, t_6),$$

on obtient maintenant :

$$R_i^n = u_t(ih, nk) + \frac{k}{2}u_{tt}(\xi_1, t_1) + \alpha u_x(ih, nk) + \alpha \frac{h^2}{12}(u_{xxx}(\xi_5, t_5) + \mu u_{xxx}(\xi_6, t_6)) \\ - \mu u_{xx}(ih, nk) - \mu \frac{h^2}{24}(u_{xxxx}(\xi_3, t_3) + \mu u_{xxxx}(\xi_4, t_4)),$$

On en déduit que

$$|R_i^n| \leq C_3(k + h^2),$$

où  $C_3 = \max(\frac{1}{2}\|u_{tt}\|_\infty, \frac{1}{8}\|u_{xxx}\|_\infty, \frac{1}{12}\|u_{xxxx}\|_\infty)$ .

(b) Le schéma s'écrit maintenant :

$$u_i^{n+1} = \tilde{a}u_i^n + \tilde{b}u_{i+1}^n + \tilde{c}u_{i-1}^n, \text{ avec } \tilde{a} = 1 - \frac{2\mu k}{h^2}, \tilde{b} = \frac{\mu k}{h^2} - \frac{\alpha k}{h} \text{ et } \tilde{c} = \frac{\mu k}{h^2} + \frac{\alpha k}{h}.$$

Remarquons que l'on a bien :  $\tilde{a} + \tilde{b} + \tilde{c} = 1$ . Pour que  $u_i^{n+1}$  soit combinaison convexe de  $u_i^n$ ,  $u_{i+1}^n$  et  $u_{i-1}^n$ , il faut et il suffit donc que  $\tilde{a} \geq 0$ ,  $\tilde{b} \geq 0$ , et  $\tilde{c} \leq 0$ . L'inégalité  $\tilde{c} \geq 0$  est toujours vérifiée. Les deux conditions qui doivent être vérifiées par  $h$  et  $k$  s'écrivent donc :

i.  $\tilde{a} \geq 0$ , i.e.  $1 - \frac{2\mu k}{h^2} \geq 0$ , soit encore

$$k \leq \frac{h^2}{2\mu}.$$

ii.  $\tilde{b} \geq 0$  i.e.  $\frac{\mu k}{h^2} - \frac{\alpha k}{h} \geq 0$ , soit encore

$$h \leq \frac{\mu}{2\alpha}.$$

Le schéma centré est donc stable sous les deux conditions suivantes :

$$h \leq \frac{\mu}{2\alpha} \text{ et } k \leq \frac{1}{2\mu}h^2. \quad (2.68)$$

Pour obtenir une borne d'erreur, on procède comme pour le schéma (2.33) : on soustrait la définition de l'erreur de consistance au schéma numérique, et on obtient :

$$e_i^{n+1} = \tilde{a}e_i^n + \tilde{b}e_{i+1}^n + \tilde{c}e_{i-1}^n + kR_i^n. \quad (2.69)$$

Par le même raisonnement que pour le schéma décentré, on obtient donc que si  $e_i^0 = 0$ , on a  $|e_i^n| \leq C_4(k + h^2)$ , avec  $C_4 = TC_3$ .

### Exercice 27 page 73

1. On admettra que la solution de (2.34) existe et qu'elle est assez régulière. Soient  $M \in \mathbb{N}^*$  et  $N \in \mathbb{N}^*$ , et soient  $k$  le pas de temps, choisi tel que  $Mk = T$  et  $h$  le pas espace, choisi tel que  $Nh = 1$ . On applique un schéma d'Euler explicite en temps, et un schéma de différences finies centré en espace, on obtient donc :

$$u_j^{n+1} = k \left[ \frac{1}{k}u_j^n - \frac{1}{2h}(u_{j+1}^n - u_{j-1}^n) + \frac{\varepsilon}{h^2}(u_{j+1}^n + u_{j-1}^n - 2u_j^n) \right] \quad (2.70)$$

On tient compte des conditions aux limites et des conditions initiales en posant :

$$\begin{cases} u_0^n = u_{N+1}^n = 0, \\ u_j^0 = u_0(jh). \end{cases}$$

On a, par développement de Taylor :

$$u(x+h, t) = u(x, t) + hu_x(x, t) + \frac{h^2}{2}u_{xx}(x, t) + \frac{h^3}{6}u^{(3)}(x, t) + \frac{h^4}{24}u^{(4)}(\alpha, t), \\ u(x-h, t) = u(x, t) - hu_x(x, t) + \frac{h^2}{2}u_{xx}(x, t) - \frac{h^3}{6}u^{(3)}(x, t) + \frac{h^4}{24}u^{(4)}(\beta, t), \\ u(x, t+k) = u(x, t) + ku_t(x, t) + \frac{k^2}{2}u_{tt}(x, \tau_k), \tau_k \in [t, t+k].$$

De ces développements de Taylor, il ressort que l'erreur de consistance vérifie  $|R| \leq C(k + h^2)$ , où  $C$  ne dépend que de  $u$ . Le schéma est donc explicite d'ordre 1 en temps et 2 en espace.

Cherchons alors les conditions pour que :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Par définition,

$$\|u^n\|_\infty = \max_{j=1,\dots,N} |u_j^n|.$$

On essaye d'abord de vérifier que :  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ , c'est-à-dire :

$$\max_{j=1,\dots,N} |u_j^{n+1}| \leq \max_{j=1,\dots,N} |u_j^n|,$$

On veut donc montrer que

$$\begin{cases} \max_{j=1,\dots,N} u_j^{n+1} \leq \max_{j=1,\dots,N} u_j^n, \\ \min_{j=1,\dots,N} u_j^{n+1} \geq \min_{j=1,\dots,N} u_j^n. \end{cases}$$

On peut réécrire le schéma (2.70) :

$$u_j^{n+1} = u_j^n \left(1 - \frac{2\varepsilon k}{h^2}\right) + u_{j+1}^n \left(-\frac{k}{2h} + \frac{k\varepsilon}{h^2}\right) + u_{j-1}^n \left(\frac{\varepsilon k}{h^2} + \frac{k}{2h}\right).$$

Posons :

$$M^n = \max_{j=1,\dots,N} u_j^n$$

Supposons que  $k$  et  $h$  vérifient :

$$1 \geq \frac{2\varepsilon k}{h^2} \text{ et } \frac{k\varepsilon}{h^2} - \frac{k}{2h} \geq 0,$$

ce qui s'écrit encore :

$$\begin{cases} \frac{k}{h^2} \leq \frac{1}{2\varepsilon} \\ k \leq \frac{2\varepsilon}{h}, \end{cases} \quad (2.71)$$

on a alors :

$$u_j^{n+1} \leq M^n \left(1 - \frac{2\varepsilon k}{h^2}\right) + M^n \left(-\frac{k}{2h} + \frac{k\varepsilon}{h^2}\right) + M^n \left(\frac{\varepsilon k}{h^2} + \frac{k}{2h}\right) \quad \forall j = 1, \dots, N,$$

et donc :

$$M^{n+1} \leq M^n.$$

Posons maintenant :

$$m^n = \min_{j=1,\dots,N} u_j^n.$$

Si  $k$  et  $h$  satisfont les conditions (2.71), on obtient de la même manière

$$m^{n+1} \geq m^n$$

On a ainsi montré que :

$$\|u^{n+1}\|_\infty \leq \|u^n\|_\infty.$$

On a de même :

$$\|u^n\|_\infty \leq \|u^{n-1}\|_\infty.$$

⋮

$$\|u^1\|_\infty \leq \|u^0\|_\infty.$$



En sommant ces inégalités, on obtient :

$$\|u^n\|_\infty \leq \|u^0\|_\infty.$$

Donc, sous les conditions (2.71), on a  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$  et donc  $\|u^n\|_\infty \leq \|u^0\|_\infty$ , pour tout  $n = 1, \dots, N$ .

Pour étudier la convergence du schéma, on va tenter de majorer l'erreur de discrétisation :

$$e_j^n = \bar{u}_j^n - u_j^n,$$

où  $u_j^n$  est calculé par (2.70), et  $\bar{u}_j^n$  est la solution du problème (2.34) en  $x_j = jh$  et  $t_n = nk$ .

On a donc, par définition de l'erreur de consistance,

$$\frac{1}{k}(\bar{u}_j^{n+1} - \bar{u}_j^n) + \frac{1}{2h}(\bar{u}_{j+1}^n - \bar{u}_{j-1}^n) - \frac{\varepsilon}{h^2}(-2\bar{u}_j^n + \bar{u}_{j+1}^n + \bar{u}_{j-1}^n) = R_j^n$$

où  $|R_i^n| \leq C(k + h^2)$

ce qui entraîne :

$$\frac{1}{k}(e_j^{n+1} - e_j^n) + \frac{1}{2h}(e_j^n - e_{j-1}^n) - \frac{\varepsilon}{h^2}(-2e_j^n + e_{j+1}^n + e_{j-1}^n) = R_j^n$$

soit encore :

$$e_j^{n+1} = \left(1 - \frac{2\varepsilon k}{h^2}\right) e_j^n + \left(-\frac{k}{2h} + \frac{k\varepsilon}{h^2}\right) e_{j+1}^n + \left(\frac{\varepsilon k}{h^2} + \frac{k}{2h}\right) e_{j-1}^n + kR_j^n.$$

de même que précédemment, on obtient sous les conditions (2.71)

$$\begin{aligned} |e_j^{n+1}| &\leq \|e^n\|_\infty + C(k + h^2)k \\ &\vdots \\ |e_j^n| &\leq \|e^{n-1}\|_\infty + C(k + h^2)k \\ &\vdots \\ |e_j^1| &\leq \|e^0\|_\infty + C(k + h^2)k. \end{aligned}$$

Et donc en sommant ces inégalités :

$$\|e^n\|_\infty \leq \|e^0\|_\infty + nCk(k + h^2)$$

Si à  $t = 0$  on a  $\|e^0\|_\infty = 0$ , alors :

$$\|e^{n+1}\|_\infty \leq CMk(k + h^2) = T(k + h^2).$$

Et donc sous les conditions (2.71) on a  $\|e^n\|_\infty$  qui tend vers 0 lorsque  $k, h \rightarrow 0$ , ce qui prouve que le schéma est convergent.

### Exercice 28 page 73

1. Notons  $R_i^{(n)}$  l'erreur de consistance en  $(x_i, t_n)$ . Pour le schéma (2.37), on a donc par définition :

$$\begin{aligned} R_i^{(n)} &= \frac{\bar{u}_i^{(n+1)} - \bar{u}_i^{(n)}}{k} + \frac{1}{h^2}(2\bar{u}_i^{(n+1)} - \bar{u}_{i-1}^{(n+1)} - \bar{u}_{i+1}^{(n+1)}) - \bar{u}_i^{n+1} \\ &= \tilde{R}_i^{(n)} + \hat{R}_i^n, \end{aligned}$$

où

$$\tilde{R}_i^n = \frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} - u_t(x_i, t_n) \text{ est l'erreur de consistance en temps}$$

et

$$\hat{R}_i^n = \frac{1}{h^2}(2\bar{u}_i^{n+1} - \bar{u}_{i-1}^{n+1} - \bar{u}_{i+1}^{n+1}) - (u_{xx}(x_i, t_n)) \text{ est l'erreur de consistance en espace.}$$

On a vu (voir (1.30)) que

$$\left| \hat{R}_i^n \right| \leq \frac{h^2}{12} \sup_{[0,1]} \left| \frac{\partial^4 u}{\partial x^4}(\cdot, t_n) \right|, \forall i \in \{1, \dots, N\}$$

Effectuons maintenant un développement de Taylor en fonction du temps d'ordre 2 :

$$u(x_i, t_{n+1}) = u(x_i, t_n) + ku_t + \frac{k^2}{2}u_{tt}(x_i, \xi_n)$$

avec  $\xi_n \in [t_n, t_{n+1}]$ . Donc  $\frac{u(x_i, t_{n+1}) - u(x_i, t_n)}{k} - u_t = \frac{k}{2}u_{tt}(x_i, \xi_n)$ . Comme  $\xi_n \in [0, T]$ , et  $u_{tt}$  admet un maximum (à  $x_i$  fixé) dans  $[0, T]$  (qui est compact), on a donc

$$\left| \tilde{R}_i^n \right| \leq \frac{k}{2} \max_{[0, T]} |u_{tt}(x_i, \cdot)|.$$

Par conséquent,

$$|R_i^n| = \left| \tilde{R}_i^n + \hat{R}_i^n \right| \leq \left| \tilde{R}_i^n \right| + \left| \hat{R}_i^n \right| \leq \frac{k}{2} \max_{[0, T]} |u_{tt}(x_i, \cdot)| + \frac{h^2}{12} \max_{[0, 1]} \left| \frac{\partial^4 u}{\partial x^4}(\cdot, t_{n+1}) \right|.$$

Donc  $|R_i^n| \leq C(k + h^2)$  avec

$$C = \frac{1}{2} \max \left( \|u_{tt}\|_{L^\infty([0, 1] \times [0, T])}; \frac{1}{6} \left\| \frac{\partial^4 u}{\partial x^4} \right\|_{L^\infty([0, 1] \times [0, T])} \right).$$

Le calcul de l'erreur de consistance pour le schéma (2.38) s'effectue de manière semblable.

2. Le schéma (2.37) est complètement implicite alors que le schéma (2.38) ne l'est que partiellement, puisque le terme de réaction est pris à l'instant  $n$ . Le schéma (2.37) s'écrit :  $AU^{n+1} = U^n$  avec  $U^{n+1} = (U_1^{n+1}, \dots, U_N^{n+1})$ ,  $U^n = (U_1^n, \dots, U_N^n)$ , et

$$A = \begin{pmatrix} 1 + 2\lambda - k & -\lambda & 0 & \dots & 0 \\ -\lambda & 1 + 2\lambda - k & -\lambda & \ddots & 0 \\ 0 & & & & \\ \vdots & & & & 0 \\ 0 & 0 & -\lambda & & 1 + 2\lambda - k \end{pmatrix}$$

où  $\lambda = \frac{k}{h^2}$ . Notons que par définition,  $A$  est symétrique. De même, le schéma (2.38) s'écrit :  $BU^{n+1} = U^n$  avec

$$B = \frac{1}{1+k} \begin{pmatrix} 1 + 2\lambda & -1 & 0 & \dots & 0 \\ -1 & 1 + 2\lambda & & \ddots & 0 \\ 0 & & & & -1 \\ \vdots & & & & \\ 0 & 0 & -1 & & 1 + 2\lambda \end{pmatrix}$$

On a donc  $A = \lambda A_h$ , où  $A_h$  est définie en (1.26) page 13, avec  $c_i = \frac{1-k}{\lambda}$ , et  $B = \frac{\lambda}{k+1} A_h$  avec  $c_i = \frac{1}{\lambda}$ . Dans les deux cas, les matrices sont donc s.d.p. en vertu de la proposition 1.3 page 13. Notons que l'hypothèse  $k \in ]0, 1[$  est nécessaire dans le cas du premier schéma, pour assurer la positivité de  $c_i$ .

3. Le schéma (2.38) s'écrit

$$(1+k)u_i^n = u_i^{n+1} + \lambda(2u_i^{n+1} - u_{i+1}^{n+1} - u_{i-1}^{n+1}).$$

On montre facilement par récurrence que  $\max_{j=1, \dots, N} u_j^n \leq (1+k)^n \max_{j=1, \dots, N} u_j^0$ , (voir preuve de la stabilité  $L^\infty$  d'Euler implicite page 69) et que  $\min_{j=1, \dots, N} u_j^{(n)} \geq (1+k)^n \min_{j=1, \dots, N} u_j^{(0)}$ . On en déduit que

$$\|u^{(n)}\|_\infty \leq (1+k)^n \|u_0\|_\infty$$

Or  $(1+k)^n \leq (1+k)^{T/k}$  car  $kn \leq T$ . Or

$$\begin{aligned} (1+k)^{T/k} &= \exp\left(\frac{T}{k} \ln(1+k)\right) \\ &\leq \exp\left(\frac{T}{k}\right) = e^T \end{aligned}$$

On en déduit le résultat, avec  $C_1(T) = e^T$ . De même, pour le schéma (2.37), on montre par récurrence que :

$$\|u^{(n)}\|_\infty \leq \frac{1}{(1-k)^n} \|u^{(0)}\|.$$

Mais pour  $k \in ]0, \alpha[$ , avec  $\alpha \in ]0, 1[$ , on a :

$$\frac{1}{(1-k)} \leq 1 + \beta k, \text{ avec } \beta = \frac{1}{(1-\alpha)}.$$

On en déduit par un calcul similaire au précédent que

$$(1-k)^{T/k} \leq e^{\beta T},$$

d'où le résultat avec  $C_2(T, \alpha) = e^{\beta T}$ .

4. Par définition de l'erreur de consistance, on a pour le schéma (2.37)

$$\frac{\bar{u}_j^{(n+1)} - \bar{u}_j^n}{k} - \frac{\bar{u}_{j+1}^{(n+1)} + \bar{u}_{j-1}^{n+1} - 2\bar{u}_j^{(n+1)}}{h^2} - \bar{u}_j^{n+1} = R_i^{(n,1)}$$

et donc, en notant  $e_j^{(n)} = \bar{u}_j^n - u_j^{(n)}$  l'erreur de discrétisation en  $(x_j, t_n)$ , on a :

$$e_j^{(n+1)}(1 + 2\lambda - k) - \lambda e_{j-1}^{(n+1)} - \lambda e_{j+1}^{(n+1)} = e_j^{(n)} + kR_j^{(n,1)}$$

On obtient donc, de manière similaire à la question 3 :

(en considérant  $e_{j_0}^{(n+1)} = \max e_j^{(n+1)}$  puis  $e_{j_0}^{(n+1)} = \min e_j^{(n+1)}$ )

$$\frac{1}{1-k} \|e^{(n+1)}\| \leq \|e^{(n)}\| + kC(k + h^2).$$

Par récurrence sur  $n$ , on obtient alors

$$\|e^{(n)}\|_\infty \leq \left(\frac{1}{1-k}\right)^n [kC(k + h^2) + \|e^0\|_\infty]$$

d'où

$$\|e^{(n)}\|_\infty \leq C_2(T, \alpha)(TC(k + h^2) + \|e^0\|_\infty).$$

De même, pour le schéma (2.38), on écrit l'erreur de consistance :

$$\frac{\bar{u}_j^{(n+1)} - \bar{u}_j^n}{k} - \frac{u_{j+1}^{(n+1)} + \bar{u}_{j-1}^{n+1} - 2\bar{u}_j^{(n+1)}}{h^2} - \bar{u}_j^n = R_j^{(n,2)}$$

et donc :

$$e_j^{(n+1)}(1 + 2\lambda) - \lambda e_{j-1}^{(n+1)} - \lambda e_{j+1}^{(n+1)} = e_j^{(n)}(1 + k) + kR_j^{(n,2)}.$$

Par des raisonnements similaires à ceux de la question 3 on obtient alors :

$$\|e_j^{(n)}\| \leq (1+k)^n (\|e^{(0)}\| + kC(k + h^2))$$

d'où

$$\|e^{(n)}\|_\infty \leq C_1(T)(\|e^{(0)}\| + kC(k + h^2)).$$

### Exercice 30 page 74

1. On s'intéresse ici à l'ordre du schéma au sens des différences finies. On suppose que  $u \in C^4([0, 1] \times [0, T])$  est solution de (2.40) et on pose

$$\bar{u}_j^n = u(jh, nk), \quad j = 0, \dots, N, \quad k = 0, \dots, M.$$

L'erreur de consistance est définie par :

$$R_j^n = \frac{\bar{u}_j^{n+1} - \bar{u}_j^{n-1}}{2k} - \frac{\bar{u}_{j-1}^n - 2\bar{u}_j^n + \bar{u}_{j+1}^n}{h^2}, j = 1, \dots, N-1, k = 1, \dots, M-1.$$

On cherche une majoration de  $R_j^n$  en utilisant des développements de Taylor. Soit  $j \in \{1, \dots, N-1\}$ ,  $k \in \{1, \dots, M-1\}$ . Il existe  $(\xi_i, t_i) \in [0, 1] \times [0, T]$ ,  $i = 1, \dots, 4$ , t.q. :

$$\bar{u}_j^{n+1} = \bar{u}_j^n + ku_t(jh, nk) + \frac{k^2}{2}u_{tt}(jh, nk) + \frac{k^3}{6}u_{ttt}(\xi_1, t_1),$$

$$\bar{u}_j^{n-1} = \bar{u}_j^n - ku_t(jh, nk) + \frac{k^2}{2}u_{tt}(jh, nk) - \frac{k^3}{6}u_{ttt}(\xi_2, t_2),$$

$$\bar{u}_{j-1}^n = \bar{u}_j^n - hu_x(jh, nk) + \frac{h^2}{2}u_{xx}(jh, nk) - \frac{h^3}{6}u_{xxx}(jh, nk) - \frac{h^4}{24}u_{xxxx}(\xi_3, t_3),$$

$$\bar{u}_{j+1}^n = \bar{u}_j^n + hu_x(jh, nk) + \frac{h^2}{2}u_{xx}(jh, nk) + \frac{h^3}{6}u_{xxx}(jh, nk) + \frac{h^4}{24}u_{xxxx}(\xi_4, t_4).$$

On en déduit :

$$R_j^n = u_t(jh, nk) + \frac{k^2}{12}(u_{ttt}(\xi_1, t_1) + u_{ttt}(\xi_2, t_2)) - u_{xx}(jh, nk) - \frac{h^2}{24}(u_{xxxx}(\xi_3, t_3) + u_{xxxx}(\xi_4, t_4)),$$

et donc, comme  $u$  est solution de (2.40),

$$|R_j^n| \leq C_1(k^2 + h^2),$$

où  $C_1$  ne dépend que de  $u$ . Le schéma (2.41) est donc consistant d'ordre 2.

2. Pour étudier la stabilité au sens de Von Neumann, on "oublie" les conditions aux limites dans (2.40). Plus précisément, on s'intéresse à (2.40) avec  $x \in \mathbb{R}$  (au lieu de  $x \in ]0, 1[$ ) et on remplace les conditions aux limites par des conditions de périodicité (exactement comme on l'a vu au paragraphe 2.2.6 page 65). Enfin, on prend une condition initiale de type "mode de Fourier", avec  $p \in \mathbb{R}$  arbitraire, et  $u_0$  défini par :

$$u_0(x) = e^{ipx}, x \in \mathbb{R}.$$

La solution exacte est alors :

$$u(x, t) = e^{-p^2 t} e^{ipx}, x \in \mathbb{R}, t \in \mathbb{R}_+,$$

c'est-à-dire

$$u(\cdot, t) = e^{-p^2 t} u_0, t \in \mathbb{R}_+.$$

Le facteur d'amplification est donc, pour tout  $t \in \mathbb{R}_+$ , le nombre  $e^{-p^2 t}$ . Ce facteur est toujours, en module, inférieur à 1. On va maintenant chercher la solution du schéma numérique sous la forme :

$$u_j^n = \xi_n e^{ipjh}, j \in \mathbb{Z}, n \in \mathbb{N}, \quad (2.72)$$

où  $\xi_0$  et  $\xi_1 \in \mathbb{R}$  sont donnés (ils donnent  $u_j^0$  et  $u_j^1$  pour tout  $j \in \mathbb{Z}$ ) et  $\xi_n \in \mathbb{R}$  est à déterminer de manière à ce que la première équation de (2.41) soit satisfaite.

Ce facteur  $\xi_n$  va dépendre de  $k$ ,  $h$  et  $p$ . Pour  $k$  et  $h$  donnés, le schéma est stable au sens de Von Neumann si, pour tout  $p \in \mathbb{R}$ , la suite  $(\xi_n)_{n \in \mathbb{N}}$  est bornée. Dans le cas contraire, le schéma est (pour ces valeurs de  $k$  et  $h$ ) dit instable au sens de Von Neumann.

Un calcul immédiat donne que la famille des  $u_j^n$ , définie par (2.72), est solution de la première équation si et seulement si la suite  $(\xi_n)_{n \in \mathbb{N}}$  vérifie (on rappelle que  $\xi_0$  et  $\xi_1$  sont donnés) :

$$\frac{\xi_{n+1} - \xi_{n-1}}{2k} = \frac{2}{h^2}(\cos ph - 1)\xi_n, n \geq 2,$$

ou encore, en posant

$$\alpha = \frac{4k}{h^2}(\cos ph - 1) (\leq 0),$$

$$\xi_{n+1} - \alpha\xi_n - \xi_{n-1} = 0, \quad n \geq 2 \quad (2.73)$$

En excluant le cas  $\alpha = -2$  (qui correspond, pour  $k$  et  $h$  donnés, à des valeurs de  $p$  très particulières), la solution de (2.73) est

$$\xi_n = Ar_1^n + Br_2^n, \quad A \geq 0, \quad (2.74)$$

où  $A$  et  $B$  sont déterminés par  $\xi_0$  et  $\xi_1$  (de sorte que  $\xi_0 = A + B$ ,  $\xi_1 = Ar_1 + Br_2$ ) et  $r_1, r_2$  sont les deux racines distinctes de :

$$r^2 - \alpha r - 1 = 0. \quad (2.75)$$

Les nombres  $r_1$  et  $r_2$  sont réels et comme  $r_1 r_2 = 1$ , l'un de ces nombres est, en module, supérieur à 1. Ceci montre que  $(\xi_n)_n$  est une suite non bornée (sauf pour des choix très particuliers de  $\xi_0$  et  $\xi_1$ , ceux pour lesquels  $\xi_1 = \xi_0 r_2$  où  $r_2$  est la racine de (2.75) de module inférieur à 1). Ce schéma est donc instable au sens de Von Neumann, pour tout  $k > 0$  et  $h > 0$ .

3. On reprend les notations de la question 1. On s'intéresse maintenant à la quantité  $S_j^n$  (qui est toujours l'erreur de consistance) :

$$S_j^n = \frac{\bar{u}_j^{n+1} - u_j^{n-1}}{2k} - \frac{\bar{u}_{j-1}^n - (\bar{u}_j^{n+1} + \bar{u}_j^{n-1}) + \bar{u}_{j+1}^n}{h^2}, \quad j = 1, \dots, N-1, \quad k = 0, \dots, M-1.$$

En reprenant la technique de la question 1, il existe  $(\xi_i, t_i)$ ,  $i = 1, \dots, 6$  t.q.

$$S_j^n = \frac{h^2}{12} (u_{ttt}(\xi_1, t_1) + u_{ttt}(\xi_2, t_2)) - \frac{h^2}{24} (u_{xxx}(\xi_3, t_3) - u_{xxx}(\xi_4, t_4)) + \frac{k^2}{2h^2} h_{tt}(\xi_5, t_5) + \frac{k^2}{2h^2} u_{tt}(\xi_6, t_6).$$

Ce qui donne, avec  $C_2$  ne dépendant que de  $u$ ,

$$|S_j^n| \leq C_2 \left( h^2 + k^2 + \frac{k^2}{h^2} \right), \quad j = 1, \dots, N-1, \quad k = 0, \dots, M-1.$$

Le schéma est donc consistant quand  $h \rightarrow 0$  avec  $\frac{k}{h} \rightarrow 0$ .

4. On reprend la méthode développée à la question 2, la suite  $(\xi_n)_n$  doit maintenant vérifier la relation suivante (avec  $\xi_0, \xi_1$  donnés).

$$\frac{\xi_{n+1} - \xi_{n-1}}{2k} = \frac{2 \cos(ph)}{h^2} \xi_n - \frac{\xi_{n-1} + \xi_{n+1}}{h^2}, \quad n \geq 2$$

c'est à dire :

$$\xi_{n+1} \left( \frac{1}{2k} + \frac{1}{h^2} \right) - \frac{2 \cos(ph)}{h^2} \xi_n + \xi_{n-1} \left( \frac{1}{h^2} - \frac{1}{2k} \right) = 0, \quad n \geq 2.$$

L'équation caractéristique est maintenant :

$$p(r) = ar^2 + br + c = 0,$$

avec

$$a = \frac{1}{2k} + \frac{1}{h^2}, \quad b = -\frac{2 \cos(ph)}{h^2} \quad \text{et} \quad c = \frac{1}{h^2} - \frac{1}{2k}.$$

Pour montrer la stabilité au sens de Von Neumann, il suffit d'après (2.74) de montrer que les deux racines du polynôme  $p$  sont de module inférieur ou égal à 1. On note  $r_1$  et  $r_2$  ces deux racines (qui peuvent être confondues) et on distingue 2 cas :

1. 1er cas : Les racines de  $p$  ne sont pas réelles. Dans ce cas, on a  $|r_1| = |r_2| = \gamma$  et

$$\gamma = \left| \frac{c}{a} \right| < 1,$$

car  $k > 0$ .

2. 2ème cas : Les racines de  $p$  sont réelles. Dans ce cas, on remarque que

$$r_1 r_2 = \frac{c}{a} < 1,$$

et l'une des racines, au moins, est donc entre  $-1$  et  $1$  (strictement). De plus on a  $p(1) = \frac{2}{h^2} - \frac{2 \cos ph}{h^2} \geq 0$

et  $p(-1) = \frac{2}{h^2} + \frac{2 \cos ph}{h^2} \geq 0$ , l'autre racine est donc aussi entre  $-1$  et  $1$  (au sens large).

On en déduit que le schéma (2.42) est stable au sens de Von Neumann.

**Exercice 32 page 77 : Discrétisation d'un problème parabolique non linéaire**

On se propose, dans cet exercice, de montrer l'existence d'une solution faible au problème (2.48)-(2.50), à partir de l'existence de la solution approchée donnée par un schéma numérique. L'inconnue de ce problème est la fonction  $u$  de  $]0, 1[ \times ]0, T[$  dans  $\mathbb{R}$ , elle doit être solution des équations suivantes :

$$\frac{\partial u}{\partial t}(x, t) - \frac{\partial^2 \varphi(u)}{\partial x^2}(x, t) = v(x, t), \quad x \in ]0, 1[, \quad t \in ]0, T[, \quad (2.76)$$

$$\frac{\partial \varphi(u)}{\partial x}(0, t) = \frac{\partial \varphi(u)}{\partial x}(1, t) = 0, \quad t \in ]0, T[, \quad (2.77)$$

$$u(x, 0) = u_0(x), \quad x \in ]0, 1[, \quad (2.78)$$

où  $\varphi, v, T, u_0$  sont donnés et sont t.q.

1.  $T > 0, v \in L^\infty(]0, 1[ \times ]0, T[)$ ,
2.  $\varphi$  croissante, lipschitzienne de  $\mathbb{R}$  dans  $\mathbb{R}$ ,
3.  $u_0 \in L^\infty(]0, 1[)$  et  $\varphi(u_0)$  lipschitzienne de  $]0, 1[$  dans  $\mathbb{R}$ .

Un exemple important est donné par  $\varphi(s) = \alpha_1 s$  si  $s \leq 0$ ,  $\varphi(s) = 0$  si  $0 \leq s \leq L$  et  $\varphi(s) = \alpha_2(s - L)$  si  $s \geq L$ , avec  $\alpha_1, \alpha_2$  et  $L$  donnés dans  $\mathbb{R}_+^*$ . Noter pour cet exemple que  $\varphi' = 0$  sur  $]0, L[$ .

Les ensembles  $]0, 1[$  et  $D = ]0, 1[ \times ]0, T[$  sont munis de leur tribu borélienne et de la mesure de Lebesgue sur cette tribu.

On appelle "solution faible" de (2.48)-(2.50) une solution de :

$$u \in L^\infty(]0, 1[ \times ]0, T[), \quad (2.79)$$

$$\int_D (u(x, t) \frac{\partial \psi}{\partial t}(x, t) + \varphi(u(x, t)) \frac{\partial^2 \psi}{\partial x^2}(x, t) + v(x, t) \psi(x, t)) dx dt + \int_{]0, 1[} u_0(x) \psi(x, 0) dx = 0, \quad (2.80)$$

$$\forall \psi \in C_T^{2,1}(\mathbb{R}^2),$$

où  $\psi \in C_T^{2,1}(\mathbb{R}^2)$  signifie que  $\psi$  est une fonction de  $\mathbb{R}^2$  dans  $\mathbb{R}$  deux fois continûment dérivable par rapport à  $x$ , une fois continûment dérivable par rapport à  $t$  et t.q.  $\frac{\partial \psi}{\partial x}(0, t) = \frac{\partial \psi}{\partial x}(1, t) = 0$ , pour tout  $t \in [0, T]$  et  $\psi(x, T) = 0$  pour tout  $x \in [0, 1]$ .

**Question 1 (Solution classique versus solution faible)**

Soit  $u \in C^2(\mathbb{R}^2, \mathbb{R})$ ; notons  $u$  sa restriction à  $D = ]0, 1[ \times ]0, T[$ ; notons que l'on a bien  $u \in L^\infty(]0, 1[ \times ]0, T[)$ . Supposons que  $u$  satisfait (2.48)-(2.50), et montrons qu'alors  $u$  vérifie (2.52). Soit  $\psi \in C_T^{2,1}(\mathbb{R}^2)$ . Multiplions (2.48) par  $\psi$  et intégrons sur  $D$ . On obtient :

$$\int_D \frac{\partial u}{\partial t}(x, t) \psi(x, t) dx dt - \int_D \frac{\partial^2 \varphi(u)}{\partial x^2}(x, t) \psi(x, t) dx dt = \int_D v(x, t) \psi(x, t) dx dt. \quad (2.81)$$

Par intégration par parties, il vient :

$$\int_D \frac{\partial u}{\partial t}(x, t) \psi(x, t) dx dt = \int_0^1 u(x, T) \psi(x, T) dx - \int_0^1 u(x, 0) \psi(x, 0) dx - \int_D u(x, t) \frac{\partial \psi}{\partial t}(x, t) dx dt.$$

Comme  $\psi \in C_T^{2,1}(\mathbb{R}^2)$  on a donc  $\psi(x, T) = 0$  pour tout  $x \in [0, 1]$  et comme  $u$  vérifie (2.50), on a  $u(x, 0) = u_0(x)$ . On en déduit que

$$\int_D \frac{\partial u}{\partial t}(x, t) \psi(x, t) dx dt = - \int_0^1 u_0(x) \psi(x, 0) dx - \int_D \frac{\partial \psi}{\partial t}(x, t) u(x, t) dx dt. \quad (2.82)$$

Intégrons par parties le deuxième terme de (2.81) :

$$\begin{aligned} \int_D \frac{\partial^2 \varphi(u)}{\partial x^2}(x, t) \psi(x, t) dx dt &= \int_0^T \left[ \frac{\partial \varphi(u)}{\partial x}(1, t) \psi(1, t) - \frac{\partial \varphi(u)}{\partial x}(0, t) \psi(0, t) \right] dt \\ &\quad - \int_D \frac{\partial \varphi(u)}{\partial x}(x, t) \frac{\partial \psi(u)}{\partial x}(x, t) dx dt. \end{aligned}$$

et comme  $u$  vérifie (2.49), on a

$$\frac{\partial \varphi(u)}{\partial x}(0, t) = \frac{\partial \varphi(u)}{\partial x}(1, t) = 0, \quad t \in ]0, T[.$$

En tenant compte de ces relations et en ré-intégrant par parties, on obtient :

$$\int_D \frac{\partial^2 \varphi(u)}{\partial x^2}(x, t) \psi(x, t) dx dt = - \int_D \varphi(u)(x, t) \frac{\partial^2 \psi(u)}{\partial x^2}(x, t) dx dt. \quad (2.83)$$

En remplaçant dans (2.82) et (2.83) dans (2.81), on obtient (2.48).

Réciproquement, supposons que  $u$  satisfait (2.52), et soit  $\psi$  continûment différentiable à support compact dans  $D$ . En intégrant (2.52) par parties et en tenant compte que  $\psi$  et toutes ses dérivées sont nulles au bord de  $D$ , on obtient :

$$\int_D \left[ -\frac{\partial u}{\partial t}(x, t) + \frac{\partial^2 \varphi(u)}{\partial x^2}(x, t) - v(x, t) \right] \psi(x, t) dx dt = 0, \quad \forall \psi \in C_c^\infty(D).$$

Comme  $u$  est régulière, ceci entraîne que l'équation (2.48) est donc satisfaite par  $u$ .

On prend ensuite  $\psi \in C_T^{2,1}(\mathbb{R}^2)$ , et on intègre (2.52) par parties. En tenant compte du fait que  $\psi(x, T) = 0$ , pour tout  $x$  et  $\frac{\partial \psi}{\partial x}(0, t) = \frac{\partial \psi}{\partial x}(1, t) = 0$ , pour tout  $t$ , on obtient :

$$\begin{aligned} & - \int_0^1 u(x, 0) \psi(x, 0) dx - \int_D \frac{\partial u}{\partial t}(x, t) \psi(x, t) dx dt + \int_0^T \frac{\partial \varphi(u)}{\partial x}(1, t) \psi(1, t) dt \\ & - \int_0^T \frac{\partial \varphi(u)}{\partial x}(0, t) \psi(0, t) dt + \int_D \frac{\partial^2 \varphi(u)}{\partial x^2} \psi(x, t) dx dt + \int_0^1 u_0(x) \psi(x, 0) dx = 0. \end{aligned}$$

En regroupant et en utilisant le fait que  $u$  satisfait (2.48), on obtient :

$$\int_0^1 (u_0(x) - u(x, 0)) \psi(x, 0) dx + \int_0^T \frac{\partial \varphi}{\partial x}(1, t) \psi(1, t) dt - \int_0^T \frac{\partial \varphi}{\partial x}(0, t) \psi(0, t) dt = 0.$$

En choisissant successivement une fonction  $\psi$  nulle en  $x = 0$  et  $x = 1$  puis nulle en  $x = 1$  et  $t = T$  et enfin nulle en  $x = 0$  et  $t = T$ , on obtient que  $u$  satisfait la condition initiale (2.50) et les conditions aux limites (2.49), ce qui conclut la question.

### Question 2 (Existence et unicité de la solution approchée)

Soit  $n \in \{0, \dots, M-1\}$ . On suppose connu  $\{u_i^n, i = 1, \dots, N\}$ . On va prouver dans cette question l'existence et l'unicité de  $\{u_i^{n+1}, i = 1, \dots, N\}$  vérifiant (2.54) (avec  $u_0^{n+1} = u_1^{n+1}$ ,  $u_{N+1}^{n+1} = u_N^{n+1}$ ).

1. L'application  $s \mapsto s$  est strictement croissante, et par hypothèse sur  $\varphi$ , l'application  $s \mapsto a\varphi(s)$  est croissante. La somme d'une fonction strictement croissante et d'une fonction croissante est strictement croissante. D'autre part, comme  $\varphi$  est croissante, pour tout  $\varphi(s) \leq \varphi(0)$ ,  $\forall s \leq 0$ , et donc  $\lim_{s \rightarrow -\infty} g_a(s) = -\infty$ . De même,  $\varphi(s) \geq \varphi(0)$ ,  $\forall s \geq 0$ , et donc  $\lim_{s \rightarrow +\infty} g_a(s) = +\infty$ . La fonction  $g_a$  est continue et prend donc toutes les valeurs de l'intervalle  $] -\infty, +\infty[$ . Comme elle est strictement croissante, elle est bijective.
2. L'équation (4.5) s'écrit encore :

$$g_a(u_i) = \frac{k}{h^2} (\bar{w}_{i-1} + \bar{w}_{i+1}) + u_i^n + kv_i^n, \quad \text{pour tout } i = 1, \dots, N,$$

avec  $a = \frac{k}{h^2}$ . Par la question précédente, il existe donc un unique  $u_i$  qui vérifie cette équation ; il suffit alors de poser  $\varphi(u_i) = w_i$  pour déterminer de manière unique la solution de (2.55)–(2.56).

3. Soit  $\bar{w}^1$  et  $\bar{w}^2 \in \mathbb{R}^N$  et soit  $w^1 = F(\bar{w}^1)$  et  $w^2 = F(\bar{w}^2)$ . Par définition de  $F$ , on a :

$$u_i^1 - u_i^2 + \frac{2k}{h^2} (w_i^1 - w_i^2) = \frac{k}{h^2} ((\bar{w}_{i-1}^1 + \bar{w}_{i+1}^1) - (\bar{w}_{i-1}^2 + \bar{w}_{i+1}^2)), \quad \text{pour tout } i = 1, \dots, N. \quad (2.84)$$

Comme  $\varphi$  est monotone, le signe de  $w_i^1 - w_i^2 = \varphi(u_i^1) - \varphi(u_i^2)$  est le même que celui de  $u_i^1 - u_i^2$ , et donc

$$|u_i^1 - u_i^2 + \frac{2k}{h^2} (w_i^1 - w_i^2)| = |u_i^1 - u_i^2| + \frac{2k}{h^2} |w_i^1 - w_i^2|. \quad (2.85)$$

Et comme  $\varphi$  est lipschitzienne de rapport  $L$ , on a

$$|w_i^1 - w_i^2| = |\varphi(u_i^1) - \varphi(u_i^2)| \leq L|u_i^1 - u_i^2|,$$

d'où :

$$|u_i^1 - u_i^2| \geq \frac{1}{L}|w_i^1 - w_i^2|. \quad (2.86)$$

On déduit donc de (2.84),(2.85) et(2.86) que

$$\frac{1}{L}|w_i^1 - w_i^2| + \frac{2k}{h^2}|w_i^1 - w_i^2| \leq \frac{k}{h^2}(|\bar{w}_{i-1}^1 - \bar{w}_{i-1}^2| + |\bar{w}_{i+1}^1 - \bar{w}_{i+1}^2|), \text{ pour tout } i = 1, \dots, N.$$

On a donc

$$|w_i^1 - w_i^2| \leq \frac{1}{1 + \frac{h^2}{2kL}} \max_{i=1, \dots, N} |\bar{w}_i^1 - \bar{w}_i^2|, \text{ pour tout } i = 1, \dots, N.$$

d'où on déduit que  $\|w^1 - w^2\|_\infty \leq C\|\bar{w}^1 - \bar{w}^2\|_\infty$  avec  $C = \frac{1}{1 + \frac{h^2}{2kL}} < 1$ . L'application  $F$  est donc bien strictement contractante.

4. Soit  $\{u_i^{n+1}, i = 1, \dots, N\}$  est solution de (2.54) et  $w_i = \varphi(u_i^{n+1})$  pour  $i \in \{1, \dots, N\}$ , alors on remarque que  $(u_i^{n+1})_{i=1, \dots, N}$  et  $(w_i)_{i=1, \dots, N}$  vérifient (2.55)–(2.56) avec  $\bar{w}_i = w_i$  pour  $i = 1, \dots, N$ . On en déduit que  $w = F(w)$ .

5. Soit  $w = (w_i)_{i=1, \dots, N}$  t.q.  $w = F(w)$ . Montrer que pour tout

Par définition de  $F$ , on a  $F(w) = \tilde{w}$  avec  $(\tilde{u}, \tilde{w}) \in \mathbb{R}^N \times \mathbb{R}^N$ ,  $\tilde{u} = (\tilde{u}_i)_{i=1, \dots, N}$ ,  $\tilde{w} = (\tilde{w}_i)_{i=1, \dots, N}$ , t.q. :

$$\varphi(\tilde{u}_i) = \tilde{w}_i, \text{ pour tout } i \in \{1, \dots, N\},$$

$$\tilde{u}_i + \frac{2k}{h^2}\tilde{w}_i = \frac{k}{h^2}(w_{i-1} + w_{i+1}) + u_i^n + kv_i^n, \text{ pour tout } i = 1, \dots, N. \quad (2.87)$$

Comme  $F(w) = w$ , on a donc  $\tilde{w}_i = w_i$  et on obtient l'existence de  $u_i^{n+1} = \tilde{u}_i$  tel que  $w_i = \varphi(u_i^{n+1})$  pour  $i = 1, \dots, N$ . Il suffit alors de remplacer  $w_i$  et  $\tilde{w}_i$  par  $\varphi(u_i^{n+1})$  dans (2.87) pour conclure que  $\{u_i^{n+1}, i = 1, \dots, N\}$  est solution de (2.54).

6. On vient de montrer dans les questions précédentes que  $\{u_i^{n+1}, i = 1, \dots, N\}$  est solution de (2.54) si et seulement si  $w$  défini par  $w_i = \varphi(u_i^{n+1})$  est solution de  $w = F(w)$ , où  $F$  est définie par (2.55) – (2.56). Comme  $F$  est une application strictement croissante, il existe un unique point fixe  $w = F(w)$ . Donc par définition de  $F$  il existe une unique famille  $\{u_i^{n+1}, i = 1, \dots, N\}$  solution de (2.54).

**Question 3** (Estimation  $L^\infty$  ( $]0, 1[ \times ]0, T[$ ) sur  $u$ )

La relation à démontrer par récurrence est clairement vérifiée au rang  $n = 0$ , par définition de  $A$ . Supposons qu'elle soit vraie jusqu'au rang  $n$ , et démontrons-la au rang  $n + 1$ . La relation (2.54) s'écrit encore :

$$u_i^{n+1} = u_i^n + \frac{k}{h^2}(\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1})) + \frac{k}{h^2}(\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1})) + kv_i^n, \quad i = 1, \dots, N, \quad n = 0, \dots, M - 1,$$

Supposons que  $i$  est tel que  $u_i^{n+1} = \min_{j=1, \dots, N} u_j^{n+1}$ . Comme  $\varphi$  est croissante, on a dans ce cas :  $\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1}) \geq 0$  et  $\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}) \geq 0$ , et on en déduit que  $\min_{j=1, \dots, N} u_j^{n+1} \geq u_i^n - kB$  d'où, par hypothèse de récurrence,  $\min_{j=1, \dots, N} u_j^{n+1} \geq -A - nkB - kB$ . Un raisonnement similaire en considérant maintenant  $i$  tel que  $u_i^{n+1} = \max_{j=1, \dots, N} u_j^{n+1}$  conduit à :  $\max_{j=1, \dots, N} u_j^{n+1} \leq u_i^n + kB \leq A + nkB + kB$ . On a donc bien :  $-A - (n + 1)kB \leq u_i^{n+1} \leq A + (n + 1)kB$ , pour tout  $i = 1, \dots, N$  et tout  $n = 0, \dots, M$ .

On en déduit alors que  $\|u^n\|_{L^\infty(]0,1])} \leq c_{u_0, v, T}$ , avec  $c_{u_0, v, T} = A + BT$ .

**Question 4** (Estimation de la dérivée p.r. à  $x$  de  $\varphi(u)$ )

En multipliant (2.54) par  $u_i^{n+1}$  et en sommant sur  $i$ , on obtient  $A_n + B_n = C_n$ , avec

$$A_n = \sum_{i=1}^N \frac{u_i^{n+1} - u_i^n}{k} u_i^{n+1}, \quad B_n = - \sum_{i=1}^N \frac{\varphi(u_{i-1}^{n+1}) - 2\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})}{h^2} u_i^{n+1} \text{ et } C_n = \sum_{i=1}^N v_i^n u_i^{n+1}.$$

En utilisant l'inégalité  $a^2 - ab = \frac{a^2}{2} - \frac{b^2}{2}$ , on obtient :

$$A_n \geq \alpha_{n+1} - \alpha_n, \text{ avec } \alpha_n = \frac{1}{2k} \sum_{i=1}^N (u_i^n)^2.$$



En développant  $B_n$ , on obtient :

$$B_n = -\frac{1}{h^2} \left( \sum_{i=1}^N (\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1})) u_i^{n+1} + \sum_{i=1}^N (-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})) u_i^{n+1} \right).$$

Par un changement d'indice sur les sommes, on obtient alors :

$$B_n = -\frac{1}{h^2} \left( \sum_{i=0}^{N-1} (\varphi(u_i^{n+1}) - \varphi(u_{i+1}^{n+1})) u_{i+1}^{n+1} - \sum_{i=1}^N (-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})) u_i^{n+1} \right).$$

En tenant compte du fait que  $u_0^{n+1} = u_1^{n+1}$ ,  $u_{N+1}^{n+1} = u_N^{n+1}$ , pour tout  $n = 0, \dots, M-1$ , on obtient alors que :

$$B_n = \frac{1}{h^2} \left( \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1})) (u_{i+1}^{n+1} - u_i^{n+1}) \right).$$

En utilisant le caractère lipschitzien de  $\varphi$ , on obtient la minoration suivante :

$$B_n \geq \frac{1}{Lh^2} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2.$$

Enfin, on majore  $C_n$  :

$$C_n \leq \frac{Bc_{u_0,v,T}}{h}.$$

L'égalité  $A_n + B_n = C_n$  entraîne donc :

$$\alpha_{n+1} - \alpha_n + \frac{1}{Lh^2} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq \frac{Bc_{u_0,v,T}}{h}.$$

En sommant pour  $n = 0$  à  $M-1$ , et en notant que  $\alpha_M \geq 0$ , on obtient alors :

$$\frac{1}{Lh^2} \sum_{n=0}^{M-1} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq \frac{Bc_{u_0,v,T}}{h} + \alpha_0.$$

Il reste à remarquer que  $\alpha_0 \leq \frac{h}{2k} c_{u_0,v,T}^2$  pour conclure que :

$$\sum_{n=0}^{M-1} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))^2 \leq C_1 \frac{h}{k}, \text{ avec } C_1 = Lc_{u_0,v,T} \left( B + \frac{1}{2} c_{u_0,v,T} \right).$$

**Question 5** (Estimation de la dérivée p.r. à  $t$  de  $\varphi(u)$ )

Multiplicons (2.54) par  $\varphi(u_i^{n+1}) - \varphi(u_i^n)$  et sommons pour  $i = 1, \dots, N$ . On obtient :

$$A_n + B_n = C_n, \tag{2.88}$$

$$\text{avec } A_n = \sum_{i=1}^N \frac{u_i^{n+1} - u_i^n}{k} (\varphi(u_i^{n+1}) - \varphi(u_i^n)), \quad B_n = -\sum_{i=1}^N \frac{\varphi(u_{i-1}^{n+1}) - 2\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1}))}{h^2} (\varphi(u_i^{n+1}) - \varphi(u_i^n))$$

et  $C_n = \sum_{i=1}^N v_i^n (\varphi(u_i^{n+1}) - \varphi(u_i^n))$ .

En utilisant le caractère lipschitzien de  $\varphi$ , on obtient la minoration suivante :

$$A_n \geq \frac{1}{Lk} \sum_{i=1}^{N-1} (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2. \tag{2.89}$$

En développant  $B_n$ , on obtient :

$$B_n = -\frac{1}{h^2} \left( \sum_{i=1}^N (\varphi(u_{i-1}^{n+1}) - \varphi(u_i^{n+1})) (\varphi(u_i^{n+1}) - \varphi(u_i^n)) + \sum_{i=1}^N (-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1})) (\varphi(u_i^{n+1}) - \varphi(u_i^n)) \right).$$

Par un changement d'indice sur les sommes, on obtient alors :

$$B_n = -\frac{1}{h^2} \left( \sum_{i=0}^{N-1} (\varphi(u_i^{n+1}) - \varphi(u_{i+1}^{n+1}))(\varphi(u_{i+1}^{n+1}) - \varphi(u_{i+1}^n)) + \sum_{i=1}^N (-\varphi(u_i^{n+1}) + \varphi(u_{i+1}^{n+1}))(\varphi(u_i^{n+1}) - \varphi(u_i^n)) \right).$$

En tenant compte du fait que  $u_0^{n+1} = u_1^{n+1}$ ,  $u_{N+1}^{n+1} = u_N^{n+1}$ , pour tout  $n = 0, \dots, M-1$ , on obtient alors que :

$$B_n = \frac{1}{h^2} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1}))((\varphi(u_{i+1}^{n+1}) - \varphi(u_i^{n+1})) - (\varphi(u_{i+1}^n) - \varphi(u_i^n))).$$

En utilisant à nouveau la relation  $a(a-b) \geq \frac{a^2}{2} - \frac{b^2}{2}$ , on obtient :

$$B_n \geq \beta_{n+1} - \beta_n, \text{ avec } \beta_n = \frac{1}{2h^2} \sum_{i=1}^{N-1} (\varphi(u_{i+1}^n) - \varphi(u_i^n))^2 \quad (2.90)$$

Enfin, on majore  $C_n$  par :

$$C_n \leq \frac{1}{2Lk} \sum_{i=1}^N (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2 + C \sum_{i=1}^N k \geq C \frac{k}{h}. \quad (2.91)$$

En utilisant (2.88), (2.89), (2.90) et (2.91), on obtient :

$$\frac{1}{2Lk} \sum_{i=1}^N (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2 + \beta_{n+1} - \beta_n \leq C \frac{k}{h}. \quad (2.92)$$

En sommant sur  $n$ , on obtient d'une part, en utilisant le fait que  $\beta_n \geq 0$  :

$$\sum_{n=0}^{M-1} \sum_{i=1}^N (\varphi(u_i^{n+1}) - \varphi(u_i^n))^2 \leq 2LC \frac{k}{h} + 2L\beta_0 k. \quad (2.93)$$

d'autre part, en utilisant que le fait que le premier terme est positif, on obtient par (2.92) une majoration sur  $\beta_M$ , et donc sur  $\beta_n$  pour tout  $n \leq M$  :

$$\beta_n \leq \frac{C}{h} + \beta_0. \quad (2.94)$$

Il ne reste donc plus qu'à majorer  $\beta_0$  pour obtenir (2.58) et (2.59). Par définition, on a

$$\beta_0 = \sum_{i=1}^{N-1} \frac{\varphi(u_i^0) - \varphi(u_{i+1}^0)}{2h^2}.$$

En utilisant le fait que  $\varphi$  est lipschitzienne et que la différence entre  $u_i^0$  et  $u_{i+1}^0$  est en  $h$ , on obtient (2.59) à partir de (2.93) et (2.58) à partir de (2.94).

**Question 6** Par définition de la fonction  $u_h$ , et grace au résultat de la question 3, on a :

$$\begin{aligned} \sup_{\substack{x \in [(i-1)h, ih[ \\ t \in [nk, (n+1)k[}} u_h(x, t) &\leq \frac{t - nk}{k} \|u_h^{(n+1)}\|_\infty + \frac{(n+1)k - t}{k} \|u_h^{(n)}\|_\infty \\ &\leq c_{u_0, v, T}. \end{aligned}$$

ce qui prouve que la suite  $(u_h)_{M \in \mathbb{N}^*}$  est bornée dans  $L^\infty([0, 1[ \times ]0, T[)$ . Comme  $\varphi$  est continue, on en déduit immédiatement que  $(\varphi(u_h))_{M \in \mathbb{N}^*}$  est bornée dans  $L^\infty([0, 1[ \times ]0, T[)$

# Chapitre 3

## Méthodes variationnelles

### 3.1 Exemples de problèmes variationnels

#### 3.1.1 Le problème de Dirichlet

Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^d$ ,  $d \geq 1$ . On considère le problème suivant :

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

où  $f \in C(\bar{\Omega})$  et  $\Delta u = \partial_1^2 u + \partial_2^2 u$ , où l'on désigne par  $\partial_i^2 u$  la dérivée partielle d'ordre 2 par rapport à la  $i$ -ème variable.

**Définition 3.1** On appelle solution classique de (3.1) une fonction  $u \in C^2(\bar{\Omega})$  qui vérifie (3.1).

Soit  $u \in C^2(\bar{\Omega})$  une solution classique de (3.1), et soit  $\varphi \in C_c^\infty(\Omega)$ , où  $C_c^\infty(\Omega)$  désigne l'ensemble des fonctions de classe  $C^\infty$  à support compact dans  $\Omega$ . On multiplie (3.1) par  $\varphi$  et on intègre sur  $\Omega$  (on appellera par la suite  $\varphi$  "fonction test") : on a donc :

$$\int_{\Omega} -\Delta u(x)\varphi(x)dx = \int_{\Omega} f(x)\varphi(x)dx.$$

Notons que ces intégrales sont bien définies, puisque  $\Delta u \in C(\Omega)$  et  $f \in C(\Omega)$ . Par intégration par parties (formule de Green), on a :

$$\begin{aligned} \int_{\Omega} -\Delta u(x)\varphi(x)dx &= -\sum_{i=1}^d \int_{\Omega} \partial_i^2 u(x)\varphi(x)dx \\ &= \sum_{i=1}^d \int_{\Omega} \partial_i u(x)\partial_i \varphi(x)dx + \sum_{i=1}^d \int_{\partial\Omega} \partial_i u \cdot n_i(s)\varphi(s)d\gamma(s) \end{aligned}$$

où  $n_i$  désigne la  $i$ -ème composante du vecteur unitaire normal à la frontière  $\partial\Omega$  de  $\Omega$ , et extérieur à  $\Omega$ , et  $d\gamma$  désigne le symbole d'intégration sur  $\partial\Omega$ . Comme  $\varphi$  est nulle sur  $\partial\Omega$ , on obtient :

$$\sum_{i=1}^d \int_{\Omega} \partial_i u(x)\partial_i \varphi(x)dx = \int_{\Omega} f(x)\varphi(x)dx.$$

ce qui s'écrit encore :

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x)dx = \int_{\Omega} f(x)\varphi(x)dx. \quad (3.2)$$

Donc toute solution classique de (3.1) satisfait (3.2)

Prenons maintenant comme fonction test  $\varphi$ , non plus une fonction de  $C_c^\infty(\Omega)$ , mais une fonction de  $H_0^1(\Omega)$ . On rappelle que l'espace  $H_0^1(\Omega)$  est défini comme l'adhérence de  $C_c^\infty(\Omega)$  dans  $H^1(\Omega) = \{u \in L^2(\Omega); Du \in L^2(\Omega)\}$ , où  $Du$  désigne la dérivée faible de  $u$ , voir par exemple [1]. On rappelle que l'espace  $H^1(\Omega)$  muni du produit scalaire

$$(u, v)_{H^1} = \int_{\Omega} u(x)v(x)dx + \sum_{i=1}^d \int_{\Omega} D_i u(x)D_i v(x)dx \quad (3.3)$$

est un espace de Hilbert. Les espaces  $H^1(\Omega)$  et  $H_0^1(\Omega)$  font partie des espaces dits “de Sobolev” (voir [1] pour une introduction).

Si  $\varphi \in H_0^1(\Omega)$ , par définition, il existe  $(\varphi_n)_{n \in \mathbb{N}} \subset C_c^\infty(\Omega)$  telle que

$$\varphi_n \rightarrow \varphi \text{ dans } H^1 \text{ lorsque } n \rightarrow +\infty,$$

Soit encore

$$\|\varphi_n - \varphi\|_{H^1} = \|\varphi_n - \varphi\|_{L^2}^2 + \sum \|D_i \varphi_n - D_i \varphi\|_{L^2}^2 \rightarrow 0 \text{ lorsque } n \rightarrow +\infty.$$

Pour chaque fonction  $\varphi_n \in C_c^\infty(\Omega)$  on a par (3.2) :

$$\sum_{i=1}^N \int_{\Omega} \partial_i u(x) \partial_i \varphi_n(x) dx = \int_{\Omega} f(x) \varphi_n(x) dx, \quad \forall n \in \mathbb{N}.$$

Or la  $i$ -ème dérivée partielle  $\partial_i \varphi_n = \frac{\partial \varphi_n}{\partial x_i}$  converge vers  $D_i \varphi$  dans  $L^2$  donc dans  $L^2$  faible lorsque  $n$  tend vers  $\infty$ , et  $\varphi_n$  tend vers  $\varphi$  dans  $L^2(\Omega)$ . On a donc :

$$\int_{\Omega} \partial_i u(x) \partial_i \varphi_n(x) dx \rightarrow \int_{\Omega} \partial_i u(x) D_i \varphi(x) dx \text{ lorsque } n \rightarrow +\infty$$

et

$$\int_{\Omega} f(x) \varphi_n(x) dx \rightarrow \int_{\Omega} f(x) \varphi(x) dx \text{ lorsque } n \rightarrow +\infty.$$

L'égalité (3.1.1) est donc vérifiée pour toute fonction  $\varphi \in H_0^1(\Omega)$ . Montrons maintenant que si  $u$  est solution classique (3.1) alors  $u \in H_0^1(\Omega)$ . En effet, si  $u \in C^2(\Omega)$ , alors  $u \in C(\bar{\Omega})$  et donc  $u \in L^2(\Omega)$  ; de plus  $\partial_i u \in C(\bar{\Omega})$  donc  $\partial_i u \in L^2(\Omega)$ . On a donc bien  $u \in H^1(\Omega)$ . Il reste à montrer que  $u \in H_0^1(\Omega)$ . Pour cela on rappelle (ou on admet . . .) les théorèmes de trace suivant :

**Théorème 3.2 (Existence de l'opérateur trace)** *Soit  $\Omega$  un ouvert (borné ou non borné) de  $\mathbb{R}^d$ ,  $d \geq 1$ , de frontière  $\partial\Omega$  lipschitzienne, alors l'espace  $C_c^\infty(\bar{\Omega})$  des fonctions de classe  $C^\infty$  et à support compact dans  $\bar{\Omega}$  est dense dans  $H^1(\Omega)$ . On peut donc définir par continuité l'application “trace”, qui est linéaire continue de  $H^1(\Omega)$  dans  $L^2(\partial\Omega)$ , définie par :*

$$\gamma(u) = u|_{\partial\Omega} \text{ si } u \in C_c^\infty(\bar{\Omega})$$

et par

$$\gamma(u) = \lim_{n \rightarrow +\infty} \gamma(u_n) \text{ si } u \in H^1(\Omega), \quad u = \lim_{n \rightarrow +\infty} u_n, \text{ où } (u_n)_{n \in \mathbb{N}} \subset C_c^\infty(\bar{\Omega}).$$

Dire que l'application (linéaire)  $\gamma$  est continue est équivalent à dire qu'il existe  $C \in \mathbb{R}_+$  tel que

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C \|u\|_{H^1(\Omega)} \text{ pour tout } u \in H^1(\Omega). \quad (3.4)$$

Notons que  $\gamma(H^1(\Omega)) \subset L^2(\partial\Omega)$ , mais  $\gamma(H^1(\Omega)) \neq L^2(\partial\Omega)$ . On note  $H^{1/2}(\Omega) = \gamma(H^1(\Omega))$ .

Remarquons que si  $\Omega$  est un ouvert borné, alors  $\bar{\Omega}$  est compact et donc toutes les fonctions  $C^\infty$  sont à support compact dans  $\bar{\Omega}$ .

**Théorème 3.3 (Noyau de l'opérateur trace)** *Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^d$  de frontière  $\partial\Omega$  lipschitzienne, et  $\gamma$  l'opérateur trace défini par le théorème (3.2). Alors*

$$\text{Ker } \gamma = H_0^1(\Omega).$$

Si  $u \in C^2(\bar{\Omega})$  est une solution classique de (3.1), alors  $\gamma(u) = u|_{\partial\Omega} = 0$  donc  $u \in \text{Ker } \gamma$ , et par le théorème 3.3, ceci prouve que  $u \in H_0^1(\Omega)$ .

Nous avons ainsi montré que toute solution classique de (3.1) vérifie  $u \in H_0^1(\Omega)$  et l'égalité (3.2). Cette remarque motive l'introduction de solutions plus générales, qui permettent de s'affranchir de la régularité  $C^2$ , et qu'on appellera “solutions faibles”.

**Définition 3.4 (Formulation faible)** Soit  $f \in L^2(\Omega)$ , on dit que  $u$  est solution faible de (3.1) si  $u$  est solution de

$$\begin{cases} u \in H_0^1(\Omega), \\ \sum_{i=1}^N \int_{\Omega} D_i u(x) D_i \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx, \forall \varphi \in H_0^1(\Omega). \end{cases} \quad (3.5)$$

**Définition 3.5 (Formulation variationnelle)** Soit  $f \in L^2(\Omega)$ ; on dit que  $u$  est solution variationnelle de (3.1) si  $u$  est solution du problème de minimisation suivant :

$$\begin{cases} u \in H_0^1(\Omega) \\ J(u) \leq J(v) \quad \forall v \in H_0^1(\Omega) \end{cases} \quad (3.6)$$

avec  $J(v) = \frac{1}{2} \int_{\Omega} \nabla v(x) \cdot \nabla v(x) dx - \int_{\Omega} f(x) v(x) dx,$

où on a noté :

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx = \sum_{i=1}^d \int_{\Omega} D_i u(x) D_i \varphi(x) dx.$$

On cherche à montrer l'existence et l'unicité de la solution de (3.5) et (3.6). Pour cela, on utilise le théorème de Lax-Milgram, qu'on rappelle ici :

**Théorème 3.6 (Lax-Milgram)** Soit  $H$  un espace de Hilbert, soit  $a$  une forme bilinéaire continue coercive sur  $H$  et  $T \in H'$ . Il existe un unique élément  $u$  tel que

$$\begin{cases} u \in H, \\ a(u, v) = T(v). \quad \forall v \in H. \end{cases} \quad (3.7)$$

De plus, si  $a$  est symétrique,  $u$  est l'unique solution du problème de minimisation suivant :

$$\begin{cases} u \in H, \\ J(u) \leq J(v), \end{cases} \quad (3.8)$$

où  $J$  est définie de  $H$  dans  $\mathbb{R}^N$  par :

$$J(v) = \frac{1}{2} a(v, v) - T(v). \quad (3.9)$$

**Démonstration :**

- Si  $a$  est symétrique l'existence et l'unicité de  $u$  est immédiate par le théorème de représentation de Riesz (car dans ce cas  $a$  est un produit scalaire, et la forme linéaire définie par  $\varphi \mapsto \int_{\Omega} f(x) \varphi(x) dx$  est continue pour la norme associée à ce produit scalaire.).
- Si  $a$  est non symétrique, on considère l'application de  $H$  dans  $H$ , qui à  $u$  associe  $Au$ , défini par :

$$(Au, v) = a(u, v) \quad \forall v \in H.$$

L'application qui à  $u$  associe  $Au$  est linéaire continue, et

$$(Au, v) \leq a(u, v) \leq M \|u\| \|v\|$$

car  $a$  est continue. D'autre part, par le théorème de représentation de Riesz, on a existence et unicité de  $\psi \in H$  tel que  $T(v) = (\psi, v)$ , pour tout  $v \in H$ . Donc  $u$  est solution de  $a(u, v) = T(v), \forall v \in H$  si et seulement si  $Au = \psi$ . Pour montrer l'existence et l'unicité de  $u$ , il faut donc montrer que  $A$  est bijectif.

Montrons d'abord que  $A$  est injectif. On suppose que  $Au = 0$ . On a  $(Au, u) \geq \alpha \|u\|^2$  par coercitivité de  $a$  et comme  $\|Au\| \|v\| \geq (Au, v)$ , on a donc :

$$\|Au\| \geq \alpha \|u\|,$$

En conclusion, si  $Au = 0 \Rightarrow u = 0$ .

Montrons maintenant que  $A$  est surjectif. On veut montrer que  $AH = H$ . Pour cela, on va montrer que  $AH$  est fermé et  $AH^\top = \{0\}$ . Soit  $w \in \overline{AH}$ ; il existe alors une suite  $(v_n)_{n \in \mathbb{N}} \subset H$  telle que  $Av_n \rightarrow w$  dans  $H$ . Montrons que la suite  $(v_n)_{n \in \mathbb{N}}$  converge dans  $H$ . On a :

$$\|Av_n - Av_m\| = \|A(v_n - v_m)\| \geq \alpha \|v_n - v_m\|_H$$

donc la suite  $(v_n)_{n \in \mathbb{N}}$  est de Cauchy. On en déduit qu'elle converge vers un certain  $v \in H$ . Comme  $A$  est continue, on a donc :  $Av_n \rightarrow Av$  dans  $H$ , et donc  $w = Av \in AH$ .

Montrons maintenant que

$$AH^\top = \{0\}$$

Soit  $v_0 \in AH^\top$ , comme  $a$  est coercive, on a :

$$\alpha \|v_0\|^2 \leq a(v_0, v_0) = (Av_0, v_0) = 0,$$

on en déduit que  $v_0 = 0$ , ce qui prouve que  $AH^\top = \{0\}$ .

Pour conclure la preuve du théorème, il reste à montrer que si  $a$  est symétrique, le problème de minimisation (3.8) est équivalent au problème (3.7) Soit  $u \in H$  solution unique de (3.7); montrons que  $u$  est solution de (3.8). Soit  $w \in H$ , on va montrer que  $J(u+w) \geq J(u)$ .

$$\begin{aligned} J(u+w) &= \frac{1}{2}a(u+w, u+w) - T(u+w) \\ &= \frac{1}{2}a(u, u) + \frac{1}{2}[a(u, w) + a(w, u)] + \frac{1}{2}a(w, w) - T(u) - T(w) \\ &= \frac{1}{2}a(u, u) + \frac{1}{2}a(w, w) + a(u, w) - T(u) - T(w) \\ &= J(u) + \frac{1}{2}a(w, w) \geq J(u) + \frac{\alpha}{2}\|w\|^2 \end{aligned}$$

Donc  $J(u+w) > J(u)$  sauf si  $w = 0$ .

Réciproquement, supposons maintenant que  $u$  est solution du problème de minimisation (3.8) et montrons que  $u$  est solution du problème (3.7). Soit  $w \in H$  et  $t > 0$ . On a :  $J(u+tw) - J(u) \geq 0$  et  $J(u-tw) - J(u) \geq 0$  car  $u$  minimise  $J$ . On en déduit que :

$$ta(u, w) + \frac{1}{2}t^2a(w, w) \geq 0 \text{ et } -ta(u, w) + \frac{1}{2}t^2a(w, w) \geq 0$$

Comme  $t$  est strictement positif, on peut diviser ces deux inégalités par  $t$  :

$$a(u, w) + \frac{1}{2}ta(w, w) \geq 0 \text{ et } -a(u, w) + \frac{1}{2}ta(w, w) \geq 0$$

On fait alors tendre  $t$  vers 0 et on obtient  $a(u, w) = 0$  pour tout  $w \in H$ , ce qui montre que  $u$  est bien solution du problème (3.7). ■

Montrons qu'on peut appliquer le théorème de Lax Milgram pour les problèmes (3.5) et (3.6).

**Proposition 3.7 (Existence et unicité de la solution de (3.1))** Si  $f \in L^2(\Omega)$ , il existe une unique  $u \in H_0^1(\Omega)$  solution de (3.5) et (3.6).

**Démonstration :** Montrons que les hypothèses du théorème de Lax Milgram sont vérifiées. L'espace  $H = H_0^1(\Omega)$  est un espace de Hilbert. La forme bilinéaire  $a$  est définie par :

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx \left( = \sum_{i=1}^N \int_{\Omega} D_i u(x) D_i v(x) dx \right),$$

et la forme linéaire  $T$  par :

$$T(v) = \int_{\Omega} f(x)v(x)dx.$$

Montrons que  $T \in H'$  ; en effet, la forme  $T$  est linéaire, et on a :

$$T(v) \leq \|f\|_{L^2} \|v\|_{L^2} \leq \|f\|_{L^2} \|v\|_{H^1}.$$

On en déduit que  $T$  est une forme linéaire continue sur  $H_0^1(\Omega)$ , ce qui est équivalent à dire que  $T \in H^{-1}(\Omega)$  (dual topologique de  $H_0^1(\Omega)$ ).

Montrons maintenant que  $a$  est bilinéaire, continue et symétrique. La continuité de  $a$  se démontre en écrivant que

$$\begin{aligned} a(u, v) &= \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx \leq \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} \\ &\leq \|u\|_{H^1} \|v\|_{H^1} \end{aligned}$$

Les caractères bilinéaire et symétrique sont évidents. Montrons maintenant que  $a$  est coercitive : en effet,

$$a(v, v) = \int_{\Omega} \nabla v(x) \cdot \nabla v(x) dx = \sum_{i=1}^N \int_{\Omega} D_i v(x) D_i v(x) dx \geq \frac{1}{\text{diam}(\Omega)^2 + 1} \|v\|_{H^1}^2,$$

par l'inégalité de Poincaré (voir note page 12 page 24. Comme  $T \in H'$  et comme  $a$  est linéaire, continue, coercitive donc le théorème de Lax Milgram s'applique : on en conclut qu'il existe une unique fonction  $u \in H_0^1(\Omega)$  solution de (3.5) et comme  $a$  est symétrique,  $u$  est l'unique solution du problème de minimisation associée. ■

**Définition 3.8 (Solution forte dans  $H^2$ )** Soit  $f \in L^2(\Omega)$ , on dit que  $u$  est solution forte de (3.1) dans  $H^2$  si  $u \in H^2(\Omega) \cap H_0^1(\Omega)$  vérifie  $-\Delta u = f$  dans  $L^2(\Omega)$ .

Remarquons que si  $u$  est solution forte  $C^2$  de (3.1), alors  $u$  est solution forte  $H^2$ . De même, si  $u$  est solution forte  $H^2$  de (3.1) alors  $u$  est solution faible de (3.1). Les réciproques sont fausses. On admettra le théorème (difficile) de régularité, qui s'énonce de la manière suivante :

**Théorème 3.9 (Régularité)** Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^d$ . On suppose que  $\Omega$  a une frontière de classe  $C^2$ , ou que  $\Omega$  est convexe à frontière lipschitzienne. Si  $f \in L^2(\Omega)$  et si  $u \in H_0^1(\Omega)$  est solution faible de (3.1), alors  $u \in H^2(\Omega)$ . De plus, si  $f \in H^m(\Omega)$  alors  $u \in H^{m+2}(\Omega)$

**Remarque 3.10 (Différences entre les méthodes de discrétisation)** Lorsqu'on adopte une discrétisation par différences finies, on a directement le problème (3.1). Lorsqu'on adopte une méthode de volumes finis, on discrétise le "bilan" obtenu en intégrant (3.1) sur chaque maille. Lorsqu'on utilise une méthode variationnelle, on discrétise la formulation variationnelle (3.6) dans le cas de la méthode de Ritz, la formulation faible (3.5) dans le cas de la méthode de Galerkin, voir section 3.2.

Remarquons également que dans la formulation faible, (3.5), les conditions aux limites de Dirichlet homogènes  $u = 0$  sont prises en compte dans l'espace  $u \in H_0^1(\Omega)$ , et donc également dans l'espace d'approximation  $H_N$ . Pour le problème de Neumann homogène, les conditions aux limites ne sont pas explicites dans l'espace fonctionnel, voir à ce sujet l'exercice 42 page 119.

### 3.1.2 Problème de Dirichlet non homogène

On se place ici en dimension 1 d'espace,  $d = 1$ , et on considère le problème suivant :

$$\begin{cases} u'' = f & \text{sur } ]0, 1[ \\ u(0) = a, \\ u(1) = b, \end{cases} \quad (3.10)$$

où  $a$  et  $b$  sont des réels donnés. Ces conditions aux limites sont dites de type Dirichlet non homogène ; comme  $a$  et  $b$  ne sont pas forcément nuls, on cherche une solution dans  $H^1(\Omega)$  et non plus dans  $H_0^1(\Omega)$ . Cependant, pour se ramener à l'espace  $H_0^1(\Omega)$  (en particulier pour obtenir que le problème est bien posé grâce au théorème de Lax Milgram et à la coercivité de la forme bilinéaire  $a(u, v) = \int_{\Omega} \nabla u(x) \nabla v(x) dx$  sur  $H_0^1(\Omega)$ ), on va utiliser une technique dite de "relèvement". On pose :  $u = u_0 + \tilde{u}$  où  $u_0$  est définie par :

$$u_0(x) = a + (b - a)x.$$

On a en particulier  $u_0(0) = a$  et  $u_0(1) = b$ . On a alors  $\tilde{u}(0) = 0$  et  $\tilde{u}(1) = 0$ . La fonction  $\tilde{u}$  vérifie donc le système :

$$\begin{cases} -\tilde{u}'' = f, \\ \tilde{u}(0) = 0, \\ \tilde{u}(1) = 0, \end{cases}$$

dont on connaît la formulation faible, et dont on sait qu'il est bien posé (voir paragraphe 3.1.1 page 97). Donc il existe un unique  $u \in H^1(\Omega)$  vérifiant  $u = u_0 + \tilde{u}$ , où  $\tilde{u} \in H_0^1(\Omega)$  est l'unique solution du problème

$$\int_0^1 \tilde{u}' v' = \int_0^1 f v \quad \forall v \in H_0^1(]0, 1[)$$

De manière plus générale, soit  $u_1 \in H_{a,b}^1(]0, 1[) = \{v \in H^1; v(0) = a \text{ et } v(1) = b\}$ , et soit  $\bar{u} \in H_0^1(]0, 1[)$  l'unique solution faible du problème :

$$\begin{cases} -\bar{u}'' = u_1'' + f, \\ \bar{u}(0) = 0, \\ \bar{u}(1) = 0. \end{cases}$$

Alors  $\bar{u} + u_1$  est l'unique solution faible de (3.10), c'est-à-dire la solution du problème

$$\begin{cases} u \in H_{a,b}^1(]0, 1[), \\ \int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \forall v \in H_0^1(]0, 1[). \end{cases}$$

**Remarque 3.11** Il est facile de montrer que  $u$  ne dépend pas du relèvement choisi (voir exercice 36 page 118).

Considérons maintenant le cas de la dimension 2 d'espace :  $d = 2$ .

Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^d$ , considère le problème :

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

Pour se ramener au problème de Dirichlet homogène, on veut construire un relèvement, c'est à dire une fonction  $u_0 \in H^1(\Omega)$  t.q.  $\gamma(u_0) = g$  où  $\gamma$  est l'application trace. On ne peut plus le faire de manière explicite comme en dimension 1. En particulier, on rappelle qu'en dimension 2, l'espace  $H^1(\Omega)$  n'est pas inclus dans l'espace  $C(\bar{\Omega})$  des fonctions continues, contrairement au cas de la dimension 1. Mais si on a  $g \in H^{1/2}(\partial\Omega)$ , on sait qu'il existe  $u_0 \in H^1(\Omega)$  tel que  $g = \gamma(u_0)$ . On cherche donc  $u$  sous la forme  $u = \tilde{u} + u_0$  avec  $\tilde{u} \in H_0^1(\Omega)$  et  $u_0 \in H^1(\Omega)$  telle que  $\gamma(u_0) = g$ . Soit  $v \in H_0^1(\Omega)$ ; on multiplie (3.11) par  $v$  et on intègre sur  $\Omega$  :

$$\int_{\Omega} -\Delta u(x)v(x)dx = \int_{\Omega} f(x)v(x)dx,$$

c'est-à-dire :

$$\int_{\Omega} \nabla u(x)\nabla v(x)dx = \int_{\Omega} f(x)v(x)dx.$$

Comme  $u = u_0 + \tilde{u}$ , on a donc :

$$\begin{cases} \tilde{u} \in H_0^1(\Omega), \\ \int_{\Omega} \nabla \tilde{u}(x)\nabla v(x)dx = \int_{\Omega} f(x)v(x)dx - \int_{\Omega} \nabla u_0(x)\nabla v(x)dx, \forall v \in H_0^1(\Omega) \end{cases} \quad (3.12)$$

En dimension 2, il n'est pas toujours facile de construire le relèvement  $u_0$ . Il est donc usuel, dans la mise en oeuvre des méthodes d'approximation (par exemple par éléments finis), de servir de la formulation suivante, qui est équivalente à la formulation (3.12) :

$$\begin{cases} u \in \{v \in H^1(\Omega); \gamma(v) = g \text{ sur } \partial\Omega\} \\ \int_{\Omega} \nabla u(x)\nabla v(x)dx = \int_{\Omega} f(x)v(x)dx \quad \forall v \in H_0^1(\Omega). \end{cases} \quad (3.13)$$



### 3.1.3 Problème avec conditions aux limites de Fourier

On considère ici le problème de diffusion avec conditions aux limites de type “Fourier” (ou “Robin” dans la littérature anglo-saxonne).

$$\begin{cases} -\Delta u = f \text{ dans } \Omega, \\ \nabla u \cdot \mathbf{n} + \lambda u = 0 \text{ sur } \partial\Omega, \end{cases} \quad (3.14)$$

où :

1.  $\Omega$  est un ouvert borné de  $\mathbb{R}^d$ ,  $d = 1, 2$  ou  $3$ , et  $\partial\Omega$  sa frontière,
2.  $f \in C^2(\bar{\Omega})$ ,
3.  $\mathbf{n}$  est le vecteur unitaire normal à  $\partial\Omega$ , extérieur à  $\Omega$ ,
4.  $\lambda(x) > 0, \forall x \in \partial\Omega$ , est un coefficient qui modélise par exemple un transfert thermique à la paroi.

Supposons qu’il existe  $u \in C^2(\bar{\Omega})$  vérifiant (3.14). Soit  $\varphi \in C^\infty(\bar{\Omega})$  une “fonction test”. On multiplie formellement (3.14) par  $\varphi$  et on intègre sur  $\Omega$ .

On obtient :

$$-\int_{\Omega} \Delta u(x) \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx.$$

Par intégration par parties, on a alors

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx - \int_{\partial\Omega} \nabla u(x) \cdot \mathbf{n}(x) \varphi(x) d\gamma(x) = \int_{\Omega} f(x) \varphi(x) dx.$$

Notons que la fonction  $\varphi$  qui n’est pas à support compact, et que la condition aux limites :

$$\nabla u \cdot \mathbf{n} = -\lambda u$$

va donc intervenir dans cette formulation. En remplaçant on obtient :

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx + \int_{\partial\Omega} \lambda u(x) \varphi(x) d\gamma(x) = \int_{\Omega} f(x) \varphi(x) dx, \quad \forall \varphi \in C^\infty(\bar{\Omega}).$$

Par densité de  $C^\infty(\bar{\Omega})$  dans  $H^1(\Omega)$ , on a donc également

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx + \int_{\partial\Omega} \lambda u(x) \varphi(x) d\gamma(x) = \int_{\Omega} f(x) \varphi(x) dx, \quad \forall \varphi \in H^1(\Omega).$$

**Définition 3.12 (Solution faible)** On dit que  $u$  est solution faible de (3.14) si  $u$  est solution de :

$$\begin{cases} u \in H^1(\Omega), \\ \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \lambda(x) u(x) v(x) d\gamma(x) = \int_{\Omega} f(x) v(x) dx \quad \forall v \in H^1(\Omega). \end{cases} \quad (3.15)$$

On peut remarquer que sous les hypothèses :

$$f \in L^2(\Omega), \quad \lambda \in L^\infty(\partial\Omega),$$

toutes les intégrales de (3.15) sont bien définies. (On rappelle que si  $\varphi \in L^2(\Omega)$  et  $\psi \in L^2(\Omega)$ , alors  $\varphi\psi \in L^1(\Omega)$ ). Pour vérifier que le problème (3.15) est bien posé, on a envie d’appliquer le théorème de Lax-Milgram. Définissons pour cela  $a : H^1(\Omega) \times H^1(\Omega) \rightarrow \mathbb{R}$  par :

$$a(u, v) = \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \lambda(x) u(x) v(x) d\gamma(x). \quad (3.16)$$

Il est facile de voir que  $a$  est une forme bilinéaire symétrique. On peut donc lui associer une forme quadratique définie par :

$$E(v) = \frac{1}{2} \int_{\Omega} \nabla v(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \lambda(x) v^2(x) d\gamma(x) - \int_{\Omega} f(x) v(x) dx. \quad (3.17)$$

**Définition 3.13 (Solution variationnelle)** On dit que  $u$  est solution variationnelle de (3.14) si  $u$  vérifie :

$$\begin{cases} u \in H^1(\Omega), \\ E(u) \leq E(v), \forall v \in H^1(\Omega), \end{cases} \quad (3.18)$$

où  $E$  est défini par (3.17).

**Lemme 3.14** On suppose que  $\lambda \in L^\infty(\partial\Omega)$ . Alors la forme bilinéaire définie par (3.16) est continue sur  $H^1(\Omega) \times H^1(\Omega)$ .

**Démonstration :** On a :

$$\begin{aligned} a(u, v) &= \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx + \int_{\partial\Omega} \lambda(x) u(x) v(x) d\gamma(x) \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|\lambda\|_{L^\infty(\partial\Omega)} \|u\|_{L^2(\partial\Omega)} \|v\|_{L^2(\partial\Omega)}. \end{aligned}$$

Or par le théorème de trace (théorème 3.2), et plus particulièrement grâce à la continuité de la trace (3.4), on a

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

On en déduit que

$$a(u, v) \leq M \|u\|_{H^1} \|v\|_{H^1}$$

avec  $M = 1 + C^2 \|\lambda\|_{L^\infty(\partial\Omega)}$ . Donc  $a$  est bilinéaire continue ■

**Lemme 3.15** Soit  $\lambda \in L^\infty(\partial\Omega)$  tel qu'il existe  $\underline{\lambda} > 0$  tel que  $\lambda(x) \geq \underline{\lambda}$  p.p. sur  $\partial\Omega$ . Alors la forme bilinéaire  $a$  définie par (3.16) est coercitive :

Montrons qu'il existe  $\alpha > 0$  tel que  $a(v, v) \geq \alpha \|v\|^2$ , pour tout  $v \in H^1$  où

$$a(v, v) = \int_{\Omega} \nabla v(x) \cdot \nabla v(x) dx + \int_{\Omega} \alpha(x) v^2(x) d\gamma(x).$$

Attention, comme  $v \in H^1(\Omega)$  et non  $H_0^1(\Omega)$ , on ne peut pas écrire l'inégalité de Poincaré, qui nous permettrait de minorer  $\int_{\Omega} \nabla v(x) \cdot \nabla v(x) dx$ . On va montrer l'existence de  $\alpha$  par l'absurde. On suppose que  $a$  n'est pas coercive.

Dans ce cas : c'est-à-dire que :

$$\forall \alpha > 0, \exists v \in H^1(\Omega); a(v, v) < \alpha \|v\|^2.$$

On a donc en particulier, en prenant  $\alpha = \frac{1}{n}$  :

$$\forall n \in \mathbb{N}, \exists v_n \in H^1(\Omega); a(v_n, v_n) < \frac{1}{n} \|v_n\|_{H^1}^2.$$

Dans cette dernière assertion, on peut prendre  $v_n$  de norme 1, puisque l'inégalité est homogène de degré 2. On a donc :

$$\forall n \in \mathbb{N}, \exists v_n \in H^1(\Omega); \|v_n\|_{H^1(\Omega)} = 1; a(v_n, v_n) < \frac{1}{n}.$$

Or, par le théorème de Rellich, toute suite bornée  $(v_n)_{n \in \mathbb{N}}$  de  $H^1(\Omega)$ , est relativement compacte dans  $L^2(\Omega)$ . Comme on a  $\|v_n\|_{H^1(\Omega)} = 1$ , il existe donc une sous-suite encore notée  $(v_n)_{n \in \mathbb{N}} \subset H^1(\Omega)$  telle que  $v_n$  converge vers  $v$  dans  $L^2(\Omega)$  lorsque  $n$  tend vers  $+\infty$ .

De plus, comme :

$$a(v_n, v_n) = \int_{\Omega} \nabla v_n(x) \cdot \nabla v_n(x) dx + \int_{\partial\Omega} v_n(x) v_n(x) d\gamma(x) < \frac{1}{n} \rightarrow 0 \text{ lorsque } n \rightarrow +\infty,$$

On en déduit que, chaque terme étant positif :

$$\int_{\Omega} \nabla v_n(x) \cdot \nabla v_n(x) dx \rightarrow_{n \rightarrow +\infty} 0 \quad (3.19)$$

et

$$\int_{\partial\Omega} v_n(x)v_n(x)dx \rightarrow_{n \rightarrow +\infty} 0 \quad (3.20)$$

On a donc :  $\nabla v_n \rightarrow 0$  dans  $L^2(\Omega)$  lorsque  $n \rightarrow +\infty$ . On en déduit que

$$\int_{\Omega} \partial_i v_n(x)\varphi dx \rightarrow 0 \text{ lorsque } n \rightarrow +\infty, \text{ pour } i = 1, \dots, d.$$

Donc par définition de la dérivée faible (voir note page 23), on a aussi

$$\int_{\Omega} v_n(x)\partial_i \varphi(x)dx \rightarrow 0 \text{ lorsque } n \rightarrow +\infty.$$

Comme  $v_n \rightarrow v$  dans  $L^2(\Omega)$  lorsque  $n \rightarrow +\infty$ , on peut passer à la limite ci-dessus et écrire que  $\int_{\Omega} v(x)\partial_i \varphi(x) = 0$ . On en déduit que la dérivée faible  $D_i v$  existe et est nulle dans  $\Omega$ . La fonction  $v$  est donc constante par composante connexe. Mais par (3.20), on a  $v = 0$  sur  $\partial\Omega$ , et la trace d'une fonction constante est la constante elle-même. On a donc

$$v = 0 \text{ dans } \Omega.$$

On a ainsi montré que

$$D_i v_n \rightarrow D_i v \text{ et } v_n \rightarrow v = 0 \text{ dans } L^2(\Omega) \text{ lorsque } n \rightarrow +\infty.$$

Donc  $v_n \rightarrow 0$  dans  $H^1(\Omega)$  lorsque  $n \rightarrow +\infty$ , ce qui contredit le fait que  $\|v_n\|_{H^1(\Omega)} = 1$ . On a ainsi montré la coercivité de  $a$ . ■

**Proposition 3.16** Soit  $f \in L^2(\Omega)$  et  $\lambda \in L^\infty(\Omega)$  t.q.  $\lambda \geq \underline{\lambda}$  p.p. avec  $\underline{\lambda} > 0$  alors il existe une unique  $u$  solution de (3.15) qui est aussi l'unique solution de (3.18).

### 3.1.4 Condition de Neumann

Considérons maintenant le problème (3.14) avec  $\lambda = 0$ , on obtient le problème :

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

qu'on appelle problème de Dirichlet avec conditions de Neumann homogènes. En intégrant la première équation du système, il est facile de voir qu'une condition nécessaire d'existence d'une solution de (3.1.4) est que :

$$\int_{\Omega} -\Delta u(x)dx = \int_{\partial\Omega} \frac{\partial u}{\partial n}(x)dx = \int_{\Omega} f(x)dx = 0$$

Si la condition aux limites de Neumann est non-homogène :  $\frac{\partial u}{\partial n} = g$ , la condition de compatibilité devient

$$\int_{\Omega} f(x)dx + \int_{\partial\Omega} g(x)d\gamma(x) = 0.$$

Remarquons que si  $\alpha = 0$ , la forme bilinéaire est

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

et que celle-ci n'est pas coercive sur  $H^1(\Omega)$ . De fait, il est clair que la solution de (3.1.4) n'est pas unique, puisque si  $u$  est solution de (3.1.4) alors  $u + c$  est aussi solution, pour tout  $c \in \mathbb{R}$ . Pour éviter ce problème on va chercher les solutions de (3.1.4) à moyenne nulle. On cherche donc à résoudre (3.1.4) dans l'espace

$$H = \{v \in H^1(\Omega); \int_{\Omega} v(x)dx = 0\}$$

On admettra que  $a$  est coercive sur  $H$  (ceci est vrai grâce à l'inégalité de Poincaré–Wirtinger<sup>1</sup>. Le problème

$$\begin{cases} u \in H, \\ a(u, v) = \int f v \quad \forall v \in H, \end{cases}$$

admet donc une unique solution.

### 3.1.5 Formulation faible et formulation variationnelle.

Nous donnons ici un exemple de problème pour lequel on peut établir une formulation faible, mais pas variationnelle. On se place en une dimension l'espace  $N = 1$ , et on considère  $\Omega = ]0, 1[$  et  $f \in L^2(]0, 1[)$ . On s'intéresse au problème suivant (dit "d'advection diffusion") :

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

Cherchons une formulation faible. Par la même méthode qu'au paragraphe 3.1.1, on choisit  $v \in H_0^1(\Omega)$ , on multiplie (3.1.5) par  $v$  et on intègre par parties :

$$\int_{\Omega} u'(x)v'(x)dx + \int_{\Omega} u'(x)v(x)dx = \int_{\Omega} f(x)v(x)dx.$$

Il est donc naturel de poser :

$$a(u, v) = \int_{\Omega} u'(x)v'(x)dx + \int_{\Omega} u'(x)v(x)dx, \text{ et } T(v) = \int_{\Omega} f(x)v(x)dx.$$

Il est évident que  $T$  est une forme linéaire continue sur  $H_0^1(\Omega)$  (c'est à dire  $T \in H^{-1}(\Omega)$ ) et que la forme  $a$  est bilinéaire continue, mais pas symétrique. De plus elle est coercive : En effet, on a :

$$\begin{aligned} a(u, u) &= \int_{\Omega} u'^2(x)dx + \int_{\Omega} u'(x)u(x)dx \\ &= \int_{\Omega} u'^2(x)dx + \int_{\Omega} \frac{1}{2}(u^2)'(x)dx \end{aligned}$$

Or, comme  $u \in H_0^1(\Omega)$ , on a  $u = 0$  sur  $\partial\Omega$  et donc  $\int_{\Omega} (u^2)'(x)dx = u^2(1) - u^2(0) = 0$ . On en déduit que :

$a(u, u) = \int_0^1 (u')^2$ , et par l'inégalité de Poincaré (voir page 24), on conclut que  $a$  est coercive sur  $H_0^1(\Omega)$ . On en déduit par le théorème de Lax Milgram, l'existence et l'unicité de  $u$  solution du problème :

$$\begin{cases} u \in H_0^1(]0, 1[) \\ \int_0^1 (u'(x)v'(x) + u'(x)v(x))dx = \int_0^1 f(x)v(x)dx. \end{cases}$$

## 3.2 Méthodes de Ritz et Galerkin

### 3.2.1 Principe général de la méthode de Ritz

On se place sous les hypothèses suivantes :

$$\begin{cases} H \text{ est un espace Hilbert} \\ a \text{ est une forme bilinéaire continue coercitive et symétrique} \\ T \in H' \end{cases} \quad (3.21)$$

<sup>1</sup>L'inégalité de Poincaré–Wirtinger s'énonce de la façon suivante : soit  $\Omega$  un ouvert borné de  $\mathbb{R}^d$  de frontière lipschitzienne, alors il existe  $C \in \mathbb{R}_+$ , ne dépendant que de  $\Omega$ , tel que pour tout  $u \in H^1(\Omega)$ , on a :

$$\|u\|_{L^2(\Omega)}^2 \leq C \|u\|_{H^1(\Omega)}^2 + 2(m(\Omega))^{-1} \left( \int_{\Omega} u(x)dx \right)^2$$

On cherche à calculer  $u \in H$  telle que :

$$a(u, v) = T(v), \quad \forall v \in H,$$

ce qui revient à calculer  $u \in H$  solution du problème de minimisation (3.8), avec  $J$  définie par (3.9). L'idée de la méthode de Ritz<sup>2</sup> est de remplacer  $H$  par un espace  $H_N \subset H$  de dimension finie (où  $\dim H_N = N$ ), et de calculer  $u_N$  solution de

$$\begin{cases} u_N \in H_N \\ J(u_N) \leq J(v), \quad \forall v \in H_N, \end{cases} \quad (3.22)$$

en espérant que  $u_N$  soit "proche" (en un sens à définir) de  $u$ .

**Théorème 3.17** *Sous les hypothèses (3.21), si  $H_N$  est un s.e.v. de  $H$  et  $\dim H_N < +\infty$  alors le problème (3.22) admet une unique solution.*

**Démonstration :** Puisque  $H_N$  est un espace de dimension finie inclus dans  $H$ , c'est donc aussi un Hilbert. On peut donc appliquer le théorème de Lax Milgram, et on en déduit l'existence et l'unicité de  $u_N \in H_N$  solution de (3.22), qui est aussi solution de :

$$\begin{cases} u_N \in H_N, \\ a(u_N, v) = T(v), \quad \forall v \in H_N. \end{cases}$$

■

Nous allons maintenant exposer une autre méthode de démonstration du théorème 3.17, qui a l'avantage d'être constructive, et qui nous permet d'introduire les idées principales des méthodes numériques envisagées plus loin. Comme l'espace  $H_N$  considéré dans le théorème 3.22 est de dimension  $N$ , il existe une base  $(\phi_1, \dots, \phi_N)$  de  $H_N$ .

Si  $u \in H_N$ , on peut donc développer  $u = \sum_{i=1}^N u_i \phi_i$ . On note :

$$U = (u_1, \dots, u_N) \in \mathbb{R}^N$$

L'application  $\xi$  qui à  $u$  associe  $U$  est une bijection de  $H_N$  dans  $\mathbb{R}^N$ . Posons  $j = J \circ \xi^{-1}$ . On a donc :

$$j(U) = J(u).$$

Or :

$$\begin{aligned} J(u) &= \frac{1}{2} a \left( \sum_{i=1}^N u_i \phi_i, \sum_{i=1}^N u_i \phi_i \right) - T \left( \sum_{i=1}^N u_i \phi_i \right) \\ &= \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N u_i u_j a(\phi_i, \phi_j) - \sum_{i=1}^N u_i T(\phi_i). \end{aligned}$$

On peut donc écrire  $J(u)$  sous la forme :

$$J(u) = \frac{1}{2} U^t \mathcal{K} U - U^t \mathcal{G} = j(U),$$

où  $\mathcal{K} \in M^{N,N}(\mathbb{R})$  est définie par  $\mathcal{K}_{ij} = a(\phi_i, \phi_j)$ , et où  $\mathcal{G}_i = T(\phi_i)$ . Chercher  $u_N$  solution de (3.22) est donc équivalent à chercher  $U$  solution de :

$$\begin{cases} U \in \mathbb{R}^N, \\ j(U) \leq j(V), \quad \forall V \in \mathbb{R}^N. \end{cases} \quad (3.23)$$

où

$$j(V) = \frac{1}{2} V^t \mathcal{K} V - V^t \mathcal{G}. \quad (3.24)$$

Il est facile de vérifier que la matrice  $\mathcal{K}$  est symétrique définie positive. Donc  $j$  est une fonctionnelle quadratique sur  $\mathbb{R}^N$ , et on a donc existence et unicité de  $U \in \mathbb{R}^N$  tel que  $j(U) \leq j(V) \quad \forall V \in \mathbb{R}^N$ . La solution du problème de minimisation (3.23) est aussi la solution du système linéaire  $\mathcal{K}U = \mathcal{G}$ ; on appelle souvent  $\mathcal{K}$  la matrice de rigidité.

<sup>2</sup>Walter Ritz, né le 22 février 1878 à Sion et mort le 7 juillet 1909 à Göttingen, est un physicien suisse. Il a inventé la méthode dite "de Ritz" dans le cadre du calcul des valeurs propres de l'opérateur bi-harmonique

**Proposition 3.18 (Existence et unicité de la solution du problème de minimisation)** . Soit  $j = \mathbb{R}^N \rightarrow \mathbb{R}$  définie par (3.24). Il existe une unique  $u \in \mathbb{R}^N$  solution du problème de minimisation (3.23).

**Démonstration :** Ceci est une conséquence du résultat général de minimisation dans  $\mathbb{R}^N$  (voir cours de licence). ■

### Résumé sur la technique de Ritz.

1. On se donne  $H_N \subset H$ .
2. On trouve une base de  $H_N$ .
3. On calcule la matrice de rigidité  $\mathcal{K}$  et le second membre  $\mathcal{G}$ . Les coefficients de  $\mathcal{K}$  sont donnés par  $\mathcal{K}_{ij} = a(\phi_i, \phi_j)$ .
4. On minimise  $j$  par la résolution de  $\mathcal{K}V = \mathcal{G}$ .
5. On calcule la solution approchée :  $u^{(N)} = \sum_{i=1}^N u_i \phi_i$ .

On appelle  $H_N$  l'espace d'approximation. Le choix de cet espace sera fondamental pour le développement de la méthode d'approximation. Le choix de  $H_N$  est formellement équivalent au choix de la base  $(\phi_i)_{i=1 \dots N}$ . Pourtant, le choix de cette base est capital même si  $u^{(N)}$  ne dépend que du choix de  $H_N$  et pas de la base.

**Choix de la base** Un premier choix consiste à choisir des bases indépendantes de  $N$  c'est à dire  $\{ \text{base de } H_{N+1} \} = \{ \text{base de } H_N \} \cup \{ \phi_{N+1} \}$ . Les bases sont donc emboîtées les unes dans les autres. Considérons par exemple  $H = H^1(]0, 1[)$ , et l'espace d'approximation :

$$H_N = \text{Vect}\{1, X, \dots, X^{N-1}\}$$

Les fonctions de base sont donc  $\phi_i = X^{i-1}$ ,  $i = 1, \dots, N$ . On peut remarquer que ce choix de base amène à une méthode d'approximation qui donne des matrices pleines. Or, on veut justement éviter les matrices pleines, car les systèmes linéaires associés sont coûteux (en temps et mémoire) à résoudre.

Le choix idéal serait de choisir une base  $(\phi_i)_{i=1, \dots, N}$  de telle sorte que

$$\begin{aligned} a(\phi_i, \phi_j) &= \lambda_i \delta_{ij} \\ \text{où } \delta_{ij} &= \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases} \end{aligned} \quad (3.25)$$

On a alors  $\mathcal{K} = \text{diag}(\lambda_1, \dots, \lambda_N)$ , et on a explicitement :  $u^{(N)} = \sum_{i=1}^N \frac{T(\phi_i)}{a(\phi_i, \phi_i)} \phi_i$ . Considérons par exemple le problème de Dirichlet (3.1) Si  $\phi_i$  est la  $i$ -ème fonction propre de l'opération  $-\Delta$  avec conditions aux limites de Dirichlet associée à  $\lambda_i$ , on obtient bien la propriété souhaitée. Malheureusement, il est rare que l'on puisse connaître explicitement les fonctions de base  $\phi_i$ .

Un deuxième choix consiste à choisir des bases dépendantes de  $N$ . Mais dans ce cas, la base de  $H_N$  n'est pas incluse dans celle de  $H_{N+1}$ . La technique des éléments finis qu'on verra au chapitre suivant, est un exemple de ce choix. Dans la matrice  $\mathcal{K}$  obtenue est creuse (c'est à dire qu'un grand nombre de ses coefficients sont nuls). Par exemple, pour des éléments finis appliqués à un opérateur du second ordre, on peut avoir un nombre de coefficients non nuls de l'ordre de  $O(N)$ .

**Convergence de l'approximation de Ritz** Une fois qu'on a calculé  $u_N$  solution de (3.23), il faut se préoccuper de savoir si  $u^{(N)}$  est une bonne approximation de  $u$  solution de (3.2.1), c'est à dire de savoir si

$$u^{(N)} \rightarrow u \text{ lorsque } N \rightarrow +\infty$$

Pour vérifier cette convergence, on va se servir de la notion de consistance.

**Définition 3.19 (Consistance)** Sous les hypothèses (3.21), on dit que l'approximation de Ritz définie par l'espace  $H_N \subset H$  avec  $\dim H_N = N < +\infty$  est consistante si  $d(H, H_N)$  tend vers 0 lorsque  $N \rightarrow +\infty$ , c'est à dire  $d(u, H_N) \rightarrow_{N \rightarrow +\infty} 0, \forall u \in H$  ou encore  $\inf_{v \in H_N} \|u - v\| \rightarrow_{N \rightarrow +\infty} 0, \forall u \in H$ .

L'autre notion fondamentale pour prouver la convergence est la stabilité, elle même obtenue grâce à la propriété de coercivité de  $a$ . Par stabilité, on entend estimation a priori sur la solution approchée  $u^{(N)}$  (avant même de savoir si elle existe), où  $u^{(N)}$  est solution de (3.23) ou encore de :

$$\begin{cases} a(u^{(N)}, v) = T(v) \quad \forall v \in H_N \\ u^{(N)} \in H_N \end{cases} \quad (3.26)$$

On a l'estimation a priori suivante sur  $u_N$  :

**Proposition 3.20 (Stabilité)** *Sous les hypothèses du théorème 3.21, on a :*

$$\|u^{(N)}\|_H \leq \frac{\|T\|_{H'}}{\alpha}.$$

**Démonstration :**

Le caractère coercif de  $a$  nous permet d'écrire :

$$\alpha \|u^{(N)}\|^2 \leq a(u^{(N)}, u^{(N)}).$$

Or comme  $u^{(N)}$  est solution de (3.26), on a :

$$a(u^{(N)}, u^{(N)}) = T(u^{(N)}).$$

Comme  $T$  est linéaire continue, on obtient

$$T(u^{(N)}) \leq \|T\|_{H'} \|u^{(N)}\|_H.$$

**Théorème 3.21 (Lemme de Céa<sup>3</sup>)** *Soit  $H$  un espace de Hilbert réel, et  $a$  une forme bilinéaire continue symétrique coercive. Soit  $T$  une forme linéaire continue et  $T \in H'$ , et soit  $M > 0$  et  $\alpha > 0$  tels que  $a(u, v) \leq M \|u\|_H \|v\|_H$  et  $a(u, u) \geq \alpha \|u\|_H^2$ . Soit  $u \in H$  l'unique solution du problème suivant :*

$$\begin{cases} u \in H, \\ a(u, v) = T(v), \forall v \in H. \end{cases} \quad (3.27)$$

Soit  $H_N \subset H$  tel que  $\dim H_N = N$ , et soit  $u^{(N)} \in H_N$  l'unique solution de

$$\begin{cases} u^{(N)} \in H_N, \\ a(u^{(N)}, v) = T(v), \forall v \in H_N. \end{cases} \quad (3.28)$$

Alors

$$\|u - u^{(N)}\| \leq \sqrt{\frac{M}{\alpha}} d(u, H_N) \quad (3.29)$$

où

$$d(u, H_N) = \inf_{v \in H_N} d(u, v).$$

**Démonstration :**

Etape 1 : On va montrer que  $u^{(N)}$  est la projection de  $u$  sur  $H_N$  pour le produit scalaire  $(\cdot, \cdot)_a$  induit par  $a$ , défini de  $H \times H$   $(u, v)_a = a(u, v)$ . On note  $\|u\|_a = \sqrt{a(u, u)}$ , la norme induite par le produit scalaire  $a$ . La norme  $\|\cdot\|_a$  est équivalente à la norme  $\|\cdot\|_H$ , en effet, grâce à la coercivité et la continuité de la forme bilinéaire  $a$ , on peut écrire :

$$\alpha \|u\|_H^2 \leq \|u\|_a^2 \leq M \|u\|_H^2$$

Donc  $(H, \|\cdot\|_a)$  est un espace de Hilbert. Soit  $u$  la solution de (3.27), et soit  $v = P_{H_N} u$  la projection orthogonale de  $u$  sur  $H_N$  relative au produit scalaire  $a(\cdot, \cdot)$ . Par définition de la projection orthogonale, on a donc

$$v - u \in H_N^\perp$$

Soit encore  $a(v - u, w) = 0, \forall w \in H_N$ . On en déduit que  $a(v, w) = a(u, w) = T(w), \forall w \in H$ , et donc que  $v = u^{(N)}$ . On a donc montré que  $u^{(N)}$  est la projection orthogonale de  $u$  sur  $H_N$ , c'est-à-dire  $u^{(N)} = P_{H_N} u$ .

Etape 2 : On va établir une estimation de la norme de la différence entre  $u$  et  $u_N$  ; par définition de  $P_{H_N}$ , on a :

$$\|u - P_{H_N} u\|_a^2 \leq \|u - v\|_a^2, \forall v \in H_N,$$

ce qui s'écrit (puisque  $P_{H_N} u = u^{(N)}$ ) :

$$a(u - u^{(N)}, u - u^{(N)}) \leq a(u - v, u - v), \quad \forall v \in H_N$$

Par coercivité et continuité de la forme bilinéaire  $a$ , on a donc :

$$\alpha \|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)}, u - u^{(N)}) \leq a(u - v, u - v) \leq M \|u - v\|_H^2, \forall v \in H_N.$$

On en déduit que :

$$\|u - u^{(N)}\| \leq \sqrt{\frac{M}{\alpha}} \|u - v\|, \forall v \in H_N.$$

En passant à l'inf sur  $v$ , on obtient alors :

$$\|u - u^{(N)}\| \leq \sqrt{\frac{M}{\alpha}} \inf_{v \in H_N} \|u - v\|$$

Ce qui est exactement (3.29). ■

### 3.2.2 Méthode de Galerkin

On se place maintenant sous les hypothèses suivantes :

$$\begin{cases} H \text{ espace de Hilbert,} \\ a : \text{ forme bilinéaire continue et coercive, } T \in H'. \end{cases} \quad (3.30)$$

Remarquons que maintenant,  $a$  n'est pas nécessairement symétrique, les hypothèses (3.30) sont donc plus générales que les hypothèses (3.21). On considère le problème

$$\begin{cases} u \in H \\ a(u, v) = T(v), \quad v \in H. \end{cases} \quad (3.31)$$

Par le théorème de Lax Milgram, il y a existence et unicité de  $u \in H$  solution de (3.31).

Le principe de la méthode de Galerkin<sup>4</sup> est similaire à celui de la méthode de Ritz. On se donne  $H_N \subset H$ , tel que  $\dim H_N < +\infty$ , et on cherche à résoudre le problème approché :

$$(P_N) \begin{cases} u^{(N)} \in H_N, \\ a(u^{(N)}, v) = T(v), \forall v \in H_N. \end{cases} \quad (3.32)$$

Par le théorème de Lax-Milgram, on a immédiatement :

**Théorème 3.22** *Sous les hypothèses, si  $H_N \subset H$  et  $\dim H_N = N$ , il existe un unique  $u^{(N)} \in H_N$  solution de (3.32).*

Comme dans le cas de la méthode de Ritz, on va donner une autre méthode, constructive, de démonstration de l'existence et unicité de  $u_N$  qui permettra d'introduire la méthode de Galerkin. Comme  $\dim H_N = N$ , il existe une base  $(\phi_1 \dots \phi_N)$  de  $H_N$ . Soit  $v \in H_N$ , on peut donc développer  $v$  sur la base :

$$v = \sum_{i=1}^N v_i \phi_i,$$

<sup>4</sup>Boris Grigoryevich Galerkin, né le 20 février 1871 à Polotsk (Biélorussie) et mort le 12 juillet 1945, est un mathématicien et un ingénieur russe réputé pour ses contributions à l'étude des treillis de poutres et des plaques élastiques. Son nom reste lié à une méthode de résolution approchée des structures élastiques, qui est l'une des bases de la méthode des éléments finis.



et identifier  $v$  au vecteur  $(v_1, \dots, v_N) \in \mathbb{R}^N$ . En écrivant que  $u$  satisfait (3.32) pour tout  $v = \phi_i = 1, N$  :

$$a(u, \phi_i) = T(\phi_i), \forall i = 1, \dots, N,$$

et en développant  $u$  sur la base  $(\phi_i)_{i=1, \dots, N}$ , on obtient :

$$\sum_{j=1}^N a(\phi_j, \phi_i) u_j = T(\phi_i), \forall i = 1, \dots, N.$$

On peut écrire cette dernière égalité sous forme d'un système linéaire :  $\mathcal{K}U = \mathcal{G}$ ,

$$\mathcal{K}_{ij} = a(\phi_j, \phi_i) \text{ et } \mathcal{G}_i = T(\phi_i), \text{ pour } i, j = 1, \dots, N.$$

La matrice  $\mathcal{K}$  n'est pas en général symétrique.

**Proposition 3.23** *Sous les hypothèses du théorème 3.22 le système linéaire (3.2.2) admet une solution .*

**Démonstration :** On va montrer que  $\mathcal{K}$  est inversible en vérifiant que son noyau est réduit à  $\{0\}$ . Soit  $w \in \mathbb{R}^N$  tel que  $\mathcal{K}w = 0$ . Décomposons  $w$  sur le  $N$  base  $(\phi_1, \dots, \phi_N)$  de  $H_N$  : On a donc :  $\sum_{j=1}^N a(\phi_j, \phi_i) w_j = 0$ . Multiplions cette relation par  $w_i$  et sommons pour  $i = 1$  à  $N$ , on obtient :

$$\sum_{i=1}^N \sum_{j=1}^N a(\phi_j, \phi_i) w_j w_i = 0.$$

Soit encore :  $a(w, w) = 0$ . Par coercitivité de  $a$ , ceci entraîne que  $w = 0$ . On en déduit que  $w_i = 0, \forall i = 1, \dots, N$ , ce qui achève la preuve. ■

**Remarque 3.24** *Si  $a$  est symétrique, la méthode de Galerkin est équivalente à celle de Ritz.*

En résumé, la méthode de Galerkin comporte les mêmes étapes que la méthode de Ritz, c'est à dire :

1. On se donne  $H_N \subset H$
2. On trouve une base de  $H_N$
3. On calcule  $\mathcal{K}$  et  $\mathcal{G}$
4. On résout  $\mathcal{K}U = \mathcal{G}$
5. On écrit  $u^{(N)} = \sum_{i=1}^N u_i \phi_i$ .

La seule différence est que l'étape 4 n'est pas issue d'un problème de minimisation. Comme pour la méthode de Ritz, il faut se poser la question du choix du sous espace  $H_N$  et de sa base, ainsi que de la convergence de l'approximation de  $u$  solution de (3.31) par  $u^{(N)}$  obtenue par la technique de Galerkin. En ce qui concerne le choix de la base  $\{\phi_1, \dots, \phi_N\}$ , les possibilités sont les mêmes que pour la méthode de Ritz, voir paragraphe 3.2.1. De même, la notion de consistance est identique à celle donnée pour la méthode de Ritz (voir définition 3.19) et la démonstration de stabilité est identique à celles effectuée pour la méthode de Ritz ; voir proposition 3.20 page 109. On peut alors établir le théorème de convergence :

**Théorème 3.25** *Sous les hypothèses du théorème (3.22), si  $u$  est la solution de (3.31) et  $u_N$  la solution de (3.32), alors*

$$\|u - u^{(N)}\|_H \leq \frac{M}{\alpha} d(u, H_N), \quad (3.33)$$

où  $M$  et  $\alpha$  sont tels que :  $\alpha \|v\|^2 \leq a(v, v) \leq M \|v\|^2$  pour tout  $v$  dans  $H$  (les réels  $M$  et  $\alpha$  existent en vertu de la continuité et de la coercitivité de  $a$ ).

**Démonstration :** Comme la forme bilinéaire  $a$  est coercive de constante  $\alpha$ , on a :

$$\alpha \|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)}, u - u^{(N)})$$

On a donc, pour tout  $v \in H$  :

$$\alpha \|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)}, u - v) + a(u - u^{(N)}, v - u^{(N)})$$

Or  $a(u - u^{(N)}, v - u^{(N)}) = a(u, v - u^{(N)}) - a(u^{(N)}, v - u^{(N)})$  et par définition de  $u$  et  $u^{(N)}$ , on a :

$$a(u, v - u^{(N)}) = T(v - u^{(N)})$$

$$a(u^{(N)}, v - u^{(N)}) = T(v - u^{(N)})$$

On en déduit que :

$$\alpha \|u - u^{(N)}\|_H^2 \leq a(u - u^{(N)}, u - v), \forall v \in H_N,$$

et donc, par continuité de la forme bilinéaire  $a$  :

$$\alpha \|u - u^{(N)}\|_H^2 \leq M \|u - u^{(N)}\|_H \|u - v\|_H.$$

On obtient donc :

$$\|u - u^{(N)}\|_H \leq \frac{M}{\alpha} \|u - v\|_H, \forall v \in H_N,$$

ce qui entraîne (3.33). ■

**Remarque 3.26** *On peut remarquer que l'estimation (3.33) obtenue dans le cadre de la méthode de Galerkin est moins bonne que l'estimation (3.29) obtenue dans le cadre de la méthode de Ritz. Ceci est moral, puisque la méthode de Ritz est un cas particulier de la méthode de Galerkin.*

Grâce au théorème 3.25, on peut remarquer que  $u^{(N)}$  converge vers  $u$  dans  $H$  lorsque  $N$  tend vers  $+\infty$  dès que  $d(u, H_N) \rightarrow 0$  lorsque  $N \rightarrow +\infty$ . C'est donc là encore une propriété de consistance dont nous avons besoin. La propriété de consistance n'est pas toujours facile à montrer directement. On utilise alors la caractérisation suivante :

**Proposition 3.27 (Caractérisation de la consistance)** *Soit  $V$  un sous espace vectoriel de  $H$  dense dans  $H$ . On suppose qu'il existe une fonction  $r_N : V \rightarrow H_N$  telle que*

$$\|v - r_N(v)\|_H \rightarrow_{N \rightarrow +\infty} 0,$$

alors

$$d(u, H_N) \rightarrow_{N \rightarrow +\infty} 0$$

**Démonstration :** Soit  $v \in V$ , et  $w = r_N(v)$ . Par définition, on a

$$\begin{aligned} d(u, H_N) &\leq \|u - r_N(v)\|_H \\ &\leq \|u - v\|_H + \|v - r_N(v)\|_H \end{aligned}$$

Comme  $V$  est dense dans  $H$ , pour tout  $\varepsilon > 0$ , il existe  $v \in V$ , tel que  $\|u - v\|_H \leq \varepsilon$ . Choisissons  $v$  qui vérifie cette dernière inégalité. Par hypothèse sur  $r_N$  :

$$\forall \varepsilon > 0, \exists N_0/N \geq N_0 \text{ alors } \|v - r_N(v)\|_H \leq \varepsilon.$$

Donc si  $N \geq N_0$ , on a  $d(u, H_N) \leq 2\varepsilon$ . On en déduit que  $d(u, H_N) \rightarrow 0$  quand  $N \rightarrow +\infty$ . ■

### 3.2.3 Méthode de Petrov-Galerkin

La méthode de Petrov-Galerkin s'apparente à la méthode de Galerkin. On cherche toujours à résoudre :

$$\begin{cases} a(u, v) = T(v), & \forall v \in H, \\ u \in H \end{cases}$$

Mais on choisit maintenant deux sous-espaces  $H_N$  et  $V_N$  de  $H$ , tous deux de même dimension finie :

$$\dim H_N = \dim V_N = N.$$

On cherche une approximation de la solution du problème dans l'espace  $H_N$ , et on choisit comme fonction test les fonctions de base de  $V_N$ . On obtient donc le système :

$$\begin{cases} u \in H_N \\ a(u, v) = T(v) & \forall v \in V_N \end{cases}$$

On appelle  $H_N$  l'espace d'approximation, et  $V_N$  l'espace des fonctions test. Si  $(\phi_1, \dots, \phi_N)$  est une base de  $H_N$  et  $(\psi_1, \dots, \psi_N)$  une base de  $V_N$ , en développant  $u^{(N)}$  sur la base de  $(\phi_1, \dots, \phi_N)$ .  $u^{(N)} = \sum u_j \phi_j$ , et en écrivant (3.2.3) pour  $v = \phi_j$ , on obtient :

$$\begin{cases} u \in H_N \\ a(u, \psi_i) = T(\psi_i), & \forall i = 1, \dots, N. \end{cases}$$

Le système à résoudre est donc :

$$\begin{cases} u^{(N)} = \sum_{i=1}^N u_i \phi_i, \\ \mathcal{K}U = \mathcal{G}, \end{cases}$$

avec  $\mathcal{K}_j = a(\phi_j, \psi_i)$  et  $\mathcal{G}_i = T(\psi_i)$ , pour  $i = 1, \dots, N$ .

## 3.3 La méthode des éléments finis

La méthode des éléments finis est une façon de choisir les bases des espaces d'approximation pour les méthodes de Ritz et Galerkin.

### 3.3.1 Principe de la méthode

On se limitera dans le cadre de ce cours à des problèmes du second ordre. L'exemple type sera le problème de Dirichlet (3.1), qu'on rappelle ici :

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

et l'espace de Hilbert sera l'espace de Sobolev  $H^1(\Omega)$  ou  $H_0^1(\Omega)$ .

On se limitera à un certain type d'éléments finis, dits "de Lagrange". Donnons les principes généraux de la méthode.

**Éléments finis de Lagrange** Soit  $\Omega \subset \mathbb{R}^2$  (ou  $\mathbb{R}^3$ ), Soit  $H$  l'espace fonctionnel dans lequel on recherche la solution (par exemple  $H_0^1(\Omega)$  s'il s'agit du problème de Dirichlet (3.1)). On cherche  $H_N \subset H = H_0^1(\Omega)$  et les fonctions de base  $\phi_1, \dots, \phi_N$ . On va déterminer ces fonctions de base à partir d'un découpage de  $\Omega$  en un nombre fini de cellules, appelés, "éléments". la procédure est la suivante :

1. On construit un "maillage"  $\mathcal{T}$  de  $\Omega$  (en triangles ou rectangles) que l'on appelle éléments  $\mathcal{K}$ .
2. Dans chaque élément, on se donne des points que l'on appelle "noeuds".
3. On définit  $H_N$  par :

$$H_N = \{u : \Omega \rightarrow \mathbb{R} / u|_K \in P_k, \forall K \in \mathcal{T}\} \cap H$$

où  $P_k$  désigne l'ensemble des polynômes de degré inférieur ou égal à  $k$ . Le degré des polynômes est choisi de manière à ce que  $u$  soit entièrement déterminée par ses valeurs aux noeuds. Pour une méthode d'éléments finis de type Lagrange, les valeurs aux noeuds sont également les "degrés de liberté", c.à.d. les valeurs qui déterminent entièrement la fonction recherchée.

4. On construit une base  $\{\phi_i \dots \phi_N\}$  de  $H_N$  tel que le support de  $\phi_i$  soit “le plus petit possible”. Les fonctions  $\phi_i$  sont aussi appelées fonctions de forme.

**Remarque 3.28 (Eléments finis non conformes)** Notons qu’on a introduit ici une méthode éléments finis conforme, c’est-à-dire que l’espace d’approximation  $H_N$  est inclus dans l’espace  $H$ . Dans une méthode non conforme, on n’aura plus  $H_N \subset H$ , et par conséquent, on devra aussi construire une forme bilinéaire approchée  $a_T$ ; on pourra voir à ce sujet l’exercice 46 page 122 où on exprime la méthode des volumes finis comme une méthode éléments finis non conformes.

**Exemple en dimension 1** Soit  $\Omega = ]0, 1[ \subset \mathbb{R}$  et soit  $H = H_0^1(]0, 1[)$ ; on cherche un espace  $H_N$  d’approximation de  $H$ . Pour cela, on divise l’intervalle  $]0, 1[$  en  $N$  intervalles de longueur  $h = \frac{1}{N+1}$ . On pose  $x_i = i$ ,  $i = 0, N + 1$ . Les étapes 1. à 4. décrites précédemment donnent dans ce cas :

1. Construction des éléments On a construit  $n + 1$  éléments  $K_i = ]x_i, x_{i+1}[$ ,  $i = 0, \dots, N$ .
2. Noeuds : On a deux noeuds par élément, ( $x_i$  et  $x_{i+1}$  sont les noeuds de  $K_i$ ,  $i = 0, \dots, N$ ) Le fait que  $H_N \subset H_0^1(]0, 1[)$  impose que les fonctions de  $H_N$  soient nulles en  $x_0 = 0$  et  $x_{N+1} = 1$ . On appelle  $x_1, \dots, x_N$  les noeuds libres et  $x_0, x_{N+1}$  les noeuds liés. Les degrés de liberté sont donc les valeurs de  $u$  en  $x_1, \dots, x_N$ . Aux noeuds liés, on a  $u(x_0) = u(x_{N+1}) = 0$ .
3. Choix de l’espace On choisit comme espace de polynôme :  $P_1 = \{ax + b, a, b \in \mathbb{R}\}$  et on pose :

$$H_N = \{u : \Omega \rightarrow \mathbb{R} \text{ t.q. } u|_{K_i} \in P_1, \forall i = 1 \dots N, u \in C(\bar{\Omega}) = C([0, 1]) \text{ et } u(0) = u(1) = 0\}.$$

Rappelons que  $H = H_0^1(]0, 1[) \subset C([0, 1])$ . Avec le choix de  $H_N$ , on a bien  $H_N \subset H$ .

4. Choix de la base de  $H_N$ .

Si on prend les fonctions de “type 1” de la méthode de Ritz, on choisit les fonctions décrites sur la figure 3.1. On a donc  $H_1 = \text{Vect}\{\phi_1\}$ ,  $H_3 = \text{Vect}\{\phi_1, \phi_2, \phi_3\}$ , et  $H_7 = \text{Vect}\{\phi_1, \phi_2, \phi_3, \phi_4, \phi_5, \phi_6, \phi_7\}$ , où  $\text{Vect}$  désigne le sous espace engendré par la famille considérée. Avec ce choix, on a donc  $H_1 \subset H_3 \subset H_7$ .

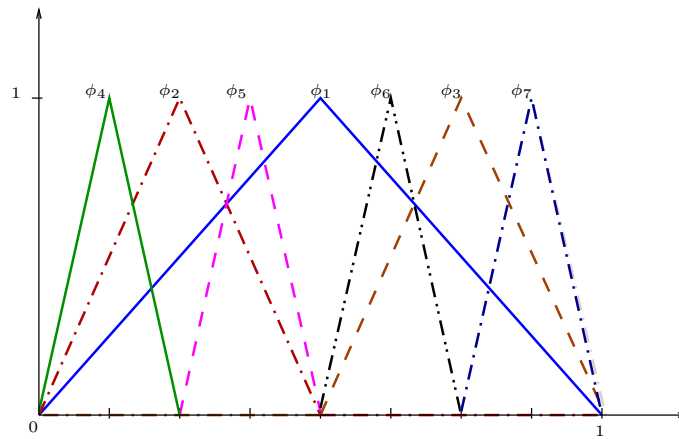


FIG. 3.1 – Fonctions de forme de type 1 (espaces emboîtés)

Si maintenant on choisit des fonctions de forme de “type 2” on peut définir  $\phi_i$  pour  $i = 1$  à  $N$  par :

$$\left\{ \begin{array}{l} \phi_i : \text{ affine par morceaux, continue} \\ \text{supp}(\phi_i) = [x_{i-1}, x_{i+1}] \\ \phi_i(x_i) = 1 \\ \phi_i(x_{i-1}) = \phi_i(x_{i+1}) = 0 \end{array} \right. \quad (3.34)$$

Il est facile de voir que  $\phi_i \in H_N$  et que  $\{\phi_1, \dots, \phi_N\}$  engendre  $H_N$ , c’est à dire que pour tout  $u \in H_N$ , il existe  $(u_1, \dots, u_N) \in \mathbb{R}^N$  tel que  $u = \sum_{i=1}^N u_i \phi_i$ . On a représenté sur la figure 3.2 les fonctions de base

obtenue pour  $H_3$  (à gauche) et  $H_7$  (à droite). On peut remarquer que dans ce cas, les espaces d'approximation ne sont plus inclus les uns dans les autres.

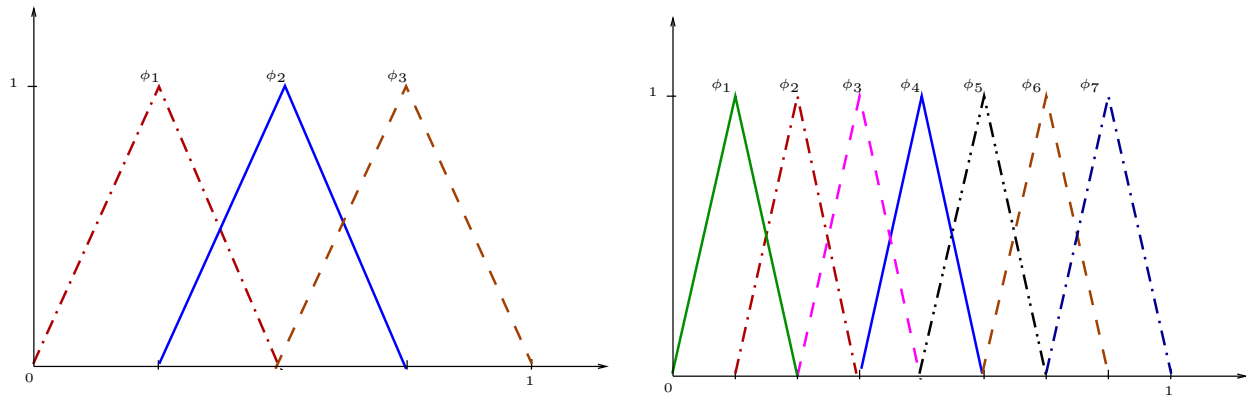


FIG. 3.2 – Fonctions de forme de type 2 (fonction P1) en une dimension d'espace

**Exemple en dimension 2** Soit  $\Omega$  un ouvert polygonal de  $\mathbb{R}^2$ , et  $H = H_0^1(\Omega)$ . Les étapes de construction de la méthode des éléments finis sont encore les mêmes.

1. **Eléments** : on choisit des triangles.
2. **Noeuds** : on les place aux sommets des triangles. Les noeuds  $x_i \in \Omega$  (intérieurs à  $\Omega$ ) sont libres, et les noeuds  $x_i \in \partial\Omega$  (sur la frontière de  $\Omega$ ) sont liés. On notera  $\Sigma$  l'ensemble des noeuds libres,  $\Sigma_F$  l'ensemble des noeuds liés, et,  $\Sigma = \Sigma_I \cup \Sigma_F$ .
3. **Espace d'approximation** L'espace des polynômes est l'ensemble des fonctions affines, noté  $P_1$ . Une fonction  $p \in P_1$  est de la forme :

$$p : \mathbb{R}^2 \rightarrow \mathbb{R}, \\ x = (x_1, x_2) \mapsto a_1x_1 + a_2x_2 + b,$$

avec  $(a_1, a_2, b) \in \mathbb{R}^3$ . L'espace d'approximation  $H_N$  est donc défini par :

$$H_N : \{u \in C(\bar{\Omega}); u|_K \in P_1, \forall K, \text{ et } u(x_i) = 0, \forall x_i \in \Sigma_F\}$$

4. **Base de  $H_N$**  : On choisit comme base de  $H_N$  la famille de fonctions  $\{\phi_i\}_{i=1, \dots, N}$ , où  $N = \text{card}(\Sigma_I)$ , où  $\phi_i$  est définie, pour  $i = 1$  à  $N$ , par :

$$\begin{cases} \phi_i \text{ est affine par morceaux,} \\ \phi_i(x_i) = 1, \\ \phi_i(x_j) = 0, \quad \forall j \neq i. \end{cases} \quad (3.35)$$

La fonction  $\phi_i$  associée au noeud  $x_i$  a donc l'allure présentée sur la figure 3.3. Le support de chaque fonction  $\phi_i$  (c'est à dire l'ensemble des points où  $\phi_i$  est non nulle), est constitué de l'ensemble des triangles dont  $x_i$  est un sommet.

**En résumé** Les questions à se poser pour construire une méthode d'éléments finis sont donc :

1. La construction du maillage.
2. Un choix cohérent entre éléments, noeuds et espace des polynômes.
3. La construction de l'espace d'approximation  $H_N$  et de sa base  $\{\phi_i\}_{i=1 \dots N}$ .
4. La construction de la matrice de rigidité  $\mathcal{K}$  et du second membre  $\mathcal{G}$ .
5. L'évaluation de  $d(u, H_N)$  en vue de l'analyse de convergence.

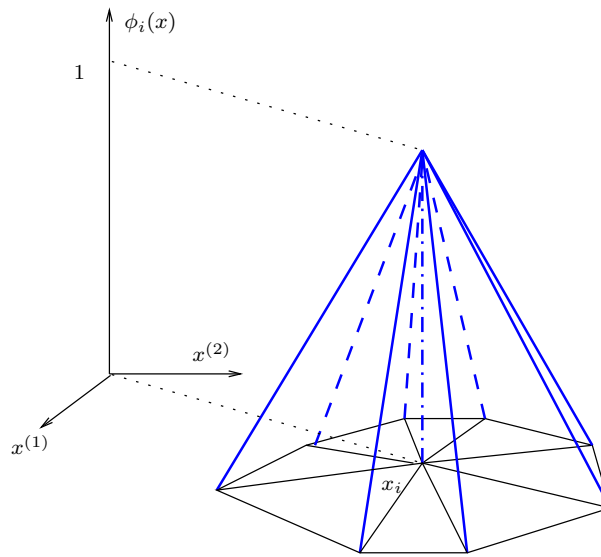


FIG. 3.3 – Fonction de forme de type 2 (fonction P1) en deux dimensions d’espace

### 3.3.2 Construction du maillage, de l’espace $H_N$ et de sa base $\phi_N$

#### Construction des éléments

Soit  $\Omega \in \mathbb{R}^2$  un ouvert borné polygonal. On construit un maillage de  $\Omega$  en divisant  $\bar{\Omega}$  en parties fermées  $\{K_\ell\}_{\ell=1,\dots,L}$  où  $L$  est le nombre d’éléments.

Les principes pour la construction du maillage sont :

- Eviter les angles trop grands ou trop petits. On préférera par exemple les triangles de gauche plutôt que ceux de droite dans la figure 3.4.

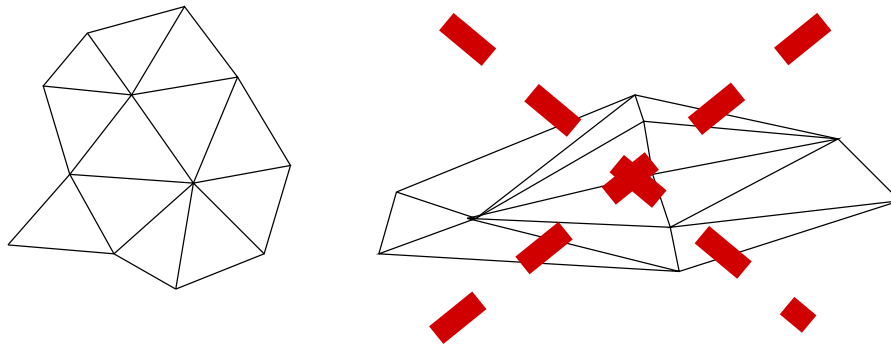


FIG. 3.4 – Exemple de triangles “bons” (à gauche) et “mauvais” (à droite)

- Mettre beaucoup d’éléments là où  $u$  varie rapidement (ceci ne peut se faire que si on connaît a priori les zones de de variation rapide, ou si on a les moyens d’évaluer l’erreur entre la solution exacte du problème et la solution calculée et de remailler les zones où celle-ci est jugée trop grande.
- On peut éventuellement mélanger des triangles et des rectangles, mais ceci n’est pas toujours facile.

Il existe un très grand nombre de logiciels de maillages en deux ou trois dimensions d’espace. On pourra pour s’en convaincre utiliser le moteur de recherche google sur internet avec les mots clés : “mesh 2D structured”, “mesh 2D unstructured”, “mesh 3D structured”, “mesh 3D unstructured”. Le mot “mesh” est le terme anglais pour maillage, les termes 2D et 3D réfèrent à la dimension de l’espace physique. Le terme “structured” (structuré en français) désigne des maillages que dont on peut numéroter les éléments de façon cartésienne, le terme “unstructured” (non structuré) désigne tous les autres maillages. L’avantage des maillages “structurés” est qu’ils nécessitent une base de données beaucoup plus simple que les maillages non structurés, car on peut connaître tous les noeuds voisins à partir du numéro global d’un noeud d’un maillage structuré, ce qui n’est pas le cas dans un maillage non

structuré (voir paragraphe suivant pour la numérotation des noeuds). La figure 3.5 montre un exemple de maillage

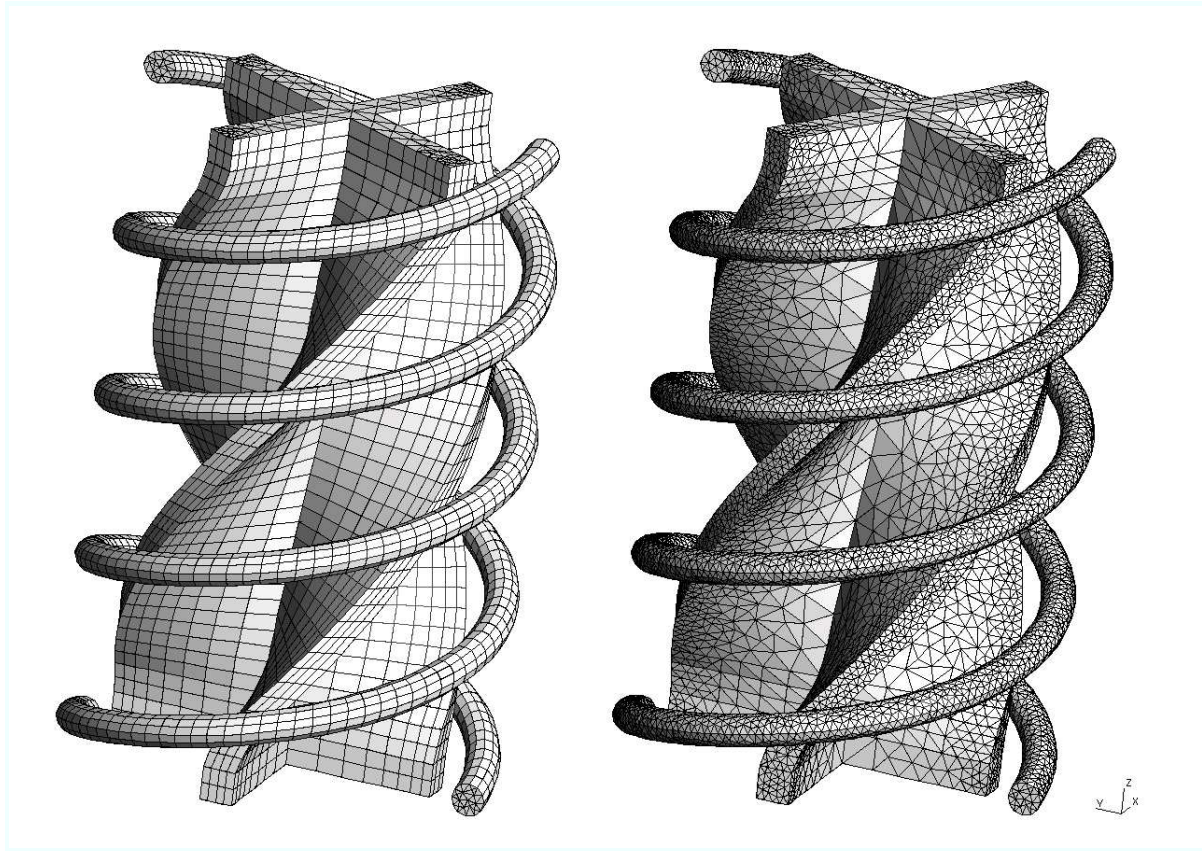


FIG. 3.5 – Exemple de maillage structuré (à gauche) et non-structuré (à droite) d’une surface

de surface structuré ou non-structuré, pris sur le site web du logiciel Gmsh : a three-dimensional finite element mesh generator with built-in pre- and post-processing facilities, développé par C. Geuzaine and J.-F. Remacle (<http://www.geuz.org/gmsh/>).

### Choix des noeuds

On se donne une famille  $\{S_i\}_{i=1,\dots,M}$  de  $M$  points de  $\bar{\Omega}$ , de composantes  $(x_i, y_i)$ , pour  $i = 1, \dots, M$ . Le maillage éléments finis est défini par éléments  $\{K_\ell\}_{\ell=1,\dots,L}$  et les noeuds  $\{S_i\}_{i=1,\dots,M}$ . Ces éléments et noeuds ne peuvent bien sûr pas être choisis indépendamment. Dans le cas général, on choisit tous les éléments de même type (par exemple, des triangles) et on se donne un nombre fixe de noeuds par élément, ce qui détermine le nombre total de noeuds. Chaque noeud appartient donc à plusieurs éléments. Dans le cas d’un maillage structuré tel que celui qu’on a décrit dans la figure 1.5 page 27, une numérotation globale des noeuds est suffisante pour retrouver les éléments dont font partie ce noeud, ainsi que tous les voisins du noeud. Par contre, dans le cas d’un maillage non structuré (un maillage en triangles, par exemple), on aura besoin d’une numérotation locale des noeuds c’est à dire une numérotation des noeuds de chaque élément, pour  $k = 1, \dots, N_\ell$ , où  $N_\ell$  est le nombre de noeuds par élément ; on aura également besoin d’une numérotation globale des noeuds, et d’une table de correspondance, l’une qui donne pour chaque élément, les numéros dans la numérotation globale des noeuds qui lui appartiennent.

$$i_r^\ell = \text{notation globale } (\ell, r) \text{ r-ième noeud de l'élément } \ell$$

### Amélioration de la précision

On a vu aux paragraphes précédents que l’erreur entre la solution exacte  $u$  recherchée et la solution  $u(N)$  obtenue par la méthode de Ritz ou de Galerkin est majorée par une constante fois la distance entre  $H$  et  $H_N$ . On a donc

intéret à ce que cette distance soit petite. Pour ce faire, il paraît raisonnable d'augmenter la dimension de l'espace  $H_N$ . Pour cela, on a deux possibilités :

- augmenter le nombre d'éléments : on augmente alors aussi le nombre global de noeuds, mais pas le nombre local.
- augmenter le degré des polynômes : on augmente alors le nombre de noeuds local, donc on augmente aussi le nombre global de noeuds, mais pas le nombre d'éléments. Ce deuxième choix (augmentation du degré des polynômes) ne peut se faire que si la solution est suffisamment régulière ; si la solution n'est pas régulière, on n'arrivera pas à diminuer  $d(H, H_N)$  en augmentant le degré des polynômes.

### 3.4 Exercices

**Exercice 33** ( Fonctions  $H^1$  en une dimension d'espace )

Montrer que si  $u \in H^1(]0, 1[)$ , alors  $u$  est continue. En déduire que  $H^2(]0, 1[) \subset C^1([0, 1])$ .

**Exercice 34** (Minimisation de la semi-norme) *Suggestions en page 125, corrigé en page 126*

Soit  $\Omega$  un ouvert borné de  $\mathbb{R}^n$ . On suppose que sa frontière est de classe  $C^1$  par morceaux. Etant donné une fonction  $u_0 \in H^1(\Omega)$ , on désigne par  $u_0 + H_0^1(\Omega)$  l'ensemble  $\{u_0 + v, v \in H_0^1(\Omega)\}$ .

1. Montrer qu'il existe une unique fonction  $u \in u_0 + H_0^1(\Omega)$  tel que :

$$|u|_{1,\Omega} = \inf_{v \in u_0 + H_0^1(\Omega)} |v|_{1,\Omega}.$$

2. Caractériser  $u$  comme étant la solution d'un problème aux limites.

**Exercice 35** (Formulation faible pour le problème de Dirichlet en 1D) *Corrigé en page 127*

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse au problème suivant :

$$-u''(x) = f(x), \quad x \in ]0, 1[, \quad (3.36)$$

$$u(0) = 0, \quad u(1) = 0. \quad (3.37)$$

Donner une formulation faible et une formulation variationnelle de (3.37).

**Exercice 36** (Relèvement) *Corrigé en page 129*

Soient  $a$  et  $b \in \mathbb{R}$ , et  $f \in C(\mathbb{R}, \mathbb{R})$ .

1. Soient  $u_0$  et  $u_1$  définies de  $[0, 1]$  dans  $\mathbb{R}$  par  $u_0(x) = a + (b - a)x$  et  $u_1(x) = a + (b - a)x^2$ . Montrer qu'il existe un unique  $\tilde{u}$  (resp.  $\bar{u}$ ) tel que  $u = u_0 + \tilde{u}$  (resp.  $v = u_1 + \bar{u}$ ) soit solution de (3.10). Montrer que  $u = v$ .
2. Mêmes questions en supposant maintenant que  $u_0$  et  $u_1$  sont des fonctions de  $C^2([0, 1])$  telles que  $u_0(0) = u_1(0) = a$  et  $u_0(1) = u_1(1) = b$ .

**Exercice 37** (Relèvement en une dimension d'espace) *Suggestions en page 125*

Ecrire une formulation faible pour laquelle on puisse appliquer le théorème de Lax Milgram, dans le cas du problème suivant :

$$\begin{cases} -u''(x) = f(x), & x \in [0, 1] \\ u'(0) = 0 \\ u(1) = 1. \end{cases} \quad (3.38)$$

**Exercice 38** (Conditions aux limites de Fourier et Neumann) *Corrigé en page 130*

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse au problème suivant :

$$\begin{cases} -u_{xx}(x) + u(x) = f(x), & x \in ]0, 1[, \\ u'(0) - u(0) = 0, \quad u'(1) = -1. \end{cases} \quad (3.39)$$

Donner une formulation faible et une formulation variationnelle de (3.39) ; y-a-t-il existence et unicité des solutions faibles de (3.39) ?



**Exercice 39 (Conditions aux limites de Fourier et Neumann, bis)**

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse au problème suivant :

$$\begin{cases} -u''(x) - u'(x) + u(x) = f(x), & x \in ]0, 1[, \\ u(0) + u'(0) = 0, & u(1) = 1 \end{cases} \quad (3.40)$$

1. Donner une formulation faible du problème de la forme

$$\begin{cases} \text{Trouver } u \in H^1(]0, 1[); u(1) = 1, \\ a(u, v) = T(v), \forall v \in H. \end{cases} \quad (3.41)$$

où  $H = \{v \in H^1(]0, 1[); v(1) = 0\}$ ,  $a$  et  $T$  sont respectivement une forme bilinéaire sur  $H^1(]0, 1[)$  et une forme linéaire sur  $H^1(]0, 1[)$ , à déterminer.

2. Y-a-t-il existence et unicité de solutions de cette formulation faible ?

**Exercice 40 (Conditions mixtes) Suggestions en page 126, corrigé en page 131**

Soit  $\Omega$  un ouvert borné  $\mathbb{R}^d$ ,  $d = 1$  ou  $2$ , de frontière  $\partial\Omega = \Gamma_0 \cup \Gamma_1$ , avec  $\Gamma_0 \cap \Gamma_1 = \emptyset$ ; on suppose que la mesure  $d - 1$  dimensionnelle de  $\Gamma_0$  est non nulle, et soit  $f \in L^2(\Omega)$ . On s'intéresse ici au problème suivant :

$$\begin{cases} -\Delta u(x) = f(x), & x \in \Omega, \\ u(x) = 0, & x \in \Gamma_0, \\ \nabla u(x) \cdot \mathbf{n}(x) = 0, & x \in \Gamma_1, \end{cases} \quad (3.42)$$

où  $\mathbf{n}$  est la normale unitaire à  $\partial\Omega$  extérieure à  $\Omega$ .

Donner une formulation faible et une formulation variationnelle de (3.42) telle qu'on puisse appliquer le lemme de Lax-Milgram. (On rappelle que l'inégalité de Poincaré donnée en bas de page 12 page 24 pour les fonctions de  $H_0^1(\Omega)$  est encore valable pour les fonctions de  $H^1(\Omega)$  dont la trace est nulle sur un sous-ensemble de  $\partial\Omega$  de mesure  $((d - 1)$ -dimensionnelle) non nulle.)

**Exercice 41 (Problème elliptique pour un problème avec conditions mixtes) Corrigé en page 132**

Soit  $\Omega$  un ouvert borné  $\mathbb{R}^d$ ,  $d = 1$  ou  $2$ , de frontière  $\partial\Omega = \Gamma_0 \cup \Gamma_1$ , avec  $\Gamma_0 \cap \Gamma_1 = \emptyset$ ; on suppose que la mesure  $d - 1$  dimensionnelle de  $\Gamma_0$  est non nulle. On s'intéresse ici au problème suivant :

$$\begin{cases} -\operatorname{div}(p(x)\nabla u(x)) + q(x)u(x) = f(x), & x \in \Omega, \\ u(x) = g_0(x), & x \in \Gamma_0, \\ p(x)\nabla u(x) \cdot \mathbf{n}(x) + \sigma u(x) = g_1(x), & x \in \Gamma_1, \end{cases} \quad (3.43)$$

où :

$$\begin{aligned} f &\in L^2(\Omega), \\ p &\in L^\infty(\Omega), \text{ est telle qu'il existe } \alpha > 0 \text{ t.q. } p(x) \geq \alpha \text{ p.p.} \\ q &\in L^\infty(\Omega), q \geq 0, \\ \sigma &\in \mathbb{R}_+, \\ g_0 &\in L^2(\Gamma_0) \text{ est telle qu'il existe } \tilde{g} \in H^1(\Omega) \text{ t.q. } \gamma(\tilde{g})|_{\Gamma_0} = g_0 \\ g_1 &\in L^2(\Gamma_1), \end{aligned}$$

$\mathbf{n}$  est la normale unitaire à  $\partial\Omega$  extérieure à  $\Omega$ .

1. Donner une formulation faible et une formulation variationnelle de (3.43) telle qu'on puisse appliquer le lemme de Lax-Milgram.

2. On suppose dans cette question que  $p \in C^1(\overline{\Omega})$ ,  $q \in C(\overline{\Omega})$ ,  $g_0 \in C(\Gamma_0)$  et  $g_1 \in C(\Gamma_1)$ . Soit  $u \in C^2(\overline{\Omega})$ . Montrer que  $u$  est solution faible si et seulement si  $u$  est une solution classique de (3.43).

**Exercice 42 (Problème de Neumann homogène) Suggestions en page 125**

On considère le problème suivant :

$$-\Delta u + u = f \text{ dans } \Omega, \quad (3.44)$$

$$\nabla u \cdot \mathbf{n} = 0 \text{ sur } \Omega \quad (3.45)$$

où  $\Omega$  est un ouvert borné de  $\mathbb{R}^d$  de frontière régulière de classe  $C^2$  et de normale unitaire extérieure  $\mathbf{n}$ , et  $f \in L^2(\Omega)$ .

1. Montrer que si  $u$  est une solution régulière de (3.45) et  $\varphi \in C^\infty(\mathbb{R}^d)$ , alors

$$\int_{\Omega} (\nabla u \cdot \nabla \varphi + u \varphi) dx = \int_{\Omega} f \varphi dx$$

En déduire que  $u$  est solution de la formulation faible suivante :

$$u \in H^1(\Omega), \quad (3.46)$$

$$\int_{\Omega} (\nabla u \cdot \nabla v + uv) dx = \int_{\Omega} f v dx, \quad \forall v \in H^1(\Omega). \quad (3.47)$$

2. Montrer que si  $u$  est une solution régulière de classe  $C^2$  de (3.47) et  $\varphi \in C^\infty(\mathbb{R}^d)$ , alors  $u$  est solution de (3.45).

**Exercice 43 ( Un problème de réaction-diffusion )** *Suggestions en page 125, corrigé en page 133*

Soit  $\alpha$  un réel positif ou nul, et soit  $f$  une fonction continue. On considère le problème suivant :

$$\begin{aligned} -u''(x) + \alpha u(x) &= f(x), \quad x \in ]0, 1[, \\ u'(0) &= u(0), \\ u'(1) &= 0. \end{aligned} \quad (3.48)$$

où  $u''$  désigne la dérivée seconde de  $u$  par rapport à  $x$ .

1. Ecrire une formulation variationnelle de (4.28).

2. On considère le problème

$$u \in H^1(]0, 1[) \\ \int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) = \int_0^1 f(x)v(x) dx, \quad \forall v \in H^1(]0, 1[). \quad (3.49)$$

2.a Déterminer le problème aux limites dont la formulation faible est (4.29).

2.b. Montrer que si  $\alpha > 0$ , (4.29) admet une unique solution.

3. Montrer que ceci est encore vrai pour  $\alpha = 0$  en appliquant l'inégalité de Poincaré à la fonction  $v - v(0)$ .

**Exercice 44 (Condition inf-sup)** *Corrigé en page 134*

Soit  $V$  un espace de Hilbert réel de produit scalaire  $(\cdot; \cdot)$  induisant une norme  $\|\cdot\|$ . On se donne  $a(\cdot; \cdot)$  une forme bilinéaire continue sur  $V \times V$ , avec  $M$  comme constante de continuité. Soit  $L$  une forme linéaire continue sur  $V$ . On suppose de plus qu'il existe une solution  $u \in V$  au problème suivant :

$$a(u, v) = L(v), \quad \forall v \in V. \quad (3.50)$$

Soit  $V_h$  un sous-espace de  $V$  de dimension finie. On suppose qu'il existe  $\beta_h \in R_+$  telle que :

$$\inf_{(v_h \in V_h, \|v_h\|=1)} \left( \sup_{(w_h \in V_h, \|w_h\|=1)} (a(v_h; w_h)) \right) \geq \beta_h \quad (3.51)$$

On cherche alors  $u_h$  solution de :

$$\begin{aligned} u_h &\in V_h, \\ a(u_h, v_h) &= L(v_h), \quad \forall v_h \in V_h \end{aligned} \quad (3.52)$$

1. Montrer que le problème (3.52) admet une unique solution.

2. Soit  $u$  la solution de (3.50) et  $u_h$  la solution de (3.52). Montrer que :

$$\|u - u_h\| \leq \left(1 + \frac{M}{\beta_h}\right) \inf_{v_h \in V_h} \|u - v_h\|. \quad (3.53)$$

**Exercice 45 (Condition inf-sup pour un problème mixte)**

Soient  $V$  et  $Q$  deux espaces de Hilbert, on note  $(\cdot, \cdot)_V$ ,  $\|\cdot\|_V$  et  $(\cdot, \cdot)_Q$ ,  $\|\cdot\|_Q$  leurs produits scalaires et normes respectives, et on considère le problème suivant :

$$\begin{cases} \text{Trouver } u \in V, p \in Q, \text{ tels que} \\ a(u, v) + b(v, p) = (f, v)_H, \forall v \in V, \\ b(u, q) = (g, q)_Q, \forall q \in Q. \end{cases} \quad (3.54)$$

où  $a$  est une forme bilinéaire continue et coercive sur  $V$  et  $b$  est une application bilinéaire continue de  $V \times Q$  dans  $\mathbb{R}$ .

Pour  $(u, p)$  et  $(v, q)$  éléments de  $V \times Q$ , on pose :

$$\begin{aligned} B(u, p; v, q) &= a(u, v) + b(v, p) + b(u, q), \\ F(v, q) &= (f, v)_H + (g, q)_Q. \end{aligned}$$

et on munit  $V \times Q$  d'une norme notée  $\|(\cdot, \cdot)\|$ , définie par  $\|(v, q)\| = \|v\|_V + \|q\|_Q$  pour  $(v, q) \in V \times Q$ .

1. Montrer que  $B$  est une forme bilinéaire continue sur  $V \times Q$ .
2. Montrer que le problème (3.54) est équivalent au problème :

$$\begin{cases} \text{Trouver } (u, p) \in V \times Q, \text{ tels que} \\ B(u, p; v, q) = F(v, q), \forall (v, q) \in V \times Q. \end{cases} \quad (3.55)$$

On considère maintenant des espaces d'approximation (par exemple construits par éléments finis). Soient donc  $(V_n)_{n \in \mathbb{N}}$  et  $(Q_n)_{n \in \mathbb{N}}$  des espaces de Hilbert de dimension finie tels que  $V_n \subset V$  et  $Q_n \subset Q$ , pour tout  $n \in \mathbb{N}$ .

3. On suppose dans cette question que la condition suivante (dite condition "inf-sup") est satisfaite :

$$\text{Il existe } \beta \in \mathbb{R}_+^* \text{ (indépendant de } n) \text{ tel que } \inf_{q \in Q_n} \sup_{\substack{w \in V_n, \\ \|w\|_V \neq 0}} \frac{b(w, q)}{\|w\|_V} \geq \beta \|q\|_Q. \quad (3.56)$$

- (a) Montrer qu'il existe  $\alpha \in \mathbb{R}_+^*$  tel que :

$$\text{Pour tout } q \in Q_n \text{ et } v \in V_n, B(v, q; v, -q) \geq \alpha \|v\|_V^2. \quad (3.57)$$

- (b) Soit  $(v, q) \in V_n \times Q_n$ , montrer qu'il existe  $w \in V_n$  tel que  $\|w\|_V = \|q\|_Q$  et  $b(w, q) \geq \beta \|q\|_Q^2$ .  
Montrer que pour ce choix de  $w$ , on a :

$$B(v, q; w, 0) \geq -M \|v\|_V \|w\|_V + \beta \|q\|_Q^2,$$

où  $M$  est la constante de continuité de  $a$ .

- (c) En déduire qu'il existe des réels positifs  $C_1$  et  $C_2$  indépendants de  $n$  tels que

$$B(v, q; w, 0) \geq -C_1 \|v\|_V^2 + C_2 \|q\|_Q^2.$$

(On pourra utiliser, en le démontrant, le fait que pour tout  $a_1 \geq 0$ ,  $a_2 \geq 0$  et  $\epsilon > 0$ , on a  $a_1 a_2 \leq \frac{1}{\epsilon} a_1^2 + \epsilon a_2^2$ .)

- (d) Soit  $\gamma \in \mathbb{R}_+^*$ . Montrer que si  $\gamma$  est suffisamment petit, on a :

$$B(v, q; v + \gamma w, -q) \geq C_3 [\|v\|_V^2 + \|q\|_Q^2].$$

et

$$\|(v + \gamma w, -q)\| \leq C_4 \|(v, q)\|,$$

où  $C_3$  et  $C_4$  sont deux réels positifs qui ne dépendent pas de  $n$ .

- (e) En déduire que la condition suivante (dite de stabilité) est satisfaite :

Il existe  $\delta \in \mathbb{R}_+^*$  (indépendant de  $n$ ) tel que pour tout  $(u, p) \in V_n \times Q_n$ ,

$$\sup_{\substack{(v, q) \in V_n \times Q_n, \\ \|(v, q)\| \neq 0}} \frac{B((u, p); (v, q))}{\|(v, q)\|} \geq \delta \|(u, p)\|. \quad (3.58)$$

4. On suppose maintenant que la condition (3.58) est satisfaite.

(a) Montrer que pour tout  $p \in Q$ ,

$$\sup_{\substack{(v,q) \in V_n \times Q_n, \\ \|(v,q)\| \neq 0}} \frac{b(v,p)}{\|(v,q)\|} \geq \delta \|p\|_Q.$$

(b) En déduire que pour tout  $p \in Q$ ,

$$\sup_{\substack{(v,q) \in V_n \times Q_n, \\ \|(v,q)\| \neq 0}} \frac{b(v,p)}{\|v\|_V} \geq \delta \|p\|_Q.$$

5. Déduire des questions précédentes que la condition (3.56) est satisfaite si et seulement si la condition (3.58) est satisfaite.

**Exercice 46 (Volumes finis vus comme des éléments finis non conformes)** *Suggestions en page 126, corrections en page 136*

Soit un ouvert borné polygonal de  $\mathbb{R}^2$ , et  $\mathcal{T}$  un maillage admissible au sens des volumes finis (voir page 1.5.2 page 28) de  $\Omega$ .

1. Montrer que la discrétisation par volumes finis de (3.1) se ramène à chercher  $(u_K)_{K \in \mathcal{T}}$ , qui vérifie :

$$\sum_{\sigma \in \mathcal{E}_{\text{int}}, \sigma = K|L} \tau_\sigma (u_L - u_K) + \sum_{\sigma \in \mathcal{E}_{\text{ext}}, \sigma \in \mathcal{E}_K} \tau_\sigma u_K = m(K) f_K \quad (3.59)$$

où  $\mathcal{E}_{\text{int}}$  représente l'ensemble des arêtes internes (celles qui ne sont pas sur le bord)  $\mathcal{E}_{\text{ext}}$  l'ensemble des arêtes externes (celles qui sont sur le bord), et

$$\tau_\sigma = \begin{cases} \frac{m(\sigma)}{d_{K,\sigma} + d_{L,\sigma}} & \text{si } \sigma \in \mathcal{E}_{\text{int}}, \sigma = K|L, \\ \frac{m(\sigma)}{d_{K,\sigma}} & \text{si } \sigma \in \mathcal{E}_{\text{ext}}, \sigma \in \mathcal{E}_K, \end{cases} \quad (3.60)$$

(voir figure 1.6 page 28).

2. On note  $H_{\mathcal{T}}(\Omega)$  le sous-espace de  $L^2(\Omega)$  formé des fonctions constantes par maille (c.à.d. constantes sur chaque élément de  $\mathcal{T}$ ). Pour  $u \in H_{\mathcal{T}}(\Omega)$ , on note  $u_K$  la valeur de  $u$  sur  $K$ . Montrer que  $(u_K)_{K \in \mathcal{T}}$  est solution de (3.60) si et seulement si  $u \in H_{\mathcal{T}}(\Omega)$  est solution de :

$$\begin{cases} u \in H_{\mathcal{T}}(\Omega), \\ a_{\mathcal{T}}(u, v) = T_{\mathcal{T}}(v), \forall v \in H_{\mathcal{T}}(\Omega), \end{cases} \quad (3.61)$$

où  $a_{\mathcal{T}}$  est une forme bilinéaire sur  $H_{\mathcal{T}}(\Omega)$  (à déterminer), et  $T_{\mathcal{T}}$  est une forme linéaire sur  $H_{\mathcal{T}}(\Omega)$  (à déterminer).

**Exercice 47 (Discrétisation du bi-laplacien)** *Corrigé en page ??.*

La modélisation d'une poutre en charge encastree à ses deux extrémités amène à s'intéresser au problème d'ordre 4 suivant (dit problème "biharmonique") :

$$\begin{aligned} u^{(4)}(x) &= f(x), \quad x \in ]0, 1[, \\ u(0) &= 0, \quad u'(0) = 0, \\ u(1) &= 0, \quad u'(1) = 0. \end{aligned} \quad (3.62)$$

où  $u^{(4)}$  désigne la dérivée quatrième de  $u$  par rapport à  $x$ , et  $f$  est une fonction continue.

**Le problème continu**

1. On suppose (dans cette question seulement) que  $f \equiv 1$ . Calculer la solution exacte  $\bar{u}$  de (3.62), et la représenter graphiquement (grossièrement).

2. Soit  $H^2(\]0, 1[)$  l'ensemble des fonctions de carré intégrable dont les dérivées (faibles) première et seconde sont également de carré intégrable :

$$H^2(\]0, 1[) = \{u \in L^2(\]0, 1[), u' \in L^2(\]0, 1[) \text{ et } u'' \in L^2(\]0, 1[).\}$$

On rappelle également que les fonctions de  $H^1(\]0, 1[)$  sont continues sur  $[0, 1]$ .

2.1 Montrer que  $H^2(\]0, 1[) \subset C^1([0, 1])$ .

On définit alors :

$$H_0^2(\]0, 1[) = \{u \in H^2(\]0, 1[); u(0) = u(1) = 0, u'(0) = u'(1) = 0\}.$$

2.2 Montrer que si  $u \in C^4([0, 1])$  est solution de (3.62), alors  $u$  est solution de :

$$\begin{aligned} u \in H_0^2(\]0, 1[) \\ \int_0^1 u''(x)v''(x) dx = \int_0^1 f(x)v(x) dx, \forall v \in H_0^2(\]0, 1[), \end{aligned} \quad (3.63)$$

2.3 Montrer que réciproquement, si  $u$  est solution de (3.63) et  $u \in C^4([0, 1])$ , alors  $u$  est solution de (3.62).

On admettra que le problème (3.63) admet une solution unique.

On cherche maintenant à trouver une solution approchée de la solution de (3.62) ou (3.63).

**Discretisation par différences finies**

3. Soit  $M > 2$  et  $h = \frac{1}{M+1}$ . On construit une subdivision de  $[0, 1]$ , notée  $(y_k)_{k=0, \dots, M+1}$ , définie par :  $y_i = ih$  pour  $i = 0, \dots, M+1$ .

On note  $u_i$  l'inconnue discrète associée au point  $y_i$ ,  $i = 0, \dots, M+1$ .

3.1 Soient  $\varphi \in C^5(\mathbb{R})$ ,  $x \in \mathbb{R}$  et  $h > 0$ , écrire les développements limités en  $x$  de  $\varphi(x+2h)$ ,  $\varphi(x+h)$ ,  $\varphi(x-h)$ , et  $\varphi(x-2h)$  à l'ordre 5 en  $h$ .

3.2 Par une combinaison linéaire adéquate, en déduire pour  $i = 2, \dots, M-1$ , une approximation par différences finies de  $u^{(4)}(y_i)$  en fonction de  $u_{i-2}$ ,  $u_{i-1}$ ,  $u_i$ ,  $u_{i+1}$  et  $u_{i+2}$  qui soit consistante d'ordre 2.

3.3 Ecrire un schéma de différences finies consistant pour la discrétisation de (3.62) sous la forme

$$(\delta^{(4)}u)_i = f(y_i), \quad i = 2, \dots, M-1, \quad (3.64)$$

$$u_0 = u_1 = 0, \quad (3.65)$$

$$u_M = u_{M+1} = 0, \quad (3.66)$$

où  $(\delta^{(4)}u)_i$  est l'approximation consistante de  $u^{(4)}(y_i)$  construite avec les inconnues discrètes  $u_{i-2}$ ,  $u_{i-1}$ ,  $u_i$ ,  $u_{i+1}$  et  $u_{i+2}$  à la question 3.2.

Ecrire le schéma sous forme matricielle, et commenter la structure de la matrice du système linéaire à résoudre.

3.4 Soit  $\delta^{(4)}u \in \mathbb{R}^{M-2}$  dont les composantes sont les valeurs  $(\delta^{(4)}u)_i$  pour  $i = 2, \dots, M-1$ . Notons  $(\delta^{(2)}u)_i = \frac{1}{h}(u_{i+1} + u_{i-1} - 2u_i)$  la discrétisation habituelle de  $u''(y_i)$  par différences finies. A-t'on  $(\delta^{(4)}u)_i = (\delta^{(2)}(\delta^{(2)}u))_i$  pour tout  $i = 2, \dots, M-1$  ?

Dans toute la suite, on considère le maillage suivant : pour  $N > 2$  et  $h = 1/N$ , on définit les  $N$  mailles  $(K_i)_{i=1, \dots, N}$  par  $K_i = ]x_{i-1/2}, x_{i+1/2}[$ , avec  $x_{i+1/2} = ih$  pour  $i = 0, \dots, N$ , et on note  $x_i = (i-1/2)h$  pour  $i = 1, \dots, N$  les centres des mailles. On pose également  $x_0 = 0$  et  $x_{N+1} = 1$ . On définit également des mailles "décalées"  $K_{i+1/2} = ]x_i, x_{i+1}[$ , pour  $i = 0, \dots, N$ .

**Discrétisation par un schéma volumes finis**

4. Soit  $(u_i)_{i=1,N} \in \mathbb{R}^N$  et  $u$  la fonction de  $]0, 1[$  dans  $\mathbb{R}$ , constante par morceaux, définie par

$$u(x) = u_i \text{ pour tout } x \in K_i = ]x_{i-1/2}, x_{i+1/2}[.$$

On définit  $D_h u$  comme la fonction constante par morceaux sur les mailles décalées, définie par

$$D_h u(x) = \begin{cases} D_{i+1/2} u = \frac{1}{h}(u_{i+1} - u_i) \text{ pour tout } x \in K_{i+1/2} = ]x_i, x_{i+1}[ \text{ pour } i = 1, \dots, N-1, \\ D_{1/2} u = \frac{2}{h} u_1 \text{ pour tout } x \in K_{1/2} = ]x_0, x_1[, \\ D_{N+1/2} u = -\frac{2}{h} u_N \text{ pour tout } x \in K_{N+1/2} = ]x_N, x_{N+1}[. \end{cases}$$

Enfin on définit  $D_h^2 u$  comme la fonction constante par morceaux sur les mailles  $K_i$ , définie par :

$$D_h^2 u(x) = D_i^2 u = \frac{1}{h}(D_{i+1/2} u - D_{i-1/2} u) \text{ pour tout } x \in K_i = ]x_{i-1/2}, x_{i+1/2}[ \text{ pour } i = 1, \dots, N.$$

4.a. Calculer  $D_i^2 u$  en fonction des valeurs  $u_j$ ,  $j = 1, \dots, N$ .

4.b En déduire, pour  $k = 1, \dots, N$ , l'expression de la fonction  $D_h^2 \chi_k$ , où  $\chi_k \in \mathbb{R}^N$  est la fonction caractéristique de la maille  $K_k$ , c.à.d :

$$\chi_k(x) = \begin{cases} 1 & \text{si } x \in K_k \\ 0 & \text{sinon.} \end{cases}$$

On note  $H_h$  l'espace des fonctions constantes sur les mailles  $K_i$ ,  $i = 1, \dots, N$ , et  $H_{h,0}$  les fonctions de  $H_h$  nulles sur les mailles 1 et  $N$ . On considère le schéma numérique défini par la forme faible discrète suivante :

$$\text{Trouver } u \in H_{h,0} = \{u \in H_h; u_1 = u_N = 0\}, \quad (3.67)$$

$$\int_0^1 D_h^2 u(x) D_h^2 v(x) dx = \int f(x) v(x) dx, \forall v \in H_{h,0}. \quad (3.68)$$

4.c En prenant les fonctions caractéristiques des mailles  $K_i$  comme fonctions tests dans (3.68), montrer que le schéma (3.67)-(3.68) s'écrit aussi :

$$u \in H_{h,0} \quad (3.69)$$

$$F_{i+1/2}(D_h^2 u) - F_{i-1/2}(D_h^2 u) = \int_{K_i} f(x) dx, i = 1, \dots, N, \quad (3.70)$$

$$F_{i+1/2}(D_h^2 u) = \frac{1}{h}(D_{i+1}^2 u - D_i^2 u), i = 1, \dots, N-1, \quad (3.71)$$

$$F_{1/2}(D_h^2 u) = -\frac{3}{h} D_1^2 u \text{ et } F_{N+1/2}(D_h^2 u) = -\frac{3}{h} D_N^2 u. \quad (3.72)$$

$$(3.73)$$

Expliquer pourquoi ce schéma peut prétendre à l'appellation "volumes finis".

**Quelques propriétés du schéma volumes finis**

On se place ici sous les hypothèses et notations de la discrétisation par volumes finis.

**5. EXISTENCE ET UNICITÉ DE LA SOLUTION DISCRÈTE.**

5.a Montrer que

$$\forall u \in H_{h,0}, \int_0^1 u(x) D_h^2 u(x) dx = \int_0^1 D_h u(x) D_h u(x) dx.$$

5.b Soit  $u \in H_{h,0}$ ; montrer que si  $D_i u = 0$  pour tout  $i = 1, \dots, N$  alors  $u \equiv 0$ .

5.c En déduire que si  $f = 0$ , et si  $u$  est solution de (3.67)-(3.68) alors  $u = 0$ .

5.d En déduire l'existence et l'unicité de  $u$  solution de (3.67)-(3.68).

6. STABILITÉ.

6.1 (Poincaré discret sur  $u$ ). Soit  $u \in H_h$ . Montrer que  $\|u\|_{L^2(\Omega)} \leq \|D_h u\|_{L^2(\Omega)}$ .

6.2 (Poincaré discret sur  $D_h u$ ). Soit  $u \in H_{h,0}$ . Montrer que  $\|D_h u\|_{L^2(\Omega)} \leq \|D_h^2 u\|_{L^2(\Omega)}$ .

6.3 (Estimation a priori sur la solution). Soit  $u \in H_{h,0}$  solution de (3.67)-(3.68). Montrer que  $\|u\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)}$ .

### Discrétisation par un schéma éléments finis non conformes

7. On considère maintenant les fonctions de forme  $\phi_k$  des éléments finis  $P_1$  associées aux noeuds  $x_k, k = 1, \dots, N$ .

7.1. Donner l'expression des fonctions de forme  $\phi_k$  pour  $k = 1, \dots, N$ .

Soit  $V_h$  l'espace engendré par les fonctions  $\phi_1, \dots, \phi_N$  et  $V_{h,0}$  l'espace engendré par les fonctions  $\phi_2, \dots, \phi_{N-1}$ .

Pour  $\tilde{v} \in V_{h,0}$ , on définit  $\tilde{D}_h^2 \tilde{v}$  comme la fonction de  $H_h$  définie par :

$$\tilde{D}_h^2 \tilde{v}(x) = -\frac{1}{h} \sum_{i=2}^{N-1} \tilde{v}(x_i) \int_0^1 \phi'_i(x) \phi'_k(x) dx, \text{ pour tout } x \in K_k, k = 1, \dots, N.$$

On considère alors le schéma suivant pour l'approximation de (3.63).

$$\text{Trouver } \tilde{u} \in V_{h,0}, \tag{3.74}$$

$$\int_0^1 \tilde{D}_h^2 \tilde{u}(x) \tilde{D}_h^2 \tilde{v}(x) dx = \int f(x) \tilde{v}(x) dx, \forall \tilde{v} \in V_{h,0}. \tag{3.75}$$

7.2 Expliquer pourquoi le schéma (3.74)-(3.75) peut prétendre à l'appellation "éléments finis non conformes".

7.3 Soit  $(u_i)_{i=2, \dots, N-1} \in \mathbb{R}^{N-2}$  et soient  $u = \sum_{k=2}^{N-1} u_i \chi_k \in H_{h,0}$  et  $\tilde{u} = \sum_{k=2}^{N-1} u_i \phi_k \in V_{h,0}$ .

(i) Montrer que  $\tilde{u}'$  est une fonction constante par morceaux sur les mailles décalées et comparer  $\tilde{u}'$  à  $D_h u$ .

(ii) Calculer  $\tilde{D}_h^2 \tilde{u}$  et comparer  $\tilde{D}_h^2 \tilde{u}$  à  $D_h^2 u$ .

## 3.5 Suggestions pour les exercices

### Exercice 34 page 118

Rappel ; par définition, l'ensemble  $u_0 + H_0^1$  est égal à l'ensemble  $\{v = u_0 + w, w \in H_0^1\}$

1. Montrer que le problème s'écrit sous la forme :  $J(u) \leq J(v), \forall v \in H$ , ou  $H$  est un espace de Hilbert, avec  $J(v) = a(v, v)$ , où  $a$  est une forme bilinéaire symétrique définie positive.

2. Prendre une fonction test à support compact dans la formulation faible.

### Exercice 37 page 118

Considérer les espaces

$$H_{1,1}^1 = \{v \in H^1(]0, 1[); v(1) = 1\} \quad \text{et}$$

$$H_{1,0}^1 = \{v \in H^1(]0, 1[); v(1) = 0\}.$$

### Exercice 42 page 119

1. On rappelle que l'espace des fonctions de  $C^\infty(\mathbb{R}^d)$  restreintes à  $\Omega$  est dense dans  $H^1(\Omega)$ .

2. On admettra que l'image de  $H^1(\Omega)$  par l'application trace est dense dans  $L^2(\partial\Omega)$ .

**Exercice 40 page 119**

Considérer comme espace de Hilbert l'ensemble  $\{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$ .

**Exercice 46 page 122**

1. Intégrer l'équation sur la maille et approcher les flux sur les arêtes par des quotients différentiels.
2. Pour montrer que (3.59) entraîne (3.61), multiplier par  $v_K$ , où  $v \in H_T(\Omega)$ , et développer. Pour montrer la réciproque, écrire  $u$  comme combinaison linéaire des fonctions de base de  $H_T(\Omega)$ , et prendre pour  $v$  la fonction caractéristique de la maille  $K$ . (3.59)

**3.6 Corrigés des exercices****Exercice 34 page 118**

1. Par définition, on sait que  $|u|_{1,\Omega} = \left( \int_{\Omega} \sum_{i=1}^N |\partial_i u(x)|^2 dx \right)^{\frac{1}{2}}$ , où  $\partial_i u$  désigne la dérivée partielle de  $u$  par rapport à sa  $i$ -ème variable. Attention ceci  $|\cdot|_{1,\Omega}$  définit une semi-norme et non une norme sur l'espace  $H^1(\Omega)$ . Cependant sur  $H_0^1(\Omega)$  c'est bien une norme, grâce à l'inégalité de Poincaré. On rappelle que  $H_0^1(\Omega) = \text{Ker}(\gamma) = \{u \in H^1(\Omega) \text{ tel que } \gamma(u) = 0\}$ , où  $\gamma$  est l'opérateur de trace linéaire et continu de  $H^1(\Omega)$  dans  $L^2(\Omega)$  (voir théorème 3.2 page 98). Le problème consiste à minimiser  $\left( \int_{\Omega} \sum_{i=1}^N |\partial_i u|^2 dx \right)^{\frac{1}{2}}$  sur  $u_0 + H_0^1(\Omega)$ . Tentons de nous ramener à minimiser une certaine fonctionnelle sur  $H_0^1(\Omega)$ . Soit  $v \in u_0 + H_0^1(\Omega)$ . Alors  $v = u_0 + w$  avec  $w \in H_0^1(\Omega)$ , et donc :

$$\begin{aligned} |v|_{1,\Omega}^2 &= |u_0 + w|_{1,\Omega}^2 \\ &= \int_{\Omega} \sum_{i=1}^N |\partial_i(u_0 + w)|^2 dx \\ &= \int_{\Omega} \sum_{i=1}^N \left| (\partial_i u_0)^2 + (\partial_i w)^2 + 2(\partial_i u_0)(\partial_i w) \right| dx \\ &= |u_0|_{1,\Omega}^2 + |w|_{1,\Omega}^2 + 2 \int_{\Omega} \sum_{i=1}^N |\partial_i u_0 \partial_i w| dx \end{aligned}$$

Ainsi chercher à minimiser  $|v|_{1,\Omega}$  sur  $u_0 + H_0^1(\Omega)$  revient à minimiser  $J$  sur  $H_0^1(\Omega)$ , où  $J$  est définie par :

$$J(w) = \inf_{H_0^1(\Omega)} 2 \left( \int_{\Omega} \sum_{i=1}^N |\partial_i u_0 \partial_i w| dx + \frac{1}{2} \int_{\Omega} \sum_{i=1}^N (\partial_i w)^2 dx \right).$$

Pour montrer l'existence et l'unicité du minimum de  $J$ , nous allons mettre ce problème sous une forme faible, puis utiliser le théorème de Lax-Milgram pour en déduire l'existence et l'unicité d'une solution faible, et finalement conclure que la fonctionnelle  $J(w)$  admet un unique inf. On pose :

$$a(w, v) = \int_{\Omega} \sum_{i=1}^N (\partial_i w \partial_i v) dx, \quad \forall w, v \in H_0^1(\Omega)$$

et

$$L(v) = - \int_{\Omega} \sum_{i=1}^N (\partial_i u_0 \partial_i v) dx, \quad \forall v \in H_0^1(\Omega)$$

Voyons si les hypothèses de Lax-Milgram sont vérifiées. La forme  $a(w, v)$  est clairement symétrique, on peut changer l'ordre de  $w$  et de  $v$  dans l'expression sans changer la valeur de l'intégrale. La forme  $a(w, v)$  est bilinéaire.



En effet, elle est linéaire par rapport au premier argument, puisque :  $\forall u, v, w \in H_0^1(\Omega)$  et  $\forall \lambda, \mu \in \mathbb{R}$ , on a :  $a(\lambda w + \mu u, v) = \lambda a(w, v) + \mu a(u, v)$ . Ainsi par symétrie, elle est aussi linéaire par rapport au second argument. Donc elle est bien bilinéaire. Pour montrer que la forme  $a(u, v)$  est continue, on utilise la caractérisation de la continuité des applications bilinéaires. On va donc montrer l'existence de  $C \in \mathbb{R}_+$  tel que  $|a(u, v)| \leq C \|u\|_{H^1} \|v\|_{H^1}$  pour tous  $u, v \in H_0^1(\Omega)$ . Or, par l'inégalité de Cauchy-Schwarz, on a :

$$\begin{aligned} |a(u, v)| &= \left| \int_{\Omega} \sum_{i=1}^N (\partial_i u \partial_i v) dx \right| \\ &\leq \left( \int_{\Omega} \sum_{i=1}^N (\partial_i u)^2 dx \right)^{\frac{1}{2}} \left( \int_{\Omega} \sum_{i=1}^N (\partial_i v)^2 dx \right)^{\frac{1}{2}} \\ &\leq \|u\|_{H^1} \|v\|_{H^1}. \end{aligned}$$

La forme  $a$  est donc bien continue. Montrons alors qu'elle est coercive, c'est-à-dire qu'il existe  $\alpha > 0$  tel que  $a(v, v) \geq \alpha \|v\|_{H^1}^2$  pour tout  $v \in H_0^1(\Omega)$ .

$$\begin{aligned} a(v, v) &= \int_{\Omega} \sum_{i=1}^N (\partial_i v(x))^2 dx \\ &= \int_{\Omega} \nabla v(x) \cdot \nabla v(x) dx \\ &\geq \frac{1}{1 + \text{diam}(\Omega)^2} \|v\|_{H^1}^2, \end{aligned}$$

grâce à l'inégalité de Poincaré, qu'on rappelle ici :

$$\|v\|_{L^2(\Omega)} \leq c(\Omega) \|\nabla v\|_{L^2(\Omega)}, \quad \forall v \in H_0^1(\Omega). \quad (3.76)$$

Donc  $a$  est bien une forme bilinéaire, symétrique, continue et coercive. Par le même genre de raisonnement, on montre facilement que  $L$  est linéaire et continue. Ainsi toutes les hypothèses de Lax-Milgram sont vérifiées, donc le problème :

$$\text{Trouver } u \in H_0^1(\Omega) \text{ tel que } a(u, v) = L(v) \text{ pour tout } v \in H_0^1(\Omega)$$

a une unique solution dans  $H_0^1(\Omega)$ . De plus, comme  $a$  est symétrique, la fonctionnelle  $J$  admet un unique minimum.

2. On va maintenant caractériser  $u$  comme étant la solution d'un problème aux limites. Soit  $\varphi \in D(\Omega)$ , donc  $\varphi$  est à support compact dans  $\Omega$ . On a :

$$a(u, \varphi) = L(\varphi) \quad \forall \varphi \in D(\Omega),$$

et donc :

$$\int_{\Omega} \nabla u(x) \nabla(x) \varphi dx = - \int_{\Omega} \nabla u_0(x) \nabla(x) \varphi(x) dx.$$

Comme  $u$  et  $u_0 \in H^1(\Omega)$ , et comme  $\varphi$  est régulière, on peut intégrer par parties ; en remarquant que  $\varphi$  est nulle sur  $\partial\Omega$ , on a donc :

$$- \int_{\Omega} \Delta u(x) \varphi(x) dx = \int_{\Omega} \Delta u_0(x) \varphi(x) dx.$$

On en déduit que  $-\Delta u = \Delta u_0$ . Comme  $u \in H_0^1(\Omega)$ , ceci revient à résoudre le problème aux limites  $\tilde{u} = u - u_0 \in H^1(\Omega)$ , tel que  $-\Delta \tilde{u} = 0$  dans  $\Omega$  et  $\tilde{u} = u_0$  sur  $\partial\Omega$ .

### Exercice 35 page 118 (Formulation faible du problème de Dirichlet)

Soit  $\varphi \in C_c^\infty([0, 1])$ , on multiplie la première équation de (3.37), on intègre par parties et on obtient :

$$\int_0^1 u'(x) \varphi'(x) dx = \int_0^1 f(x) \varphi(x) dx. \quad (3.77)$$

Pour trouver une formulation faible (ou variationnelle) il faut commencer par trouver un espace de Hilbert pour les fonctions duquel (3.77) ait un sens, et qui soit compatible avec les conditions aux limites. Comme  $f \in L^2(]0, 1[)$ , le second membre de (3.77) est bien défini dès que  $\varphi \in L^2(]0, 1[)$ .

De même, le premier membre de (3.77) est bien défini dès que  $u' \in L^2(]0, 1[$  et  $\varphi' \in L^2(]0, 1[)$ .

Comme de plus, on doit avoir  $u = 0$  en 0 et en 1, il est naturel de choisir  $H = H_0^1(]0, 1[) \stackrel{\text{def}}{=} \{u \in L^2(]0, 1[; Du \in L^2(]0, 1[) \text{ et } u(0) = u(1) = 0\}$

(Rappelons qu'en une dimension d'espace  $H^1(]0, 1[) \subset C([0, 1])$  et donc  $u(0)$  et  $u(1)$  sont bien définis).

Une formulation faible naturelle est donc :

$$\begin{cases} u \in H = \{u \in H_0^1(\Omega); v(0) = v(1) = 0\}, \\ a(u, v) = T(v), \forall v \in H, \end{cases}$$

où  $a(u, v) = \int_0^1 u'(x)v'(x)dx$  et  $T(v) = \int_0^1 f(x)v(x)dx$ .

La formulation variationnelle associée (notons que  $a$  est clairement symétrique), s'écrit :

$$\begin{cases} \text{Trouver } u \in H, \\ J(u) = \min_{v \in H} J(v) \end{cases}$$

avec  $J(v) = \frac{1}{2}a(v, v) - T(v)$

Le fait que  $a$  soit une forme bilinéaire continue symétrique et coercive et que  $T \in H'$  a été prouvé (dans le cas plus général de la dimension quelconque) lors de la démonstration de la proposition 3.7 page 100.

**Exercice 36 page 118**

1. Comme  $f \in C(\mathbb{R}, \mathbb{R})$ , et comme  $-u'' = f$ , on a  $u \in C^2(\mathbb{R}, \mathbb{R})$ . Or  $u_0 \in C^2(\mathbb{R}, \mathbb{R})$  et  $u_0'' = 0$ ; de même,  $u_1 \in C^2(\mathbb{R}, \mathbb{R})$  et  $u_1'' = 2(b-a)$ .

Les fonctions  $\tilde{u}$  et  $\bar{u}$  doivent donc vérifier :

$$\begin{cases} -\tilde{u}'' = f \\ \tilde{u}(0) = 0 \\ \tilde{u}(1) = 0. \end{cases}$$

et

$$\begin{cases} -\bar{u}'' = f + 2(b-a) \\ \bar{u}(0) = 0 \\ \bar{u}(1) = 0. \end{cases}$$

Donc  $\tilde{u}$  est l'unique solution du problème

$$\begin{cases} \tilde{u} \in H_0^1(\Omega) \\ a(u, \varphi) = \tilde{T}(\varphi), \forall \varphi \in H_0^1(\Omega), \end{cases}$$

avec  $a(u, \varphi) = \int_0^1 u'(x)\varphi'(x)dx$  et  $\tilde{T}(\varphi) = \int_0^1 f(x)\varphi(x)dx$ , et  $\bar{u}$  est l'unique solution du problème.

$$\begin{cases} \bar{u} \in H_0^1(\Omega) \\ a(\bar{u}, \varphi) = \bar{T}(\varphi), \forall \varphi \in H_0^1(\Omega), \end{cases}$$

avec  $\bar{T}(\varphi) = \int_0^1 (f(x) + 2(b-a))\varphi(x)dx$ .

Montrons maintenant que  $u = v$ . Remarquons que  $w = u - v$  vérifie

$$\begin{cases} w'' = 0 \\ w(0) = w(1) = 0 \end{cases}$$

ce qui prouve que  $w$  est solution de

$$\begin{cases} w \in H_0^1(\Omega) \\ a(w, \varphi) = 0, \forall \varphi \in H_0^1(\Omega), \end{cases}$$

ce qui prouve que  $w = 0$ .

2. Le même raisonnement s'applique pour  $u_0$  et  $u_1 \in C^2([0, 1])$  tel que

$$u_0(0) = u_1(0) = a \text{ et } u_1(0) = u_1(1) = b.$$

**Exercice 37 page 118**

On introduit les espaces :

$$H_{1,1}^1 = \{v \in H^1(]0, 1[); v(1) = 1\}$$

$$H_{1,0}^1 = \{v \in H^1(]0, 1[); v(1) = 0\}$$

Soit  $u_0 : ]0, 1[ \rightarrow \mathbb{R}$ , définie par  $u_0(x) = x$ . On a bien  $u_0(1) = 1$ , et  $u_0 \in H_{1,1}^1$ .

Cherchons alors  $u$  sous la forme  $u = u_0 + \tilde{u}$ , avec  $\tilde{u} \in H_{1,0}^1$ .

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) + u'(0)v(0) = \int_0^1 f(x)v(x)dx, \forall v \in H_{1,0}^1.$$

Comme  $v(1) = 0$  et  $u'(0) = 0$ , on obtient donc :

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx,$$

ou encore :

$$\int_0^1 \tilde{u}'(x)v'(x)dx = \int_0^1 f(x)v(x)dx - \int_0^1 u_0'(x)v'(x)dx = \int_0^1 f(x)v(x)dx - \int_0^1 v'(x)dx.$$

car  $u_0' = 1$ .

**Exercice 38 page 118**

1. Soit  $v \in C_c^\infty([0, 1])$ , on multiplie la première équation de (3.40), on intègre par parties et on obtient :

$$\int_0^1 u'(x)v'(x)dx - u'(1)v(1) + u'(0)v(0) + \int_0^1 u(x)v(x)dx = \int_0^1 f(x)v(x)dx.$$

En tenant compte des conditions aux limites sur  $u$  en 0 et en 1, on obtient :

$$\int_0^1 u'(x)v'(x)dx + \int_0^1 u(x)v(x)dx + u(0)v(0) = \int_0^1 f(x)v(x)dx - v(1). \quad (3.78)$$

Pour trouver une formulation faible (ou variationnelle) il faut commencer par trouver un espace de Hilbert pour les fonctions duquel (3.78) ait un sens, et qui soit compatible avec les conditions aux limites. Comme  $f \in L^2(]0, 1[)$ , le second membre de (3.78) est bien défini dès que  $v \in L^2(]0, 1[)$ .

De même, le premier membre de (3.78) est bien défini dès que  $u \in H^1(]0, 1[)$  et  $v \in H^1(]0, 1[)$   $\stackrel{\text{def}}{=} \{u \in L^2(]0, 1[); Du \in L^2(]0, 1[)$ . Il est donc naturel de choisir  $H = H(]0, 1[)$ . On obtient ainsi la formulation faible suivante :

$$\begin{cases} u \in H = \{u \in H(\Omega)\}, \\ a(u, v) = T(v), \forall v \in H, \end{cases}$$

où  $a(u, v) = \int_0^1 u'(x)v'(x)dx + \int_0^1 u(x)v(x)dx + u(0)v(0)$  et  $T(v) = \int_0^1 f(x)v(x)dx - v(1)$ .

La formulation variationnelle associée (notons que  $a$  est clairement symétrique), s'écrit :

$$\begin{cases} \text{Trouver } u \in H, \\ J(u) = \min_{v \in H} J(v) \end{cases}$$

avec  $J(v) = \frac{1}{2}a(u, v) - T(v)$

Pour montrer l'existence et l'unicité des solutions de (3.78), on cherche à appliquer le théorème de Lax–Milgram. On remarque d'abord que  $T$  est bien une forme linéaire sur  $H$ , et que de plus, par l'inégalité de Cauchy–Schwarz, :

$$|T(v)| = \left| \int_0^1 f(x)v(x)dx \right| + |v(1)| \leq \|f\|_{L^2(]0, 1[)} \|v\|_{L^2(]0, 1[)} + |v(1)|. \quad (3.79)$$

Montrons maintenant que  $|v(1)| \leq 2\|v\|_{H^1(]0, 1[)}$ . Ce résultat est une conséquence du théorème de trace, voir cours d'EDP. Dans le cas présent, comme l'espace est de dimension 1, la démonstration est assez simple en remarquant que comme  $v \in H^1(]0, 1[)$ , on peut écrire que  $v$  est intégrale de sa dérivée. On a en particulier :

$$v(1) = v(x) + \int_x^1 v'(t)dt,$$

et donc par l'inégalité de Cauchy–Schwarz,

$$|v(1)| = |v(x)| + \int_x^1 |v'(t)|dt \leq |v(x)| + \|v'\|_{L^2(]0, 1[)}.$$

En intégrant cette inégalité entre 0 et 1 on obtient :

$$|v(1)| \leq \|v(x)\|_{L^1(]0, 1[)} + \|v'\|_{L^2(]0, 1[)}.$$

Or  $\|v\|_{L^1(]0, 1[)} \leq \|v(x)\|_{L^2(]0, 1[)}$ . De plus

$$\|v\|_{L^2(]0, 1[)} + \|v'\|_{L^2(]0, 1[)} \leq 2 \max(\|v(x)\|_{L^2(]0, 1[)}, \|v'\|_{L^2(]0, 1[)})$$

on a donc

$$\begin{aligned} \left( \|v\|_{L^2(]0, 1[)} + \|v'\|_{L^2(]0, 1[)} \right)^2 &\leq 4 \max(\|v(x)\|_{L^2(]0, 1[)}^2, \|v'\|_{L^2(]0, 1[)}^2) \\ &\leq 4(\|v\|_{L^2(]0, 1[)}^2 + \|v'\|_{L^2(]0, 1[)}^2). \end{aligned}$$

On en déduit que

$$|v(1)| \leq \|v\|_{L^2(]0,1[)} + \|v'\|_{L^2(]0,1[)} \leq 2\|v\|_{H^1(]0,1[)}.$$

En reportant dans (3.79), on obtient :

$$|T(v)| \leq (\|f\|_{L^2(]0,1[)} + 2)\|v\|_{H^1(]0,1[)}$$

ce qui montre que  $T$  est bien continue.

Remarquons que le raisonnement effectué ci-dessus pour montrer que  $|v(1)| \leq 2\|v\|_{H^1(]0,1[)}$  s'applique de la même manière pour montrer que

$$|v(a)| \leq 2\|v\|_{H^1(]0,1[)} \text{ pour tout } a \in [0, 1]. \quad (3.80)$$

Ceci est une conséquence du fait que  $H^1(]0, 1[)$  s'injecte continûment dans  $C([0, 1])$ .

Il est clair que  $a$  est une forme bilinéaire symétrique (notons que le caractère symétrique n'est pas nécessaire pour l'application du théorème de Lax–Milgram). Montrons que  $a$  est continue. On a :

$$\begin{aligned} |a(u, v)| &\leq \int_0^1 |u'(x)v'(x)| dx + \int_0^1 |u(x)| |v(x)| dx + |u(0)| |v(0)| \\ &\leq \|u'\|_{L^2(]0,1[)} \|v'\|_{L^2(]0,1[)} + \|u\|_{L^2(]0,1[)} \|v\|_{L^2(]0,1[)} + |u(0)| |v(0)| \end{aligned}$$

Grâce à (3.80), on en déduit que

$$\begin{aligned} |a(u, v)| &\leq \|u'\|_{L^2(]0,1[)} \|v'\|_{L^2(]0,1[)} + \|u\|_{L^2(]0,1[)} \|v\|_{L^2(]0,1[)} + 4\|u\|_{H^1(]0,1[)} \|v\|_{H^1(]0,1[)} \\ &\leq 6\|u\|_{H^1(]0,1[)} \|v\|_{H^1(]0,1[)}. \end{aligned}$$

On en déduit que  $a$  est continue. Soit  $u \in H^1(]0, 1[)$ , Par définition de  $a$ , on a :

$$\begin{aligned} a(u, u) &= \int_0^1 (u'(x))^2 dx + \int_0^1 (u(x))^2 dx + u(0)^2 \\ &\geq \|u\|_{H^1(]0,1[)}^2. \end{aligned}$$

Ceci prouve que la forme  $a$  est coercive. Par le théorème de Lax–Milgram, on en déduit l'existence et l'unicité des solutions faibles de (3.78).

### Exercice 40 page 119

Soit  $\varphi \in H = \{v \in H^1(\Omega) : u = 0 \text{ sur } \Gamma_0\}$ .

Multiplions la première équation de (3.42) par  $\varphi \in H$ . On obtient :

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx + \int_{\Gamma_0 \cup \Gamma_1} \nabla u \cdot n(x) \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx$$

et comme  $\nabla u \cdot n = 0$  sur  $\Gamma_1$  et  $\varphi = 0$  sur  $\Gamma_0$ , on obtient donc

$$\int_{\Omega} \nabla u(x) \cdot \nabla \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx.$$

On obtient donc la formulation faible.

$$\begin{cases} \text{Trouver } u \in H; \\ \int_{\Omega} \nabla u(x) \cdot \nabla v(x) dx = \int_{\Omega} f(x) u(x) dx. \end{cases}$$

Notons que cette formulation ne diffère de la formulation faible du problème (3.1) que par la donnée de la condition aux limites de Dirichlet sur  $\Gamma_0$  et non  $\partial\Omega$  dans l'espace  $H$ . La condition de Neumann homogène est implicitement prise en compte dans la formulation faible.

La démonstration du fait que cette formulation satisfait les hypothèses du théorème de Lax–Milgram est similaire à celle de la proposition 3.7 en utilisant, pour la coercivité, le fait que les fonctions à trace nulle sur une partie du bord de  $\Omega$  (de mesure non nulle) vérifient encore l'inégalité de Poincaré.

**Exercice 41 page 119**

1. Multiplions la première équation de (3.43) par  $\varphi \in C^\infty(\Omega)$  et intégrons sur  $\Omega$ . Par la formule de Green, on obtient :

$$\int_{\Omega} p(x) \nabla u(x) \cdot \nabla \varphi(x) dx - \int_{\partial\Omega} p(x) \nabla u(x) \cdot n(x) \varphi(x) dx + \int_{\Omega} q(x) u(x) \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx.$$

En tenant compte des conditions aux limites sur  $u$  et en prenant  $\varphi$  nulle sur  $\Gamma_0$ , on obtient alors :

$$a(u, \varphi) = T(\varphi)$$

avec :

$$a(u, \varphi) = \int_{\Omega} (p(x) \nabla u(x) \cdot \nabla \varphi(x) + q(x) u(x) \varphi(x)) dx + \int_{\Gamma_1} \sigma(x) u(x) \varphi(x) d\gamma(x), \quad (3.81)$$

et

$$T(\varphi) = \int_{\Omega} f(x) \varphi(x) dx + \int_{\Gamma_1} g_1(x) \varphi(x) d\gamma(x). \quad (3.82)$$

Pour assurer la condition aux limites de type Dirichlet non homogène, on choisit donc  $u \in H_{\Gamma_0, g_0}^1(\Omega) = \{u \in H^1(\Omega); u = g_0 \text{ sur } \Gamma_0\}$ , qu'on peut aussi décomposer en :  $u = \tilde{u} + u_0$  avec  $\tilde{u} \in H_{\Gamma_0, g_0}^1(\Omega)$  ("relèvement" de  $u$ ) et  $u_0 \in H_{\Gamma_0}^1(\Omega) = \{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$ . Une formulation faible naturelle est alors :

$$\begin{cases} u \in H_{\Gamma_0, g_0}^1(\Omega) \\ a(u, v) = T(v), \forall v \in H_{\Gamma_0, 0}^1(\Omega), \end{cases}$$

ou encore :

$$\begin{cases} u = u_0 + \tilde{u} \\ \tilde{u} \in H_{\Gamma_0, 0}^1(\Omega) \\ a(\tilde{u}, v) = T(v) - T(u_0), \forall v \in H_{\Gamma_0, 0}^1(\Omega), \end{cases} \quad (3.83)$$

L'espace  $H = H_{\Gamma_0, 0}^1(\Omega)$  muni de la norme  $H^1$  est un espace de Hilbert. Il est facile de montrer que l'application  $a$  définie de  $H \times H$  dans  $\mathbb{R}$  est bilinéaire. Montrons qu'elle est continue ; soient  $(u, v) \in H \times H$ , alors

$$a(u, v) \leq \|p\|_{L^\infty(\Omega)} \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} + \|q\|_{L^\infty(\Omega)} \|u\|_{L^2(\Omega)} \|v\|_{L^2(\Omega)} + \sigma \|\gamma(u)\|_{L^2(\Omega)} \|\gamma(v)\|_{L^2(\partial\Omega)}.$$

Par le théorème de trace, il existe  $C_\Omega$  ne dépendant que de  $\Omega$  tel que

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leq C_\Omega \|u\|_{H^1(\Omega)} \text{ et } \|\gamma(v)\|_{L^2(\partial\Omega)} \leq C_\Omega \|v\|_{H^1(\Omega)}.$$

On en déduit que

$$a(u, v) \leq (\|p\|_{L^\infty(\Omega)} + \|q\|_{L^\infty(\Omega)} + \sigma C_\Omega^2) \|u\|_{H^1(\Omega)} \|v\|_{H^1(\Omega)},$$

ce qui montre que  $a$  est continue. La démonstration de la coercivité de  $a$  est similaire à la démonstration du lemme 3.15 page 104. Enfin, il est facile de voir que  $T$  définie par (3.82) est une forme linéaire. On en déduit que le théorème de Lax-Milgram s'applique.

2. On a déjà vu à la question précédente que si  $u$  est solution de (3.83), alors  $u$  est solution de (3.43). Il reste à démontrer la réciproque. Soit donc  $u$  solution de (3.83), et soit  $\varphi \in C_c^\infty(\Omega) \subset H$ . En utilisant la formule de Green, et en notant que  $\varphi$  est nulle sur  $\partial\Omega$ , on obtient :

$$\int_{\Omega} (-\operatorname{div}(p \nabla u)(x) + q(x) u(x) - f(x)) \varphi(x) dx = 0, \forall \varphi \in C_0^\infty(\Omega).$$

Comme  $u \in C^2(\bar{\Omega})$ , on en déduit que :

$$-\operatorname{div}(p \nabla u)(x) + q(x) u(x) - f(x) = 0, \forall x \in \Omega.$$

Comme  $u \in H_{\Gamma_0, g_0}^1$  et  $u \in C^2(\bar{\Omega})$ , on a aussi  $u = g_0$  sur  $\Gamma_0$ . Prenons maintenant  $\varphi \in H_{\Gamma_0, 0}^1$  on a :

$$\int_{\Omega} p(x) \nabla u(x) \cdot \nabla \varphi(x) dx + \int_{\Omega} q(x) u(x) \varphi(x) dx + \int_{\Gamma_1} \sigma(x) u(x) \varphi(x) d\gamma(x) = \int_{\Omega} f(x) dx + \int_{\Gamma_1} g(x) \varphi(x) d\gamma(x).$$

Par intégration par parties, il vient donc :

$$\begin{aligned} \int_{\Omega} -\operatorname{div}(p(x)\nabla u(x))\varphi(x)dx + \int_{\Gamma_1} p(x)\nabla u(x)\cdot n(x)\varphi(x)dx + \int_{\Gamma_1} \sigma(x)u(x)\varphi(x)d\gamma(x) + \int_{\Omega} q(x)u(x)\varphi(x)dx = \\ = \int_{\Omega} f(x)\varphi(x)dx + \int_{\Gamma_1} g_1(x)\varphi(x)d\gamma(x). \end{aligned}$$

Or on a montré que  $-\operatorname{div}(p\nabla u) + qu = 0$ . On a donc :

$$\int_{\Gamma_1} (p(x)\nabla u(x)\cdot n(x) + \sigma u(x) - g_1(x))\varphi(x)d\gamma(x) = 0, \quad \forall \varphi \in H_{\Gamma_0, g_0}^1.$$

On en déduit que :

$$p\nabla u \cdot n + \sigma u - g_1 = 0 \text{ sur } \Gamma_1.$$

Donc  $u$  vérifie bien (3.43).

### Exercice 43 page 120

1. Soit  $v$  une fonction suffisamment régulière, on multiplie la première équation de (4.28) par  $v$  et on intègre sur  $]0, 1[$ . En effectuant des intégrations par parties et en tenant compte des conditions aux limites, on obtient :

$$\int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) = \int_0^1 f(x)v(x) dx.$$

Pour que les intégrales aient un sens, il suffit de prendre  $u, v \in H^1(]0, 1[)$ , auquel cas les fonctions sont continues et donc les valeurs  $u(0)$  et  $v(0)$  ont aussi un sens. On en déduit qu'une formulation faible est

$$\begin{aligned} u \in H^1(]0, 1[) \\ \int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) = \int_0^1 f(x)v(x) dx, \quad \forall v \in H^1(]0, 1[). \end{aligned}$$

On en déduit que la formulation variationnelle est

$$\begin{aligned} \text{Trouver } u \in H^1(]0, 1[); \\ J(u) = \min_{v \in H} J(v), \end{aligned}$$

avec  $J(v) = \frac{1}{2}a(v, v) - T(v)$ , où  $a$  est la forme bilinéaire définie par  $a(u, v) = \int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0)$  et  $T$  la forme linéaire continue définie par  $T(v) = \int_0^1 f(x)v(x) dx$ .

2. Supposons  $u$  régulière, et prenons d'abord  $v \in C_c^1(]0, 1[)$ . On a alors

$$\int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) = \int_0^1 f(x)v(x) dx$$

et donc en intégrant par parties :

$$\int_0^1 (-u''(x) + \alpha u(x) - f(x))v(x) dx = 0.$$

Comme ceci est vrai pour toute fonction  $v \in C_c(]0, 1[)$ , on en déduit que  $-u''(x) + \alpha u(x) = f(x)$ ,  $x \in ]0, 1[$ .

Prenons maintenant  $v \in H^1(]0, 1[)$ , en intégrant par parties et tenant compte de ce qui précède, on obtient

$$(-u'(0) + u(0))v(0) + u'(1)v(1) = 0.$$

Comme ceci est vrai pour toute fonction  $v \in H^1(]0, 1[)$ , on en déduit que  $u$  vérifie (4.28).

2.b. On peut appliquer le lemme de Lax Milgram ; en effet,

- la forme linéaire  $T$  est continue car  $|T(v)| = |\int_0^1 f(x)v(x) dx| \leq \|f\|_{L^2} \|v\|_{L^2}$  par l'inégalité de Cauchy-Schwarz, et donc  $|T(v)| \leq C \|v\|_{L^2}$  avec  $C = \|f\|_{L^2}$ .
- la forme bilinéaire  $a$  (qui est évidemment symétrique, ce qui n'est d'ailleurs pas nécessaire pour appliquer Lax-Milgram) est continue ; en effet :

$$|a(u, v)| \leq \|u'\|_{L^2} \|v'\|_{L^2} + \alpha \|u\|_{L^2} \|v\|_{L^2} + |u(0)| |v(0)|;$$

or pour tout  $x \in ]0, 1[$   $v(0) = v(x) + \int_0^x v'(t)dt$  et donc par inégalité triangulaire et par Cauchy-Schwarz, on obtient que  $|v(0)| \leq |v(x)| + \|v'\|_{L^2}$ . En intégrant cette inégalité entre 0 et 1, on obtient

$$|v(0)| \leq \|v\|_{L^1} + \|v'\|_{L^2} \leq \|v\|_{L^2} + \|v'\|_{L^2} \leq 2\|v\|_{H^1}.$$

La meme inégalité est évidemment vraie pour  $u(0)$ . On en déduit que :

$$\begin{aligned} a(u, v) &\leq \|u'\|_{L^2}\|v'\|_{L^2} + \alpha\|u\|_{L^2}\|v\|_{L^2} + 4\|u\|_{H^1}\|v\|_{H^1} \\ &\leq \|u\|_{H^1}\|v\|_{H^1} + \alpha\|u\|_{H^1}\|v\|_{H^1} + 4\|u\|_{H^1}\|v\|_{H^1} \\ &\leq (5 + \alpha)\|u\|_{H^1}\|v\|_{H^1}, \end{aligned}$$

ce qui prouve que  $a$  est continue.

Montrons maintenant que  $a$  est coercive. Dans le cas où  $\alpha > 0$ , ceci est facile à vérifier, car on a

$$a(u, u) = \int_0^1 u'(x)^2 dx + \alpha \int_0^1 u(x)^2 dx + u(0)^2 \geq \min(\alpha, 1)\|u\|_{H^1}^2.$$

Par le lemme de Lax -Milgram, on peut donc conclure à l'existence et l'unicité de la solution de (4.29). 3. Dans le cas où  $\alpha = 0$ , on applique l'inégalité de Poincaré à la fonction  $w = u - u(0)$ , ce qui est licite car  $w(0) = 0$ ; on a donc :  $\|w\|_{L^2} \leq \|w'\|_{L^2}$ , et donc  $\|u'\|_{L^2} \geq \|u - u(0)\|_{L^2}$ . On en déduit que  $a(u, u) \geq \|u - u(0)\|_{L^2}^2 + u(0)^2 \geq \frac{1}{2}\|u\|_{L^2}^2$ .

On écrit alors que

$$\begin{aligned} a(u, u) &= \frac{1}{2}a(u, u) + \frac{1}{2}a(u, u) \\ &\geq \frac{1}{2}\|u'\|_{L^2}^2 + \frac{1}{4}\|u\|_{L^2}^2 \\ &\geq \frac{1}{4}\|u\|_{H^1}^2, \end{aligned}$$

ce qui montre que la forme bilinéaire  $a$  est encore coercive.

### Corrigé de l'exercice 44 page 120

1. Pour montrer que le problème (3.52) admet une unique solution, on aimerait utiliser le théorème de Lax-Milgram. Comme  $V_h \subset V$  un Hilbert, que  $a$  une forme bilinéaire continue sur  $V \times V$ , et que  $L$  est une forme linéaire continue sur  $V$ , il ne reste qu'à montrer la coercivité de  $a$  sur  $V_h$ . Mais la condition (3.51) page 120 n'entraîne pas la coercivité de  $a$  sur  $V_h$ . Il suffit pour s'en convaincre de considérer la forme bilinéaire  $a(u, v) = u_1u_2 - v_1v_2$  sur  $V_h = \mathbb{R}^2$ , et de vérifier que celle-ci vérifie la condition (3.51) sans être pour autant coercive. Il faut trouver autre chose...

On utilise le théorème représentation de F. Riesz, que l'on rappelle ici : Soit  $H$  un espace de Hilbert et  $T$  une forme linéaire continue sur  $H$ , alors il existe un unique  $u_T \in H$  tel que  $T(v) = (u_T, v) \forall v \in H$ . Soit  $A$  l'opérateur de  $V_h$  dans  $V_h$  défini par  $a(u, v) = (Au, v)$  pour tout  $v \in V_h$ . Comme  $L$  est une forme linéaire continue sur  $V_h \subset V$ , par le théorème de Riesz, il existe un unique  $\psi \in V_h$  tel que  $L(v) = (\psi, v)$ , pour tout  $v \in V_h$ . Le problème (3.52) s'écrit donc

$$\text{Trouver } u \in V_h \text{ tel que } (Au, v) = (\psi, v), \text{ pour tout } v \in V_h.$$

Si  $A$  est bijectif de  $V_h$  dans  $V_h$ , alors  $u = A^{-1}\psi_u$  est donc la solution unique de (3.52). Comme  $V_h$  est de dimension finie, il suffit de montrer que  $A$  est injectif. Soit donc  $w \in V_h$  tel que  $Aw = 0$ , on a dans ce cas  $\|Aw\| = 0$  et donc

$$\sup_{(v \in V_h, \|v\|=1)} a(w, v) = 0.$$

Or par la condition (3.51), on a

$$\inf_{w \in V_h, w \neq 0} \sup_{(v \in V_h, \|v\|=1)} a(w, v) \geq \beta_h > 0.$$



On en déduit que  $w = 0$ , donc que  $A$  est bijectif et que le problème (3.52) admet une unique solution. On peut remarquer de plus que si  $A$  est inversible,

$$\inf_{v \in V_h, \|v\|=1} \sup_{v \in V_h, \|v\|=1} a(w, v) = \|A^{-1}\|^{-1}, \quad (3.84)$$

et donc si (3.51) est vérifiée, alors

$$\|A^{-1}\| \leq \frac{1}{\beta_h} \quad (3.85)$$

En effet, par définition,

$$\begin{aligned} \|A^{-1}\|^{-1} &= \left( \sup_{v \in V_h, v \neq 0} \frac{\|A^{-1}v\|}{\|v\|} \right)^{-1} \\ &= \inf_{v \in V_h, v \neq 0} \frac{\|v\|}{\|A^{-1}v\|} \\ &= \inf_{f \in V_h, v \neq 0} \frac{\|Af\|}{\|f\|} \\ &= \inf_{f \in V_h, \|f\|=1} \|Af\| \\ &= \inf_{f \in V_h, \|f\|=1} \sup_{w \in V_h, \|w\|=1} (Af, w). \end{aligned}$$

2. Soit  $v \in V_h, v \neq 0$ ; par l'inégalité triangulaire, on a :

$$\|u - u_h\| \leq \|u - v\| + \|v - u_h\|. \quad (3.86)$$

Mais grâce à (3.85), on a :

$$\begin{aligned} \|v - u_h\| &= \|A^{-1}A(v - u_h)\| \\ &\leq \frac{1}{\beta_h} \|A(v - u_h)\| \\ &\leq \frac{1}{\beta_h} \sup_{w \in V_h, \|w\|=1} a(v - u_h, w) \\ &\leq \frac{1}{\beta_h} \sup_{w \in V_h, \|w\|=1} (a(v, w) - a(u_h, w)) \\ &\leq \frac{1}{\beta_h} \sup_{w \in V_h, \|w\|=1} (a(v, w) - a(u, w)), \end{aligned}$$

car  $a(u_h, w) = L(w) = a(u, w)$ . On a donc

$$\begin{aligned} \|v - u_h\| &\leq \frac{1}{\beta_h} \sup_{w \in V_h, \|w\|=1} a(v - u, w) \\ &\leq \frac{1}{\beta_h} \sup_{w \in V_h, \|w\|=1} M \|v - u\| \|w\| \\ &\leq \frac{M}{\beta_h} \|v - u\|. \end{aligned}$$

En reportant dans (3.86), il vient alors :

$$\|u - u_h\| \leq \|u - v\| + \frac{M}{\beta_h} \|v - u\|, \quad \forall v \in V_h,$$

et donc

$$\|u - u_h\| \leq \left(1 + \frac{M}{\beta_h}\right) \inf_{v \in V_h} \|u - v\|.$$

**Exercice 46 page 122**

1. Soit  $K$  un volume de contrôle du maillage volumes finis. On intègre (3.1) sur  $K$  et en utilisant la formule de Stokes, on obtient :

$$\sum_{\sigma \in \mathcal{E}_K} \int_{\sigma} \nabla u(x) \cdot n_{K,\sigma} d\gamma(x) = m(K) f_K,$$

avec les notations du paragraphe 1.1.2 page 10.

On approche cette équation par :

$$\sum_{\sigma \in \mathcal{E}_K} F_{K,\sigma} = m(K) f_K,$$

où  $F_{K,\sigma}$  est le flux numérique à travers  $\sigma$ , qu'on approche par :

$$F_{K,\sigma} = \begin{cases} \frac{m(\sigma)}{d_{K,\sigma} + d_{L,\sigma}} (u_K - u_L) & \text{si } \sigma \in \mathcal{E}_{int} \cap \mathcal{E}_K, \\ \frac{m(\sigma)}{d_{K,\sigma}} u_K & \text{si } \sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K. \end{cases}$$

On obtient donc bien le schéma (3.59) - (3.60)

2. Soit  $v = (v_K)_{K \in \mathcal{T}} \in H_{\mathcal{T}}(\Omega)$  une fonction constante par volumes de contrôle.

On multiplie l'équation (3.59) par  $V_K$  et on somme sur  $K$ . On obtient :

$$\sum_{K \in \mathcal{T}} \left( \sum_{\substack{\sigma \in \mathcal{E}_{int} \\ \sigma = K|L}} \tau_{\sigma} (u_K - u_L) v_K + \sum_{\sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K} \tau_{\sigma} u_K V_K \right) = \sum_K m(K) f_K v_K.$$

Remarquons maintenant que le premier membre de cette égalité est aussi égal, en sommant sur les arêtes du maillage à :

$$\sum_{\substack{\sigma \in \mathcal{E} \\ \sigma = K|L}} (\tau_{\sigma} (u_K - u_L) v_K) + \tau_{\sigma} (u_L - u_K) v_L + \sum_{\tau \in \mathcal{E}_{ext}} \tau_{\sigma} u_{K_{\sigma}} v_{K_{\sigma}}$$

où  $K_{\sigma}$  désigne le volume de contrôle dont  $\sigma$  est une arête (du bord) dans la deuxième sommation. On obtient donc :

$$a_{\mathcal{T}}(u, v) = T_{\mathcal{T}}(V), \forall v \in H_{\mathcal{T}}(\Omega), \quad (3.87)$$

avec :

$$a_{\mathcal{T}}(u, v) = \sum_{\substack{\sigma \in \mathcal{E} \\ \sigma = K|L}} \tau_{\sigma} (u_K - u_L) (v_K - v_L) + \sum_{\sigma \in \mathcal{E}_{ext}} u_{K_{\sigma}} v_{K_{\sigma}} \text{ et } T_{\mathcal{T}}(v) = \sum_K m(K) f_K v_K.$$

On a donc montré que si  $u = (u_K)_{K \in \mathcal{T}}$  la solution de (3.59) - (3.60), alors  $u$  est solution de (3.87). Montrons maintenant la réciproque. Soit  $1_K$  la solution caractéristique du volume de contrôle  $K$ , définie par

$$1_K(x) = \begin{cases} 1 & \text{si } x \in K \\ 0 & \text{sinon,} \end{cases}$$

Prenons  $v = 1_K$  dans (3.87), on obtient alors

$$\sum_{\sigma \in \mathcal{E}_{int}} \tau_{\sigma} (u_K - u_L) + \sum_{\sigma \in \mathcal{E}_{ext} \cap \mathcal{E}_K} \tau_{\sigma} u_K = m(K) f_K.$$

On retrouve donc bien (3.59).

Notons qu'en faisant ceci, on a introduit une discrétisation de la formulation faible (3.5) page 99 par une méthode de discrétisation non conforme, puisque  $H_{\mathcal{T}} \not\subset H^1(\Omega)$ .

# Chapitre 4

## Eléments finis de Lagrange

La construction d'une méthode d'éléments finis nécessite la donnée d'un maillage, de noeuds et d'un espace de polynômes, qui doivent être choisis de manière cohérente. Les éléments finis de type Lagrange font intervenir comme "degrés de liberté" (c.à.d. les valeurs qui permettent de déterminer entièrement une fonction) les valeurs de la fonction aux noeuds. Ils sont très largement utilisés dans les applications. Il existe d'autres familles d'éléments finis, comme par exemple les éléments finis de type Hermite qui font également intervenir les valeurs des dérivées directionnelles. Dans le cadre de ce cours, nous n'aborderons que les éléments finis de type Lagrange, et nous renvoyons aux ouvrages cités en introduction pour d'autres éléments.

### 4.1 Espace d'approximation

#### 4.1.1 Cohérence "locale"

Soit  $\mathcal{T}$  un maillage de  $\Omega$ , pour tout élément  $K$  de  $\mathcal{T}$ , on note  $\Sigma_K$  l'ensemble des noeuds de l'élément. On suppose que chaque élément a  $N_\ell$  noeuds  $K : \Sigma_K = \{a_1, \dots, a_{N_\ell}\}$ , qui ne sont pas forcément ses sommets. On note  $P$  un espace de dimension finie constitué de polynômes, qui définit la méthode d'éléments finis choisie.

**Définition 4.1 (Unisolvanse, élément fini de Lagrange)** Soit  $K$  un élément et  $\Sigma_K = (a_i)_{i=1, \dots, N_\ell}$  un ensemble de noeuds de  $K$ . Soit  $P$  un espace de polynômes de dimension finie. On dit que le triplet  $(K, \Sigma_K, P)$  est un élément fini de Lagrange si  $\Sigma_K$  est  $P$ -unisolvant, c'est à dire si pour tout  $(\alpha_1, \dots, \alpha_{N_\ell}) \in \mathbb{R}^{N_\ell}$ , il existe un unique élément  $f \in P$  tel que  $f(a_i) = \alpha_i \quad \forall i = 1 \dots N_\ell$ . Pour  $i = 1, \dots, N_\ell$ , on appelle degré de liberté la forme linéaire  $\zeta_i$  définie par  $\zeta_i(p) = p(a_i)$ , pour tout  $p \in P$ . La propriété d'unisolvanse équivaut à dire que la famille  $(\zeta_i)_{i=1, \dots, N_\ell}$  forme une base de  $P'$  (espace dual de  $P$ ).

La  $P$ -unisolvanse revient à dire que toute fonction de  $P$  est entièrement déterminée par ses valeurs aux noeuds.

**Exemple : l'élément fini de Lagrange  $P_1$**  Prenons par exemple, en dimension 1, l'élément  $K = [a_1, a_2]$ , avec  $\Sigma_K = \{a_1, a_2\}$ , et  $P = P_1$  (ensemble des polynômes de degré inférieur ou égal à 1). Le triplet  $(K, \Sigma_K, P)$  est unisolvant s'il existe une unique fonction  $f$  de  $P$  telle que :

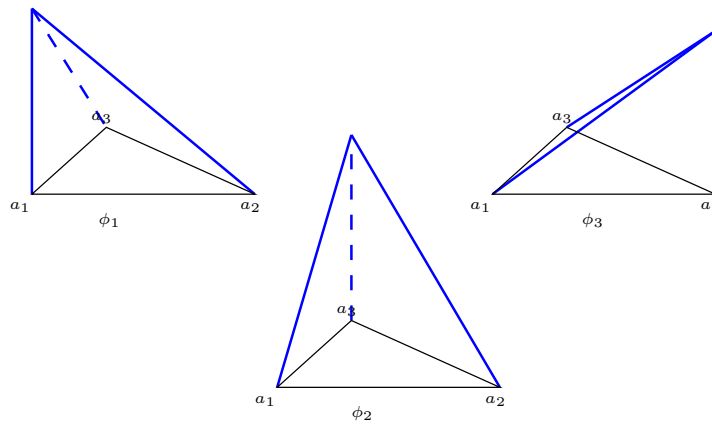
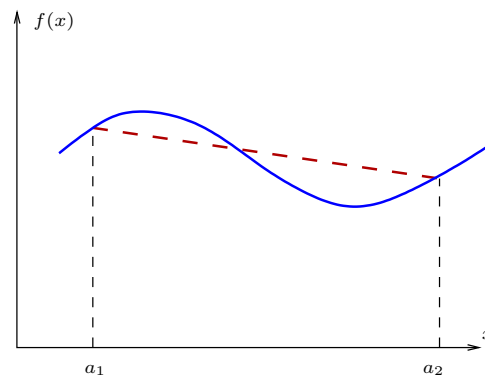
$$\begin{cases} f(a_1) = \alpha_1 \\ f(a_2) = \alpha_2 \end{cases}$$

Or toute fonction  $f$  de  $P$  s'exprime sous la forme  $f(x) = \lambda x + \mu$  et le système

$$\begin{cases} \lambda a_1 + \mu = \alpha_1 \\ \lambda a_2 + \mu = \alpha_2 \end{cases}$$

détermine  $\lambda$  et  $\mu$  de manière unique.

De même si on considère le cas  $d = 2$ . On prend comme élément  $K$  un triangle et comme noeuds les trois sommets,  $a_1, a_2, a_3$  du triangle. Soit  $P = P_1 = \{f : \mathbb{R} \rightarrow \mathbb{R}; f(x) = \lambda x_1 + \mu x_2 + \nu\}$  l'ensemble des fonctions affines. Alors le triplet  $(K, \Sigma_K, P)$  est un élément fini de Lagrange car  $f \in P$  est entièrement déterminée par  $f(a_1), f(a_2)$  et  $f(a_3)$ .

FIG. 4.1 – fonctions de base locales pour l'élément fini de Lagrange  $P_1$  en dimension 2FIG. 4.2 – Interpolée  $P_1$  sur  $[a_1, a_2]$  (en trait pointillé) d'une fonction régulière (en trait continu)

**Définition 4.2 (Fonctions de base locales)** Si  $(K, \Sigma_K, P)$  est un élément fini de Lagrange, alors toute fonction  $f$  de  $P$  peut s'écrire :

$$f = \sum_{i=1}^{N_\ell} f(a_i) f_i$$

avec  $f_i \in P$  et  $f_i(a_j) = \delta_{ij}$ . Les fonctions  $f_i$  sont appelées fonctions de base locales.

Pour l'élément fini de Lagrange  $P_1$  en dimension 2 considéré plus haut, les fonctions de base locales sont décrites sur la figure 4.1

**Définition 4.3 (Interpolée)** Soit  $(K, \Sigma_K, P)$  un élément fini de Lagrange, et soit  $v \in C(K, \mathbb{R})$ . L'interpolée de  $v$  est la fonction  $\Pi v \in P$  définie par :

$$\Pi v = \sum_{i=1}^{N_\ell} v(a_i) f_i$$

On montre sur la figure 4.2 un exemple d'interpolée pour l'élément fini de Lagrange  $P_1$  en dimension 1. L'étude de  $\|v - \Pi v\|$  va nous permettre d'établir une majoration de l'erreur de consistance  $d(u, H_N)$ .

**Remarque 4.4** Pour que le triplet  $(K, \Sigma_K, P)$  soit un élément fini de Lagrange, il faut, mais il ne suffit pas, que  $\dim P = \text{card} \Sigma_K$ . Par exemple si  $P = P_1$  et qu'on prend comme noeuds du triangle deux sommets et le milieu de l'arête joignant les deux sommets, (voir figure 4.3),  $(K, \Sigma_K, P)$  n'est pas un élément fini de Lagrange.

**Proposition 4.5 (Critère de détermination)** Soit  $(K, \Sigma, P)$  un triplet constitué d'un élément, d'un ensemble de noeuds et d'un espace de polynômes, tel que :

$$\dim P = \text{card} \Sigma = N_\ell \quad (4.1)$$

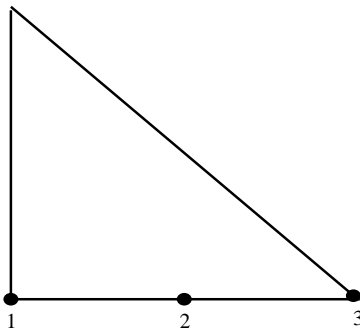


FIG. 4.3 – Exemple de triangle à trois noeuds qui n'est pas un élément fini de Lagrange)

Alors

$$\text{si } \exists! f \in P; f = 0 \text{ sur } \Sigma \quad (4.2)$$

ou si

$$\forall i \in \{1 \dots N_\ell\} \exists f_i \in P \quad f_i(a_j) = \delta_{ij} \quad (4.3)$$

alors  $(K, \Sigma, P)$  est un élément fini de Lagrange.

**Démonstration :** Soit :

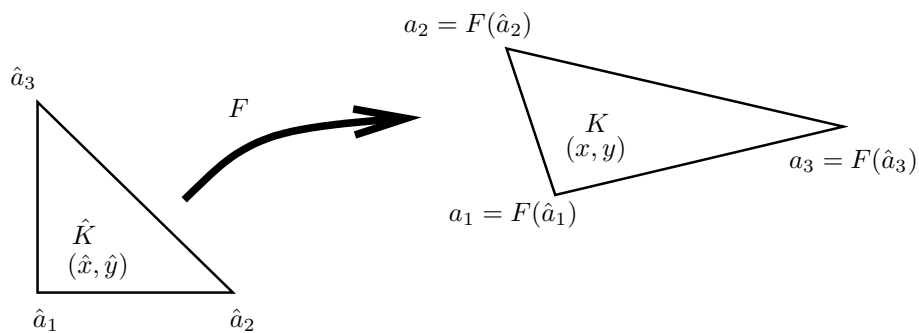
$$\phi : P \rightarrow \mathbb{R}^{N_\ell}$$

$$f \mapsto (f(a_i))_{i=1, N_\ell}.$$

L'application  $\phi$  est linéaire de  $P$  dans  $\mathbb{R}^{N_\ell}$ , et, par hypothèse  $\text{card } s\Sigma = \dim P$ . Donc  $\phi$  est une application linéaire continue de  $P$  dans  $\mathbb{R}^{N_\ell}$ , avec  $\dim P = \dim(\mathbb{R}^{N_\ell}) = N_\ell$ . Si  $(K, \Sigma, P)$  vérifie la condition (4.2) alors  $\phi$  est injective. En effet, si  $\phi(f) = 0$ , alors  $f(a_i) = 0, \forall i = 1, \dots, N_\ell$ , et donc par hypothèse,  $f = 0$ . Donc  $\phi$  est une application linéaire,  $\phi$  est injective de  $P$  dans  $\mathbb{R}^{N_\ell}$  avec  $\dim P = N_\ell$ . On en déduit que  $\phi$  est bijective. Donc toute fonction de  $P$  est entièrement déterminée par ses valeurs aux noeuds :  $(K, \Sigma, P)$  est donc un élément fini de Lagrange.

On montre facilement que si la condition (4.3) est vérifiée alors  $\phi$  est surjective. Donc  $\phi$  est bijective, et  $(K, \Sigma, P)$  est un élément fini de Lagrange. ■

**Proposition 4.6** Soit  $(\bar{K}, \bar{\Sigma}, \bar{P})$ , un élément fini de Lagrange, où  $\bar{\Sigma}$  est l'ensemble des noeuds de  $\bar{K}$  et  $\bar{P}$  un espace de fonctions de dimension finie, et soit  $F$  une bijection de  $\bar{K}$  dans  $K$ , où  $K$  est une maille d'un maillage éléments finis. On pose  $\Sigma = F(\bar{\Sigma})$  et  $P = \{f : K \rightarrow \mathbb{R}; f \circ F \in \bar{P}\}$  (voir figure 4.4). Alors le triplet  $(K, \Sigma, P)$  est un élément fini de Lagrange.

FIG. 4.4 – Transformation  $F$

**Démonstration :** Supposons que les hypothèses de la proposition sont réalisées. On veut donc montrer que  $(\Sigma, P)$  est unisolvant. Soit  $\Sigma = (a_1, \dots, a_{N_\ell})$ , et soit  $(\alpha_1, \dots, \alpha_{N_\ell}) \in \mathbb{R}^{N_\ell}$ . On veut montrer qu'il existe une unique fonction  $f \in P$  telle que

$$f(a_i) = \alpha_i, \quad \forall i = 1, \dots, N_\ell.$$

Or par hypothèse,  $(\bar{\Sigma}, \bar{P})$  est unisolvant. Donc il existe une unique fonction  $\bar{f} \in \bar{P}$  telle que

$$\bar{f}(\bar{a}_i) = \alpha_i, \quad \forall i = 1, \dots, N_\ell,$$

(où  $(\bar{a}_i)_{i=1, \dots, N_\ell}$  désignent les noeuds de  $\bar{K}$ ). Soit  $F$  la bijection de  $\bar{K}$  sur  $K$ , on pose  $f = \bar{f} \circ F^{-1}$ . Or par hypothèse,  $a_i = F(\bar{a}_i)$ . On a donc :  $f(a_i) = \bar{f} \circ F^{-1}(a_i) = \bar{f}(\bar{a}_i) = \alpha_i$ . On a ainsi montré l'existence de  $f$  telle que  $f(a_i) = \alpha_i$ .

Montrons maintenant que  $f$  est unique. Supposons qu'il existe  $f$  et  $g \in P$  telles que :

$$f(a_i) = g(a_i) = \alpha_i, \quad \forall i = 1, \dots, N_\ell.$$

Soit  $h = f - g$  on a donc :

$$h(a_i) = 0 \quad \forall i = 1 \dots N_\ell.$$

On a donc  $h \circ F(\bar{a}_i) = h(a_i) = 0$ . Or  $h \circ F \in \bar{P}$ , et comme  $(\bar{\Sigma}, \bar{P})$  est unisolvant, on en déduit que  $h \circ F = 0$ . Comme, pour tout  $x \in K$ , on a  $h(x) = h \circ F \circ F^{-1}(x) = h \circ F(F^{-1}(x)) = 0$ , on en conclut que  $h = 0$ . ■

**Définition 4.7 (Eléments affine-équivalents) .** Sous les hypothèses de la proposition 4.6, si la bijection  $F$  est affine, on dit que les éléments finis  $(\bar{K}, \bar{\Sigma}, \bar{P})$  et  $(K, \Sigma, P)$  sont affine-équivalents.

**Remarque 4.8** Soient  $(\bar{K}, \bar{\Sigma}, \bar{P})$  et  $(K, \Sigma, P)$  deux éléments finis affine-équivalents. Si les fonctions de base locales de  $(\bar{K}, \bar{\Sigma}, \bar{P})$ . (resp. de  $(K, \Sigma, P)$ ) sont affines, alors celles de  $K$  (resp.  $\bar{K}$ ) le sont aussi, et on a :

$$\begin{cases} \bar{f}_i &= f_i \circ F, \\ f_i &= \bar{f}_i \circ F^{-1}, \end{cases} \quad i = 1, \dots, \text{card}\Sigma$$

La preuve de cette remarque fait l'objet de l'exercice 54.

**Proposition 4.9 (Interpolation)** Sous les hypothèses de la proposition 4.10 page 141, soient  $\Pi_{\bar{K}}$  et  $\Pi_K$  les opérateurs d'interpolation respectifs sur  $\bar{K}$  et  $K$ , voir définition 4.3 page 138. Soient  $v \in C(K, \mathbb{R})$ ,  $\Pi_{\bar{K}}v$  et  $\Pi_Kv$  les interpolées respectives de  $v$  sur  $(\bar{K}, \bar{P})$  et  $(K, P)$ , alors on a :

$$\Pi_Kv \circ F = \Pi_{\bar{K}}(v \circ F)$$

**Démonstration :** Remarquons tout d'abord que  $\Pi_Kv \circ F$  et  $\Pi_{\bar{K}}(v \circ F)$  sont toutes deux des fonctions définies de  $\bar{K}$  à valeurs dans  $\mathbb{R}$ , voir figure 4.5. Remarquons ensuite que, par définition de l'interpolée,  $\Pi_Kv \in P$ . Comme

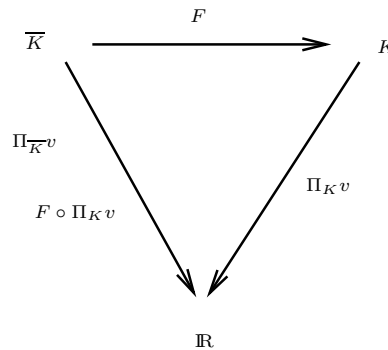


FIG. 4.5 – Opérateurs d'interpolation  $\Pi_{\bar{K}}$  et  $\Pi_K$

$(\bar{K}, \bar{\Sigma}, \bar{P})$  est l'élément de référence, on a donc :

$$\Pi_K v \circ F \in \bar{P}$$

On a aussi, par définition de l'interpolée :  $\Pi_{\bar{K}}(v \circ F) \in \bar{P}$ . On en déduit que  $\Pi_K v \circ F$  et  $\Pi_{\bar{K}}(v \circ F)$  sont toutes deux des fonctions de  $\bar{P}$ . Comme l'élément  $(\bar{K}, \bar{P}, \bar{\Sigma})$  est unisolvant (car c'est un élément fini de Lagrange), toute fonction de  $\bar{P}$  est uniquement déterminée par ses valeurs aux noeuds de  $\bar{\Sigma}$ . Pour montrer l'égalité de  $\Pi_K v \circ F$  et  $\Pi_{\bar{K}}(v \circ F)$ , il suffit donc de montrer que :

$$\Pi_{\bar{K}}(v \circ F)(\bar{a}_i) = \Pi_K v \circ F(\bar{a}_i), \quad i = 1, \dots, N_\ell,$$

où  $N_\ell = \text{card} \bar{\Sigma}$ . Décomposons  $\Pi_{\bar{K}}(v \circ F)$  sur les fonctions de base locales  $(\bar{f}_j), j = 1, \dots, N_\ell$ . On obtient :

$$\Pi_{\bar{K}}(v \circ F)(\bar{a}_i) = \sum_{j=1}^{N_\ell} v \circ F(\bar{a}_j) \bar{f}_j(\bar{a}_i).$$

On a donc :

$$\Pi_{\bar{K}}(v \circ F)(\bar{a}_i) = v \circ \left( \sum_{j=1}^{N_\ell} F(\bar{a}_j) \bar{f}_j \right) (\bar{a}_i) = v \circ F(\bar{a}_i) = v(a_i).$$

Mais on a aussi :

$$\Pi_K v \circ F(\bar{a}_i) = \Pi_K v(F(\bar{a}_i)) = \Pi_K v(a_i) = v(a_i).$$

D'où l'égalité. ■

## 4.1.2 Construction de $H_N$ et conformité

Nous allons considérer deux cas : le cas où l'espace  $H$  est l'espace  $H^1$  tout entier, et le cas où l'espace  $H$  est l'espace  $H_0^1$

**Cas  $H = H^1(\Omega)$**

Plaçons-nous ici dans le cas où  $H = H^1(\Omega)$ , où  $\Omega \subset \mathbb{R}^d$  est un ouvert borné polygonal (si  $d = 2$ , polyédrique si  $d = 3$ ). Soit  $\mathcal{T}$  un maillage éléments finis, avec  $\mathcal{T} = (K_\ell)_{\ell=1, \dots, L}$ , où les éléments finis  $K_\ell$  sont fermés et tels que  $\bigcup_{\ell=1}^L K_\ell = \bar{\Omega}$ . Soit  $\mathcal{S} = (S_i)_{i=1, \dots, M}$  l'ensemble des noeuds du maillage éléments finis, avec  $S_i \in \bar{\Omega}$ ,  $\forall i = 1, \dots, M$ . On cherche à construire une méthode d'éléments finis de Lagrange ; donc à chaque élément  $K_\ell$ ,  $\ell = 1, \dots, L$ , est associé un ensemble de noeuds  $\Sigma_\ell = \mathcal{S} \cap K_\ell$ , et un espace  $P_\ell$  de polynômes. On veut que chaque triplet  $(K_\ell, \Sigma_\ell, P_\ell)$  soit un élément fini de Lagrange. On définit les fonctions de base globales  $(\phi_i)_{i=1, \dots, M}$ , par :

$$\phi_i|_{K_\ell} \in P_\ell \quad \forall i = 1, \dots, M; \quad \forall \ell = 1; \dots, L, \quad (4.4)$$

et

$$\phi_i(S_j) = \delta_{ij} \quad \forall i = 1, \dots, M, \quad \forall j = 1, \dots, M. \quad (4.5)$$

Chaque fonction  $\phi_i$  est définie de manière unique, grâce au caractère unisolvant de  $(K_\ell, \Sigma_\ell, P_\ell)$ ,  $\ell = 1, \dots, M$ . On pose  $H_N = \text{Vect}(\phi_1, \dots, \phi_M)$ . Pour obtenir une méthode d'éléments finis conforme, il reste à s'assurer que  $H_N \subset H^1$ .

Une manière de construire l'espace  $H_N$  est de construire un maillage à partir d'un élément de référence, grâce à la proposition suivante, qui se déduit facilement de la proposition 4.6 page 139

**Proposition 4.10 (Elément fini de référence)** *Soit  $\mathcal{T}$  un maillage constitué d'éléments  $K$ . On appelle élément fini de référence un élément fini de Lagrange  $(\bar{K}, \bar{\Sigma}, \bar{P})$ , où  $\bar{\Sigma}$  est l'ensemble des noeuds de  $\bar{K}$  et  $\bar{P}$  un espace de fonctions, de dimension finie, tel que, pour tout autre élément  $K \in \mathcal{T}$ , il existe une bijection  $F : \bar{K} \rightarrow K$  telle que  $\Sigma = F(\bar{\Sigma})$  et  $P = \{f : K \rightarrow \mathbb{R}; f \circ F \in \bar{P}\}$  (voir figure 4.4). Le triplet  $(K, \Sigma, P)$  est un élément fini de Lagrange.*

**Proposition 4.11 (Critère de conformité, cas  $H^1$ )** Soit  $\Omega$  un ouvert polygonal (ou polyédrique) de  $\mathbb{R}^d$ ,  $d = 2$  ou  $3$ . Soit  $\mathcal{T} = (K_\ell)_{\ell=1,\dots,L}$ , un maillage éléments finis de  $\Omega$ ,  $\mathcal{S} = (S_i)_{i=1,\dots,M}$  l'ensemble des noeuds de maillage. On se place sous les hypothèses de la proposition 4.10; soient  $(\phi_i)_{i=1,\dots,M}$  les fonctions de base globales, vérifiant (4.4) et (4.5), et on suppose de plus que les hypothèses suivantes sont vérifiées :

$$\text{Pour toute arête (ou face si } d = 3) \epsilon = K_{\ell_1} \cap K_{\ell_2}, \text{ on a : } \Sigma_{\ell_1} \cap \epsilon = \Sigma_{\ell_2} \cap \epsilon \text{ et } P_{\ell_1}|_\epsilon = P_{\ell_2}|_\epsilon, \quad (4.6)$$

où  $P_{\ell_1}|_\epsilon$  (resp.  $P_{\ell_2}|_\epsilon$ ) désigne l'ensemble des restrictions des fonctions de  $P_{\ell_1}$  (resp.  $P_{\ell_2}$ ) à  $\epsilon$ ,

$$\text{Si } \epsilon \text{ est un côté de } K_\ell, (\Sigma_\ell \cap \epsilon, P_\ell|_\epsilon) \text{ est unisolvant.} \quad (4.7)$$

Alors on a :  $H_N \subset C(\bar{\Omega})$  et  $H_N \subset H^1(\Omega)$ . On a donc ainsi construit une méthode d'éléments finis conformes. (Notons que les côtés de  $K_\ell$  sont des arêtes en 2D et des faces en 3D.)

**Démonstration :** Pour montrer que  $H_N \subset C(\bar{\Omega})$  et  $H_N \subset H^1(\Omega)$ , il suffit de montrer que pour chaque fonction de base globale  $\phi_i$ , on a  $\phi_i \in C(\bar{\Omega})$  et  $\phi_i \in H^1(\Omega)$ . Or par hypothèse, (4.4), chaque fonction  $\phi_i$  est polynômiale par morceaux. De plus, grâce à l'hypothèse (4.6), on a raccord des polynômes sur les interfaces des éléments, ce qui assure la continuité de  $\phi_i$ . Il reste à montrer que  $\phi_i \in H^1(\Omega)$  pour tout  $i = 1, \dots, M$ . Comme  $\phi_i \in C(\bar{\Omega})$ , il est évident que  $\phi_i \in L^2(\Omega)$  (car  $\Omega$  est un ouvert borné, donc  $\phi_i \in L^\infty(\Omega) \subset L^2(\Omega)$ ).

Montrons maintenant que les dérivées faibles  $D_j \phi_i$ ,  $j = 1, \dots, d$ , appartiennent à  $L^2(\Omega)$ . Par définition, la fonction  $\phi_i$  admet une dérivée faible dans  $L^2(\Omega)$  s'il existe une fonction  $\psi_{i,j} \in L^2(\Omega)$  telle que :

$$\int_{\Omega} \phi_i(x) \partial_j \varphi(x) dx = - \int_{\Omega} \psi_{i,j}(x) \varphi(x) dx, \quad (4.8)$$

pour toute fonction  $\varphi \in C_c^1(\Omega)$  (on rappelle que  $C_c^1(\Omega)$  désigne l'ensemble des fonctions de classe  $C^1$  à support compact, et que  $\partial_j$  désigne la dérivée classique par rapport à la  $j$ -ème variable). Or, comme  $\bar{\Omega} = \bigcup_{\ell=1}^L K_\ell$ , on a :

$$\int_{\Omega} \phi_i(x) D_j \varphi(x) dx = \sum_{\ell=1}^L \int_{K_\ell} \phi_i(x) D_j \varphi(x) dx.$$

Sur chaque élément  $K_\ell$ , la fonction  $\phi_i$  est polynômiale. On peut donc appliquer la formule de Green, et on a :

$$\int_{K_\ell} \phi_i(x) \partial_j \varphi(x) dx = \int_{\partial K_\ell} \phi_i(x) \varphi(x) n_j(x) d\gamma(x) - \int_{K_\ell} \partial_j \phi_i(x) \varphi(x) dx,$$

où  $n_j(x)$  est la  $j$ -ième composante du vecteur unitaire normal à  $\partial K_\ell$  en  $x$ , extérieur à  $K_\ell$ . Mais, si on note  $\mathcal{E}_{int}$  l'ensemble des arêtes intérieures du maillage (i.e. celles qui ne sont pas sur le bord), on a :

$$\begin{aligned} X &= \sum_{\ell=1}^L \int_{\partial K_\ell} \phi_i(x) \varphi(x) n_j(x) d\gamma(x) = \int_{\partial \Omega} \phi_i(x) \varphi(x) n_j(x) d\gamma(x) \\ &+ \sum_{\epsilon \in \mathcal{E}_{int}} \int [(\phi_i(x) \varphi(x) n_j(x))|_{K_{\ell_1}} + (\phi_i(x) \varphi(x) n_j(x))|_{K_{\ell_2}}] d\gamma(x). \end{aligned}$$

où  $K_{\ell_1}$  et  $K_{\ell_2}$  désignent les deux éléments dont  $\epsilon$  est l'interface.

Comme  $\varphi$  est à support compact,

$$\int_{\partial \Omega} \phi_i(x) \varphi(x) n_j(x) d\gamma(x) = 0.$$

Comme  $\phi_i$  et  $\varphi$  sont continues et comme  $n_j(x)|_{K_{\ell_1}} = -n_j(x)|_{K_{\ell_2}}$  pour tout  $x \in \epsilon$ , on en déduit que  $X = 0$ . En reportant dans (4.1.2), on obtient donc que :

$$\int_{\Omega} \phi_i(x) \partial_j \varphi(x) dx = - \sum_{\ell=1}^L \int_{K_\ell} \partial_j \phi_i(x) \varphi(x) dx.$$

Soit  $\psi_{i,j}$  la fonction de  $\Omega$  dans  $\mathbb{R}$  définie presque partout par

$$\psi_{i,j} \Big|_{K_\ell} = -\partial_j \phi_i.$$

Comme  $\partial_j \phi_i$  est une fonction polynômiale par morceaux, on a  $\psi_{i,j} \in L^2(\Omega)$  qui vérifie (4.8), ce qui termine la démonstration. ■



Cas  $H = H_0^1(\Omega)$

Plaçons-nous maintenant dans le cas où  $H = H_0^1(\Omega)$ . On décompose alors l'ensemble  $\mathcal{S}$  des noeuds du maillage :

$$\mathcal{S} = \mathcal{S}_{int} \cup \mathcal{S}_{ext}$$

où

$$\mathcal{S}_{int} = \{S_i, i = 1, \dots, N\} \subset \Omega$$

est l'ensemble des noeuds intérieurs à  $\Omega$  et

$$\mathcal{S}_{ext} = \{S_i, i = N + 1, \dots, M\} \subset \partial\Omega$$

est l'ensemble des noeuds de la frontière. Les fonctions de base globales sont alors les fonctions  $\phi_i, i = 1, \dots, N$  telles que

$$\phi_i|_{K_\ell} \in P_\ell, \forall i = 1, \dots, N, \quad \forall \ell = 1, \dots, L \quad (4.9)$$

$$\phi_i(S_j) = \delta_{ij}, \forall j = 1, \dots, N, \quad (4.10)$$

et on pose là encore  $H_N = \text{Vect}\{\phi_1, \dots, \phi_N\}$ . On a alors encore le résultat suivant :

**Proposition 4.12 (Critère de conformité, cas  $H_0^1$ )** Soit  $\Omega$  un ouvert polygonal (ou polyédrique) de  $\mathbb{R}^d, d = 2$  ou  $3$ . Soit  $\mathcal{T} = (K_\ell)_{\ell=1, \dots, L}$  un maillage éléments finis de  $\Omega$ ,  $\mathcal{S} = (S_i)_{i=1, \dots, M} = \mathcal{S}_{int} \cup \mathcal{S}_{ext}$  l'ensemble des noeuds du maillage. On se place sous les hypothèses de la proposition 4.6. On suppose que les fonctions de base globale  $(\phi_i)_{i=1, \dots, M}$  vérifient (4.9) et (4.10), et que les conditions (4.6) et (4.7) sont vérifiées. Alors on a :  $H_N \subset C(\bar{\Omega})$  et  $H_N \subset H_0^1(\Omega)$

**Démonstration :** La preuve de cette proposition est laissée à titre d'exercice. ■

**Remarque 4.13 (Eléments finis conformes dans  $H^2(\Omega)$ )** On a construit un espace d'approximation  $H_N$  inclus dans  $C(\bar{\Omega})$ . En général, on n'a pas  $H_N \subset C^1(\bar{\Omega})$ , et donc on n'a pas non plus  $H_N \subset H^2(\Omega)$  (en dimension 1 d'espace,  $H^2(\Omega) \subset C^1(\bar{\Omega})$ ). Même si on augmente le degré de l'espace des polynômes, on n'obtiendra pas l'inclusion  $H_N \subset C^1(\bar{\Omega})$ . Si on prend par exemple les polynômes de degré 2 sur les éléments, on n'a pas de condition pour assurer le raccord, des dérivées aux interfaces. Pour obtenir ce raccord, les éléments finis de Lagrange ne suffisent pas : il faut prendre des éléments de type Hermite, pour lesquels les degrés de liberté ne sont plus seulement les valeurs de la fonction aux noeuds, mais aussi les valeurs de ses dérivées aux noeuds. Les éléments finis de Hermite seront par exemple bien adaptés à l'approximation des problèmes elliptiques d'ordre 4, dont un exemple est l'équation :

$$\Delta^2 u = f \text{ dans } \Omega$$

où  $\Omega$  est un ouvert borné de  $\mathbb{R}^2$ ,  $\Delta^2 u = \Delta(\Delta u)$ , et avec des conditions aux limites adéquates, que nous ne détaillerons pas ici. On peut, en fonction de ces conditions aux limites, trouver un espace de Hilbert  $H$  et une formulation faible de (4.13), qui s'écrit :

$$\begin{cases} \int_{\Omega} \Delta u(x) \Delta \varphi(x) dx = \int_{\Omega} f(x) \varphi(x) dx \\ u \in H, \forall \varphi \in H. \end{cases}$$

Pour que cette formulation ait un sens, il faut que  $\Delta u \in L^2(\Omega)$  et  $\Delta \varphi \in L^2(\Omega)$ , et donc que  $H \subset H^2(\Omega)$ . Pour construire une approximation par éléments finis conforme de ce problème, il faut donc choisir  $H_N \subset H^2(\Omega)$ , et le choix des éléments finis de Hermite semble donc indiqué.

## 4.2 Exemples

Pour chaque méthode d'élément fini de Lagrange, on définit :

1. un élément de référence  $\bar{K}$
2. des fonctions de base locales sur  $\bar{K}$
3. une bijection  $F_\ell$  de  $K$  sur  $K_\ell$ , pour  $\ell = 1, \dots, L$ , où  $L$  est le nombre d'éléments du maillage.

### 4.2.1 Élément fini de Lagrange $P1$ sur triangle ( $d = 2$ )

Le maillage du domaine est constitué de  $L$  triangles  $(K_\ell)_{\ell=1,\dots,L}$ , et les polynômes d'approximation sont de degré 1.

**Élément fini de référence** : on choisit le triangle  $\bar{K}$  de sommets  $(0, 0)$ ,  $(1, 0)$  et  $(0, 1)$ , et  $\bar{P} = \{\psi : K \rightarrow \mathbb{R}(x, y) \mapsto ax + by + c, (a, b, c) \in \mathbb{R}^3\}$ .

**Proposition 4.14 (Unisolvance)** Soit  $\bar{\Sigma} = (\bar{a}_i)_{i=1,2,3}$  avec  $\bar{a}_1 = (0, 0)$ ,  $\bar{a}_2 = (1, 0)$  et  $\bar{a}_3 = (0, 1)$ , et

$$\bar{P} = \{\psi; K \rightarrow \mathbb{R}; (x, y) \mapsto a + bx + cy, (a, b, c) \in \mathbb{R}^3\}$$

Alors le couple  $(\bar{\Sigma}, \bar{P})$  est unisolvant.

**Démonstration** : Soit  $(\alpha_1, \alpha_2, \alpha_3) \in \mathbb{R}^3$ , et  $\psi \in \bar{P}$ . On suppose que  $\psi(\bar{a}_i) = \alpha_i$ ,  $i = 1, 2, 3$ . La fonction  $\psi$  est de la forme  $\psi(x, y) = a + bx + cy$  et on a donc :

$$\begin{cases} a = \alpha_1 \\ a + b = \alpha_2 \\ a + c = \alpha_3 \end{cases}$$

d'où  $c = \alpha_3 - \alpha_1$ ,  $b = \alpha_2 - \alpha_1$  et  $b_2 = \alpha_3 - \alpha_2$ . La connaissance de  $\psi$  aux noeuds  $(\bar{a}_i)_{i=1,2,3}$  détermine donc entièrement la fonction  $\psi$ . ■

**Fonctions de bases locales.**

Les fonctions de base locales sur l'élément fini de référence  $\bar{K}$  sont définies par  $\bar{\phi}_i \in \bar{P}$   $\bar{\phi}_i(\bar{a}_j) = \delta_{ij}$ , ce qui détermine les  $\bar{\phi}_i$ ; de manière unique, comme on vient de le voir. Et on a donc

$$\begin{cases} \bar{\phi}_1(\bar{x}, \bar{y}) = 1 - \bar{x} - \bar{y} \\ \bar{\phi}_2(\bar{x}, \bar{y}) = \bar{x} \\ \bar{\phi}_3(\bar{x}, \bar{y}) = \bar{y}. \end{cases}$$

**Transformation  $F_\ell$**

On construit ici une bijection affine qui transforme  $\bar{K}$  le triangle de référence en un autre triangle  $K$  du maillage. On cherche donc  $\ell : \bar{K} \rightarrow K$ , telle que

$$F_\ell(\bar{a}_i) = a_i \quad i = 1, \dots, 3$$

où  $\Sigma = (a_i)_{i=1,2,3}$  est l'ensemble des sommets de  $K$ . Notons  $(x_i, y_i)$  les coordonnées de  $a_i$ ,  $i = 1, 2, 3$ . Comme  $F_\ell$  est une fonction affine de  $\mathbb{R}^2$  dans  $\mathbb{R}^2$ , elle s'écrit sous la forme.

$$F_\ell(\bar{x}, \bar{y}) = (\beta_1 + \gamma_1 \bar{x} + \delta_1 \bar{y}, \beta_2 + \gamma_2 \bar{x} + \delta_2 \bar{y})$$

et on cherche  $\beta_i, \gamma_i, \delta_i$ ,  $i = 1, 2$  tels que :

$$\begin{cases} F_\ell((0, 0)) = (x_1, y_1) \\ F_\ell((1, 0)) = (x_2, y_2) \\ F_\ell((0, 1)) = (x_3, y_3). \end{cases}$$

Une résolution de système élémentaire amène alors à :

$$F_\ell(\bar{x}, \bar{y}) = \begin{pmatrix} x_1 + (x_2 - x_1)\bar{x} + (x_3 - x_1)\bar{y} \\ y_1 + (y_2 - y_1)\bar{x} + (y_3 - y_1)\bar{y} \end{pmatrix}$$

D'après la remarque 4.8 page 140, si on note  $\bar{\phi}_k$ ,  $k = 1, 2, 3$  les fonctions de base locales de l'élément de référence  $(\bar{K}, \bar{\Sigma}, \bar{P})$ , et  $\phi_k^{(\ell)}$ ,  $k = 1, 2, 3$  les fonctions de base locales de l'élément  $(K_\ell, \Sigma_\ell, P_\ell)$ , on a  $\phi_k^{(\ell)} = \bar{\phi}_k \circ F_\ell^{-1}$

Si on note maintenant  $(\phi_i)_{i=1,\dots,N}$  les fonctions de base globales, on a :

$$\phi_i \Big|_{K_\ell} = \phi_k^{(\ell)},$$

où  $i = ng(\ell, k)$  est l'indice du  $k$ -ième noeud de l'élément  $\ell$  dans la numérotation globale. Notons que l'élément fini de Lagrange ainsi défini vérifie les critères de cohérence 4.6 page 142 et (4.7) page 142. Pour compléter la définition de l'espace d'approximation  $H_N$ , il ne reste qu'à déterminer les "noeuds liés", de la façon dont on a traité le cas de l'espace  $H_0^1(\Omega)$ .

Il faut également insister sur le fait que cet élément est très souvent utilisé, en raison de sa facilité d'implantation et de la structure creuse des systèmes linéaires qu'il génère. Il est particulièrement bien adapté lorsqu'on cherche des solutions dans l'espace  $H^1(\Omega)$ . Il se généralise facilement en trois dimensions d'espace, où on utilise alors des tétraèdres, avec toujours comme espace de polynôme l'espace des fonctions affines.

### 4.2.2 Élément fini triangulaire $P_2$

Comme le titre du paragraphe l'indique, on considère un maillage triangulaire, et un espace de polynômes de degré 2 pour construire l'espace d'approximation.

Élément fini de référence On choisit comme élément fini de référence le triangle de sommets  $(0,0)$ ,  $(1,0)$  et  $(0,1)$ , voir Figure 4.6 et on prend pour  $\Sigma$  :

$$\bar{\Sigma} = \left\{ (0,0), (1,0), (0,1), \left(\frac{1}{2}, \frac{1}{2}\right), \left(0, \frac{1}{2}\right), \left(\frac{1}{2}, 0\right) \right\}$$

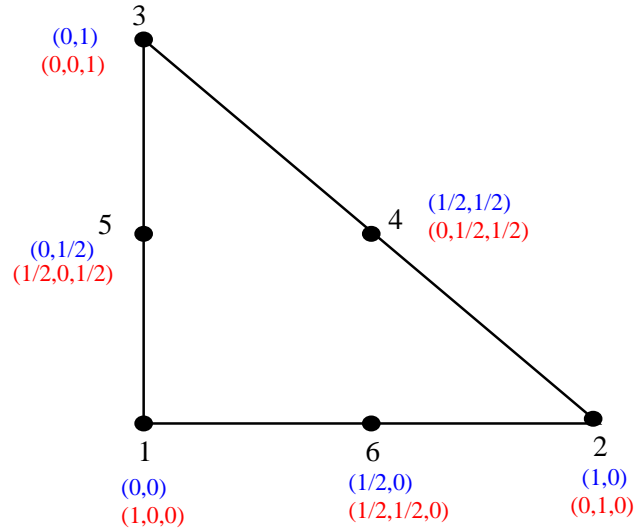


FIG. 4.6 – Élément de référence pour les éléments finis  $P_2$ , avec coordonnées cartésiennes et barycentriques des nœuds

Fonctions de base locales Les fonctions de base locales sont définies à partir des coordonnées barycentriques. On rappelle que les coordonnées barycentriques d'un point  $\mathbf{x}$  du triangle  $K$  de sommets  $a_1, a_2$  et  $a_3$  sont les réels  $\lambda_1, \lambda_2, \lambda_3$  tels que :

$$\mathbf{x} = \lambda_1 a_1 + \lambda_2 a_2 + \lambda_3 a_3.$$

Dans le cas du triangle de référence  $\bar{K}$  de sommets  $(0,0)$ ,  $(1,0)$  et  $(0,1)$ , les coordonnées barycentriques d'un point  $\{\mathbf{x}\}$  de coordonnées cartésiennes  $x$  et  $y$  sont donc :  $\lambda_1 = 1 - x - y$ ,  $\lambda_2 = x$ ,  $\lambda_3 = y$ . Par définition, on a  $\sum_{i=1}^3 \lambda_i = 1$  et  $\lambda_i \geq 0$  (car le triangle  $K$  est l'enveloppe convexe de l'ensemble de ses sommets). On peut alors déterminer les fonctions de base en fonction des coordonnées barycentriques des six nœuds de  $\bar{K}$  exprimés par leurs coordonnées barycentriques :  $a_1 = (1,0,0)$ ,  $a_2 = (0,1,0)$ ,  $a_3 = (0,0,1)$ ,  $a_4 = (0, \frac{1}{2}, \frac{1}{2})$ ,  $a_5 = (\frac{1}{2}, 0, \frac{1}{2})$ ,  $a_6 = (\frac{1}{2}, \frac{1}{2}, 0)$ . Les fonctions de base sont telles que  $\phi_i \in P_2$ , et  $\phi_i(a_j) = \delta_{ij}$ ,  $\forall i = 1, \dots, 6$ , for all  $j = 1, \dots, 6$ . Commençons par  $\phi_1$  ; on veut  $\phi_1(a_1) = 1$ , et  $\phi_1(a_i) = 0$ ,  $\forall i = 2, \dots, 6$ . La fonction  $\phi_1$  définie par

$$\phi_1(x, y) = 2\lambda_1\left(\lambda_1 - \frac{1}{2}\right)$$

convient, et comme le couple  $(\bar{\Sigma}, P_2)$  est unisolvant, c'est la seule fonction qui convient. Par symétrie, on définit

$$\phi_2(x, y) = 2\lambda_2\left(\lambda_2 - \frac{1}{2}\right),$$

et

$$\phi_3(x, y) = 2\lambda_3\left(\lambda_3 - \frac{1}{2}\right).$$

Les fonctions de base associées aux nœuds  $a_4, a_5, a_6$  sont alors

$$\phi_4(x, y) = 4\lambda_2\lambda_3,$$

$$\begin{aligned}\phi_5(x, y) &= 4\lambda_1\lambda_3, \\ \text{et } \phi_6(x, y) &= 4\lambda_1\lambda_2.\end{aligned}$$

Il est facile de voir que ces fonctions forment une famille libre d'éléments de  $P_2$  et comme  $\text{card } \bar{\Sigma} = \text{card } P_2$ , le couple  $(\bar{\Sigma}, P_2)$  est bien unisolvant.

**Transformation  $F_\ell$**  La bijection  $F_\ell$  qui permet de passer de l'élément fini de référence  $\bar{K}$  à l'élément  $K_\ell$  a déjà été vue dans le cas de l'élément fini  $P_1$  c'est la fonction affine définie par :

$$F_\ell(x, y) = \begin{pmatrix} x_1 + (x_2 - x_1)x + (x_3 - x_1)y \\ y_1 + (y_2 - y_1)x + (y_3 - y_1)y \end{pmatrix}$$

où  $(x_i, y_i), i = 1, 2, 3$  sont les coordonnées respectives des trois sommets du triangle  $K_\ell$ . Comme cette transformation est affine, les coordonnées barycentriques restent inchangées par cette transformation.

On peut montrer (ce n'est pas facile) que l'erreur d'interpolation  $\|u - u_N\|_{H^1}$  est contrôlée, en éléments finis  $P_1$  et  $P_2$  par les inégalités suivantes :

$$P_1 : \text{ si } u \in H^2(\Omega), \text{ on a } \|u - u_N\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}$$

$$P_2 : \text{ si } u \in H^3(\Omega), \text{ on a } \|u - u_N\|_{H^1(\Omega)} \leq Ch^2\|u\|_{H^3(\Omega)}.$$

On peut généraliser les éléments finis  $P_1$  et  $P_2$  aux éléments finis  $P_k$  sur triangles, pour  $k \geq 1$ . On prend toujours le même élément de référence, dont on divise chaque côté en  $k$  intervalles. Les extrémités de ces intervalles sont les noeuds du maillage. On a donc  $3k$  noeuds, qu'on peut repérer par leurs coordonnées barycentriques, qui prennent les valeurs  $0, \frac{1}{k}, \frac{2}{k}, \dots, 1$ . On peut montrer que si  $u \in H^{k+1}$ , alors

$$\|u_N - u\|_{H^1(\Omega)} \leq Ch^k\|u\|_{H^{k+1}(\Omega)}$$

### 4.2.3 Éléments finis sur quadrangles

#### Le cas rectangulaire

On prend comme élément fini de référence le carré  $\bar{K} = [-1, 1] \times [-1, 1]$ , et comme noeuds les coins de ce carré :

$$a_1 = (1, -1), a_2 = (1, 1), a_3 = (-1, 1), \text{ et } a_4 = (-1, -1).$$

On prend comme espace de polynômes

$$P = \{f : \bar{K} \rightarrow \mathbb{R}; f \in Q_1\}$$

où  $Q_1 = \{f : \mathbb{R}^2 \rightarrow \mathbb{R}; f(x, y) = a + bx + cy + dxy, (a, b, c, d) \in \mathbb{R}^4\}$  Le couple  $(\Sigma, P)$  est unisolvant. Les fonctions de base locales sont les fonctions :

$$\phi_1(x, y) = -\frac{1}{4}(x+1)(y-1)$$

$$\phi_2(x, y) = \frac{1}{4}(x+1)(y+1)$$

$$\phi_3(x, y) = -\frac{1}{4}(x-1)(y+1)$$

$$\phi_4(x, y) = \frac{1}{4}(x-1)(y-1).$$

La transformation  $F_\ell$  permet de passer de l'élément de référence carré  $\bar{K}$  à un rectangle quelconque du maillage  $K_\ell$ . Si on considère un rectangle  $K_\ell$  parallèle aux axes, dont les noeuds sont notés  $(x_1, y_1), (x_2, y_1), (x_2, y_2), (x_1, y_2)$ , les noeuds du rectangle  $K_\ell$ , la bijection  $F_\ell$  s'écrit :

$$F_\ell(x, y) = \frac{1}{2} \begin{pmatrix} (x_2 - x_1)x + x_2 + x_1 \\ (y_2 - y_1)y + y_2 + y_1 \end{pmatrix}.$$

Considérons maintenant le cas d'un maillage quadrangulaire quelconque. Dans ce cas, on choisit toujours comme élément de référence le carré unité. La transformation  $F_\ell$  qui transforme l'élément de référence en un quadrangle  $K_\ell$  est toujours affine, mais par contre, les composantes de  $F_\ell((x, y))$  dépendent maintenant de  $x$  et de  $y$  voir exercice 53 page 166. En conséquence, le fait que  $f \in Q_1$  n'entraîne plus que  $f \circ F_\ell \in Q_1$ . Les fonctions de base seront donc des polynômes  $Q_1$  sur l'élément de référence  $\bar{K}$ , mais pas sur les éléments "courants"  $K_\ell$ .

**Eléments finis d'ordre supérieur** Comme dans le cas d'un maillage triangulaire, on peut choisir un espace de polynômes d'ordre supérieur,  $Q_k$ , pour les fonctions de base de l'élément de référence  $\bar{K} = [-1, 1] \times [-1, 1]$ . On choisit alors comme ensemble de noeuds :  $\bar{\Sigma} = \bar{\Sigma}_k = \{(x, y) \in \bar{K}, (x, y) \in \{-1, -1 + \frac{1}{k}, -1 + \frac{2}{k}, \dots, 1\}^2\}$ . On peut montrer facilement que  $(\bar{\Sigma}_k, Q_k)$  est unisolvant. Là encore, si la solution exacte de problème continu est suffisamment régulière, on peut démontrer l'estimation d'erreur suivante (voir [3]) :

$$\|u - u_N\|_{H^1(\Omega)} \geq C \|u\|_{H^{k+1}(\Omega)} h^k.$$

Exprimons par exemple l'espace des polynômes  $Q_2$ . On a :

$$Q_2 = \{f : \mathbb{R} \rightarrow \mathbb{R}; f(x) = a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2 + a_7xy^2 + a_8x^2y + a_9x^2y^2, a_i \in \mathbb{R}, i = 1, \dots, 9\}$$

L'espace  $Q_2$  comporte donc neuf degrés de liberté. On a donc besoin de neuf noeuds dans  $\bar{\Sigma}$  pour que le couple  $(\bar{\Sigma}, Q_2)$  soit unisolvant (voir exercice 58 page 167). On peut alors utiliser comme noeuds sur le carré de référence  $[-1, 1] \times [-1, 1]$  :

$$\bar{\Sigma} = \{(-1, -1), (-1, 0), (-1, 1), (0, -1), (0, 0), (0, 1), (1, -1), (1, 0), (1, 1)\}$$

En général, on préfère pourtant supprimer le noeud central (0,0) et choisir :

$$\Sigma^* = \bar{\Sigma} \setminus \{(0, 0)\}.$$

Il faut donc un degré de liberté en moins pour l'espace des polynômes. On définit alors :

$$Q_2^* = \{f : \mathbb{R} \rightarrow \mathbb{R}; f(x) = a_1 + a_2x + a_3y + a_4xy + a_5x^2 + a_6y^2 + a_7xy^2 + a_8x^2y, a_i \in \mathbb{R}, i = 1, \dots, 8\}$$

Le couple  $(\Sigma^*, Q_2^*)$  est unisolvant (voir exercice 59 page 167), et on peut montrer que l'élément  $Q_2^*$  est aussi précis (et plus facile à mettre en oeuvre que l'élément  $Q_2$ ).

### 4.3 Construction du système linéaire

On construit ici le système linéaire pour un problème à conditions aux limites mixtes de manière à envisager plusieurs types de conditions aux limites. Soit  $\Omega$  un ouvert polygonal <sup>1</sup>, on suppose que  $\partial\Omega = \Gamma_0 \cup \Gamma_1$  avec  $\text{mes}(\Gamma_0) \neq 0$ . On va imposer des conditions de Dirichlet sur  $\Gamma_0$  et des conditions de Fourier sur  $\Gamma_1$ ; c'est ce qu'on appelle des conditions "mixtes". On se donne donc des fonctions  $p : \Omega \rightarrow \mathbb{R}$ ,  $g_0 : \Gamma_0 \rightarrow \mathbb{R}$  et  $g_1 : \Gamma_1 \rightarrow \mathbb{R}$ , et on cherche à approcher  $u$  solution de :

$$\begin{cases} -\text{div}(p(x)\nabla u(x)) + q(x)u(x) = f(x), x \in \Omega, \\ u = g_0 \text{ sur } \Gamma_0, \\ p(x)\nabla u(x) \cdot \mathbf{n}(x) + \sigma u(x) = g_1(x), x \in \Gamma_1, \end{cases} \quad (4.11)$$

où  $\mathbf{n}$  désigne le vecteur unitaire normal à  $\partial\Omega$  extérieure à  $\Omega$ . Pour assurer l'existence et unicité du problème (4.11), (voir exercice 41), on se place sous les hypothèses suivantes :

$$\begin{cases} p(x) \geq \alpha > 0, p.p. x \in \Omega \\ q \geq 0 \\ \sigma \geq 0 \\ \text{mes}(\Gamma_0) > 0. \end{cases} \quad (4.12)$$

Pour obtenir une formulation variationnelle, on introduit l'espace

$$H_{\Gamma_0, g_0}^1 = \{u \in H^1(\Omega); u = g_0 \text{ sur } \Gamma_0\}$$

et l'espace vectoriel associé :

$$H = H_{\Gamma_0, 0}^1 = \{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$$

<sup>1</sup>Dans le cas où la frontière  $\partial\Omega$  de  $\Omega$  n'est pas polygonale mais courbe, il faut considérer des éléments finis dits "isoparamétriques" que nous verrons plus loin

Notons que  $H$  est un espace de Hilbert. Par contre, attention, l'espace  $H_{\Gamma_0, g_0}^1$  n'est pas un espace vectoriel. On va chercher  $u$  solution de (4.11) sous la forme  $u = \tilde{u} + u_0$ , avec  $u_0 \in H_{\Gamma_0, g_0}^1$  et  $\tilde{u} \in H_{\Gamma_0, 0}^1$ . Soit  $v \in H$ , on multiplie (4.11) par  $v$  et on intègre sur  $\Omega$ . On obtient :

$$\int_{\Omega} -\operatorname{div}(p(x)\nabla u(x))v(x)dx + \int_{\Omega} q(x)u(x)v(x)dx = \int_{\Omega} f(x)v(x)dx, \quad \forall v \in H.$$

En appliquant la formule de Green, il vient alors :

$$\int_{\Omega} p(x)\nabla u(x)\nabla v(x)dx - \int_{\partial\Omega} p(x)\nabla u(x)nv(x)d\gamma(x) + \int_{\Omega} qu(x)v(x)dx = \int_{\Omega} f(x)v(x)dx, \quad \forall v \in H.$$

Comme  $v = 0$  sur  $\Gamma_0$  on a :

$$\int_{\partial\Omega} p(x)\nabla u(x)nv(x)d\gamma(x) = \int_{\Gamma_1} p\nabla u(x)nv(x)d\gamma(x).$$

Mais sur  $\Gamma_1$ , la condition de Fourier s'écrit :  $\nabla u \cdot n = -\sigma u + g_1$ , et on a donc

$$\int_{\Omega} p(x)\nabla u(x)\nabla v(x)dx + \int_{\Gamma_1} p(x)\sigma u(x)v(x)d\gamma(x) + \int_{\Omega} qu(x)v(x)dx = \int_{\Omega} f(x)v(x)dx + \int_{\Gamma_1} g_1(x)v(x)d\gamma(x).$$

On peut écrire cette égalité sous la forme :  $a(u, v) = \tilde{T}(v)$ , avec  $a(u, v) = a_{\Omega}(u, v) + a_{\Gamma_1}(u, v)$ , où :

$$\begin{cases} a_{\Omega}(u, v) = \int_{\Omega} p(x)\nabla u(x)\nabla v(x)dx + \int_{\Omega} q(x)u(x)v(x)dx, \\ a_{\Gamma_1}(u, v) = \int_{\Gamma_1} p(x)\sigma(x)u(x)v(x)d\gamma(x), \end{cases}$$

et  $\tilde{T}(v) = T_{\Omega}(v) + T_{\Gamma_1}(v)$ , avec

$$T_{\Omega}(v) = \int_{\Omega} f(x)v(x)dx \text{ et } T_{\Gamma_1} = \int_{\Gamma_1} g_1(x)v(x)d\gamma(x).$$

On en déduit une formulation faible associée à (4.11) :

$$\begin{cases} \text{chercher } \tilde{u} \in H \\ a(u_0 + \tilde{u}, v) = \tilde{T}(v), \forall v \in H, \end{cases} \quad (4.13)$$

où  $u_0 \in H^1(\Omega)$  est un revêtement de  $g_0$ , c'est à dire une fonction de  $H^1(\Omega)$  telle que  $u_0 = g_0$  sur  $\Gamma$ . Le problème (4.13) peut aussi s'écrire sous la forme :

$$\begin{cases} \tilde{u} \in H \\ a(\tilde{u}, v) = T(v), \forall v \in H. \end{cases} \quad (4.14)$$

où  $T(v) = \tilde{T}(v) - a(u_0, v)$ . Sous les hypothèses (4.12), on peut alors appliquer le théorème de Lax Milgram (voir théorème 3.6 page 99) au problème (4.14) pour déduire l'existence et l'unicité de la solution de (4.13); notons que, comme la forme bilinéaire  $a$  est symétrique, ce problème admet aussi une formulation variationnelle :

$$\begin{cases} J(u) = \min_{v \in H_{\Gamma_0, g_0}^1} J(v), \\ J(v) = \frac{1}{2}a(v, v) + T(v), \forall v \in H_{\Gamma_0, g_0}^1. \end{cases} \quad (4.15)$$

Dans ce cas, les méthodes de Ritz et Galerkin sont équivalentes. Remarquons que l'on peut choisir  $u_0$  de manière abstraite, tant que  $u_0$  vérifie  $u_0 = g_0$  sur  $\Gamma_0$  et  $u_0 \in H^1$ . Intéressons nous maintenant à la méthode d'approximation variationnelle. On approche l'espace  $H$  par  $H_N = \operatorname{Vect}\{\phi_1, \dots, \phi_N\}$  et on remplace (4.14) par :

$$\begin{cases} \tilde{u}_N \in H_N \\ a(\tilde{u}_N, \phi_i) = T(\phi_i) - a(u_0, \phi_i), \forall i = 1, \dots, N. \end{cases} \quad (4.16)$$

On pose maintenant  $\tilde{u}_N = \sum_{j=1}^N \tilde{u}_j \phi_j$ . Le problème (4.16) est alors équivalent au système linéaire :

$$\mathcal{K}\tilde{U} = \mathcal{G},$$

avec

$$\begin{cases} \mathcal{K}_{ij} = a(\phi_j, \phi_i), \quad i, j = 1, \dots, N, \\ \tilde{U} = (\tilde{u}_1, \dots, \tilde{u}_N), \\ \mathcal{G}_i = T(\phi_i) - a(u_0, \phi_i), \quad i = 1, \dots, N. \end{cases}$$

L'implantation numérique de la méthode d'approximation nécessite donc de :

1. construire  $\mathcal{K}$  et  $\mathcal{G}$
2. résoudre  $\mathcal{K}\tilde{U} = \mathcal{G}$ .

Commençons par la construction de l'espace  $H_N$  et des fonctions de base pour une discrétisation par éléments finis de Lagrange du problème (4.14).

### 4.3.1 Construction de $H_N$ et $\Phi_i$

On considère une discrétisation à l'aide d'éléments finis de Lagrange, qu'on note :  $(K_\ell, \Sigma_\ell, P_\ell)$   $\ell = 1, \dots, L$ , où  $L$  est le nombre d'éléments. On note  $S_i$ ,  $i = 1, \dots, M$ , les noeuds du maillage, et  $\phi_1, \dots, \phi_N$ , les fonctions de base, avec  $N \leq M$ . On peut avoir deux types de noeuds :

- les noeuds libres :  $S_i \notin \Gamma_0$ . On a  $N$  noeuds libres
- les noeuds liés :  $S_i \in \Gamma_0$ . On a  $M - N$  noeuds liés.

Notons qu'on a intérêt à mettre des noeuds à l'intersection de  $\Gamma_0$  et  $\Gamma_1$  (ce seront des noeuds liés). Grâce à ceci, et à la cohérence globale et locale des éléments finis de Lagrange, on a  $H_N \subset H$ . On a donc bien des éléments finis conformes. Récapitulons alors les notations :

- $M$  : nombre de noeuds total
- $N$  : nombre de noeuds libres
- $M_0 = M - N$  : nombre de noeuds liés
- $J_0 = \{ \text{indices des noeuds liés} \} \subset \{1, \dots, M\}$ . On a  $\text{card}J_0 = M_0$
- $J = \{ \text{indices des noeuds libres} \} = \{1 \dots M\} \setminus J_0$ . On a  $\text{card}J = N$ .

Pour la programmation des éléments finis, on a besoin de connaître, pour chaque noeud (local) de chaque élément, son numéro dans la numérotation globale. Pour cela on introduit un tableau  $\text{ng}(L, N_\ell)$ , où  $L$  est le nombre d'éléments et  $N_\ell$  est le nombre de noeuds par élément. (on le suppose constant par souci de simplicité,  $N_\ell$  peut en fait dépendre de  $L$ . Exemple : triangle - quadrangle). Pour tout  $\ell \in \{1, \dots, L\}$  et tout  $r \in \{1, \dots, N_\ell\}$ ,  $\text{ng}(\ell, r)$  est alors le numéro global du  $r$ -ième noeud du  $\ell$ -ième élément. On a également besoin de connaître les coordonnées de chaque noeud. On a donc deux tableaux  $x$  et  $y$  de dimension  $M$ , où  $x(i), y(i)$  représentent les coordonnées du  $i$ -ième noeud. Notons que les tableaux  $\text{ng}$ ,  $x$  et  $y$  sont des données du mailleur (qui est un module externe par rapport au calcul éléments finis proprement dit). Pour les conditions aux limites, on se donne deux tableaux :

- $CF$  : conditions de Fourier
- $CD$  : conditions de Dirichlet

(on verra plus tard le format de ces deux tableaux)

### 4.3.2 Construction de $\mathcal{K}$ et $\mathcal{G}$

On cherche à construire la matrice  $\mathcal{K}$  d'ordre  $(N \times N)$ , définie par :

$$\mathcal{K}_{ij} = a(\phi_j, \phi_i) \quad i, j \in J$$

Ainsi que le vecteur  $\mathcal{G}$ , défini par :

$$\mathcal{G}_i = T(\phi_i) - a(u_0, \phi_i) \quad i \in J \quad \text{card}J = N$$

La première question à résoudre est le choix de  $u_0$ . En effet, contrairement au cas unidimensionnel (voir exercice 37 page 118), il n'est pas toujours évident de trouver  $u_0 \in H_{\Gamma_0, g_0}^1$ . Pour se faciliter la tâche, on commet un "crime variationnel", en remplaçant  $u_0$  par

$$u_{0,N} = \sum_{j \in J_0}^N u_0(S_j) \phi_j.$$

Notons qu'on a pas forcément :  $u_{0,N} \in H_{\Gamma_0, g_0}^1$  ; c'est en ce sens que l'on commet un "crime". Mais par contre, on a bien  $u_{0,N}(S_j) = u_0(S_j)$  pour tout  $j \in J$ . On peut voir la fonction  $u_{0,N}$  comme une approximation non conforme de  $u_0 \in H_{\Gamma_0, g_0}^1$ . On remplace donc  $\mathcal{G}_i$  par :

$$\mathcal{G}_i = T(\phi_i) - \sum_{j \in J_0} g_0(S_j) a(\phi_j, \phi_i).$$

Calculons maintenant  $a(\phi_j, \phi_i)$  pour  $j = 1, \dots, M$ , et  $i = 1, \dots, M$ . On se sert donc pour l'implémentation pratique de la méthode, des fonctions de forme associées aux noeuds "liés", même si dans l'écriture du problème discret théorique, on n'en avait pas besoin.

### Calcul de $\mathcal{K}$ et $\mathcal{G}$

1. Calcul des contributions intérieures : on initialise les coefficients de la matrice  $\mathcal{K}$  et les composantes par les contributions provenant de  $a_\Omega$  et  $T_\Omega$ .

$$\left. \begin{array}{l} \mathcal{K}_{ij} = a_\Omega(\phi_j, \phi_i) \\ \mathcal{G}_i = T_\Omega(\phi_i) \end{array} \right\} \begin{array}{l} i = 1, \dots, N, \\ j = 1, \dots, N. \end{array}$$

2. Calcul des termes de bord de Fourier. On ajoute maintenant à la matrice  $\mathcal{K}$  les contributions de bord :

$$\begin{array}{l} \mathcal{K}_{ij} \leftarrow \mathcal{K}_{ij} + a_{\Gamma_i}(\phi_j, \phi_i), \quad i = 1, \dots, N, \quad j = 1, \dots, N. \\ \mathcal{G}_i \leftarrow \mathcal{G}_i + T_{\Gamma_i}(\phi_i) \quad i = 1 \dots M. \end{array}$$

3. Calcul des termes de bord de Dirichlet. On doit tenir compte ici du relèvement de la condition de bord :

$$\mathcal{G}_i \leftarrow \mathcal{G}_i - \sum_{j \in J_0} g_0(N_i) \mathcal{K}_{ij} \quad \forall i \in J$$

Après cette affectation, les égalités suivantes sont vérifiées :

$$\begin{array}{l} \mathcal{K}_{ij} = a(\phi_j, \phi_i) \quad i, j \in J(\cup J_0) \\ \mathcal{G}_i = T(\phi_i) - a(u_{0,N}, \phi_i). \end{array}$$

Il ne reste plus qu'à résoudre le système linéaire

$$\sum_{j \in J} \mathcal{K}_{ij} \alpha_j = \mathcal{G}_i, \quad \forall i \in J. \quad (4.17)$$

4. Prise en compte des noeuds liés. Pour des questions de structure de données, on inclut en général les noeuds liés dans la résolution du système, et on résout donc le système linéaire d'ordre  $M \geq N$  suivant :

$$\sum_{j=1, \dots, N} \tilde{\mathcal{K}}_{ij} \alpha_j = \mathcal{G}_i, \quad \forall i = 1, \dots, N. \quad (4.18)$$

avec  $\tilde{\mathcal{K}}_{ij} = \mathcal{K}_{ij}$  pour  $i, j \in J$ ,  $\tilde{\mathcal{K}}_{ij} = 0$  si  $(i, j) \notin J^2$ , et  $i \neq j$ , et  $\tilde{\mathcal{K}}_{ii} = 1$  si  $i \notin J$ . Ces deux systèmes sont équivalents, puisque les valeurs aux noeuds liés sont fixées.

Si par chance on a numéroté les noeuds de manière à ce que tous les noeuds liés soient en fin de numérotation, c.à.d. si  $J = \{1, \dots, N\}$  et  $J_0 = \{N+1, \dots, M\}$ , le système (4.18) est de la forme :

$$\left( \begin{array}{c|c} \mathcal{K} & 0 \\ \hline 0 & Id_M \end{array} \right), \quad U = \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_N \\ \alpha_{N+1} \\ \vdots \\ \alpha_M \end{pmatrix}, \quad \text{et } \mathcal{G} = \begin{pmatrix} \mathcal{G}_1 \\ \vdots \\ \mathcal{G}_N \\ \mathcal{G}_{N+1} \\ \vdots \\ \mathcal{G}_M \end{pmatrix}$$

Dans le cas où la numérotation est quelconque, les noeuds liés ne sont pas forcément à la fin, et pour obtenir le système linéaire d'ordre  $M$  (4.18) (donc incluant les inconnues  $\alpha_i$ ,  $i \in J_0$ , qui n'en sont pas vraiment) on peut adopter deux méthodes :



(a) Première méthode : on force les valeurs aux noeuds liés de la manière suivante :

$$\begin{aligned}\mathcal{K}_{ii} &\leftarrow 1 \text{ pour tout } i \in J_0 \\ \mathcal{K}_{ij} &\leftarrow 0 \text{ pour tout } i \in J_0 \quad j \in \{1 \dots M\} \quad i \neq j \\ \mathcal{G}_i &\leftarrow g_0(S_i) \text{ pour tout } i \in J_0\end{aligned}$$

(b) Deuxième méthode : on force les valeurs aux noeuds liés de la manière suivante :

$$\begin{aligned}\mathcal{K}_{ii} &\leftarrow 10^{20} \quad \forall i \in J_0 \\ \mathcal{G}_i &\leftarrow 10^{20}g_0(S_i) \quad \forall i \in J_0\end{aligned}$$

La deuxième méthode permet d'éviter l'affectation à 0 de coefficients extra-diagonaux de la matrice. Elle est donc un peu moins chère en temps de calcul.

**Conclusion** Après les calculs 1, 2, 3, 4, on a obtenu une matrice  $\mathcal{K}$  d'ordre  $M \times M$  et le vecteur  $\mathcal{G}$  de  $\mathbb{R}^M$ . Soit  $\alpha \in \mathbb{R}^M$  la solution du système  $\mathcal{K}\alpha = \mathcal{G}$ . Rappelons qu'on a alors :

$$\begin{aligned}u_N &= \sum_{i=1}^N \alpha_i \phi_i, \\ &= \sum_{i \in J} \alpha_i \phi_i + \sum_{i \in J_0} \alpha_i \phi_i \\ u_N &= \tilde{u}_N + u_0\end{aligned}$$

**Remarque 4.15 (Numérotation des noeuds)** Si on utilise une méthode itérative sans préconditionnement, la numérotation des noeuds n'est pas cruciale. Elle l'est par contre dans le cas d'une méthode directe et si on utilise une méthode itérative avec préconditionnement. Le choix de la numérotation s'effectue pour essayer de minimiser la largeur de bande. On pourra à ce sujet étudier l'influence de la numérotation sur deux cas simples sur la structure de la matrice.

### 4.3.3 Calcul de $a_\Omega$ et $T_\Omega$ , matrices élémentaires.

Détaillons maintenant le calcul des contributions intérieures, c'est à dire  $a_\Omega(\phi_i, \phi_j)$   $i = 1, \dots, M$ ,  $j = 1, \dots, M$  et  $T_\Omega(\phi_i)$   $i = 1, \dots, M$ . Par définition,

$$a_\Omega(\phi_i, \phi_j) = \int_{\Omega} p(x) \nabla \phi_i(x) \nabla \phi_j(x) dx + \int_{\Omega} q(x) \phi_i(x) \phi_j(x) dx.$$

Décomposons  $\Omega$  à l'aide du maillage éléments finis.

$$\Omega = \bigcup_{\ell=1}^L K_\ell.$$

En notant  $\theta(\phi_i, \phi_j)(x) = p(x) \nabla \phi_i(x) \nabla \phi_j(x) + q(x) \phi_i(x) \phi_j(x)$ ,

On a donc :

$$a_\Omega(\phi_i, \phi_j) = \sum_{\ell=1}^L \int_{K_\ell} \theta(\phi_i, \phi_j) dx.$$

Pour  $r$  et  $s$  numéros locaux de l'élément  $K_\ell$ , on pose :

$$k_{r,s}^\ell = \int_{K_\ell} \theta(\phi_s, \phi_r) dx.$$

On va calculer  $k_{r,s}^\ell$  puis on calcule  $a_\Omega(\phi_i, \phi_j)$ , en effectuant un parcours sur les éléments, ce qui s'exprime par l'algorithme suivant :

Initialisation :  $\mathcal{K}_{ij} \leftarrow 0$ ,  $i = 1, \dots, M$ ,  $j \leq i$ .  
 Boucle sur les éléments  
 Pour  $\ell = 1$  à  $L$  faire  
 Pour  $r = 1$  à  $N_\ell$  faire

$i = \text{ng}(\ell, r)$  numéro global du noeud  $r$  de l'élément  $\ell$

Pour  $s = 1$  à  $r$  faire

calcul de  $k_{r,s}^\ell$

$j = \text{ng}(\ell, s)$

si  $i \geq j$

$$\mathcal{K}_{ij} \leftarrow \mathcal{K}_{ij} + k_{r,s}^\ell$$

sinon

$$\mathcal{K}_{ji} \leftarrow \mathcal{K}_{ji} + k_{r,s}^\ell$$

Fin pour

Fin pour

On a ainsi construit complètement la matrice de rigidité  $\mathcal{K}$ . Il reste à savoir comment calculer

$$k_{r,s}^\ell = \int_{\ell_e} \theta(\phi_s, \phi_r)(x) dx.$$

Ce calcul s'effectue sur l'élément de référence, et non sur les éléments  $K_\ell$ . On calcule ensuite la valeur de  $k_{r,s}^\ell$  par des changements de variable à l'aide de la transformation  $F_\ell$  (voir Figure 4.4 page 139). Notons :

$$F_\ell(\bar{x}, \bar{y}) = (x, y) = (a_0^\ell + a_1^\ell \bar{x} + a_2^\ell \bar{y}, b_0^\ell + b_1^\ell \bar{x} + b_2^\ell \bar{y}) \quad (4.19)$$

Notons que les coefficients  $a_i^\ell$  et  $b_i^\ell$  sont déterminés à partir des connaissances des coordonnées  $(x(i), y(i))$  où  $i = \text{ng}(\ell, r)$ . En effet, on peut déduire les coordonnées locales  $x(r), y(r)$ ,  $r = 1, N_\ell$ , des noeuds de l'élément  $\ell$ , à partir des coordonnées globales des noeuds  $(x(i), y(i))$ , et du tableau  $\text{ng}(\ell, r) = i$ . Sur l'élément courant  $K_\ell$ , le terme élémentaire  $k_{r,s}^\ell$  s'écrit donc

$$k_{r,s}^\ell = \int_{\ell} \theta(\phi_s(x, y), \phi_r(x, y)) dx dy.$$

Or,  $(x, y) = F_\ell(\bar{x}, \bar{y})$ ; donc par changement de variables, on a :

$$k_{r,s}^\ell = \int_{\bar{e}} \theta(\phi_s \circ F_\ell(\bar{x}, \bar{y}), \phi_r \circ F_\ell(\bar{x}, \bar{y})) \text{Jac}_{\bar{x}, \bar{y}}(F_\ell) d\bar{x} d\bar{y}$$

ou  $\text{Jac}_{\bar{x}, \bar{y}}(F_\ell)$  désigne le Jacobien de  $F_\ell$  en  $(\bar{x}, \bar{y})$ . Or,  $\phi_s \circ F_\ell = \bar{\phi}_s$ , et, puisque  $F_\ell$  est définie par (4.19), on a :

$$\text{Jac}(F_\ell) = \text{Det}(DF_\ell) = \begin{vmatrix} a_1^\ell & b_1^\ell \\ a_2^\ell & b_2^\ell \end{vmatrix} = |a_1^\ell b_2^\ell - a_2^\ell b_1^\ell|$$

donc  $k_{r,s}^\ell = \text{Jac}(F_\ell) \bar{k}_{r,s}$ , où

$$\bar{k}_{r,s} = \int_{\bar{e}} \theta(\bar{\phi}_s(\bar{x}, \bar{y}), \bar{\phi}_r(\bar{x}, \bar{y})) d\bar{x} d\bar{y}$$

Etudions maintenant ce qu'on obtient pour  $\bar{k}_{r,s}$  dans le cas du problème modèle (4.11), on a :

$$\bar{k}_{r,s} = \int_{\bar{e}} [p(\bar{x}, \bar{y}) \nabla \bar{\phi}_s(\bar{x}, \bar{y}) \nabla \bar{\phi}_r(\bar{x}, \bar{y}) + q(\bar{x}, \bar{y}) \bar{\phi}_s(\bar{x}, \bar{y}) \bar{\phi}_r(\bar{x}, \bar{y})] d\bar{x} d\bar{y}.$$

Les fonctions de base  $\bar{\phi}_s$  et  $\bar{\phi}_r$  sont connues; on peut donc calculer  $\bar{k}_{r,s}$  explicitement si  $p$  et  $q$  sont faciles à intégrer. Si les fonctions  $p$  et  $q$  ou les fonctions de base  $\bar{\phi}$ , sont plus compliquées, on calcule  $\bar{k}_{r,s}$  en effectuant une intégration numérique. Rappelons que le principe d'une intégration numérique est d'approcher l'intégrale d'une fonction continue donnée  $\psi$ ,

$$I = \int_{\bar{e}} \psi(\bar{x}, \bar{y}) d\bar{x} d\bar{y}, \text{ par } \tilde{I} = \sum_{i=1}^{NP_I} \omega_i(P_i) \psi(P_i),$$

où  $NP_I$  est le nombre de points d'intégration, notés  $P_i$ , qu'on appelle souvent points d'intégration de Gauss, et les coefficients  $\omega_i$  sont les poids associés. Notons que les points  $P_i$  et les poids  $\omega_i$  sont indépendants de  $\psi$ . Prenons par exemple, dans le cas unidimensionnel,  $\bar{K} = [0, 1]$ ,  $p_1 = 0$ ,  $p_2 = 1$ , et  $\omega_1 = \omega_2 = \frac{1}{2}$ . On approche alors

$$I = \int_0^1 \psi(x) dx \text{ par } \tilde{I} = \frac{1}{2}(\psi(0) + \psi(1)).$$

C'est la formule (bien connue) des trapèzes. Notons que dans le cadre d'une méthode, il est nécessaire de s'assurer que la méthode d'intégration numérique choisie soit suffisamment précise pour que :

1. le système  $\mathcal{K}\alpha = \mathcal{G}(N \times N)$  reste inversible,
2. l'ordre de convergence de la méthode reste le même.

Examinons maintenant des éléments en deux dimensions d'espace.

1. Élément fini  $P_1$  sur triangle Prenons  $NP_I = 1$  (on a donc un seul point de Gauss), choisissons  $p_1 = (\frac{1}{3}, \frac{1}{3})$ , le centre de gravité du triangle  $\bar{K}$ , et  $\omega_1 = 1$ . On approche alors

$$I = \int_{\bar{\ell}} \psi(\bar{x}) d\bar{x} \text{ par } \psi(p_1).$$

On vérifiera que cette intégration numérique est exacte pour les polynômes d'ordre 1 (exercice 57 page 167).

2.  $P_2$  sur triangles. On prend maintenant  $NP_I = 3$ , et on choisit comme points de Gauss :

$$p_1 = \left(\frac{1}{2}, 0\right), p_2 = \left(\frac{1}{2}, \frac{1}{2}\right), p_3 = \left(0, \frac{1}{2}\right)$$

et les poids d'intégration  $\omega_1 = \omega_2 = \omega_3 = \frac{1}{6}$ . On peut montrer que cette intégration numérique est exacte pour les polynômes d'ordre 2 (voir exercice 57 page 167).

Remarquons que, lors de l'intégration numérique du terme élémentaire

$$k_{r,s}^{\ell} = \int_{\bar{\ell}} [p(\bar{x}, \bar{y})(F_{\ell}(\bar{x}, \bar{y})) \nabla \bar{\phi}_r(\bar{x}, \bar{y}) \cdot \nabla \bar{\phi}_s(\bar{x}, \bar{y}) + q(\bar{x}, \bar{y})(F_{\ell}(\bar{x}, \bar{y})) \bar{\phi}_r(\bar{x}, \bar{y}) \bar{\phi}_s(\bar{x}, \bar{y})] d\bar{x} d\bar{y},$$

on approche  $k_{r,s}^{\ell}$  par

$$\bar{k}_{r,s} \simeq \sum_{i=1}^{NP_I} \omega_i [p(F_{\ell}(P_i)) \nabla \bar{\phi}_r(P_i) \cdot \nabla \bar{\phi}_s(P_i) + q(F_{\ell}(P_i)) \bar{\phi}_r(P_i) \bar{\phi}_s(P_i)].$$

Les valeurs  $\nabla \bar{\phi}_r(P_i)$ ,  $\nabla \bar{\phi}_s(P_i)$ ,  $\bar{\phi}_r(P_i)$  et  $\bar{\phi}_s(P_i)$  sont calculées une fois pour toutes, et dans la boucle sur  $\ell$ , il ne reste donc plus qu'à évaluer les fonctions  $p$  et  $q$  aux points  $F_{\ell}(P_i)$ . Donnons maintenant un résumé de la mise en oeuvre de la procédure d'intégration numérique (indépendante de  $\ell$ ). Les données de la procédure sont :

- les coefficients  $\omega_i, i = 1, \dots, NP_I$ ,
  - les coordonnées  $(xpg(i), ypg(i)), i = 1, \dots, NP_I$  des points de Gauss,
  - les valeurs de  $\phi_r, \frac{\partial \phi_r}{\partial x}$  et  $\frac{\partial \phi_r}{\partial y}$  aux points de Gauss, notées  $\phi(r, i), \phi_x(r, i)$  et  $\phi_y(r, i), r = 1 \dots N_{\ell}, i = 1, \dots, NP_I$ .
- Pour  $\ell$  donné, on cherche à calculer :

$$I = \int_{\bar{K}} p(F_{\ell}(\bar{x}, \bar{y})) \frac{\partial \phi_r}{\partial x}(\bar{x}, \bar{y}) \frac{\partial \phi_s}{\partial y}(\bar{x}, \bar{y}) d\bar{x} d\bar{y} + \int_e q(F_{\ell}(\bar{x}, \bar{y})) \phi_r(\bar{x}, \bar{y}) d\bar{x} d\bar{y} \phi_s(\bar{x}, \bar{y}).$$

On propose l'algorithme suivant :

**Initialisation** :  $I \leftarrow 0$

Pour  $i = 1$  à  $NP_I$ , faire :

$$p_i = p(F_e(P_i))$$

$$q_i = q(F_e(P_i))$$

$$I \leftarrow I + \omega_i (p_i \phi_x(r, i) \phi_y(s, i) + q_i \phi(r, i) \phi(s, i))$$

Fin pour

On procède de même pour le calcul du second membre

$$T_{\Omega}(\phi_i) = \int_{\Omega} f(x, y) \phi_i'(x, y) dx dy = \sum_{\ell=1}^L g_{\ell}, \text{ où } g_{\ell} = \int_{\ell_e} f(x, y) \phi_i(x, y) dx dy.$$

L'algorithme s'écrit :

**Initialisation de  $\mathcal{G}$  à 0** :  $\mathcal{G}_i \leftarrow 0 \quad i = 1$  à  $M$

Pour  $\ell = 1$  à  $L$

Pour  $r = 1$  à  $N_{\ell}$

$$\text{Calcul de } g_{\ell}^r = \int_{\ell_e} f(x, y) \phi_r(x, y) dx dy$$

$$i = \text{ng}(\ell, r)$$

$$\mathcal{G}_i \leftarrow \mathcal{G}_i + g_{\ell}^r$$

Fin pour

Fin pour

Il reste le calcul de  $g_\ell^r$  qui se ramène au calcul de l'élément de référence par changement de variable.

On a :

$$g_\ell^r = \int_{K_\ell} f(x, y) \phi_r(x, y) dx dy = \int_{\bar{K}} f \circ F_\ell(\bar{x}, \bar{y}) \bar{\phi}_r(\bar{x}, \bar{y}) \text{Jac}_{\bar{x}, \bar{y}}(F_\ell) d\bar{x} d\bar{y}.$$

L'intégration numérique est identique à celle effectuée pour  $\bar{k}_{r,s}$ .

#### 4.3.4 Calcul de $a_{\Gamma_1}$ et $T_{\Gamma_1}$ (contributions des arêtes de bord "Fourier").

Détaillons maintenant le calcul des contributions des bords où s'applique la condition de Fourier, c'est à dire  $a_{\Gamma_1}(\phi_i, \phi_j)$   $i = 1, \dots, M$ ,  $j = 1, \dots, M$  et  $T_{\Gamma_1}(\phi_i)$   $i = 1, \dots, M$ . Par définition,

$$a_{\Gamma_1}(\phi_i, \phi_j) = \int_{\Gamma_1} p(x) \nabla \phi_i(x) \cdot \nabla \phi_j(x) dx + \int_{\Gamma_1} q(x) \phi_i(x) \phi_j(x) dx.$$

Notons que  $a_{\Gamma_1}(\phi_i, \phi_j) = 0$  si  $\phi_i$  et  $\phi_j$  sont associées à des noeuds  $S_i, S_j$  de d'un élément sans arête commune avec les arêtes de la frontière. Soit  $L1$  le nombre d'arêtes  $\epsilon_k$ ,  $k = 1, \dots, L1$  du maillage incluses dans  $\Gamma_1$ . Rappelons que les noeuds soumis aux conditions de Fourier sont repertoriés dans un tableau  $CF$ , de dimensions  $(L1, 2)$ , qui donne les informations suivantes

1.  $CF(k, 1)$  contient le numéro  $\ell$  de l'élément  $K_\ell$  auquel appartient l'arête  $\epsilon_k$ .
2.  $CF(k, 2)$  contient le premier numéro des noeuds de l'arête  $\epsilon_k$  dans l'élément  $K_\ell$ . On suppose que la numérotation des noeuds locaux a été effectuée de manière "adroite", par exemple dans le sens trigonométrique. Dans ce cas,  $CF(k, 2)$  détermine tous les noeuds de l'arête  $\epsilon_k$  dans l'ordre, puisqu'on connaît le nombre de noeuds par arête et le sens de numérotation des noeuds. Donnons des exemples pour trois cas différents, représentés sur la figure 4.7.
  - (a) Dans le premier cas (à droite sur la figure), qui représente un élément fini  $P1$ , on a  $CF(k, 2) = 3$  et le noeud suivant sur l'arête est 1.
  - (b) Dans le second cas (au centre sur la figure), qui représente un élément fini  $P2$ , on a  $CF(k, 2) = 3$  et les noeuds suivants sur l'arête sont 4 et 5.
  - (c) Enfin dans l'élément  $P1$  "de coin" représenté à gauche sur la figure, on a  $CF(k, 1) = \ell$ ,  $CF(k', 1) = \ell$ ,  $CF(k, 2) = 1$ ,  $CF(k', 2) = 2$ .

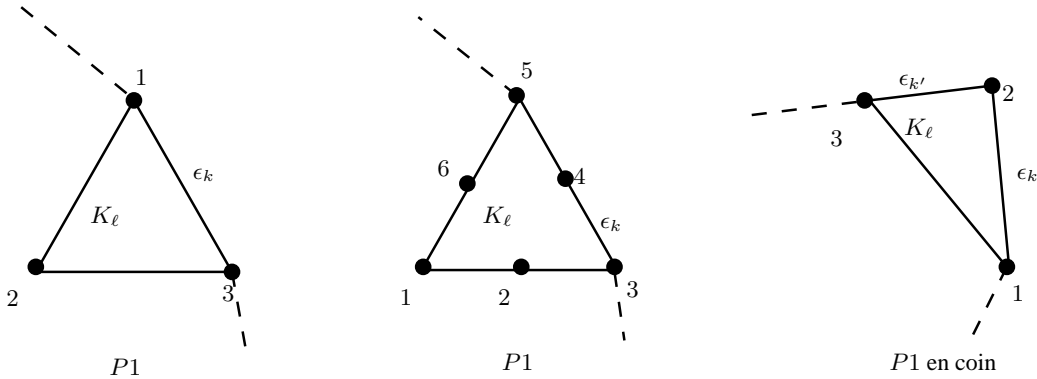


FIG. 4.7 – Exemples de numérotation d'arête du bord

Pour  $k = 1, \dots, L1$ , on note  $\hat{S}_k$  l'ensemble des noeuds locaux de  $\epsilon_k$ , donnés par  $CF(k, 2)$  en appliquant la règle *ad hoc* (par exemple le sens trigonométrique). On peut alors définir :

$$S_k = \{(r, s) \in (\hat{S}_k)^2 / r < s\}$$

L'algorithme de prise en compte des conditions de Fourier s'écrit alors :

Pour  $k = 1 \dots L1$

$\ell = CF(k, 1)$ .

Pour chaque  $(r, s) \in S_k$  faire

calcul de  $I_{rs}^\ell = \int_{C_k} p(x)\sigma(x)\phi_r^\ell(x)\phi_s^\ell(x)dx$  (éventuellement avec intégration numérique)

$i = \text{ng}(\ell, r)$

$j = \text{ng}(\ell, s)$

si  $j \leq i$

$\mathcal{K}_{ij} \leftarrow \mathcal{K}_{ij} + I_{rs}^\ell$

sinon

$\mathcal{K}_{ij} \leftarrow \mathcal{K}_{ji} + I_{rs}^\ell$

Fin si

Fin pour

Fin pour

Le calcul de  $I_{rs}^\ell$  s'effectue sur l'élément de référence (avec éventuellement intégration numérique). De même, on

a une procédure similaire pour le calcul de  $T_{\Gamma_1} = \int_{\Gamma_1} p(x)g_1(x)v(x)d\gamma(x)$ .

$$\mathcal{G}_i \leftarrow \mathcal{G}_i + \int_{\Gamma_2} p(x)g_1(x)\phi_i(x)d\gamma(x)$$

### 4.3.5 Prise en compte des noeuds liés dans le second membre

Après les calculs précédents, on a maintenant dans  $\mathcal{G}_i$  :

$$\mathcal{G}_i = \int_{\Omega} f(x)\phi_i(x)dx + \int_{\Gamma_1} p(x)g_1(x)\phi_i(x)d\gamma(x)$$

Il faut maintenant retirer du second membre, les combinaisons venant des noeuds liés :

$$\mathcal{G}_i \leftarrow \mathcal{G}_i - \sum_{j \in J_0} g_0(S_j)a(\phi_j, \phi_i)$$

où  $J_0$  est l'ensemble des indices des noeuds liés. On utilise pour cela le tableau  $CD$  qui donne les conditions, de Dirichlet, de dimension  $M_0$  où  $M_0 = \text{card}J_0$ . Pour  $i_0 = 1, \dots, M_0$ ,  $CD(i_0) = j_0 \in J_0$  est le numéro du noeud lié dans la numérotation globale. La procédure est donc la suivante.

Pour  $i_0 = 1, \dots, M_0$ , faire

$j = CD(i_0)$

$a = g_0(S_j)$

si  $(i \leq j)$

$\mathcal{G}_i \leftarrow \mathcal{G}_i - a\mathcal{K}_{ij}$

sinon

$\mathcal{G}_i \leftarrow$

$\mathcal{G}_i - a\mathcal{K}_{ji}$  sinon

Fin si

Fin pour

### 4.3.6 Stockage de la matrice $\mathcal{K}$

Remarquons que la matrice  $\mathcal{K}$  est creuse (et même très creuse), en effet  $a(\phi_j, \phi_i) = 0$  dès que

$$\text{supp}(\phi_i) \cap \text{supp}(\phi_j) = \emptyset$$

Examinons une possibilité de stockage de la matrice  $\mathcal{K}$ . Soit  $NK$  le nombre d'éléments non nuls de la matrice  $\mathcal{K}$ . On peut stocker la matrice dans un seul tableau  $KMAT$  en mettant bout à bout les coefficients non nuls de la première ligne, puis ceux de la deuxième ligne, etc... jusqu'à ceux de la dernière ligne. Pour repérer les éléments de  $\mathcal{K}$  dans le tableau  $KMAT$ , on a alors besoin de pointeurs. Le premier pointeur, nommé,  $IC$  est de dimension  $NK$ . La valeur de  $IC(k)$  est le numéro de la colonne de  $K(k)$ . On introduit alors le pointeur  $IL(\ell)$ ,  $\ell = 1, \dots, NL$ , où  $NL$  est le nombre de lignes, où  $IL(\ell)$  est l'indice dans  $KMAT$  du début de la  $\ell$ -ième ligne. L'identification entre  $KMAT$  et  $\mathcal{K}$  se fait alors par la procédure suivante :

Pour  $k = 1 \dots NK$

si  $IL(m) \leq k < IL(m+1)$  alors

$KMAT(k) = \mathcal{K}_{m, IC(k)}$

Fin si

Fin pour

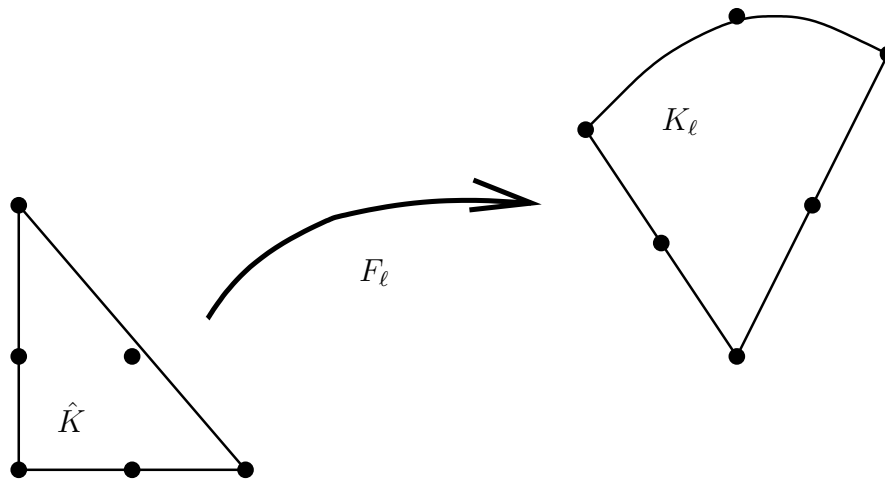


FIG. 4.8 – Transformation isoparamétrique

La matrice  $\mathcal{K}$  est symétrique définie positive, on peut donc utiliser une méthode de type gradient conjugué préconditionné (voir cours de Licence). Notons que la structure de la matrice dépend de la numérotation des noeuds. Il est donc important d'utiliser des algorithmes performants de maillage et de numérotation.

## 4.4 Eléments finis isoparamétriques

Dans le cas où  $\Omega$  est polygonal, si on utilise des éléments finis de type  $P_2$ , les noeuds de la frontière sont effectivement sur la frontière même si on les calcule à partir de l'élément fini de référence. Par contre, si le bord est courbe, ce n'est plus vrai. L'utilisation d'éléments finis "isoparamétriques" va permettre de faire en sorte que tous les noeuds frontières soient effectivement sur le bord, comme sur la figure 4.8. Pour obtenir une transformation isoparamétrique, on définit

$$F_\ell : K \rightarrow K_\ell$$

$$(\bar{x}, \bar{y}) \mapsto (x, y)$$

à partir des fonctions de base de l'élément fini de référence :

$$x = \sum_{r=1}^{N_\ell} x_r \bar{\phi}_r(\bar{x}, \bar{y}), \quad y = \sum_{r=1}^{N_\ell} y_r \bar{\phi}_r(\bar{x}, \bar{y}),$$

où  $N_\ell$  est le nombre de noeuds de l'élément et  $(x_r, y_r)$  sont les coordonnées du  $r$ -ième noeud de  $K_\ell$ . Remarquons que la transformation  $F_\ell$  isoparamétrique  $P_1$  est identique à celle des éléments finis classiques. Par contre, la transformation isoparamétrique  $P_2$  n'est plus affine, alors qu'elle l'est en éléments finis classiques. Notons que les fonctions de base locales vérifient toujours

$$\phi_r^\ell \circ F_\ell = \phi_r, \quad \forall \ell = 1, \dots, L, \quad \forall r = 1, \dots, N_\ell.$$

On peut alors se poser le problème de l'inversibilité de  $F_\ell$ . On ne peut pas malheureusement démontrer que  $F_\ell$  est inversible dans tous les cas, toutefois, cela s'avère être le cas dans la plupart des cas pratiques. L'intérêt de la transformation isoparamétrique est de pouvoir traiter les bords courbes, ainsi que les éléments finis  $Q1$  sur quadrilatères. Notons que le calcul de  $\phi_r^\ell$  est toujours inutile, car on se ramène encore à l'élément de référence.

## 4.5 Analyse d'erreur

### 4.5.1 Erreurs de discrétisation et d'interpolation

On considère toujours le problème modèle (4.11) page 147 sur lequel on a étudié la mise en oeuvre de la méthode des éléments finis. On rappelle que la formulation faible de ce problème est donnée en (4.14) page 148, et que

sous les hypothèses (4.12) page 147, le problème (4.14) admet une unique solution  $\tilde{u} \in H = H_{\Gamma_0}^1 = \{u \in H^1(\Omega); u = 0 \text{ sur } \Gamma_0\}$ . La méthode d'approximation variationnelle du problème (4.14) consiste à chercher  $\tilde{u}_N \in H_N = Vect\{\phi_1, \dots, \phi_N\}$  solution de (4.16) page 148, où les fonctions  $\phi_1, \dots, \phi_N$  sont les fonctions de base éléments finis associés aux noeuds  $x_1, \dots, x_N$ . Comme les hypothèses (3.21) page 106 sont vérifiées, l'estimation (3.29) page 109 entre  $\tilde{u}$  solution de (4.14) et  $\tilde{u}^{(N)}$  solution de (4.16) est donc vérifiée. On a donc :

$$\|\tilde{u} - \tilde{u}_N\|_{H^1} \leq \sqrt{\frac{M}{\alpha}} d(\tilde{u}, H_N),$$

où  $M$  (resp.  $\alpha$ ) est la constante de continuité (resp. de coercivité) de  $a$ . Comme  $u = u_0 + \tilde{u}$ , on a, en posant  $c = \sqrt{\frac{M}{\alpha}}$ ,

$$\|u - u_N\| \leq C \|u - w\| \forall w \in H_N, \quad (4.20)$$

où  $u_N = \tilde{u}_N + u_0$ . Notons que dans l'implantation pratique de la méthode d'éléments finis, lorsqu'on calcule  $T(v) = T(v) - a(u_0, v)$ , on remplace  $u_0$  par  $u_{0,N} \in H_N$ , donc on commet une légère erreur sur  $T$ . De plus, on calcule  $a(\phi_i, \phi_j)$  à l'aide d'intégrations numériques : l'inégalité (4.20) n'est donc vérifiée en pratique que de manière approchée. On supposera cependant, dans la suite de ce paragraphe, que les erreurs commises sont négligeables et que l'inégalité (4.20) est bien vérifiée. De la même manière qu'on a défini l'interpolée sur un élément  $K$ , (voir définition 4.3 page 138, on va maintenant définir l'interpolée sur  $H^1(\Omega)$  tout entier, de manière à établir une majoration de l'erreur de discrétisation grâce à (4.20).

**Définition 4.16 (Interpolée dans  $H_N$ )** . Soit  $u \in H^1(\Omega)$  et  $H_N = Vect\{\phi_1, \dots, \phi_N\}$  où les fonctions  $\phi_1 \dots \phi_N$  sont des fonctions de base éléments finis associées aux noeuds  $S_1 \dots S_N$  d'un maillage éléments finis de  $\Omega$ . Alors on définit l'interpolée de  $u$  dans  $H_N$ ,  $u_I \in H_N$  par :

$$u_I = \sum_{i=1}^N u(S_i) \phi_i.$$

Comme  $u_I \in H_N$ , on peut prendre  $W = u_I$  dans (4.20), ce qui fournit un majorant de l'erreur de discrétisation :

$$\|u - u_N\|_{H^1} \leq C \|u - u_I\|_{H^1}$$

On appelle erreur d'interpolation le terme  $\|u - u_I\|_{H^1}$

## 4.5.2 Erreur d'interpolation en dimension 1

Soit  $\Omega = ]0, 1[$ , on considère un maillage classique, défini par les  $N + 2$  points  $(x_i)_{i=0 \dots N+1}$ , avec  $x_0 = 0$  et  $x_{N+1} = 1$ , et on note

$$h_i = x_{i+1} - x_i, \quad i = 0, \dots, N + 1, \quad \text{et } h = \max\{h_i, i = 0, \dots, N + 1\}$$

On va montrer que si  $u \in H^2(]0, 1[)$ , alors on peut obtenir une majoration de l'erreur d'interpolation  $\|u - u_I\|_{H^1}$ . On admettra le lemme suivant (voir exercice 33 page 118) :

**Lemme 4.17** Si  $u \in H^1(]0, 1[)$  alors  $u$  est continue.

En particulier, on a donc  $H^2(]0, 1[) \subset C^1([0, 1])$ . Remarquons que ce résultat est lié à la dimension 1, voir injection de Sobolev, cours d'analyse fonctionnelle ou [1]. On va démontrer le résultat suivant sur l'erreur d'interpolation.

**Théorème 4.18 (majoration de l'erreur d'interpolation, dimension 1)** Soit  $u \in H^2(]0, 1[)$ , et soit  $u_I$  son interpolée sur  $H_N = Vect\{\phi_i, i = 1, \dots, N\}$ , où  $\phi_i$  désigne la  $i$ -ème fonction de base élément fini P1 associée au noeud  $x_i$  d'un maillage élément fini de  $]0, 1[$ . Alors il existe  $c \in \mathbb{R}$  ne dépendant que de  $u$ , tel que

$$\|u - u_I\|_{H^1} \leq Ch. \quad (4.21)$$

**Démonstration :** On veut estimer

$$\|u - u_I\|_{H^1}^2 = |u - u_I|_0^2 + |u - u_I|_1^2$$

où  $|v|_0 = \|v\|_{L^2}$  et  $|v|_1 = \|Dv\|_{L^2}$ . Calculons  $|u - u_I|_1^2$  :

$$|u - u_I|_1^2 = \int_0^1 |u' - u'_I|^2 dx = \sum_{i=0}^N \int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx.$$

Or pour  $x \in ]x_i, x_{i+1}[$  on a

$$u'_I = \frac{u(x_{i+1}) - u(x_i)}{h_i} = u'(\xi_i),$$

pour un certain  $\xi_i \in ]x_i, x_{i+1}[$ . On a donc :

$$\int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx = \int_{x_i}^{x_{i+1}} |u'(x) - u'(\xi_i)|^2 dx.$$

On en déduit que :

$$\int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx = \int_{x_i}^{x_{i+1}} \left| \int_{\xi_i}^x u''(t) dt \right|^2 dx,$$

et donc, par l'inégalité de Cauchy-Schwarz,

$$\begin{aligned} \int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx &\leq \int_{x_i}^{x_{i+1}} \int_{\xi_i}^x |u''(t)|^2 dt |x - \xi_i| dx \\ &\leq h_i \int_{x_i}^{x_{i+1}} \left( \int_{\xi_i}^x |u''(t)|^2 dt \right) dx, \end{aligned}$$

car  $|x - \xi_i| \leq h_i$ . En réappliquant l'inégalité de Cauchy-Schwarz, on obtient :

$$\int_{x_i}^{x_{i+1}} |u'(x) - u'_I(x)|^2 dx \leq h_i^2 \int_{x_i}^{x_{i+1}} |u''(t)|^2 dt$$

En sommant sur  $i$ , ceci entraîne :

$$|u - u_I|_1^2 \leq h^2 \int_0^1 |u''(t)|^2 dt. \quad (4.22)$$

Il reste maintenant à majorer  $|u - u_I|_0^2 = \int_0^1 |u - u_I|^2 dx$ . Pour  $x \in [x_i, x_{i+1}]$

$$|u(x) - u_I(x)|^2 = \left( \int_{x_i}^x (u'(t) - u'_I(t)) dt \right)^2.$$

Par l'inégalité de Cauchy-Schwarz, on a donc :

$$|u(x) - u_I(x)|^2 \leq \int_{x_i}^x (u'(t) - u'_I(t))^2 dt \underbrace{|x - x_i|}_{\leq h_i}$$

Par des calculs similaires aux précédents, on obtient donc :

$$\begin{aligned} |u(x) - u_I(x)|^2 &\leq \int_{x_i}^x h_i \left( \int_{x_i}^{x_{i+1}} |u''(t)|^2 dt \right) dx h_i \\ &\leq h_i^3 \int_{x_i}^{x_{i+1}} |u''(t)|^2 dt. \end{aligned}$$

En intégrant sur  $[x_i, x_{i+1}]$ , il vient :

$$\int_{x_i}^{x_{i+1}} |u(x) - u_I(x)|^2 dx \leq h_i^4 \int_{x_i}^{x_{i+1}} |u''(t)|^2 dt,$$

et en sommant sur  $i = 1, \dots, N$  :

$$\int_0^1 (u(x) - u_I(x))^2 dx \leq h^4 \int_0^1 |u''(t)|^2 dt.$$



On a donc :

$$|u - u_I|_0 \leq h^2 |u|_2$$

ce qui entraîne, avec (4.22) :

$$\begin{aligned} \|u - u_I\|^2 &\leq h^4 |u|_2^2 + h^2 |u|_2^2 \\ &\leq (1 + h^2) h^2 |u|_2^2 \end{aligned}$$

On en déduit le résultat annoncé. ■

On en déduit le résultat d'estimation d'erreur suivant :

**Corollaire 4.19 (Estimation d'erreur, P1, dimension 1)** Soit  $\Omega$  un ouvert polygonal convexe de  $\mathbb{R}^d$ ,  $d \geq 1$  ; soit  $f \in L^2(\Omega)$  et  $u \in H_0^1(\Omega)$  l'unique solution du problème

$$\begin{cases} u \in H_0^1(\Omega) \\ a(u, v) = \int_{\Omega} \nabla u(x) \nabla v(x) dx = \int_{\Omega} f(x) v(x) dx, \end{cases},$$

et  $u_{\mathcal{T}}$  l'approximation éléments finis P1 obtenue sur un maillage admissible  $\mathcal{T}$  de pas  $h_{\mathcal{T}} = \max_{i=1, \dots, N} \{ |h_i| \}$ . Alors il existe  $C \in \mathbb{R}$  ne dépendant que de  $\Omega$  et  $f$  tel que  $\|u - u_{\mathcal{T}}\| < Ch$ .

Ces résultats se généralisent au cas de plusieurs dimensions d'espace (voir Ciarlet), sous des conditions géométriques sur le maillage, nécessaires pour obtenir le résultat d'interpolation. Par exemple pour un maillage triangulaire en deux dimensions d'espace, intervient une condition d'angle : on demande que la famille de maillages considérée soit telle qu'il existe  $\beta > 0$  tel que  $\beta \leq \theta \leq \pi - \beta$  pour tout angle  $\theta$  du maillage.

**Remarque 4.20 (Sur les techniques d'estimation d'erreur)** Lorsqu'on a voulu montrer des estimations d'erreur pour la méthode des différences finies, on a utilisé le principe de positivité, la consistance et la stabilité en norme  $L^\infty$ . En volumes finis et éléments finis, on n'utilise pas le principe de positivité. En volumes finis, la stabilité en norme  $L^2$  est obtenue grâce à l'inégalité de Poincaré discrète, et la consistance est en fait la consistance des flux. Notons qu'en volumes finis on se sert aussi de la conservativité des flux numériques pour la preuve de convergence. Enfin, en éléments finis, la stabilité est obtenue grâce à la coercivité de la forme bilinéaire, et la consistance provient du contrôle de l'erreur d'interpolation.

Même si le principe de positivité n'est pas explicitement utilisé pour les preuves de convergence des éléments finis et volumes finis, il est toutefois intéressant de voir à quelles conditions ce principe est respecté, car il est parfois très important en pratique.

Reprenons d'abord le cas du schéma volumes finis sur un maillage  $\mathcal{T}$  admissible pour la discrétisation de l'équation (3.1).

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

Rappelons que le schéma volumes finis s'écrit :

$$\sum_{K \in \mathcal{C}} \left( \sum_{\sigma \in \xi_K \cap \xi_{int}} \tau_{K,L} (u_K - u_L) + \sum_{\sigma \in \xi_K \cap \xi_{ext}} \tau_{K,\sigma} u_K \right) = |K| f_K, \quad (4.23)$$

avec

$$\tau_{K,L} = \frac{|K| |L|}{d(x_K, x_L)} \text{ et } \tau_{K,\sigma} = \frac{|\sigma|}{d(x_K, \partial\Omega)},$$

où  $|K|$ , (resp.  $|\sigma|$ ) désigne la mesure de Lebesgue en dimension  $d$  (resp.  $d - 1$ ) de  $K$  (resp.  $\sigma$ ).

Notons que les coefficients  $\tau_{K,L}$  et  $\tau_{K,\sigma}$  sont positifs, grâce au fait que le maillage est admissible (et donc  $X_K \vec{X}_L = d(X_K, X_L) n_{\vec{K}L}$ , où  $X_K \vec{X}_L$  désigne le vecteur d'extrémités  $X_K$  et  $X_L$  et  $n_{\vec{K}L}$  la normale unitaire à  $K|L$  sortante de  $K$ ).

Notons que le schéma (4.23) s'écrit comme une somme de termes d'échange entre les mailles  $K$  et  $L$ , avec des coefficients  $\tau_{KL}$  positifs. C'est grâce à cette propriété que l'on montre facilement que le principe de positivité est vérifié. Considérons maintenant la méthode des éléments finis P1, pour la résolution du problème (3.1) sur maillage triangulaire. On sait (voir par exemple Ciarlet) que si le maillage satisfait la condition faible de Delaunay (qui stipule que la somme de deux angles opposés à une même arête doit être inférieure à  $\pi$ ), alors le principe du maximum est vérifié. Ce résultat peut se retrouver en écrivant le schéma éléments finis sous la forme d'un schéma volumes finis.

### 4.5.3 Super convergence

On considère ici un ouvert  $\Omega$  polygonal convexe de  $\mathbb{R}^d$ ,  $d \geq 1$ , et on suppose que  $f \in L^2(\Omega)$ . On s'intéresse à l'approximation par éléments finis  $P1$  de la solution  $u \in H_0^1(\Omega)$  du problème (3.5). On a vu dans le paragraphe précédent (corollaire 4.19) qu'on peut estimer l'erreur en norme  $L^2$  entre la solution exacte  $u$  et la solution approchée par éléments finis  $P1$ ; en effet, comme l'erreur d'interpolation est d'ordre  $h$ , on déduit une estimation sur l'erreur de discrétisation, également d'ordre  $h$ . En fait, si la solution  $u$  de (3.1) est dans  $H^2$ , il se produit un "petit miracle", car on peut montrer grâce à une technique astucieuse, dite "truc d'Aubin-Nitsche", que l'erreur de discrétisation en norme  $L^2$  est en fait d'ordre 2.

**Théorème 4.21 (Super convergence des éléments finis  $P1$ )** *Soit  $\Omega$  un ouvert polygonal convexe de  $\mathbb{R}^d$ ,  $d \geq 1$ ; soit  $f \in L^2(\Omega)$ ,  $u$  solution de (3.5),  $u_{\mathcal{T}}$  la solution approchée obtenue par éléments finis  $P1$ , sur un maillage éléments finis  $\mathcal{T}$ . Soit*

$$h_{\mathcal{T}} = \max_{K \in \mathcal{T}} \text{diam} K.$$

Alors il existe  $C \in \mathbb{R}$  ne dépendant que de  $\Omega$  et  $f$  tel que :

$$\|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq Ch \text{ et } \|u - u_{\mathcal{T}}\|_{L^2(\Omega)} \leq Ch^2.$$

**Démonstration :** Par le théorème de régularité 3.9 page 101, il existe  $C_1 \in \mathbb{R}_+$  ne dépendant que de  $\Omega$  tel que

$$\|u\|_{H^2(\Omega)} \leq C_1 \|f\|_{L^2(\Omega)}.$$

Grâce à ce résultat, on a obtenu (voir le théorème 4.19) qu'il existe  $C_2$  ne dépendant que de  $\Omega$ ,  $\beta$  et tel que

$$\|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq C_2 \|f\|_{L^2} h$$

Soit maintenant  $e_{\mathcal{T}} = u - u_{\mathcal{T}}$  et  $\varphi \in H_0^1(\Omega)$  vérifiant

$$\int_{\Omega} \nabla \varphi(x) \cdot \nabla \psi(x) dx = \int_{\Omega} e_{\mathcal{T}}(x) \psi(x) dx, \forall \psi \in H_0^1(\Omega). \quad (4.24)$$

On peut aussi dire que  $\varphi$  est la solution faible du problème

$$\begin{cases} -\Delta \varphi = e_{\mathcal{T}} \text{ dans } \Omega \\ \varphi = 0 \text{ sur } \partial\Omega. \end{cases}$$

Comme  $e \in L^2(\Omega)$ , par le théorème 3.9, il existe  $C_3 \in \mathbb{R}_+$  ne dépendant que  $\Omega$  tel que

$$\|\varphi\|_{H^2(\Omega)} \leq C_3 \|e_{\mathcal{T}}\|_{L^2(\Omega)}.$$

Or  $\|e_{\mathcal{T}}\|_{L^2(\Omega)}^2 = \int_{\Omega} e_{\mathcal{T}}(x) e_{\mathcal{T}}(x) dx = \int_{\Omega} \nabla \varphi(x) \cdot \nabla e(x) dx$ , en prenant  $\psi = e_{\mathcal{T}}$  dans (4.24).

Soit  $\varphi_{\mathcal{T}}$  la solution approchée par éléments finis  $P1$  du problème (4.24), c.à.d solution de :

$$\begin{cases} \varphi_{\mathcal{T}} \in V_{\mathcal{T},0} = \{v \in C(\bar{\Omega}); v|_K \in P_1, \forall K \in \mathcal{T}, v|_{\partial\Omega} = 0\} \\ \int_{\Omega} \nabla \varphi_{\mathcal{T}}(x) \cdot \nabla v(x) dx = \int_{\Omega} e(x) v(x) dx, \forall v \in V_{\mathcal{T},0} \end{cases} \quad (4.25)$$

On sait que  $u_{\mathcal{T}}$  vérifie :

$$\int_{\Omega} \nabla \varphi_{\mathcal{T}}(x) \cdot \nabla (u - u_{\mathcal{T}})(x) dx = 0;$$

on peut donc écrire que :

$$\|e_{\mathcal{T}}\|_{L^2(\Omega)}^2 = \int_{\Omega} \nabla (\varphi - \varphi_{\mathcal{T}})(x) \cdot \nabla (u - u_{\mathcal{T}})(x) dx \geq \|\varphi - \varphi_{\mathcal{T}}\|_{H^1(\Omega)} \|u - u_{\mathcal{T}}\|_{H^1(\Omega)}$$

D'après le théorème 4.19, on a :

$$\|\varphi - \varphi_{\mathcal{T}}\|_{H^1(\Omega)} \leq C_2 \|e\|_{L^2(\Omega)} h_{\mathcal{T}} \text{ et } \|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq C_2 \|f\|_{L^2(\Omega)} h_{\mathcal{T}}.$$

On en déduit que :

$$\|e_{\mathcal{T}}\|_{L^2(\Omega)} \leq C_2^2 \|f\|_{L^2(\Omega)} h_{\mathcal{T}}^2.$$

Ce qui démontre le théorème.

### 4.5.4 Traitement des singularités

Les estimations d'erreur obtenues au paragraphe précédent reposent sur la régularité  $H^2$  de  $u$ . Que se passe-t-il si cette régularité n'est plus vérifiée ? Par exemple, si le domaine  $\Omega$  possède un coin rentrant, on sait que dans ce cas, la solution  $u$  du problème (3.5) n'est plus dans  $H^2(\Omega)$ , mais dans un espace  $H^{1+s}(\Omega)$ , où  $s$  dépend de l'angle du coin rentrant. Considérons donc pour fixer les idées le problème

$$\begin{cases} -\Delta u = f \text{ dans } \Omega, \text{ où } \Omega \text{ est un ouvert} \\ u = 0 \text{ sur } \partial\Omega \text{ polygônale avec un coin rentrant.} \end{cases}$$

Pour approcher correctement la singularité, on peut raffiner le maillage dans le voisinage du coin. On peut également, lorsque cela est possible, modifier l'espace d'approximation pour tenir compte de la singularité. Dans le cas d'un polygone avec un coin rentrant par exemple, on sait trouver  $\psi \in H_0^1(\Omega)$  (et  $\psi \notin H^2(\Omega)$ ) telle que si  $u$  est solution de (3.5) avec  $f \in L^2(\Omega)$ , alors il existe un unique  $\alpha \in \mathbb{R}$  tel que  $u - \alpha\psi \in H^2(\Omega)$ .

Examinons le cas d'une approximation par éléments finis de Lagrange. Dans le cas où  $u$  est régulière, l'espace d'approximation est

$$V_{\mathcal{T}} = \text{Vect}\{\phi_i, i = 1, N_{\mathcal{T}}\},$$

où  $N_{\mathcal{T}}$  est le nombre de noeuds internes du maillage  $\mathcal{T}$  de  $\Omega$  considéré et  $(\phi_i)_{i=1, N_{\mathcal{T}}}$  la famille des fonctions de forme associées aux noeuds.

Dans le cas d'une singularité portée par la fonction  $\psi$  introduite ci-dessus, on modifie l'espace  $V$  et on prend maintenant :  $V_{\mathcal{T}} = \text{Vect}\{\phi_i, i = 1, N_{\mathcal{T}}\} \oplus \mathbb{R}\psi$  Notons que  $V_{\mathcal{T}} \subset H_0^1(\Omega)$ , car  $\psi \in H_0^1(\Omega)$ . Reprenons maintenant l'estimation d'erreur. Grâce au lemme de Céa, on a toujours

$$\|u - u_{\mathcal{T}}\|_{H^1} \leq \frac{M}{\alpha} \|u - w\|_{H^1(\Omega)}, \forall w \in V_{\mathcal{T}}.$$

On a donc également :

$$\|u - u_{\mathcal{T}}\|_{H^1} \leq \frac{M}{\alpha} \|u - \alpha\psi - w\|_{H^1(\Omega)}, \forall w \in V_{\mathcal{T}}.$$

puisque  $\alpha\psi + w \in V_{\mathcal{T}}$ . Or,  $u - \alpha\psi = \tilde{u} \in H^2(\Omega)$ .

Donc  $\|u - u_{\mathcal{T}}\|_{H^1} \leq \frac{M}{\alpha} \|\tilde{u} - w\|_{H^1(\Omega)}, \forall w \in V_{\mathcal{T}}$ .

Et grâce aux résultats d'interpolation qu'on a admis, si on note  $\tilde{u}_I$  l'interpolée de  $\tilde{u}$  dans  $V_{\mathcal{T}}$ , on a :

$$\|u - u_{\mathcal{T}}\|_{H^1(\Omega)} \leq \frac{M}{\alpha} \|\tilde{u} - \tilde{u}_I\|_{H^1(\Omega)} \leq \frac{M}{\alpha} C_2 \|\tilde{u}\|_{H^2(\Omega)} h.$$

On obtient donc encore une estimation d'erreur en  $h$ .

Examinons maintenant le système linéaire obtenu avec cette nouvelle approximation. On effectue un développement de Galerkin sur la base de  $V_{\mathcal{T}}$ . On pose

$$u_{\mathcal{T}} = \sum_{i=1, N_{\mathcal{T}}} u_i \phi_i + \gamma \psi.$$

Le problème discrétisé revient donc à chercher

$$\begin{cases} (u_s)_{i=1, N_{\mathcal{T}}} \subset \mathbb{R}^{N_{\mathcal{T}}} \text{ et } \gamma \in \mathbb{R} \quad t.q. \\ \sum_{j=1, N_{\mathcal{T}}} u_j \int_{\Omega} \nabla \phi_j(x) \cdot \nabla \phi_i(x) dx + \gamma \int_{\Omega} \nabla \psi(x) \cdot \nabla \phi_i(x) dx = \int_{\Omega} f(x) \phi_i(x) dx, \forall i = 1, N_{\mathcal{T}} \\ \sum_{j=1, N_{\mathcal{T}}} u_j \int_{\Omega} \nabla \phi_j(x) \cdot \nabla \psi(x) dx + \gamma \int_{\Omega} \nabla \psi(x) \cdot \nabla \psi(x) dx = \int_{\Omega} f(x) \psi(x) dx. \end{cases}$$

On obtient donc un système linéaire de  $N_{\mathcal{T}} + 1$  équations à  $N_{\mathcal{T}} + 1$  inconnues.

## 4.6 Exercices

**Exercice 48 (Eléments finis P1 pour le problème de Dirichlet)** *Corrigé en page 168*

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse au problème suivant :

$$\begin{aligned} -u''(x) &= f(x), \quad x \in ]0, 1[, \\ u(0) &= 0, \quad u(1) = 0. \end{aligned}$$

dont on a étudié une formulation faible à l'exercice 35.

Soient  $N \in \mathbb{N}$ ,  $h = 1/(N + 1)$  et  $x_i = ih$ , pour  $i = 0, \dots, N + 1$ , et  $K_i = [x_i, x_{i+1}]$ , pour  $i = 0, \dots, N$ . Soit  $H_N = \{v \in C([0, 1], \mathbb{R}) \text{ t.q. } v|_{K_i} \in P_1, i = 0, \dots, N, \text{ et } v(0) = v(1) = 0\}$ , où  $P_1$  désigne l'ensemble des polynômes de degré inférieur ou égal à 1.

1. Montrer que  $H_N \subset H_0^1$ .

2. Pour  $i = 1, \dots, N$ , on pose :

$$\phi_i(x) = \begin{cases} 1 - \frac{|x - x_i|}{h} & \text{si } x \in K_i \cup K_{i-1}, \\ 0 & \text{sinon.} \end{cases}$$

Montrer que  $\phi_i \in H_N$  pour tout  $i = 1, \dots, N$  et que  $H_N$  est engendré par la famille  $\{\phi_1, \dots, \phi_N\}$ .

3. Donner le système linéaire obtenu en remplaçant  $H$  par  $H_N$  dans la formulation faible. Comparer avec le schéma obtenu par différences finies.

#### Exercice 49 (Conditions aux limites de Fourier et Neumann) *Corrigé en page 169*

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse au problème :

$$\begin{aligned} -u''(x) + u(x) &= f(x), \quad x \in ]0, 1[, \\ u'(0) - u(0) &= 0, \quad u'(1) = -1. \end{aligned} \tag{4.26}$$

L'existence et l'unicité d'une solution faible ont été démontrées à l'exercice 49 page 162. On s'intéresse maintenant à la discrétisation de (4.26).

e (3.40) 1. Ecrire une discrétisation d par différences finies pour un maillage non uniforme. Ecrire le système linéaire obtenu.

2. Ecrire une discrétisation de (4.26) par volumes finis pour un maillage non uniforme. Ecrire le système linéaire obtenu.

3. Ecrire une discrétisation par éléments finis conformes de type Lagrange  $P_1$  de (4.26) pour un maillage non uniforme. Ecrire le système linéaire obtenu.

#### Exercice 50 (Conditions aux limites de Fourier et Neumann, bis)

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse au problème : :

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

1. Ecrire une discrétisation par éléments finis conformes de type Lagrange  $P_1$  pour un maillage uniforme. Ecrire le système linéaire obtenu.

2. Ecrire une discrétisation par volumes finis centrés pour un maillage uniforme. Ecrire le système linéaire obtenu.

3. Ecrire une discrétisation par différences finies centrés de pour un maillage uniforme. Ecrire le système linéaire obtenu.

4. Quel est l'ordre de convergence de chacune des méthodes étudiées aux questions précédentes ?

#### Exercice 51 (Eléments finis pour un problème avec conditions mixtes)

Soit  $f \in L^2(]0, 1[)$ . On s'intéresse ici au problème

$$\begin{aligned} -u''(x) + u(x) &= f(x), \quad x \in ]0, 1[, \\ u(0) &= 0, \\ u'(1) &= 0, \end{aligned} \tag{4.27}$$

Ce problème est un cas particulier du problème (3.43) étudié à l'exercice 41 page 119, en prenant :  $\Omega = ]0, 1[$ ,  $p \equiv 1$  et  $q \equiv 1$ ,  $\Gamma_0 = \{0\}$ ,  $\Gamma_1 = \{1\}$ ,  $g_0 \equiv 0$ ,  $g_1 \equiv 0$  et  $\sigma = 0$ .

On s'intéresse ici à la discrétisation du problème (4.27). Soient  $N \in \mathbb{N}$ ,  $h = 1/(N+1)$  et  $x_i = ih$ , pour  $i = 0, \dots, N+1$ , et  $K_i = [x_i, x_{i+1}]$ , pour  $i = 0, \dots, N$ . On cherche une solution approchée de (4.27), notée  $u_h$ , en utilisant les éléments finis  $(K_i, \{x_i, x_{i+1}\}, P_1)_{i=0}^N$ .

1. Déterminer l'espace d'approximation  $V_h$ . Montrer que les fonctions de base globales sont les fonctions  $\Phi_i$  de  $[0, 1]$  dans  $\mathbb{R}$  définies par  $\Phi_i(x) = (1 - \frac{|x-x_i|}{h})^+$ , pour  $i = 1, \dots, N+1$ .

2. Construire le système linéaire à résoudre et comparer avec les systèmes obtenus par différences finies et volumes finis.

3 A-t-on  $u'_h(1) = 0$  ?

### Exercice 52 (Éléments finis pour un problème de réaction-diffusion)

Soit  $\alpha$  un réel positif ou nul, et  $f$  est une fonction continue. On considère le problème suivant

$$\begin{aligned} -u''(x) + \alpha u(x) &= f(x), \quad x \in ]0, 1[, \\ u'(0) &= u(0), \\ u'(1) &= 0. \end{aligned} \tag{4.28}$$

où  $u''$  désigne la dérivée seconde de  $u$  par rapport à  $x$ .

Dans toute la suite, on considère une subdivision uniforme de l'intervalle  $[0, 1]$  : on note  $h = 1/N$  où  $N \geq 2$  est un entier fixé. On pose  $x_i = ih$ , pour  $i = 0, \dots, N$ .

#### 1. Formulation variationnelle

1.1 Ecrire une formulation variationnelle de (4.28).

Soit  $v$  une fonction suffisamment régulière, on multiplie la première équation de (4.28) par  $v$  et on intègre sur  $]0, 1[$ . En effectuant des intégrations par parties et en tenant compte des conditions aux limites, on obtient :

$$\int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) = \int_0^1 f(x)v(x) dx.$$

Pour que les intégrales aient un sens, il suffit de prendre  $u, v \in H^1(]0, 1[)$ , auquel cas les fonctions sont continues et donc les valeurs  $u(0)$  et  $v(0)$  ont aussi un sens. On en déduit qu'une formulation faible est

$$\begin{aligned} u &\in H^1(]0, 1[) \\ \int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) &= \int_0^1 f(x)v(x) dx, \quad \forall v \in H^1(]0, 1[). \end{aligned}$$

On en déduit que la formulation variationnelle est

$$\begin{aligned} \text{Trouver } u &\in H^1(]0, 1[); \\ J(u) &= \min_{v \in H} J(v), \end{aligned}$$

avec  $J(v) = \frac{1}{2}a(v, v) - T(v)$ , où  $a$  est la forme bilinéaire définie par  $a(u, v) = \int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0)$  et  $T$  la forme linéaire continue définie par  $T(v) = \int_0^1 f(x)v(x) dx$ .

1.2. On considère le problème

$$\begin{aligned} u &\in H^1(]0, 1[) \\ \int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) &= \int_0^1 f(x)v(x) dx, \quad \forall v \in H^1(]0, 1[). \end{aligned} \tag{4.29}$$

1.2.a Déterminer le problème aux limites dont la formulation faible est (4.29).

Supposons  $u$  régulière, et prenons d'abord  $v \in C_c^1(]0, 1[)$ . On a alors

$$\int_0^1 u'(x)v'(x) dx + \alpha \int_0^1 u(x)v(x) dx + u(0)v(0) = \int_0^1 f(x)v(x) dx$$

et donc en intégrant par parties :

$$\int_0^1 (-u''(x) + \alpha u(x) - f(x))v(x) dx = 0.$$

Comme ceci est vrai pour toute fonction  $v \in C_c(\]0, 1[)$ , on en déduit que  $-u''(x) + \alpha u(x) = f(x)$ ,  $x \in \]0, 1[$ .

Prenons maintenant  $v \in H^1(\]0, 1[)$ , en intégrant par parties et tenant compte de ce qui précède, on obtient

$$(-u'(0) + u(0))v(0) + u'(1)v(1) = 0.$$

Comme ceci est vrai pour toute fonction  $v \in H^1(\]0, 1[)$ , on en déduit que  $u$  vérifie (4.28).

1.2.b. Montrer que si  $\alpha > 0$ , (4.29) admet une unique solution.

On peut appliquer le lemme de Lax Milgram ; en effet,

- la forme linéaire  $T$  est continue car  $|T(v)| = |\int_0^1 f(x)v(x) dx| \leq \|f\|_{L^2}\|v\|_{L^2}$  par l'inégalité de Cauchy-Schwarz, et donc  $|T(v)| \leq C\|v\|_{L^2}$  avec  $C = \|f\|_{L^2}$ .
- la forme bilinéaire  $a$  (qui est évidemment symétrique, ce qui n'est d'ailleurs pas nécessaire pour appliquer Lax-Milgram) est continue ; en effet :

$$|a(u, v)| \leq \|u'\|_{L^2}\|v'\|_{L^2} + \alpha\|u\|_{L^2}\|v\|_{L^2} + |u(0)|v(0);$$

or pour tout  $x \in \]0, 1[$   $v(0) = v(x) + \int_0^x v'(t)dt$  et donc par inégalité triangulaire et par Cauchy-Schwarz, on obtient que  $|v(0)| \leq |v(x)| + \|v'\|_{L^2}$ . En intégrant cette inégalité entre 0 et 1, on obtient

$$|v(0)| \leq \|v\|_{L^1} + \|v'\|_{L^2} \leq \|v\|_{L^2} + \|v'\|_{L^2} \leq 2\|v\|_{H^1}.$$

La même inégalité est évidemment vraie pour  $u(0)$ . On en déduit que :

$$\begin{aligned} a(u, v) &\leq \|u'\|_{L^2}\|v'\|_{L^2} + \alpha\|u\|_{L^2}\|v\|_{L^2} + 4\|u\|_{H^1}\|v\|_{H^1} \\ &\leq \|u\|_{H^1}\|v\|_{H^1} + \alpha\|u\|_{H^1}\|v\|_{H^1} + 4\|u\|_{H^1}\|v\|_{H^1} \\ &\leq (5 + \alpha)\|u\|_{H^1}\|v\|_{H^1}, \end{aligned}$$

ce qui prouve que  $a$  est continue.

Montrons maintenant que  $a$  est coercive. Dans le cas où  $\alpha > 0$ , ceci est facile à vérifier, car on a

$$a(u, u) = \int_0^1 u'(x)^2 dx + \alpha \int_0^1 u(x)^2 dx + u(0)^2 \geq \min(\alpha, 1)\|u\|_{H^1}^2.$$

Par le lemme de Lax -Milgram, on peut donc conclure à l'existence et l'unicité de la solution de (4.29).

(pour les braves : on peut démontrer que ceci est encore vrai pour  $\alpha = 0$  en appliquant l'inégalité de Poincaré à la fonction  $v - v(0)$ ).

Dans le cas où  $\alpha = 0$ , on applique l'inégalité de Poincaré à la fonction  $w = u - u(0)$ , ce qui est licite car  $w(0) = 0$  ; on a donc :  $\|w\|_{L^2} \leq \|w'\|_{L^2}$ , et donc  $\|u'\|_{L^2} \geq \|u - u(0)\|_{L^2}$ . On en déduit que  $a(u, u) \geq \|u - u(0)\|_{L^2}^2 + u(0)^2 \geq \frac{1}{2}\|u\|_{L^2}^2$ .

On écrit alors que

$$\begin{aligned} a(u, u) &= \frac{1}{2}a(u, u) + \frac{1}{2}a(u, u) \\ &\geq \frac{1}{2}\|u'\|_{L^2}^2 + \frac{1}{4}\|u\|_{L^2}^2 \\ &\geq \frac{1}{4}\|u\|_{H^1}^2, \end{aligned}$$

ce qui montre que la forme bilinéaire  $a$  est encore coercive.

**2. Discrétisation par éléments finis.**

Soit  $V_h$  l'ensemble des fonctions continues sur  $[0, 1]$  dont la restriction à chaque intervalle  $[x_i, x_{i+1}]$  est affine pour  $0 \leq i \leq N - 1$ .

2.1 Quelle est la dimension de  $V_h$  ?

Une base de l'espace  $V_h$  est la famille des fonctions dites "chapeau", définies par

$$\begin{aligned}\varphi_i(x) &= \min\left(\frac{1}{h}(x - x_{i-1})^+, \frac{1}{h}(x_{i+1} - x)^+\right) \text{ pour } i = 1, \dots, N - 1, \\ \varphi_1(x) &= \frac{1}{h}(x_1 - x)^+, \\ \varphi_N(x) &= \frac{1}{h}(x - x_{N+1})^+.\end{aligned}$$

On en déduit que l'espace  $V_h$  est de dimension  $N + 1$ .

2.2. Donner la discrétisation éléments finis de (4.29).

Le problème discrétisé par éléments finis s'écrit :

$$u_h \in V_h \quad \int_0^1 u_h'(x)v_h'(x) dx + \alpha \int_0^1 u_h(x)v_h(x) dx + u_h(0)v_h(0) = \int_0^1 f(x)v_h(x) dx, \quad \forall v_h \in V_h. \quad (4.30)$$

2.3 Montrer que le problème discret ainsi obtenu admet une solution unique.

Comme on a effectué une discrétisation par éléments finis conformes, le lemme de Lax Milgram s'applique à nouveau.

2.4 Ecrire le problème discret sous la forme d'un système linéaire  $AU = b$  en explicitant les dimensions des vecteurs et matrice et en donnant leur expression.

Commençons par le second membre :  $B = (b_i)_{0 \leq i \leq N}$ , avec  $b_i = \int_0^1 f(x)\phi_i(x)dx$ .

Calculons  $A_{i,j} = A_{j,i} = a(\phi_i, \phi_j) = \int_0^1 \phi_i'(x)\phi_j'(x) dx + \alpha \int_0^1 \phi_i(x)\phi_j(x) dx + \phi_i(0)\phi_j(0)$ , pour  $i = 1, \dots, N$ . En raison de la forme des fonctions de base  $(\phi_i)_{i=0,N}$  les seuls termes non nuls sont les termes  $A_{i-1,i}$ ,  $A_{i,i}$  et  $A_{i,i+1}$ . Après calculs, on obtient :

$$A = \begin{bmatrix} \frac{1}{h} + \frac{\alpha h}{3} + 1 & -\frac{1}{h} + \frac{\alpha h}{6} & 0 & 0 & \dots & 0 \\ -\frac{1}{h} + \frac{\alpha h}{6} & \frac{2}{h} + \frac{2\alpha h}{3} & -\frac{1}{h} + \frac{\alpha h}{6} & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & & \vdots \\ \vdots & & & & & \\ 0 & \dots & 0 & -\frac{1}{h} + \frac{\alpha h}{6} & \frac{2}{h} + \frac{2\alpha h}{3} & -\frac{1}{h} + \frac{\alpha h}{6} \\ 0 & \dots & & 0 & -\frac{1}{h} + \frac{\alpha h}{6} & \frac{1}{h} + \frac{\alpha h}{3} \end{bmatrix}$$

2.5 Pour les braves : donner une borne supérieure de l'erreur  $\|u_h - u\|_{H^1(0,1)}$ .

Soit  $u \in H^1(]0, 1[)$  une solution de (4.29). Alors  $-u'' = f - \alpha u$  au sens des distributions, mais comme  $f \in L^2(]0, 1[)$  et  $u \in L^2(]0, 1[)$ , on en déduit que  $u \in H^2(]0, 1[)$ . On peut donc appliquer les résultats du cours. On a vu en cours que si  $u_I$  l'interpolée de  $u$  dans  $V_h$ , On a  $\|u - u_h\|_{L^2(0,1)} \leq C\|u - u_I\|_{L^2(0,1)}$  où  $C$  est la racine carrée du rapport de la constante de continuité sur la constante de coercivité, c.à.d.  $C = \sqrt{\frac{5+\alpha}{\min(\alpha,1)}}$ . De plus, on a aussi vu que si  $u \in H^2(]0, 1[)$ , l'erreur d'interpolation est d'ordre  $h$ ; plus précisément, on a :

$$\|u - u_I\|_{L^2(0,1)}^2 \leq (1 + h^2)h^2\|u''\|_{L^2(0,1)}^2$$

On en déduit que

$$\|u_h - u\|_{H^1(0,1)} \leq \sqrt{\frac{5 + \alpha}{\min(\alpha, 1)}} \sqrt{1 + h^2} \|u''\|_{L^2(0,1)} h.$$

**Exercice 53 (Éléments finis  $Q_1$ )** *Corrigé en page 172*

On considère le rectangle  $\Omega$  de sommets  $(-1, 0), (2, 0), (-1, 1), (2, 1)$ . On s'intéresse à la discrétisation par éléments finis de l'espace fonctionnel  $H^1(\Omega)$ .

I. On choisit de découper  $\Omega$  en deux éléments  $e_1$  et  $e_2$  définis par les quadrilatères de sommets respectifs  $M_1(-1, 1), M_2(0, 1), M_5(1, 0), M_4(-1, 0)$  et  $M_2(0, 1), M_3(2, 1), M_6(2, 0), M_5(1, 0)$ .

On prend comme noeuds les points  $M_1, \dots, M_6$  et comme espace par élément l'ensemble des polynômes  $Q_1$ . On note  $\Sigma_1 = \{M_4, M_5, M_2, M_1\}$  et  $\Sigma_2 = \{M_5, M_6, M_3, M_2\}$ .

On a donc construit la discrétisation  $\{(e_1, \Sigma_1, Q_1), (e_2, \Sigma_2, Q_1)\}$ .

I.1 Montrer que les éléments  $(e_1, \Sigma_1, Q_1)$  et  $(e_2, \Sigma_2, Q_1)$  sont des éléments finis de Lagrange.

I.2. Montrer que l'espace de dimension finie correspondant à cette discrétisation n'est pas inclus dans  $H^1(\Omega)$  (construire une fonction de cet espace dont la dérivée distribution n'est pas dans  $L^2$ ). Quelle est dans les hypothèses appelées en cours "cohérence globale" celle qui n'est pas vérifiée ?

II. On fait le même choix des éléments et des noeuds que dans I. On introduit comme élément de référence  $e$  le carré de sommets  $(\pm 1, \pm 1)$ ,  $\Sigma$  est l'ensemble des sommets de  $e$  et  $P = Q_1$ .

II.1. Quelles sont les fonctions de base locales de  $(e, \Sigma, P)$ . On note ces fonctions  $\Phi_1, \dots, \Phi_4$ .

II.2 A partir des fonctions  $\Phi_1, \dots, \Phi_4$ , construire des bijections  $F_1$  et  $F_2$  de  $e$  dans  $e_1$  et  $e_2$ . Les fonctions  $F_1$  et  $F_2$  sont elles affines ?

II.3 On note  $P_{e_i} = \{f : e_i \rightarrow \mathbb{R}, f \circ F_i|_e \in Q_1\}$ , pour  $i = 1, 2$ , où les  $F_i$  sont définies à la question précédente. Montrer que les éléments  $(e_1, \Sigma_1, P_{e_1})$  et  $(e_2, \Sigma_2, P_{e_2})$  sont des éléments finis de Lagrange et que l'espace vectoriel construit avec la discrétisation  $\{(e_1, \Sigma_1, P_{e_1}), (e_2, \Sigma_2, P_{e_2})\}$  est inclus dans  $H^1(\Omega)$  (i.e. vérifier la "cohérence globale" définie en cours). On pourra pour cela montrer que si  $S = e_1 \cap e_2 = \{(x, y); x + y = 1\}$ , alors  $\{f|_S, f \in P_{e_i}\} = \{f : S \rightarrow \mathbb{R}; f(x, y) = a + by, a, b \in \mathbb{R}\}$ .

**Exercice 54 (Éléments affine-équivalents)** *Corrigé en page 174*

Soit  $\Omega$  un ouvert polygonal de  $\mathbb{R}^2$ , et  $\mathcal{T}$  un maillage de  $\Omega$ .

Soient  $(\bar{K}, \bar{\Sigma}, \bar{P})$  et  $(K, \Sigma, P)$  deux éléments finis de Lagrange affine - équivalents. On suppose que les fonctions de base locales de  $\bar{K}$  sont affines.

Montrer que toute fonction de  $P$  est affine.

En déduire que les fonctions de base locales de  $(K, \Sigma, P)$  affines.

**Exercice 55 (Éléments finis  $P^2$  en une dimension d'espace)**

On veut résoudre numériquement le problème aux limites suivant

$$\begin{cases} -u''(x) + u(x) = x^2, & 0 < x < 1 \\ u(0) = 0, \\ u'(1) = 1. \end{cases} \quad (4.31)$$

1. Donner une formulation faible du problème (4.31)

2. Démontrer que le problème (4.31) admet une unique solution.

3. On partage l'intervalle  $]0, 1[$  en  $N$  intervalles égaux et on approche la solution par une méthode d'éléments finis de degré 2. Ecrire le système qu'il faut résoudre.

**Exercice 56 (Éléments finis  $P^1$  sur maillage triangulaire)** *Corrigé en page 175*



On veut résoudre numériquement le problème

$$\begin{aligned} -\Delta u(x, y) &= f(x, y), & (x, y) \in D &= (0, a) \times (0, b), \\ u(x, y) &= 0, & (x, y) \in \partial D, \end{aligned}$$

où  $f$  est une fonction donnée, appartenant à  $L^2(D)$ . Soient  $M, N$  deux entiers. On définit

$$\Delta x = \frac{a}{M+1}, \Delta y = \frac{b}{N+1}$$

et on pose

$$x_k = k\Delta x, 0 \leq k \leq M+1, y_l = l\Delta y, 0 \leq l \leq N+1$$

On note

$$T_{k+1/2, l+1/2}^0 \text{ le triangle de sommets } (x_k, y_l), (x_{k+1}, y_l), (x_{k+1}, y_{l+1}),$$

$$T_{k+1/2, l+1/2}^1 \text{ le triangle de sommets } (x_k, y_l), (x_k, y_{l+1}), (x_{k+1}, y_{l+1}).$$

Ecrire la matrice obtenue en discrétisant le problème avec les éléments finis triangulaires linéaires (utilisant le maillage précédent).

### Exercice 57 (Intégration numérique) Corrigé en page 177

1. Vérifier que l'intégration numérique à un point de Gauss, donné par le centre de gravité du triangle, sur l'élément fini  $P_1$  sur triangle, est exacte pour les polynômes d'ordre 1.

2. Vérifier que l'intégration numérique à trois points de Gauss définis sur le triangle de référence par  $p_1 = \left(\frac{1}{2}, 0\right)$ ,  $p_2 = \left(\frac{1}{2}, \frac{1}{2}\right)$ ,  $p_3 = \left(0, \frac{1}{2}\right)$ , avec les poids d'intégration  $\omega_1 = \omega_2 = \omega_3 = \frac{1}{6}$ , est exacte pour les polynômes d'ordre 2.

### Exercice 58 (Éléments finis $Q_2$ ) Corrigé en page 177

On note  $C$  le carré  $[0, 1] \times [0, 1]$  de sommets

$$a_1 = (0, 0), \quad a_2 = (1, 0), \quad a_3 = (1, 1), \quad a_4 = (0, 1).$$

On note

$$a_5 = (1/2, 0), \quad a_6 = (1, 1/2), \quad a_7 = (1/2, 1), \quad a_8 = (0, 1/2), \quad a_9 = (1/2, 1/2)$$

et

$$\Sigma = \{a_i, 1 \leq i \leq 8\}.$$

1. Montrer que pour tout  $p \in P_2$

$$\sum_{i=1}^4 p(a_i) - 2 \sum_{i=5}^8 p(a_i) + 4p(a_9) = 0.$$

2. En déduire une forme linéaire  $\phi$  telle que si  $p \in P = \{p \in Q_2, \phi(p) = 0\}$  et  $p(a_i) = 0$  pour  $i = 1, \dots, 8$ , alors  $p = 0$ . 3. Calculer les fonctions de base de l'élément fini  $(C, P, \Sigma)$ , avec  $\Sigma = \{a_1, \dots, a_8\}$ .

### Exercice 59 (Éléments finis $Q_2^*$ )

Soit  $C = [-1, 1] \times [-1, 1]$ . On note  $a_1, \dots, a_8$  les noeuds de  $C$ , définis par

$$\begin{aligned} a_1 &= (-1, -1), & a_2 &= (1, -1), & a_3 &= (1, 1), & a_4 &= (-1, 1), \\ a_5 &= (0, -1), & a_6 &= (1, 0), & a_7 &= (0, 1), & a_8 &= (-1, 0). \end{aligned}$$

On rappelle que  $Q_2 = \text{Vect}\{1, x, y, xy, x^2, y^2, x^2y, xy^2, x^2y^2\}$  et que  $\dim Q_2 = 9$ . On note  $Q_2^*$  l'espace de polynôme engendré par les fonctions  $\{1, x, y, xy, x^2, y^2, x^2y, xy^2\}$ .

a) Construire  $(\varphi_i^*)_{i=1, \dots, 8} \subset Q_2^*$  tel que

$$\varphi_j^*(a_i) = \delta_{ij} \quad \forall i, j = 1, \dots, 8.$$

b) Montrer que  $\Sigma$  est  $Q_2^*$ -unisolvant, avec  $\Sigma = \{a_1, \dots, a_8\}$ .

c) Soit  $S = [-1, 1] \times \{1\}$ ,  $\Sigma_S = \Sigma \cap S$ , et soit  $P$  l'ensemble des restrictions à  $S$  des fonctions de  $Q_2^*$ , i.e.  $P = \{f|_S; f \in Q_2^*\}$ . Montrer que  $\Sigma_S$  est  $P$ -unisolvant. La propriété est elle vraie pour les autres arêtes de  $C$  ?

## 4.7 Corrigés des exercices

### Exercice 48 page 161

1. Soit  $v \in H_N$ , comme  $H_N \subset C([0, 1])$ , on a  $v \in L^2([0, 1])$ . D'autre part, comme  $v|_{K_i} \in P_1$ , on a  $v|_{K_i}(x) = \alpha_i x + \beta_i$ , avec  $\alpha_i, \beta_i \in \mathbb{R}$ . Donc  $v$  admet une dérivée faible dans  $L^2([0, 1])$ , et  $Dv|_{K_i} = \alpha_i$  on a donc :

$$\|Dv\|_{L^2} \leq \max_{i=1, \dots, N} |\alpha_i| < +\infty.$$

De plus  $v(0) = v(1) = 0$ , donc  $v \in H_0^1([0, 1])$ .

On en déduit que  $H_N \subset H_0^1([0, 1])$ .

2. On a :

$$\phi_i(x) = \begin{cases} 1 - \frac{x - x_i}{h} & \text{si } x \in K_i \\ 1 + \frac{x - x_i}{h} & \text{si } x \in K_{i-1} \\ 0 & \text{si } x \in ]0, 1[ \setminus (K_i \cup K_{i-1}) \end{cases}$$

On en déduit que  $\phi_i|_{K_j} \in P_1$  pour tout  $j = 0, \dots, N$ .

De plus, les fonctions  $\phi_i$  sont clairement continues. Pour montrer que  $\phi_i \in H_N$ , il reste à montrer que  $\phi_i(0) = \phi_i(1) = 0$ . Ceci est immédiat pour  $i = 2, \dots, N-1$ , car dans ce cas  $\phi_i|_{K_0} = \phi_i|_{K_{N+1}} = 0$ . On vérifie alors facilement que  $\phi_1(0) = 1 - \frac{h}{h} = 0$  et  $\phi_N(1) = 0$ .

Pour montrer que  $H_N = \text{Vect}\{\phi_1, \dots, \phi_N\}$ , il suffit de montrer que  $\{\phi_1, \dots, \phi_N\}$  est une famille libre de  $H_N$ .

En effet, si  $\sum_{i=1}^N a_i \phi_i = 0$ , alors en particulier  $\sum_{i=1}^N a_i \phi_i(x_k) = 0$ , pour  $k = 1, \dots, N$ , et donc  $a_k = 0$  pour  $k = 1, \dots, N$ .

3. Soit  $u = \sum_{j=1}^N u_j \phi_j$  solution de

$$a(u, \phi_i) = T(\phi_i) \quad \forall i = 1, \dots, N.$$

La famille  $(u_j)_{j=1, \dots, N}$  est donc solution du système linéaire

$$\sum_{j=1}^N \mathcal{K}_{i,j} u_j = \mathcal{G}_i \quad i = 1, \dots, N$$

où  $\mathcal{K}_{i,j} = a(\phi_j, \phi_i)$  et  $\mathcal{G}_i = T(\phi_i)$ . Calculons  $\mathcal{K}_{i,j}$  et  $\mathcal{G}_i$  ; on a :

$$\mathcal{K}_{i,j} = \int_0^1 \phi_j'(x) \phi_i'(x) dx; \text{ or } \phi_i'(x) = \begin{cases} \frac{1}{h} & \text{si } x \in ]x_{i-1}, x_i[ \\ -\frac{1}{h} & \text{si } x \in ]x_i, x_{i+1}[ \\ 0 & \text{ailleurs} \end{cases}$$

On en déduit que :

$$\mathcal{K}_{i,i} = \int_0^1 (\phi_i'(x))^2 dx = 2h \frac{1}{h^2} = \frac{2}{h} \text{ pour } i = 1, \dots, N,$$

$$\mathcal{K}_{i,i+1} - \int_0^1 \phi_i'(x) \phi_{i+1}'(x) dx = -h \times \frac{1}{h^2} = -\frac{1}{h}, \text{ pour } i = 1, \dots, N-1,$$

$$\mathcal{K}_{i,i-1} = \int_0^1 \phi_i'(x) \phi_{i-1}'(x) dx = -\frac{1}{h} \text{ pour } i = 2, \dots, N,$$

$$\mathcal{K}_{i,j} = 0 \text{ pour } |i - j| > 1.$$

Calculons maintenant  $\mathcal{G}_i$  :

$$\mathcal{G}_i = \int_{x_{i-1}}^{x_{i+1}} f(x) \phi_i(x) dx$$

Si  $f$  est constante, on a alors  $\mathcal{G}_i = f \int_{x_{i-1}}^{x_{i+1}} \phi_i(x) dx = hf$ . Si  $f$  n'est pas constante, on procède à une intégration numérique. On peut, par exemple, utiliser la formule des trapèzes pour le calcul des intégrales  $\int_{x_{i-1}}^{x_i} f(x)\phi_i(x) dx$  et  $\int_{x_i}^{x_{i+1}} f(x)\phi_i(x) dx$ . On obtient alors :

$$\mathcal{G}_i = hf(x_i).$$

Le schéma obtenu est donc :

$$\begin{cases} 2u_i - u_{i-1} - u_{i+1} = h^2 f(x_i) & i = 1, \dots, N \\ u_0 = u_{N+1} = 0 \end{cases}$$

C'est exactement le schéma différences finis avec un pas constant  $h$ .

### Exercice 49 page 162

1. Soit  $(x_i)_{i=1, \dots, N+1}$  une discrétisation de l'intervalle  $[0, 1]$ , avec  $0 = x_0 < x_1 < \dots < x_i < x_{i+1} < x_N < x_{N+1} = 1$ . Pour  $i = 1, \dots, N$ , on pose  $h_{i+\frac{1}{2}} = x_{i+1} - x_i$ . L'équation (4.26) au point  $x_i$  s'écrit :

$$-u''(x_i) + u(x_i) = f(x)$$

On écrit les développements de Taylor de  $u(x_{i+1})$  et  $u(x_{i-1})$  :

$$u(x_{i+1}) = u(x_i) + h_{i+\frac{1}{2}} u'(x_i) + \frac{1}{2} h_{i+\frac{1}{2}}^2 u''(x_i) + \frac{1}{6} h_{i+\frac{1}{2}}^3 u'''(\zeta_i), \text{ avec } \zeta_i \in [x_i, x_{i+1}],$$

$$u(x_{i-1}) = u(x_i) - h_{i-\frac{1}{2}} u'(x_i) + \frac{1}{2} h_{i-\frac{1}{2}}^2 u''(x_i) - \frac{1}{6} h_{i-\frac{1}{2}}^3 u'''(\theta_i), \text{ avec } \theta_i \in [x_{i-1}, x_i],$$

En multipliant la première égalité par  $h_{i-\frac{1}{2}}$ , la deuxième par  $h_{i+\frac{1}{2}}$  et en additionnant :

$$\begin{aligned} u''(x_i) &= \frac{2}{h_{i+\frac{1}{2}} h_{i-\frac{1}{2}} (h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}})} \left[ h_{i-\frac{1}{2}} u(x_{i+1}) + h_{i+\frac{1}{2}} u(x_{i-1}) + (h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}) u(x_i) \right] \\ &\quad - \frac{1}{6} \frac{h_{i+\frac{1}{2}}^2}{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}} u'''(\zeta_i) + \frac{1}{6} \frac{h_{i-\frac{1}{2}}^2}{h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}} u'''(\theta_i). \end{aligned}$$

En posant  $\gamma_i = \frac{2}{h_{i+\frac{1}{2}} h_{i-\frac{1}{2}} (h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}})}$ , on déduit donc l'approximation aux différences finies suivante pour tous les noeuds internes :

$$\gamma_i \left[ h_{i-\frac{1}{2}} u_{i+1} + h_{i+\frac{1}{2}} u_{i-1} + (h_{i+\frac{1}{2}} + h_{i-\frac{1}{2}}) u_i \right] + u_i = f(x_i), \quad i = 1, \dots, N.$$

La condition de Fourier en 0 se discrétise par

$$\frac{u_1 - u_0}{h_{\frac{1}{2}}} - u_0 = 0,$$

et la condition de Neumann en 1 par :

$$\frac{u_{N+1} - u_N}{h_{N+\frac{1}{2}}} = -1.$$

On obtient ainsi un système linéaire carré d'ordre  $N + 1$ .

2. On prend maintenant une discrétisation volumes finis non uniforme ; on se donne  $N \in \mathbb{N}^*$  et  $h_1, \dots, h_N > 0$  t.q.  $\sum_{i=1}^N h_i = 1$ . On pose  $x_{\frac{1}{2}} = 0$ ,  $x_{i+\frac{1}{2}} = x_{i-\frac{1}{2}} + h_i$ , pour  $i = 1, \dots, N$  (de sorte que  $x_{N+\frac{1}{2}} = 1$ ),  $h_{i+\frac{1}{2}} = \frac{h_{i+1} + h_i}{2}$ , pour  $i = 1, \dots, N - 1$ , et  $f_i = \frac{1}{h_i} \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} f(x) dx$ , pour  $i = 1, \dots, N$ .

En intégrant la première équation de (4.26), et en approchant les flux  $u'(x_{i+\frac{1}{2}})$  par le flux numérique  $F_{i+\frac{1}{2}}$ , on obtient le schéma suivant :

$$F_{i+\frac{1}{2}} - F_{i-\frac{1}{2}} + h_i u_i = h_i f_i, \quad i \in \{1, \dots, N\}, \quad (4.32)$$

où  $(F_{i+\frac{1}{2}})_{i \in \{0, \dots, N\}}$  donné en fonction des inconnues discrètes  $(u_1, \dots, u_N)$  par les expressions suivantes, tenant compte des conditions aux limites :

$$F_{i+\frac{1}{2}} = -\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}}, \quad i \in \{1, \dots, N-1\}, \quad (4.33)$$

$$F_{\frac{1}{2}} = -\frac{u_1 - u_0}{\frac{x_1}{2}}, \quad (4.34)$$

$$F_{\frac{1}{2}} - u_0 = 0 \quad (4.35)$$

$$F_{N+\frac{1}{2}} = -1. \quad (4.36)$$

Notons que  $u_0$  peut être éliminé des équations (4.34) et (4.35). On obtient ainsi un système linéaire de  $N$  équations à  $N$  inconnues :

$$-\frac{u_2 - u_1}{h_{\frac{3}{2}}} + \frac{u_1}{1 - \frac{x_1}{2}} + h_1 u_1 = h_1 f_1, \quad (4.37)$$

$$-\frac{u_{i+1} - u_i}{h_{i+\frac{1}{2}}} + \frac{u_i - u_{i-1}}{h_{i-\frac{1}{2}}} + h_i u_i = h_i f_i, \quad i \in \{2, \dots, N-1\}, \quad (4.38)$$

$$-1 + \frac{u_N - u_{N-1}}{h_{N-\frac{1}{2}}} + h_N u_N = h_N f_N, \quad (4.39)$$

3. Comme pour les différences finies, on se donne  $(x_i)_{i=1, \dots, N+1}$  une discrétisation de l'intervalle  $[0, 1]$ , avec  $0 = x_0 < x_1 < \dots < x_i < x_{i+1} < x_N < x_{N+1} = 1$ . Pour  $i = 1, \dots, N$ , on pose  $h_{i+\frac{1}{2}} = x_{i+1} - x_i$  et  $K_{i+\frac{1}{2}} = [x_i, x_{i+1}]$ , pour  $i = 0, \dots, N$ . On définit l'espace d'approximation  $H_N = \{v \in C([0, 1], \mathbb{R}) \text{ t.q. } v|_{K_{i+\frac{1}{2}}} \in P_1, i = 0, \dots, N\}$ , où  $P_1$  désigne l'ensemble des polynômes de degré inférieur ou égal à 1. Remarquons que l'on a bien  $H_N \subset H$ .

Pour  $i = 1, \dots, N$ , on pose :

$$\begin{aligned} \phi_i(x) &= \frac{1}{h_{i-\frac{1}{2}}} (x - x_{i-1}) \text{ si } x \in K_{i-\frac{1}{2}}, \\ \phi_i(x) &= \frac{1}{h_{i+\frac{1}{2}}} (x_{i+1} - x) \text{ si } x \in K_{i+\frac{1}{2}}, \\ \phi_i(x) &= 0 \text{ sinon,} \end{aligned} \quad (4.40)$$

et on pose également

$$\begin{aligned} \phi_{N+1}(x) &= \frac{1}{h_{N+\frac{1}{2}}} (x - x_N) \text{ si } x \in K_{N+\frac{1}{2}}, \\ \phi_{N+1}(x) &= 0 \text{ sinon,} \end{aligned} \quad (4.41)$$

$$\begin{aligned} \phi_0(x) &= \frac{1}{h_{\frac{1}{2}}} (x_1 - x) \text{ si } x \in K_{\frac{1}{2}}, \\ \phi_0(x) &= 0 \text{ sinon,} \end{aligned} \quad (4.42)$$

On vérifie facilement que  $\phi_i \in H_N$  pour tout  $0 = 1, \dots, N+1$  et que  $H_N = \text{Vect}\{\phi_0, \dots, \phi_{N+1}\}$ .

La formulation éléments finis s'écrit alors :

$$\begin{aligned} u^{(N)} &\in H_N, \\ a(u^{(N)}, v) &= T(v), \forall v \in H_N, \end{aligned} \quad (4.43)$$

Pour construire le système linéaire à résoudre, on prend successivement  $v = \phi_i$ ,  $i = 0, \dots, N+1$  dans (4.43).

Soit  $u^{(N)} = \sum_{j=0}^{N+1} u_j \phi_j$  solution de

$$a(u^{(N)}, \phi_i) = T(\phi_i) \quad \forall i = 0, \dots, N+1.$$

La famille  $(u_j)_{j=0, \dots, N+1}$  est donc solution du système linéaire

$$\sum_{j=0}^N \mathcal{K}_{i,j} u_j = \mathcal{G}_i \quad i = 0, \dots, N+1,$$

où  $\mathcal{K}_{i,j} = a(\phi_j, \phi_i)$  et  $\mathcal{G}_i = T(\phi_i)$ . Calculons  $\mathcal{K}_{i,j}$  et  $\mathcal{G}_i$ ; on a :  $\mathcal{K}_{ij} = \int_0^1 \phi_j'(x)\phi_i'(x)dx + \int_0^1 \phi_j(x)\phi_i(x)dx$ .

Or

$$\phi_i'(x) = \begin{cases} \frac{1}{h_{i-\frac{1}{2}}} \text{ si } x \in ]x_{i-1}, x_i[ \\ -\frac{1}{h_{i+\frac{1}{2}}} \text{ si } x \in ]x_i, x_{i+1}[ \\ 0 \text{ ailleurs.} \end{cases}$$

Donc si  $1 \leq i = j \leq N$ , on a

$$\mathcal{K}_{i,i} = \int_0^1 (\phi_i'(x))^2 dx + \int_0^1 (\phi_i(x))^2 dx = \frac{1}{h_{i-\frac{1}{2}}} + \frac{1}{h_{i+\frac{1}{2}}} + \frac{h_{i-\frac{1}{2}}}{3} + \frac{h_{i+\frac{1}{2}}}{3}.$$

Si  $i = j = N + 1$ , alors

$$\mathcal{K}_{N+1,N+1} = \int_0^1 (\phi_{N+1}'(x))^2 dx + \int_0^1 (\phi_{N+1}(x))^2 dx = \frac{1}{h_{N+\frac{1}{2}}} + \frac{h_{N+\frac{1}{2}}}{3}.$$

Si  $i = j = 0$ , alors

$$\mathcal{K}_{0,0} = \int_0^1 (\phi_0'(x))^2 dx + \int_0^1 (\phi_0(x))^2 dx + \phi_0^2 = \frac{1}{h_{\frac{1}{2}}} + \frac{h_{\frac{1}{2}}}{3} + 1.$$

Si  $0 \leq i \leq N$  et  $j = i + 1$ , on a :

$$\mathcal{K}_{i,i+1} = \int_0^1 \phi_i'(x)\phi_{i+1}'(x)dx + \int_0^1 \phi_i(x)\phi_{i+1}(x)dx = -h_{i+\frac{1}{2}} \times \frac{1}{h_{i+\frac{1}{2}}^2} + \frac{h_{i+\frac{1}{2}}}{2} - \frac{h_{i+\frac{1}{2}}}{3} = -\frac{1}{h_{i+\frac{1}{2}}} + \frac{h_{i+\frac{1}{2}}}{6}.$$

La matrice étant symétrique, si  $2 \leq i \leq N + 1$  et  $j = i - 1$ , on a :

$$\mathcal{K}_{i,i-1} = \mathcal{K}_{i-1,i} = -\frac{1}{h_{i-\frac{1}{2}}} + \frac{h_{i-\frac{1}{2}}}{6}.$$

Calculons maintenant  $\mathcal{G}_i$ .

$$\mathcal{G}_i = \int_{x_{i-1}}^{x_{i+1}} f(x)\phi_i(x)dx + \phi_i(1).$$

Si  $f$  est constante, on a alors  $\mathcal{G}_i = f \int_{x_{i-1}}^{x_{i+1}} \phi_i(x)dx + \phi_i(1) = \frac{1}{2}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}})f + \phi_i(1)$ .

Si  $f$  n'est pas constante, on procède à une intégration numérique. On peut, par exemple, utiliser la formule des trapèzes pour le calcul des intégrales  $\int_{x_{i-1}}^{x_i} f(x)\phi_i(x)dx$  et  $\int_{x_i}^{x_{i+1}} f(x)\phi_i(x)dx$ . On obtient alors :

$$\mathcal{G}_i = \frac{1}{2}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}})f(x_i) + \phi_i(1).$$

Le schéma obtenu est donc :

$$\left\{ \begin{array}{l} \left( \frac{1}{h_{i-\frac{1}{2}}} + \frac{1}{h_{i+\frac{1}{2}}} + \frac{h_{i-\frac{1}{2}}}{3} + \frac{h_{i+\frac{1}{2}}}{3} \right) u_i + \left( \frac{h_{i-\frac{1}{2}}}{6} - \frac{1}{h_{i-\frac{1}{2}}} \right) u_{i-1} + \left( \frac{h_{i+\frac{1}{2}}}{6} - \frac{1}{h_{i+\frac{1}{2}}} \right) u_{i+1} \\ \quad = \frac{1}{2}(h_{i-\frac{1}{2}} + h_{i+\frac{1}{2}})f(x_i) \quad i = 1, \dots, N \\ \left( \frac{1}{h_{\frac{1}{2}}} + \frac{h_{\frac{1}{2}}}{3} + 1 \right) u_0 + \left( \frac{1}{h_{i+\frac{1}{2}}} + \frac{h_{i+\frac{1}{2}}}{6} \right) u_1 = \frac{1}{2}h_{\frac{1}{2}}f(x_0) \\ \left( \frac{1}{h_{N+\frac{1}{2}}} + \frac{h_{N+\frac{1}{2}}}{3} \right) u_{N+1} + \left( \frac{h_{N+\frac{1}{2}}}{6} - \frac{1}{h_{N+\frac{1}{2}}} \right) u_N = \frac{1}{2}h_{N+\frac{1}{2}}f(x_{N+1}) + 1. \end{array} \right.$$

## Correction de l'exercice 53 page 166

I.1. On note  $x, y$  les deux variables de  $\mathbb{R}^2$ . L'espace  $Q_1$  est l'ensemble des polynômes de la forme  $a + bx + cy + dxy$  avec  $a, b, c, d \in \mathbb{R}$ . On a donc  $\dim Q_1 = 4 = \text{Card } \Sigma_1 = \text{Card } \Sigma_2$  pour montrer que  $(e_1, \Sigma_1, Q_1)$  est un élément fini de Lagrange, il suffit de montrer que  $f \in Q_1$  et  $f|_{\Sigma_1} = 0$  implique  $f = 0$ . Soient donc  $a, b, c, d \in \mathbb{R}$ . On pose  $f(x, y) = a + bx + cy + dxy$  pour  $(x, y) \in e_1$  et on suppose que  $f|_{\Sigma_1} = 0$ , c'est à dire :  $f(-1, 1) = 0, f(0, 1) = 0, f(1, 0) = 0$  et  $f(-1, 0) = 0$ . On a donc :

$$\begin{cases} a - b + c - d = 0 \\ a + c = 0 \\ a + b = 0 \\ a - b = 0 \end{cases}$$

Les deux dernières équations donnent que  $a = b = 0$ , la troisième donne alors que  $c = 0$ , et la première donne enfin que  $d = 0$ . On a donc montré que  $f = 0$ . On en déduit que  $(e_1, \Sigma_1, Q_1)$  est un élément fini de Lagrange. Pour montrer que  $(e_2, \Sigma_2, Q_1)$  est un élément fini de Lagrange, on procède de la même façon : soient  $a, b, c, d \in \mathbb{R}$  et  $f(x, y) = a + bx + cy + dxy$  pour  $(x, y) \in e_2$ . On suppose que  $f|_{\Sigma_2} = 0$ , c'est à dire :  $f(0, 1) = 0, f(2, 1) = 0, f(2, 0) = 0$  et  $f(1, 0) = 0$ . On a donc :

$$\begin{cases} a + c = 0 \\ a + 2b + c + 2d = 0 \\ a + 2b = 0 \\ a + b = 0 \end{cases}$$

Les deux dernières équations donnent  $a = b = 0$ , la première donne alors  $c = 0$  et, finalement, la quatrième donne  $d = 0$ . On a donc montré que  $f = 0$ . On en déduit que  $(e_2, \Sigma_2, Q_1)$  est un élément fini de Lagrange.

I.2. L'espace (de dimension finie) associé à cette discrétisation est engendré par les six fonctions de base globales. On va montrer que la fonction de base associée à  $M_1$  (par exemple) n'est pas dans  $H^1(\Omega)$ . On note  $\phi_1$  cette fonction de base. On doit avoir  $\phi_1|_{e_1} \in Q_1, \phi_1|_{e_2} \in Q_1$  et  $\phi_1(M_1) = 1, \phi_1(M_i) = 0$  si  $i \neq 1$ . On en déduit que  $\phi_1 = 0$  sur  $e_2$  et  $\phi_1(x, y) = -xy$  si  $(x, y) \in e_1$ . On a bien  $\phi_1 \in L^2(\Omega)$  mais on va montrer maintenant que  $\phi_1$  n'a pas de dérivée faible dans  $L^2(\Omega)$  (et donc que  $\phi_1 \notin H^1(\Omega)$ ). On va s'intéresser à la dérivée faible par rapport à  $x$  (mais on pourrait faire un raisonnement similaire pour la dérivée faible par rapport à  $y$ ). On suppose que  $\phi_1$  a une dérivée faible par rapport à  $x$  dans  $L^2(\Omega)$  (et on va montrer que ceci mène à une contradiction). Supposons donc qu'il existe une fonction  $\psi \in L^2(\Omega)$  telle que

$$I = \int_{-1}^2 \int_0^1 \phi_1(x, y) \frac{\partial \varphi}{\partial x}(x, y) dx dy = \int_{-1}^2 \int_0^1 \psi(x, y) \varphi(x, y) dx dy, \text{ pour tout } \varphi \in C_c^\infty(\Omega). \quad (4.44)$$

Soit  $\varphi \in C_c^\infty(\Omega)$ , comme  $\phi_1$  est nulle sur  $e_2$ , on a  $I = \int \int_{e_1} \phi_1(x, y) \frac{\partial \varphi}{\partial x}(x, y) dx dy$  et donc :

$$I = \int_0^1 \left( \int_{-1}^{1-y} (-xy) \frac{\partial \varphi}{\partial x}(x, y) dx \right) dy.$$

Par intégration par parties, en tenant compte du fait que  $\varphi$  est à support compact sur  $\Omega$ , on obtient :

$$\begin{aligned} I &= \int_0^1 \left[ \int_{-1}^{1-y} y \varphi(x, y) dx - (1-y)y\varphi(1-y, y) \right] dy \\ &= \int_0^1 \int_{-1}^2 y 1_{e_1}(x, y) \varphi(x, y) dx - \int_0^1 (1-y)y\varphi(1-y, y) dy. \end{aligned}$$

En posant  $\tilde{\psi}(x, y) = -\psi(x, y) + y 1_{e_1}(x, y)$ , on a  $\tilde{\psi} \in L^2(\Omega)$  et :

$$\int_0^1 (1-y)y\varphi(1-y, y) dy = \int_{-1}^2 \int_0^1 \tilde{\psi}(x, y) \varphi(x, y) dx dy. \quad (4.45)$$

Pour aboutir à une contradiction, on va montrer que (4.45) est fautive pour certains  $\varphi \in C_c^\infty(\Omega)$ . On remarque tout d'abord qu'il existe  $\varphi \in C_c^\infty(\Omega)$  t.q.

$$\int_0^1 (1-y)y\varphi(1-y,y)dy > 0.$$

(Il suffit de choisir  $\varphi \in C_c^\infty(\Omega)$  t.q.  $\varphi \leq 0$  et  $\varphi(1-y,y) > 0$  pour  $y = \frac{1}{2}$ , par exemple.) On se donne maintenant une fonction  $\varphi \in C_c^\infty(\mathbb{R})$  t.q.  $\varphi(0) = 1$  et  $\rho = 0$  sur  $[-1, 1]^c$  et on écrit (4.45) avec  $\varphi_n$  au lieu de  $\varphi$ , où  $\varphi_n$  est définie par :

$$\varphi_n(x, y) = \varphi(x, y)\rho(n(x+y-1))$$

(noter que l'on a bien  $\varphi_n \in C_c^\infty(\Omega)$  car  $\rho \in C^\infty(\mathbb{R})$  et  $\varphi \in C_c^\infty(\Omega)$ ) On a donc

$$\int_0^1 (1-y)y\varphi_n(1-y,y)dy = \int_{-1}^2 \int_0^1 \tilde{\psi}(x,y)\varphi_n(x,y) dx dy.$$

Le terme de gauche de cette égalité est indépendant de  $n$  et non nul car  $\varphi_n(1-y,y) = \varphi(1-y,y)$  pour tout  $n$  et tout  $y \in [0, 1]$ . Le terme de droite tend vers 0 quand  $n \rightarrow \infty$  par convergence dominée car

$$\tilde{\psi}\varphi_n \rightarrow 0 \text{ p.p., et } |\tilde{\psi}\varphi_n| \leq \|\rho\|_\infty |\tilde{\psi}| |\varphi| \in L^1(\Omega).$$

Ceci donne la contradiction désirée et donc que  $\phi_1 \notin H^1(\Omega)$ . L'hypothèse non vérifiée (pour avoir la cohérence globale) est l'hypothèse (4.7). En posant  $S = \bar{e}_1 \cap \bar{e}_2$ , on a

$$\Sigma_1 \cap S = \Sigma_2 \cap S = \{M_2, M_5\},$$

et on a, bien sûr,  $\varphi_1|_S = \varphi_1|_S$  mais on remarque que  $(\{M_2, M_5\}, Q_1|_S)$  n'est pas unisolvant car  $\text{Card}(\{M_2, M_5\}) = 2$  et  $\dim(Q_1|_S) = 3$ .

II.1. Les quatre fonctions de base de  $(e, \Sigma, P)$  sont :

$$\phi_1(x, y) = \frac{1}{4}(x+1)(y+1)$$

$$\phi_2(x, y) = -\frac{1}{4}(x+1)(y-1)$$

$$\phi_3(x, y) = -\frac{1}{4}(x-1)(y+1)$$

$$\phi_4(x, y) = \frac{1}{4}(x-1)(y-1).$$

II.2.

**Construction de  $F_1$**  Pour  $(x, y) \in e$ , on pose

$$F_1(x, y) = M_1\phi_3(x, y) + M_2\phi_1(x, y) + M_5\phi_2(x, y) + M_4\phi_4(x, y),$$

ce qui donne

$$4F_1(x, y) = \begin{pmatrix} -1 \\ 1 \end{pmatrix} (1-x)(1+y) + \begin{pmatrix} 0 \\ 1 \end{pmatrix} (1+x)(1+y) + \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1+x)(1-y) + \begin{pmatrix} -1 \\ 0 \end{pmatrix} (1-x)(1-y)$$

et donc

$$4F_1(x, y) = \begin{pmatrix} -1 + 3x - y - xy \\ 2(1+y) \end{pmatrix}.$$

Pour  $y \in [-1, 1]$  fixé, la première composante de  $F_1(x, y)$  est linéaire par rapport à  $x$  et  $F_1(\cdot, y)$  est une bijection de  $[-1, 1] \times \{y\}$  dans  $[-1, \frac{1-y}{2}] \times \{\frac{1+y}{2}\}$ . On en déduit que  $F_1$  est une bijection de  $e$  dans  $e_1$ .

**Construction de  $F_2$**  Pour  $(x, y) \in e$ , on pose

$$F_2(x, y) = M_2\phi_3(x, y) + M_3\phi_1(x, y) + M_6\phi_2(x, y) + M_5\phi_4(x, y),$$

ce qui donne

$$4F_2(x, y) = \binom{0}{1} (1-x)(1+y) + \binom{2}{1} (1+x)(1+y) + \binom{2}{0} (1+x)(1-y) + \binom{1}{0} (1-x)(1-y)$$

et donc  $4F_2(x, y) = \begin{pmatrix} 5+3x-y+xy \\ 2+2y \end{pmatrix}$  Pour  $y \in [-1, 1]$  fixé, la première composante de  $F_2(x, y)$  est linéaire par rapport à  $x$  et  $F_2(\cdot, y)$  est une bijection de  $[-1, 1] \times \{y\}$  dans  $[\frac{1-y}{2}, 2] \times \{\frac{1+y}{2}\}$  On en déduit que  $F_2$  est une bijection de  $e$  dans  $e_2$ . Les fonctions  $F_1$  et  $F_2$  ne sont pas affines.

II.3. Les éléments  $(e_1, \Sigma_1, P_{e_1})$  et  $(e_2, \Sigma_2, P_{e_2})$  sont les éléments finis de Lagrange construits à partir de l'élément fini de Lagrange  $(e, \Sigma, P)$  et des bijections  $F_1$  et  $F_2$  (de  $e$  dans  $e_1$  et de  $e$  dans  $e_2$ ), voir la proposition 4.10 page 141. Pour montrer que l'espace vectoriel construit avec  $(e_1, \Sigma_1, P_{e_1})$  et  $(e_2, \Sigma_2, P_{e_2})$  est inclus dans  $H^1(\Omega)$ , il suffit de vérifier la propriété de "cohérence globale" donnée dans la proposition 4.11 page 142. On pose

$$\begin{aligned} S = \bar{e}_1 \cap \bar{e}_2 &= \{(x, y) \in \bar{\Omega}, x + y = 1\} \\ &= \{(1-y, y), y \in [0, 1]\} \end{aligned}$$

On remarque tout d'abord que  $\Sigma_1 \cap S = \Sigma_2 \cap S = \{M_2, M_5\}$ . On détermine maintenant  $P_{e_1|_S}$  et  $P_{e_2|_S}$ . Soit  $f \in P_{e_1}$ . Soit  $(x, y) \in S$  (c'est à dire  $y \in [0, 1]$  et  $x + y = 1$ ) on a  $f(x, y) = f \circ F_1(1, 2y - 1)$ . (On a utilisé ici le fait que  $F_1(\{1\} \times [-1, 1]) = 5$ ). Donc  $P_{e_1|_S}$  est l'ensemble des fonctions de  $S$  dans  $\mathbb{R}$  de la forme :  $(x, y) \mapsto g(1, 2y - 1)$ , où  $g \in Q_1$ , c'est à dire l'ensemble des fonctions de  $S$  dans  $\mathbb{R}$  de la forme :

$$(x, y) \mapsto \alpha + \beta + \gamma(2y - 1) + \delta(2y - 1),$$

avec  $\alpha, \beta, \gamma$ , et  $\delta \in \mathbb{R}$ . On en déduit que  $P_{e_1|_S}$  est l'ensemble des fonctions de  $S$  dans  $\mathbb{R}$  de la forme  $(x, y) \mapsto a + by$  avec  $a, b \in \mathbb{R}$ . On a donc  $P_{e_1|_S} = P_{e_2|_S}$ . Ceci donne la condition (4.6) page 142. Enfin, la condition (4.7) est bien vérifiée, c'est à dire  $(\Sigma_1, P_{e_1}|_S)$  est unisolvant, car un élément de  $P_{e_1|_S}$  est bien déterminé de manière unique par ses valeurs en  $(0, 1)$  et  $(1, 0)$ .

### Correction de l'exercice 54 page 166(Eléments affine-équivalents)

Si les fonctions de base de  $(\bar{K}, \bar{\Sigma}, \bar{P})$  sont affines, alors l'espace  $\bar{P}$  est constitué des fonctions affines, on peut donc écrire.

$$\bar{P} = \{\bar{f} : \bar{K} \rightarrow \mathbb{R}, \bar{x} = (\bar{x}_1, \bar{x}_2) \mapsto \bar{f}(\bar{x}) = a_1\bar{x}_1 + a_2\bar{x}_2 + b\}.$$

Comme  $(\bar{K}, \bar{\Sigma}, \bar{P})$  et  $((K, \Sigma, P))$  sont affines équivalents, on a par définition :

$$P = \{f : K \rightarrow \mathbb{R}; f = \bar{f} \circ F^{-1}, \bar{f} \in \bar{P}\},$$

où  $F$  est une fonction affine de  $\bar{K}$  dans  $K$  la fonction  $F^{-1}$  est donc aussi affine et s'écrit donc sous la forme :

$$F^{-1}(x) = F^{-1}((x_1, x_2)) = (\alpha_1 x_1 + \alpha_2 x_2 + \gamma, \beta_1 x_1 + \beta_2 x_2 + \delta)$$

Donc si  $f = \bar{f} \circ F^{-1} \in P$ , on a

$$\begin{aligned} f(x) &= \bar{f} \circ F^{-1}((x_1, x_2)) \\ &= \bar{f}[(\alpha_1 x_1 + \alpha_2 x_2 + \gamma, \beta_1 x_1 + \beta_2 x_2 + \delta)] \\ &= A_1 x_1 + A_2 x_2 + B \end{aligned}$$

où  $A_1, A_2$  et  $B \in \mathbb{R}^2$ . On en déduit que  $f$  est bien affine. L'espace  $P$  est donc constitué de fonctions affines. Pour montrer que les fonctions de base locales sont affines, il suffit de montrer que l'espace  $P$  est constitué de toutes les fonctions affines. En effet, si  $f$  est affine, i.e.  $f(x_1, x_2) = A_1 x_1 + A_2 x_2 + B$ , avec  $A_1, A_2, B \in \mathbb{R}^2$ , on montre facilement que  $\bar{f} : f \circ F \in \bar{P}$ , ce qui montre que  $f \in P$ .



**Exercice 56 page 166**

La formulation faible du problème s'écrit :

$$\begin{cases} \int_D \nabla u(x) \nabla v(x) dx = \int_D f(x) v(x) dx, \forall v \in H_0^1(\Omega) \\ u \in H_0^1(\Omega) \end{cases}$$

On note  $I = \{(k, \ell), 1 \leq k \leq M, 1 \leq \ell \leq N\}$  noter que  $\text{Card } I = MN$ . L'espace vectoriel de dimension finie dans lequel on cherche la solution approchée (en utilisant les éléments finis suggérés par l'énoncé) est donc  $H = \text{Vect } \{\phi_i, i \in I\}$ , où  $\phi_i$  est la fonction de base globale associée au noeud  $i$ . Cette solution approchée s'écrit  $u = \sum_{j \in I} u_j \phi_j$  où la famille  $\{u_j, j \in I\}$  est solution du système linéaire :

$$\sum_{j \in I} a_{ij} u_j = b_i, \forall i \in I \quad (4.46)$$

avec  $b_i = \int_D f(x, y) \phi_i(x, y) dx dy$ , pour tout  $i \in I$  et  $a_{ij} = \int_D \nabla \phi_i(x, y) \nabla \phi_j(x, y) dx dy$ , pour tout  $i, j \in I$ .

La matrice de ce système linéaire est donc donnée par le calcul de  $a_{ij}$  pour  $i, j \in I$  et un ordre de numérotation des inconnues, plus précisément, soit  $\varphi : I \rightarrow \{1, \dots, MN\}$  bijective. On note  $\psi$  la fonction réciproque de  $\varphi$ . Le système (4.46) peut alors s'écrire :

$$\sum_{n=1}^{MN} a_{i, \psi(n)} u_{\psi(n)} = b_i, \forall i \in I$$

ou encore :

$$\sum_{n=1}^{MN} a_{\psi(m), \psi(n)} u_{\psi(n)} = b_{\psi(m)}, \forall m \in \{1, \dots, MN\},$$

$\{u_j, j \in I\}$  est donc solution de (4.46) si et seulement si  $u_{\psi(n)} = \lambda_n$  pour tout  $n \in \{1, \dots, MN\}$  où  $\lambda = (\lambda_1, \dots, \lambda_{MN}) \in \mathbb{R}^{MN}$  est solution du système linéaire :

$$A\lambda = C$$

avec  $C = (C_1, \dots, C_{MN})$ ,  $C_m = b_{\psi(m)}$  pour tout  $m \in \{1, \dots, MN\}$  et  $A = (A_{m,n})_{m,n=1}^{MN} \in \mathbb{R}^{MN}$  avec  $A_{m,n} = a_{\psi(m), \psi(n)}$  pour tout  $m, n \in \{1, \dots, MN\}$ . Il reste donc à calculer  $a_{ij}$  pour  $i, j \in I$ . Un examen de support des fonctions  $\phi_i$  et  $\phi_j$  et le fait que le maillage soit à pas constant nous montrent que seuls 4 nombres différents peuvent apparaître dans la matrice :

1.  $i = j$ . On pose alors  $a_{ii} = \alpha$ .
2.  $i = (k, \ell), j = (k \pm 1, \ell)$ . On pose alors  $a_{ij} = \beta$ .
3.  $i = (k, \ell), j = (k, \ell \pm 1)$ . On pose alors  $a_{ij} = \gamma$ .
4.  $i = (k, \ell), j = (k + 1, \ell + 1)$  ou  $(k - 1, \ell - 1)$ . On pose alors  $a_{ij} = \delta$ .

En dehors des quatre cas décrits ci-dessus, on a nécessairement  $a_{ij} = 0$  (car les supports de  $\phi_i$  et  $\phi_j$  sont disjoints). Calculons maintenant  $\alpha, \beta, \gamma$  et  $\delta$ .

**Calcul de  $\beta$**  On prend ici  $i = (k, \ell)$  et  $j = (k + 1, \ell)$  On calcule tout d'abord  $\int_{T^0} \nabla \phi_i - \nabla \phi_j dx$  avec  $T^0 = T_{k+\frac{1}{2}, j+\frac{1}{2}}^0$ . Un argument d'invariance par translation permet de supposer que  $x_k = y_\ell = 0$ . On a alors

$$\phi_i(x, y) = \frac{\Delta x - x}{\Delta x} \text{ et } \phi_j(x, y) = \frac{x \Delta y - y \Delta x}{\Delta x \Delta y},$$

de sorte que

$$\nabla \phi_i(x, y) - \nabla \phi_j(x, y) = - \left( \frac{1}{\Delta x} \right)^2.$$

On a donc

$$\int_{T^0} \nabla \phi_i - \nabla \phi_j dx = - \left( \frac{1}{\Delta x} \right)^2 \frac{\Delta x \Delta y}{2} = - \frac{\Delta y}{2 \Delta x}$$

Un calcul similaire donne l'intégrale de  $\nabla\phi_i \cdot \nabla\phi_j$  sur le deuxième triangle commun aux supports de  $\phi_i$  et  $\phi_j$ . Sur ce deuxième triangle, formé par les points  $(k, \ell)$ ,  $(k+1, \ell)$  et  $(k, \ell-1)$ , noté  $T^2$ , on a

$$\phi_i(x, y) = 1 - \frac{x \Delta y - y \Delta x}{\Delta x \Delta y} \text{ et } \phi_j(x, y) = \frac{x}{\Delta x},$$

de sorte que

$$\nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) = -\left(\frac{1}{\Delta x}\right)^2 \text{ et } \int_{T^2} \nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) dx dy = -\left(\frac{1}{\Delta x}\right)^2 \frac{\Delta x \Delta y}{2} = -\frac{\Delta y}{2\Delta x}.$$

On a donc, finalement,

$$\beta = \int_D \nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) dx dy = -\frac{\Delta y}{\Delta x}.$$

**Calcul de  $\gamma$**  Le calcul de  $\gamma$  est le même que celui de  $\beta$  en changeant les rôles de  $\Delta x$  et  $\Delta y$ , on obtient donc

$$\gamma = -\frac{\Delta x}{\Delta y}$$

**Calcul de  $\delta$**  On prend ici  $i = (k, \ell)$  et  $j = (k+1, \ell+1)$ . On a donc, en notant  $T^0 = T_{k+\frac{1}{2}, \ell+\frac{1}{2}}^0$  et  $T^1 = T_{k+\frac{1}{2}, \ell+\frac{1}{2}}^1$ ,

$$\delta = \int_{T^0} \nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) dx dy + \int_{T^1} \nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) dx dy.$$

On peut supposer (par translation) que  $x_k = 0 = y_\ell$ . Sur  $T_1$ , on a alors  $\phi_i(x, y) = \frac{\Delta y - y}{\Delta y}$  et  $\phi_j(x, y) = \frac{x}{\Delta x}$  de sorte que  $\int_{T^1} \nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) dx dy = 0$  (car  $\nabla\phi_i \cdot \nabla\phi_j = 0$ ). En changeant les rôles de  $x$  et  $y$ , on a aussi  $\int_{T^0} \nabla\phi_i(x, y) \cdot \nabla\phi_j(x, y) dx dy = 0$ . On a donc  $\delta = 0$ .

**Calcul de  $\alpha$**  On prend ici  $i = j = (k, \ell)$ . On peut toujours supposer que  $x_k = y_\ell = 0$ . En reprenant les notations précédentes, on a, par raison de symétrie :

$$\alpha = \int_D \nabla\phi_i - \nabla\phi_i dx = 2 \int_{T^0} |\nabla\phi_i|^2(x, y) dx dy + 2 \int_{T^1} |\nabla\phi_i|^2(x, y) dx dy + 2 \int_{T^2} |\nabla\phi_i|^2(x, y) dx dy.$$

Sur  $T^0$ , on a  $\phi_i(x, y) = \frac{\Delta x - x}{\Delta x}$  et donc

$$\int_{T^0} |\nabla\phi_i|^2(x, y) dx dy = \left(\frac{1}{\Delta x}\right)^2 \frac{\Delta x \Delta y}{2} = \frac{1}{2} \frac{\Delta y}{\Delta x}$$

Sur  $T_1$ , on a  $\phi_1(x, y) = \frac{\Delta y - y}{\Delta y}$  et donc

$$\int_{T^2} |\nabla\phi_i|^2(x, y) dx dy = \left[ \left(\frac{1}{\Delta x}\right)^2 + \left(\frac{1}{\Delta y}\right)^2 \right] \frac{\Delta x \Delta y}{2} = \frac{1}{2} \frac{\Delta y}{\Delta x} + \frac{1}{2} \frac{\Delta x}{\Delta y}$$

On en déduit

$$\alpha = 2 \frac{\Delta x}{\Delta y} + 2 \frac{\Delta y}{\Delta x}.$$

**Exercice 57 page 167**

1. Soit  $K$  le triangle de référence, de sommets  $(0,0)$ ,  $(1,0)$  et  $(0,1)$ . On veut montrer que si  $p$  est un polynôme de degré 1, alors

$$\int \int_K p(x, y) dx dy = \int \int_K dx dy p(x_G, y_G) \quad (4.47)$$

où  $(x_G, y_G)$  est le centre de gravité de  $K$ . Comme  $K$  est le triangle de sommets  $(0,0)$ ,  $(1,0)$  et  $(0,1)$ , on a  $x_G = y_G = \frac{1}{3}$ . Pour montrer (4.47), on va le montrer pour  $p \equiv 1$ , pour  $p(x, y) = x$  et pour  $p(x, y) = y$ . On a

$$\int \int_K dx dy = \int_0^1 \int_0^{1-x} dy dx = \frac{1}{2}.$$

On a donc bien (4.47) si  $p \equiv 1$ . Et

$$\int \int_K x dx dy = \int_0^1 x \int_0^{1-x} dy dx = \int_0^1 (x - x^2) dx = \frac{1}{6}$$

Or si  $p(x, y) = x$ , on a  $p(x_G, y_G) = \frac{1}{3}$ , et donc on a encore bien (4.47). Le calcul de  $\int \int_K y dx dy$  est identique ; on a donc bien montré que l'intégration numérique à un point de Gauss est exacte pour les polynômes d'ordre 1.

2. On veut montrer que pour tout polynôme  $p$  de degré 2, on a :

$$\int \int_K p(x, y) dx dy = L(p), \text{ où on a posé } L(p) = \frac{1}{6} \left( p\left(\frac{1}{2}, 0\right) + p\left(\frac{1}{2}, \frac{1}{2}\right) + p\left(0, \frac{1}{2}\right) \right) \quad (4.48)$$

On va démontrer que (4.48) est vérifié pour tous les monômes de  $P_2$ . Si  $p \equiv 1$ , on a  $L(p) = \frac{1}{2}$ , et (4.48) est bien vérifiée. Si  $p(x, y) = x$ , on a  $L(p) = \frac{1}{6}$ , et on a vu à la question 1 que  $\int \int_K x dx dy = \frac{1}{6}$ . On a donc bien (4.48).

Par symétrie, si  $p(x, y) = y$  vérifie aussi (4.48). Calculons maintenant  $I = \int \int_K xy dx dy = \int_0^1 \int_0^{1-x} xy dy dx$ . On a donc

$$I = \int_0^1 \frac{(1-x)^2}{2} dx = \frac{1}{2} \int_0^1 (x - 2x^2 + x^3) dx = \frac{1}{24},$$

et si  $p(x, y) = xy$ , on a bien :  $L(p) = \frac{1}{6} \times \frac{1}{4}$ . Donc (4.48) est bien vérifiée. Il reste à vérifier que (4.48) est vérifiée pour  $p(x, y) = x^2$  (ou  $p(x, y) = y^2$ , par symétrie). Or,  $J = \int \int_K x^2 dx dy = \int_0^1 x^2 \int_0^{1-x} dy dx = \int_0^1 (x^2 - x^3) dx$ . Donc  $J = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$ . Et pour  $p(x, y) = x^2$ , on a bien :  $L(p) = \frac{1}{6} \left( \frac{1}{4} + \frac{1}{4} \right) = \frac{1}{12}$ .

**Exercice 58 page 167**

I. Comme  $p \in P_2$ ,  $p$  est de la forme :  $p(x, y) = a + bx + cy + dxy + \alpha x^2 + \beta y^2$ , on a par développement de Taylor (exact car  $p''' = 0$ ) :

$$2p(a_9) - p(a_6) - p(a_8) = p_{xx}(a_9) = \alpha$$

$$2p(a_9) - p(a_5) - p(a_7) = p_{yy}(a_9) = \beta$$

d'où on déduit que

$$4p(a_9) - \sum_{i=5}^8 p(a_i) = \alpha + \beta. \quad (4.49)$$

De même, on a :

$$2p(a_5) - p(a_1) - p(a_2) = \alpha$$

$$2p(a_7) - p(a_3) - p(a_4) = \alpha$$

$$2p(a_6) - p(a_2) - p(a_3) = \beta$$

$$2p(a_8) - p(a_1) - p(a_4) = \beta.$$

Ces quatre dernières égalités entraînent :

$$\sum_{i=5}^8 p(a_i) - \sum_{i=1}^4 p(a_i) = \alpha + \beta \quad (4.50)$$

De (4.49) et (4.50), on déduit que :

$$\sum_{i=1}^4 p(a_i) - 2 \sum_{i=5}^8 p(a_i) + 4p(a_9) = 0.$$

2. La question précédente nous suggère de choisir  $\phi : Q_2 \rightarrow \mathbb{R}$  définie par

$$\phi(p) = \sum_{i=1}^4 p(a_i) - 2 \sum_{i=5}^8 p(a_i) + 4p(a_9).$$

Soit  $p \in \mathcal{P}$  tel que  $p(a_i) = 0$ ,  $i = 1, \dots, 8$ . Comme  $p \in Q_2$ ,  $p$  est une combinaison linéaire des fonctions de base  $\varphi_1, \dots, \varphi_9$ , associées aux noeuds  $a_1, \dots, a_9$ , et comme  $p(a_i) = 0$ ,  $i = 1, \dots, 8$ , on en déduit que  $p = \alpha\varphi_9$ ,  $\alpha \in \mathbb{R}$ . On a donc  $\phi(p) = \alpha\phi(\varphi_9) = 4\alpha = 0$ , ce qui entraîne  $\alpha = 0$ . On a donc  $p = 0$ .

3. Calculons les fonctions de base  $\varphi_1^*, \dots, \varphi_8^*$  associées aux noeuds  $a_1, \dots, a_8$  qui définissent  $\Sigma$ . On veut que  $\varphi_i^* \in \mathcal{P}$  et  $\varphi_i^*(a_j) = \delta_{ij}$  pour  $i, j = 1, \dots, 8$ . Or  $\varphi_9(a_j) = 0 \quad \forall i = 1, \dots, 8$ , et  $\phi(\varphi_9) = 4$ . Remarquons alors que pour  $i = 1$ , à 4 on a

$$p(\varphi_i) = 1, \text{ et donc si } \varphi_i^* = \varphi_i - \frac{1}{4} \varphi_9,$$

on a  $p(\varphi_i^*) = 0$  et  $\varphi_i^*(a_j) = \delta_{ij}$  pour  $j = 1, \dots, 8$ . De même, pour  $i = 5$  à 8, on a  $p(\varphi_i) = -2$ , et donc si  $\varphi_i^* = \varphi_i + \frac{1}{2}\varphi_9$ , on a  $p(\varphi_i^*) = 0$  et  $\varphi_i^*(a_j) = \delta_{ij}$ , pour  $j = 1, \dots, 8$ . On a ainsi trouvé les fonctions de base de l'élément fini  $(C, \mathcal{P}, \Sigma)$ . Notons que cet élément fini n'est autre que l'élément fini  $(C, Q_2^*, \Sigma)$  vu en cours (voir paragraphe 4.2.3 page 147 et que  $\text{Ker } \phi = \mathcal{P} = Q_2^*$ .

# Chapitre 5

## Problèmes hyperboliques

### 5.1 Une équation de transport

L'exemple type d'équation hyperbolique est l'équation de transport. Supposons par exemple, qu'on connaisse l'emplacement d'une nappe de pétrole due au dégazement intempestif d'un supertanker au large des côtes, et qu'on cherche à prévoir son déplacement dans les heures à venir, par exemple pour la mise en oeuvre efficace de barrages. On suppose connu  $v : \mathbb{R}^2 \times \mathbb{R}_+ \rightarrow \mathbb{R}^2$ , le champ des vecteurs vitesse des courants marins, qui dépend de la variable d'espace  $x$  et du temps  $t$ ; ce champ de vecteurs est donné par exemple par la table des marées (des exemples de telles cartes de courants sont données en Figure 5.1). A  $t = 0$ , on connaît  $\rho_0(x)$  : la densité

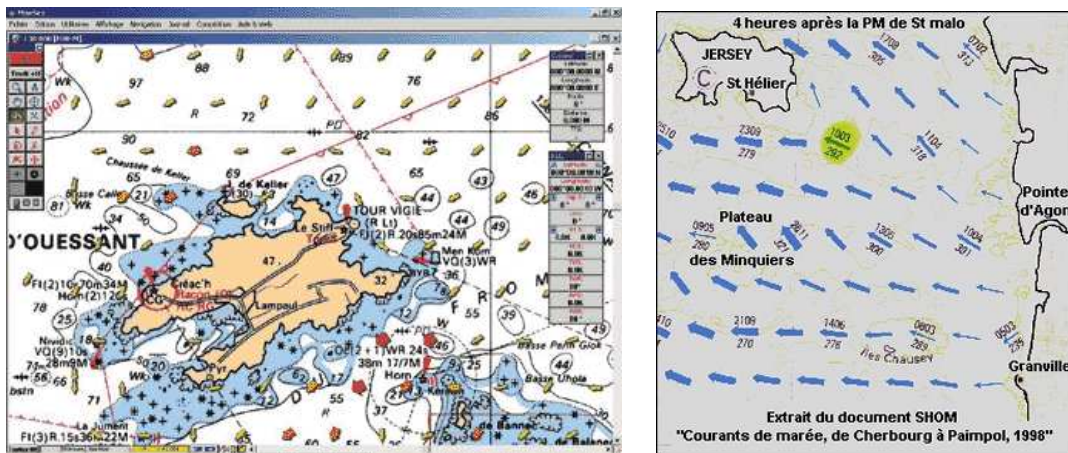


FIG. 5.1 – Exemples de cartes de courants marins au large de côtes de Bretagne (source : SHOM)

d'hydrocarbure initiale, et on cherche à calculer  $\rho(x, t)$  = densité de d'hydrocarbure au point  $x$  et au temps  $t$ . On écrit alors l'équation de conservation de la masse :

$$\rho_t + \text{div}(\rho v) = 0, \tag{5.1}$$

$$\rho_0(x) = \begin{cases} 1 & x \in A, \\ 0 & x \in A^c, \end{cases} \tag{5.2}$$

où  $A$  représente le lieu initial de la nappe de pétrole. Dans le cas d'un déplacement maritime, le vecteur  $v : \mathbb{R}^2 \times \mathbb{R}_+ \rightarrow \mathbb{R}^2$ , n'est évidemment pas constant (la marée n'est pas la même à Brest qu'à Saint Malo). De plus le déplacement de la nappe dépend également du vent, qui affecte donc le vecteur  $v$ . On supposera pourtant ici, pour simplifier l'exposé, que  $v$  soit constant en espace et en temps. Alors le problème (5.1) - (5.2) admet comme solution :

$$\rho(x, t) = \rho_0(x - vt), \tag{5.3}$$

qui exprime le transport de la nappe à la distance  $vt$  du point de départ dans la direction de  $V$ , au temps  $t$ . En fait, il est clair que (5.3) n'est pas une solution "classique" de puisque  $\rho$  n'est pas continue, et que ces dérivées en temps

ne sont donc pas définies au sens habituel. Nous verrons par la suite comment on peut donner une formulation correcte des solutions de (5.1) - (5.2). Notons que les systèmes d'équations hyperboliques sont très importants en mécanique des fluides ; les équations d'Euler, par exemple sont utilisées pour modéliser l'écoulement de l'air autour d'une aile d'avion. Dans le cadre de ce cours, nous n'étudierons que le cas des équations scalaires. Par souci de simplicité, nous n'aborderons dans le cadre de ce cours que les problèmes posés en une dimension d'espace, tout d'abord dans le cas relativement simple d'une équation linéaire (section 5.2 page 180, puis dans le cas d'une équation non linéaire (section 5.4 page 186. Par souci de clarté, les schémas numériques seront introduits pour l'équation linéaire  $u_t + u_x = 0$  (section 5.3 page 183). On donnera ensuite quelques schémas pour les équations hyperboliques non linéaires (section 5.5 page 194).

## 5.2 Equation linéaire, cas 1D

Commençons par étudier le cas d'une équation hyperbolique linéaire :

$$\begin{cases} u_t + cu_x = 0, & x \in \mathbb{R}, t > 0, \\ u(x, 0) = u_0(x), & x \in \mathbb{R}. \end{cases} \quad (5.4)$$

où la vitesse de transport  $c \in \mathbb{R}$  et la condition initiale  $u_0 : \mathbb{R} \rightarrow \mathbb{R}$  sont données. Le problème (5.4) s'appelle "problème de Cauchy". On cherche  $u : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$ , solution de ce problème. Nous commençons par une étude succincte du problème continu, pour lequel on peut trouver une solution exacte explicite.

### Solution classique et solution faible

**Définition 5.1 (Solution classique)** On dit qu'une fonction  $u : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$  est solution classique du problème (5.4) si  $u \in C^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$  et  $u$  vérifie (5.4).

Une condition nécessaire pour avoir une solution classique est que  $u_0 \in C^1(\mathbb{R})$ .

**Théorème 5.2** Si  $u_0 \in C^1(\mathbb{R})$ , alors il existe une unique solution classique du problème (5.4), qui s'écrit  $u(x, t) = u_0(x - ct)$ .

**Démonstration :** Pour montrer l'existence de la solution, il suffit de remarquer que  $u$  définie par (5.1) est de classe  $C^1$ , et que  $u_t + cu_x = 0$  en tout point. Pour montrer l'unicité de la solution, on va introduire la notion de caractéristique, qui est d'ailleurs aussi fort utile dans le cadre de la résolution numérique. Soit  $u$  solution classique de (5.4). On appelle droite caractéristique de (5.4) issue de  $x_0$  la droite d'équation  $x(t) = ct + x_0$ , qui est illustrée sur la figure 5.2. Montrons que si  $u$  est solution de (5.4), alors  $u$  est constante sur la droite  $\mathcal{D}_{x_0}$ , pour tout  $x_0 \in \mathbb{R}$ . Soit  $x_0 \in \mathbb{R}$ , et  $\varphi_{x_0}$  la fonction de  $\mathbb{R}_+$  dans  $\mathbb{R}$  définie par  $\varphi_{x_0}(t) = u(x_0 + ct, t)$ . Dérivons  $\varphi_{x_0}$  par rapport au

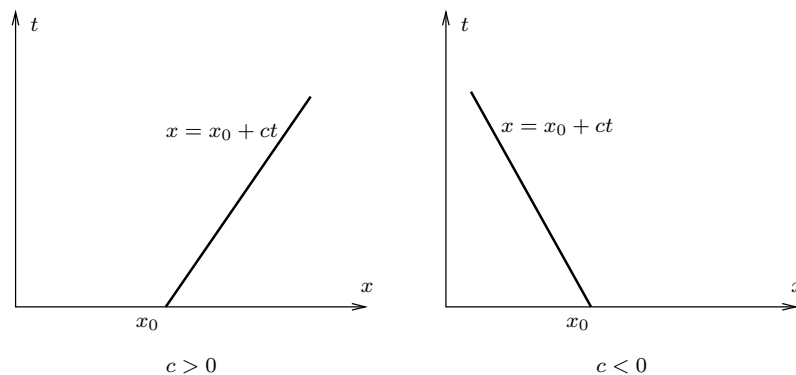


FIG. 5.2 – Droites caractéristiques, cas linéaire

temps :

$$\begin{aligned} \varphi'_{x_0}(t) &= cu_x(x_0 + ct, t) + u_t(x_0 + ct, t) \\ &= (u_t + cu_x)(x_0 + ct, t) = 0, \end{aligned}$$

car  $u$  est solution de (5.4). On en déduit que

$$\varphi_{x_0}(t) = \varphi_{x_0}(0) = u_0(x_0), \forall t \in \mathbb{R}_+.$$

On a donc  $u(x_0 + ct, t) = u_0(x_0), \forall x_0 \in \mathbb{R}$ , donc  $u$  est constante sur la droite caractéristique  $\mathcal{D}_{x_0}$ , et en posant  $x = x_0 + ct$  :

$$u(x, t) = u_0(x - ct),$$

ce qui prouve l'existence et l'unicité de (5.4). ■

**Remarque 5.3 (Terme source)** *Le modèle physique peut amener à une équation avec terme source au second membre  $f \in C(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$  :*

$$\begin{cases} u_t + cu_x = f(x, t), \\ u(x, 0) = u_0(x), \end{cases} \quad (5.5)$$

et  $u_0 \in C(\mathbb{R})$ . Ceci peut modéliser un dégazage sur un temps plus long, comme dans le cas du Prestige sur les côtes de Galice en 2003 par exemple. Pour montrer l'unicité de la solution de (5.5), on suppose que  $u$  est solution classique et on pose :  $\varphi_{x_0}(t) = u(x_0 + ct, t)$ . Par un calcul identique au précédent, on a

$$\varphi'_{x_0}(t) = f(x_0 + ct, t).$$

Donc  $\varphi_{x_0}(t) = \varphi_{x_0}(0) + \int_0^t f(x_0 + cs, s)ds$  On en déduit que :

$$u(x_0 + ct, t) = \varphi_{x_0}(0) + \int_0^t f(x_0 + cs, s)ds.$$

on pose alors :  $x = x_0 + ct$ , et on obtient :

$$u(x, t) = u_0(x - ct) + \int_0^t f(x - c(t - s), s)ds,$$

ce qui prouve l'unicité. On obtient alors l'existence en remarquant que la fonction  $u(x, t)$  ainsi définie est effectivement solution de (5.5), car elle est de classe  $C^1$  et elle vérifie  $u_t + cu_x = f$ .

Dans ce qui précède, on a fortement utilisé le fait que  $u_0$  est  $C^1$ . Ce n'est largement pas toujours le cas dans la réalité. Que faire si, par exemple,  $u_0 \in L^\infty(\mathbb{R})$  ?

**Définition 5.4 (Solution faible)** *On dit que  $u$  est solution faible de (5.4) si  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$  et  $u$  vérifie :*

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} [u(x, t)\varphi_t(x, t) + cu(x, t)\varphi_x(x, t)] dt dx + \int_{\mathbb{R}} u_0(x)\varphi(x, 0)dx = 0, \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}). \quad (5.6)$$

Notons que dans la définition ci-dessus, on note  $\mathbb{R}_+ = [0, +\infty[$ , et  $C_c^1(\mathbb{R} \times \mathbb{R}_+)$  l'ensemble des restrictions à  $\mathbb{R} \times \mathbb{R}_+$  des fonctions  $C_c^1(\mathbb{R} \times \mathbb{R})$ . On insiste sur le fait qu'on peut donc avoir  $\varphi(x, 0) \neq 0$ . Voyons maintenant les liens entre solution classique et solution faible.

**Proposition 5.5** *Si  $u$  est solution classique de (5.4) alors  $u$  est solution faible. Réciproquement, si  $u \in C^1(\mathbb{R} \times ]0, +\infty[) \cap C(\mathbb{R} \times [0, +\infty[)$  est solution classique de (5.17) alors  $u$  est solution forte de (5.4).*

La démonstration de cette proposition est effectuée dans le cadre plus général des équations hyperboliques non linéaires (voir Proposition 5.20). Notons que si on prend  $\varphi \in C_c^1(\mathbb{R} \times ]0, +\infty[, \mathbb{R})$  au lieu de  $\varphi \in C_c^1(\mathbb{R} \times [0, +\infty[, \mathbb{R})$  dans (5.6), on obtient :

$$u_t + cu_x = 0,$$

mais on ne récupère pas la condition initiale. Il est donc essentiel de prendre des fonctions test dans  $C_c^1(\mathbb{R} \times [0, +\infty[, \mathbb{R})$ .

**Théorème 5.6 (Existence et unicité de la solution faible)** *Si  $u_0 \in L_{loc}^\infty(\mathbb{R})$ , il existe une unique fonction  $u$  solution faible de (5.4).*

**Démonstration :** On va montrer que  $u(x, t) = u_0(x - ct)$  est solution faible. Comme  $u_0 \in L_{loc}^\infty(\mathbb{R})$ , on a  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$ . Soit  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$ , on veut montrer que :

$$\int \int_{\mathbb{R} \times \mathbb{R}_+} u(x, t) \varphi_t(x, t) dx dt + \int \int_{\mathbb{R} \times \mathbb{R}_+} cu(x, t) \varphi_x(x, t) dx dt + \int_{\mathbb{R}} u_0(x) \varphi(x, 0) dx = 0.$$

Posons

$$A = \int \int_{\mathbb{R} \times \mathbb{R}_+} u(x, t) \varphi_t(x, t) dx dt + \int \int_{\mathbb{R} \times \mathbb{R}_+} cu(x, t) \varphi_x(x, t) dx dt.$$

Si  $u(x, t) = u_0(x - ct)$ , on a donc :

$$A = \int \int_{\mathbb{R} \times \mathbb{R}_+} [u_0(x - ct) \varphi_t(x, t) + cu_0(x - ct) \varphi_x(x, t)] dx dt.$$

En appliquant le changement de variable  $y = x - ct$  et en utilisant le théorème de Fubini, on obtient :

$$A = \int_{\mathbb{R}} u_0(y) \int_{\mathbb{R}_+} [\varphi_t(y + ct, t) + c\varphi_x(y + ct, t)] dt dy.$$

Posons alors

$$\psi_y(t) = \varphi(y + ct, t).$$

On a donc :

$$A = \int_{\mathbb{R}} \left( u_0(y) \int_0^{+\infty} \psi'_y(t) dt \right) dy,$$

et comme  $\psi$  est à support compact sur  $[0, +\infty[$ ,

$$A = - \int_{\mathbb{R}} u_0(y) \psi_y(0) dy,$$

donc finalement :

$$A = - \int_{\mathbb{R}} u_0(y) \varphi(y, 0) dy.$$

On a ainsi démontré que la fonction  $u$  définie par  $u(x, t) = u_0(x - ct)$  est solution faible de l'équation (5.4). On a donc existence d'une solution faible. Montrons maintenant que celle-ci est unique. Soient  $u$  et  $v$  deux solutions faibles de (5.4). On pose  $w = u - v$  et on va montrer que  $w = 0$ . Par définition,  $w$  satisfait :

$$\int \int_{\mathbb{R} \times \mathbb{R}_+} w(x, t) (\varphi_t(x, t) + c\varphi_x(x, t)) dx dt = 0, \quad \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}) \quad (5.7)$$

Par le lemme 5.7 donné ci-dessous, pour toute fonction  $f \in C_c^\infty(\mathbb{R} \times \mathbb{R}_+^*, \mathbb{R})$  il existe  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$ , telle que  $\varphi_t + c\varphi_x = f$ , et on a donc par (5.7) :

$$\int \int_{\mathbb{R} \times \mathbb{R}_+} w(x, t) f(x, t) dx dt = 0, \quad \forall f \in C_c(\mathbb{R} \times \mathbb{R}_+^*, \mathbb{R})$$

Ceci entraîne que  $w = 0$  p.p. ■

**Lemme 5.7 (Résultat d'existence)** Soit  $f \in C_c(\mathbb{R} \times \mathbb{R}_+^*, \mathbb{R})$ , alors il existe

$$\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}) \text{ telle que } \varphi_t + c\varphi_x = f$$

**Démonstration :** Soit  $f \in C_c(\mathbb{R} \times \mathbb{R}_+^*, \mathbb{R})$ , et  $T > 0$  tel que  $f(x, t) = 0$  si  $t \geq T$ . On considère le problème :

$$\begin{cases} \varphi_t + c\varphi_x = f \\ \varphi(x, T) = 0 \end{cases} \quad (5.8)$$

On vérifie facilement que le problème (5.8) admet une solution classique

$$\varphi(x, t) = - \int_t^T f(x - c(s - t), s) ds$$

En effet, avec ce choix de  $\varphi$ , on a effectivement

$$\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}) \text{ et } \varphi_t + c\varphi_x = f.$$

De plus, comme  $f$  est à support compact,  $\varphi$  est à support compact. ■



**Remarque 5.8 (Sur les propriétés de la solution)** Remarquons que la solution faible de (5.4) possède les propriétés suivantes :

1. Si  $u_0 \geq 0$  p.p. alors  $u \geq 0$  p.p.,
2.  $\|u(\cdot, t)\|_{L^p(\mathbb{R})} = \|u_0(x)\|_{L^p(\mathbb{R})} \quad \forall p \quad p \in [1, +\infty]$ .

Lors de l'élaboration de schémas numériques pour la recherche d'une approximation, on s'attachera à vérifier que ces propriétés sont encore satisfaites par la solution approchée.

### 5.3 Schémas numériques pour $u_t + u_x = 0$

On considère ici le problème de transport linéaire :

$$\begin{cases} u_t + u_x = 0, \\ u(x, 0) = u_0(x), u_0 \in L^\infty(\mathbb{R}). \end{cases} \quad (5.9)$$

On sait que la solution de ce problème s'écrit :

$$u(x, t) = u_0(x - ct).$$

On rappelle que  $u$  est une solution classique si  $u \in C^1(\mathbb{R})$ , et que  $u$  est une solution faible si  $u_0 \in L^\infty(\mathbb{R})$ . On va chercher à retrouver cette solution par une approximation numérique. Notons que dans le cas linéaire, l'utilisation d'un schéma numérique, n'est évidemment pas utile, mais nous commençons par ce cas par souci pédagogique.

#### 5.3.1 Schéma explicite différences finies centrées

On effectue une discrétisation espace temps en se donnant un pas de discrétisation en espace  $h$ , et en posant :  $x_i = ih, \forall i \in \mathbb{Z}$ ; de même on se donne un pas de discrétisation en temps  $k$ , et on pose  $t_n = nk, \forall n \in \mathbb{N}$ . Ecrivons le schéma d'Euler explicite pour l'approximation de  $u_t$  et un schéma centré pour l'approximation de  $u_x$ . On approche  $u_t(x_i, t_n)$  par  $\frac{u(x_i, t_{n+1}) - u(x_i, t_n)}{k}$  et  $u_x(x_i, t_n)$  par  $\frac{u(x_{i+1}, t_n) - u(x_{i-1}, t_n)}{2h}$ . Le schéma centré s'écrit donc, en fonction des inconnues discrètes :

$$\begin{cases} \frac{u_i^{n+1} - u_i^n}{k} + \frac{u_{i+1}^n - u_{i-1}^n}{2h} = 0, \\ u_i^0 = u_0(x_i). \end{cases} \quad (5.10)$$

(où on a supposé  $u_0 \in C(\mathbb{R})$ ) Ce schéma est inconditionnellement instable, et il faut donc éviter de l'utiliser. Qu'entend-on par instable ? On peut montrer que :

1. Le schéma (5.10) ne respecte pas la positivité, car  $u_0(x) \geq 0 \forall x$  n'entraîne pas forcément  $u_i^n \geq 0$ . En effet si  $u_0$  est telle que

$$\begin{cases} u_i^0 = 0, \forall i \leq 0, \\ u_i^0 = 1, \forall i > 0. \end{cases}$$

Alors :

$$u_i^{n+1} = u_i^n - \frac{k}{2h}(u_{i+1}^n - u_{i-1}^n)$$

donne, pour  $n = 0$

$$u_0^1 = -\frac{k}{2h} < 0$$

2. Le schéma (5.10) n'est pas  $L^\infty$  stable :

$$\|u^n\|_\infty \leq C \text{ n'entraîne pas } \|u^{n+1}\|_\infty \leq C.$$

3. Le schéma (5.10) n'est pas  $L^2$  stable :

$$\|u^n\|_2 \leq C \text{ n'entraîne pas que } \|u^{n+1}\|_2 \leq C.$$

4. Le schéma n'est pas stable au sens de Von Neumann. En effet, si

$$u_0(x) = e^{ipx}, \text{ où } i^2 = -1 \text{ et } p \in \mathbb{Z},$$

la solution exacte est  $u(x, t) = e^{ip(x-t)}$ . Une discrétisation de  $u_0$  s'écrit :

$$u_j^0 = e^{ipjh} \quad j \in \mathbb{Z}.$$

On a donc :

$$\begin{aligned} u_i^1 &= u_i^0 - \frac{k}{2h}(u_{i+1}^0 - u_{i-1}^0) \\ &= e^{ipjh} - \frac{k}{2h}(e^{ip(j+1)h} - e^{ip(j-1)h}) \\ &= \mathcal{J}_{k,h} u_j^0, \end{aligned}$$

avec  $\mathcal{J}_{k,h} = 1 - \frac{ik}{h} \sin ph$ . On a donc  $|\mathcal{J}_{k,h}| > 1$  si  $\sin ph \neq 0$ , ce qui montre que le schéma n'est pas stable au sens de Von Neuman.

5. Le schéma (5.10) n'est pas convergent. En effet, on peut montrer qu'il existe  $u_0, k$  et  $h$  telle que la solution approchée  $u_{h,k} : (u_i^n)_{i \in \mathbb{Z}}^{n \in \mathbb{N}}$  ne converge pas vers  $u$  lorsque  $h$  et  $k$  tendent vers 0.

### 5.3.2 Schéma différences finies décentré amont

On utilise toujours le schéma d'Euler explicite pour la discrétisation en temps, mais on approche maintenant

$$u_x(x_i, t_n) \text{ par } \frac{u(x_i, t_n) - u(x_{i-1}, t_n)}{h_{i-1/2}}.$$

On considère de plus un pas de discrétisation variable, défini par  $h_{i-1/2} = x_i - x_{i-1}$ . Le schéma par différences

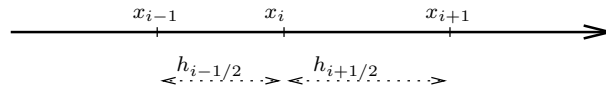


FIG. 5.3 – Maillage volumes finis

finies avec décentrement amont s'écrit :

$$\begin{cases} \frac{u_i^{n+1} - u_i^n}{k} + \frac{u_i^n - u_{i-1}^n}{h_{i-1/2}} = 0, \\ u(x, 0) = u_0(x). \end{cases} \quad (5.11)$$

**Proposition 5.9** *Le schéma (5.11) est stable sous condition de Courant-Friedrichs-Levy (CFL)*

$$k \leq h = \inf_{i \in \mathbb{Z}} h_{i-1/2} > 0. \quad (5.12)$$

c'est à dire que si  $A \leq u_i^n \leq B$ , alors  $A \leq u_i^{n+1} \leq B$ .

**Démonstration :** On a :  $u_i^{n+1} = u_i^n(1 - \alpha_i) + \alpha_i u_{i-1}^n$  avec  $\alpha_i = \frac{k}{h_{i-1/2}}$ . Donc, si la condition (5.12) est vérifiée,  $u_i^{n+1}$  est une combinaison convexe de  $u_i^n$  et  $u_{i-1}^n$ , et donc  $u_i^{n+1} \in [u_{i-1}^n, u_i^n]$ . ■

**Théorème 5.10 (Convergence du schéma (5.11))** *On suppose que  $u_0 \in C^2(\mathbb{R}, \mathbb{R})$  et que  $u_0, u'_0, u''_0$  sont bornées. Soit  $A = \inf_{x \in \mathbb{R}} u_0(x)$  et  $B = \sup_{x \in \mathbb{R}} u_0(x)$ . Alors :*

1.  $A \leq u_i^n \leq B, \forall i \in \mathbb{Z}, \forall n \in \mathbb{N}$ .
2. Soit  $\bar{u}_i^n = u(x_i, t_n)$ , à  $u$  est la solution exacte de (5.9), alors :

$$\sup_{\substack{i \in \mathbb{Z} \\ nk \leq T}} |u_i^n - \bar{u}_i^n| \leq TC_{u_0}(k + h),$$

où  $TC_{u_0} \geq 0$  ne dépend que de  $u_0$ .

**Démonstration :** le point 1 se démontre par récurrence sur  $n$  à partir de la propriété précédente. Le point 2 (estimation d'erreur) se démontre en remarquant d'abord que l'erreur de consistance

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} + \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_{i-1/2}} = R_i^n$$

vérifie :

$$|R_i^n| \leq C_{u_0}(\bar{h} + k)$$

où  $\bar{h} = \max i \in \mathbb{Z} h_{i-1/2}$ . On a donc :

$$\bar{u}_i^{n+1} = \bar{u}_i^n \left(1 - \frac{k}{h_{i-1/2}}\right) + \frac{k}{h_{i-1/2}} \bar{u}_{i-1}^n + kR_i^n,$$

or le schéma numérique s'écrit :

$$u_i^{n+1} = u_i^n \left(1 - \frac{k}{h_{i-1/2}}\right) + \frac{k}{h_{i-1/2}} u_{i-1}^n$$

Par différence, on obtient :

$$u_i^{n+1} - \bar{u}_i^{n+1} = (u_i^n - \bar{u}_i^n) \left(1 - \frac{k}{h_{i-1/2}}\right) + (u_{i-1}^n - \bar{u}_{i-1}^n) \frac{k}{h_{i-1/2}} - kR_i^n$$

et donc :

$$|u_i^{n+1} - \bar{u}_i^{n+1}| \leq |u_i^n - \bar{u}_i^n| \left(1 - \frac{k}{h_{i-1/2}}\right) + |u_{i-1}^n - \bar{u}_{i-1}^n| \frac{k}{h_{i-1/2}} + kC_{u_0}(\bar{h} + k) \quad (5.13)$$

On effectue alors l'hypothèse de récurrence :

$$\sup |u_i^n - \bar{u}_i^n| \leq (n-1)kC_{u_0}(k + \bar{h}) \quad (5.14)$$

grâce à (5.13) et (5.14), on obtient :

$$|u_i^{n+1} - \bar{u}_i^{n+1}| \leq (n-1)kC_{u_0}(k + \bar{h}) + k(C_{u_0}(k + \bar{h})).$$

Donc finalement :

$$|u_i^{n+1} - \bar{u}_i^{n+1}| \leq TC_{u_0}(k + \bar{h}).$$

**Remarque 5.11 (Décentrement) .** Pour une équation de transport telle que (5.9), le choix du décentrement est crucial. Ici, on a approché  $u_x(x_i)$  par  $\frac{u_i - u_{i-1}}{h_{i-1/2}}$ . Dans le cas où on étudie une équation de transport de type,  $u_t + cu_x = 0$ , avec  $c \in \mathbb{R}$ , le choix décentré amont sera toujours

$$\frac{u_i - u_{i-1}}{h} \text{ si } c > 0,$$

par contre, si  $c < 0$ , le choix amont donnera

$$\frac{u_i - u_{i+1}}{h}.$$

Regardons ce qui se passe si l'on effectue un "mauvais" décentrement. Considérons toujours l'équation  $u_t + u_x = 0$ . Effectuer le "mauvais décentrement" amène au schéma :

$$\frac{u_i^{n+1} - u_i^n}{k} + \frac{u_{i+1}^n - u_i^n}{h} = 0,$$

c'est à dire :

$$u_i^{n+1} = u_i^n \left(1 + \frac{k}{h}\right) - \frac{k}{h} u_{i+1}^n.$$

Examinons le comportement de la solution approchée donnée par le schéma si on prend une condition initiale  $u_0$  telle que  $u_0(x) = 0, \forall x \geq 0$ . Dans ce cas, on sait que  $u(x, t) \neq 0$  pour  $t$  assez grand, or après calculs on obtient  $u_{-1}^{n+1} = u_{-1}^n \left(1 + \frac{k}{h}\right) + 0 = u_{-1}^0 \left(1 + \frac{k}{h}\right)^n$ , alors que  $u_i^{n+1} = 0 \quad \forall i \geq 0$ . On en déduit que la solution approchée est très mauvaise.

**Remarque 5.12 (Equation non linéaire, donnée initiale)** 1. Dans le cas non linéaire, la démonstration précédente de convergence ne s'adapte pas car les solutions ne sont pas régulières.

2. On a défini (5.11) pour  $u_0 \in C(\mathbb{R})$ . Si  $u_0 \notin C(\mathbb{R})$ , on peut prendre comme donnée initiale  $u_i^0 = \int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x) dx$ .

### 5.3.3 Schéma volumes finis décentrés amont

On considère toujours le problème (5.9), avec condition initiale  $u_0 \in L^\infty(\mathbb{R})$ . On se donne une discrétisation en espace, c'est à dire un ensemble de points  $(x_{i+1/2})_{i \in \mathbb{Z}}$ , tels que  $x_{i+1/2} > x_{i-1/2}$ , et on note  $h_i = x_{i+1/2} - x_{i-1/2}$ . On approche toujours la dérivée en temps par un schéma d'Euler explicite, on intègre (5.9) sur la maille  $]x_{i-1/2}, x_{i+1/2}[$ , et on obtient :

$$\int_{x_{i-1/2}}^{x_{i+1/2}} (u_t + u_x) dx = 0.$$

En approchant  $u(x_{i+1/2})$  (resp.  $u(x_{i-1/2})$ ) par  $u_i^n$  (resp.  $u_{i-1}^n$ ) et en approchant  $u_t$  par un schéma d'Euler explicite, on obtient :

$$\begin{cases} h_i \frac{u_i^{n+1} - u_i^n}{k} + u_i^n - u_{i-1}^n = 0, \\ u_i^0 = \frac{1}{h_i} \int_{x_{i-1/2}}^{x_{i+1/2}} u_0(x) dx. \end{cases} \quad (5.15)$$

**Proposition 5.13** Soit  $(u_i^n)_{\substack{n \in \mathbb{N} \\ i \in \mathbb{Z}}}$  la solution de (5.15). Si  $k \leq h = \min h_i$  et si  $A \leq u_0(x) \leq B$ , alors  $A \leq u_i^n \leq B \quad \forall i \in \mathbb{Z}, \forall n \in \mathbb{N}$ .

La démonstration est similaire à celle de la proposition (5.9), et laissée à titre d'exercice.

**Définition 5.14 (Solution approchée)** Soit  $\mathcal{T}$  un maillage volumes finis de  $\mathbb{R}$  défini par  $\mathcal{T} = (K_i)_{i \in \mathbb{Z}}$  avec  $K_i = ]x_{i-1/2}, x_{i+1/2}[$ . On appelle solution approchée de (5.9) par le schéma (5.15) la fonction  $u_{\mathcal{T},k} : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$ , définie par

$$u_{\mathcal{T},k}(x, t) = u_i^n \text{ si } x \in K_i \text{ et } t \in [nk, nk + 1[ \quad (5.16)$$

On admettra le théorème de convergence suivant (voir aussi exercice 5.11) :

**Théorème 5.15 (Convergence du schéma 5.15)** Soit  $u_0 \in L^\infty(\mathbb{R})$ , on suppose que  $k \leq h = \inf(h_i)$ , alors  $u_{\mathcal{T},k}$  converge vers  $u$  dans  $L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$  lorsque  $h$  (et  $k$ ) tend vers 0, c'est à dire qu'on a :  $\int_C |u_{\mathcal{T},k} - u| dx dt \rightarrow 0$  pour tout compact  $C$  de  $\mathbb{R} \times \mathbb{R}_+$ , lorsque  $h$  (et  $k$ ) tend vers 0.

## 5.4 Equations hyperboliques non linéaires

On se donne  $f \in C^1(\mathbb{R}, \mathbb{R})$  et  $u_0 \in C(\mathbb{R})$  et on considère maintenant l'équation hyperbolique non linéaire :

$$\begin{cases} u_t + (f(u))_x = 0, & (x, t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x, 0) = u_0(x). \end{cases} \quad (5.17)$$

Commençons par donner la définition de solution classique de ce problème même si, comme nous le verrons après, celle-ci n'a pas grand intérêt puisque le problème (5.17) n'a pas, en général de solution classique.

**Définition 5.16 (Solution classique)** On suppose que  $u_0 \in C^1(\mathbb{R})$  et  $f \in C^2(\mathbb{R}, \mathbb{R})$ . Alors  $u$  est solution classique de (5.17) si  $u \in C^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$  et  $u$  vérifie

$$\begin{cases} (u_t + (f(u))_x)(x, t) = 0, & \forall (x, t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x, 0) = u_0(x), & \forall x \in \mathbb{R}. \end{cases}$$

Avant d'énoncer le théorème de non existence, rappelons que dans le cas d'une équation différentielle du type non linéaire,

$$\begin{cases} x'(t) = f(x(t)), & t \in \mathbb{R}_+, \\ x(0) = x_0, \end{cases}$$

si on note  $T_{\max}$  le temps d'existence de la solution, et si  $T_{\max} < +\infty$  alors  $\|x(t)\| \rightarrow +\infty$  lorsque  $t \rightarrow T_{\max}$ . Donnons maintenant la définition des courbes caractéristiques de l'équation (5.17), qui permet le lien entre les équations hyperboliques non linéaires et les équations différentielles ordinaires.

**Définition 5.17 (Courbe caractéristique)** On appelle courbe caractéristique du problème (5.17) issue de  $x_0 \in \mathbb{R}$ , la courbe définie par le problème de Cauchy suivant :

$$\begin{cases} x'(t) = f'(u(x(t), t)) \\ x(0) = x_0 \end{cases} \quad (5.18)$$

**Théorème 5.18 (Non existence)** Soit  $f \in C^1(\mathbb{R}, \mathbb{R})$ , on suppose que  $f'$  n'est pas constante, alors il existe  $u_0 \in C_c^\infty(\mathbb{R})$  telle que (5.17) n'admette pas de solution classique.

**Démonstration :** Comme  $f \in C^2(\mathbb{R}, \mathbb{R})$ , on a  $f' \in C^1(\mathbb{R}, \mathbb{R})$ , et donc le théorème de Cauchy-Lipschitz s'applique. Il existe donc une solution maximale  $x(t)$  définie sur  $[0, T_{\max}[$ , et  $x(t)$  tend vers l'infini lorsque  $t$  tend vers  $T_{\max}$  si  $T_{\max} < +\infty$ . Les quatre étapes de la démonstration sont les suivantes :

1.  $u(x(t), t) = u_0(x_0), \forall t \in [0, T_{\max}[$ , et donc que toute solution de (5.17) est constante sur les caractéristiques.
2. Les courbes caractéristiques sont des droites.
3.  $T_{\max} = +\infty$  et donc  $u(x, t) = u_0(x_0) \quad \forall t \in [0, +\infty[$ .
4. On en déduit alors qu'on n'a pas de solution classique de (5.17).

Détaillons maintenant ces étapes.

1. Soit  $\varphi$  définie par  $\varphi(t) = u(x(t), t)$ ; en dérivant  $\varphi$ , on obtient :  $\varphi'(t) = u_t(x(t), t) + u_x(x(t), t)x'(t)$ . Comme  $x$  vérifie (5.18), ceci entraîne :  $\varphi'(t) = u_t(x(t), t) + f'(u(x(t), t))u_x(x(t), t)$ , et donc

$$\varphi'(t) = (u_t + (f(u))_x)(x(t), t) = 0.$$

La fonction  $\varphi$  est donc constante, et on a :

$$u(x(t), t) = \varphi(t) = \varphi(0) = u(x(0), 0) = u(x_0, 0) = u_0(x_0), \forall t \in [0, T_{\max}[.$$

2. Comme  $u(x(t), t) = u_0(x_0), \forall t \in [0, T_{\max}[$ , on a donc  $x'(t) = f'(u_0(x_0))$ . Donc en intégrant, on obtient que le système (5.18) décrit la droite d'équation :

$$x(t) = f'(u_0(x_0))t + x_0. \quad (5.19)$$

3. Puisque  $x$  vérifie (5.19), on a donc

$$\lim_{t \rightarrow T_{\max}} |x(t)| < +\infty. \text{ On en déduit que } T = T_{\max}.$$

4. Comme  $f'$  est non constante, il existe  $v_0, v_1$  tel que  $f'(v_0) > f'(v_1)$ , et on peut construire  $u_0 \in C_c^\infty(\mathbb{R}, \mathbb{R})$  telle que  $u_0(x_0) = v_0$  et  $u_0(x_1) = v_1$ , où  $x_0$  et  $x_1$  sont donnés et  $x_0 < x_1$ , voir figure 5.4. Supposons que

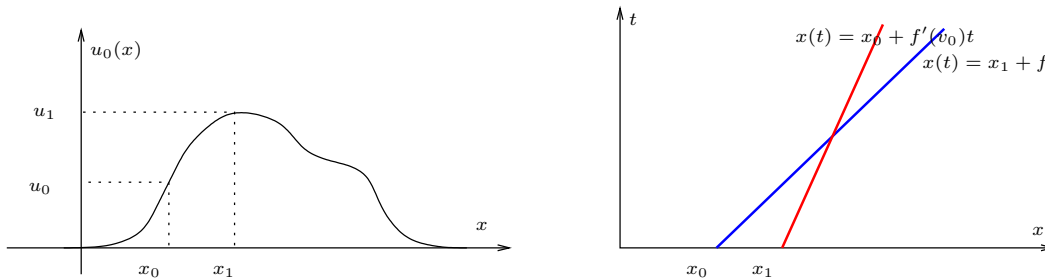


FIG. 5.4 – Droites caractéristiques, cas non linéaire

$u$  soit solution classique avec cette donnée initiale. Alors :

$$u(x_0 + f'(u_0(x_0))t, t) = u_0(x_0) = v_0 \text{ et } u(x_1 + f'(u_0(x_1))t, t) = u_0(x_1) = v_1.$$

Soit  $T$  tel que  $x_0 + f'(v_0)T = x_1 + f'(v_1)T = \bar{x}$ , c'est à dire

$$T = \frac{x_1 - x_0}{f'(v_0) - f'(v_1)}.$$

On a alors :

$$u(\bar{x}, T) = u_0(x_0) = v_0 = u_0(x_1) = v_1,$$

ce qui est impossible. On en conclut que (5.17) n'admet pas de solution classique pour cette donnée initiale.

■

**Définition 5.19 (Solution faible)** Soit  $u_0 \in L^\infty(\mathbb{R})$  et  $f \in C^1(\mathbb{R}, \mathbb{R})$ , On appelle solution faible de (5.17) une fonction  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$  telle que

$$\int \int_{\mathbb{R} \times \mathbb{R}_+} [u(x, t)\varphi_t(x, t) + f(u(x, t))\varphi_x(x, t)] dx dt + \int_{\mathbb{R}} u_0(x)\varphi(x, 0) dx = 0, \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}). \quad (5.20)$$

Donnons maintenant les liens entre solution classique et solution faible.

**Proposition 5.20** Soient  $f \in C^1(\mathbb{R}, \mathbb{R})$  et  $u_0 \in C(\mathbb{R}, \mathbb{R})$  des fonctions données.

1. Si  $u$  est solution classique de (5.17) alors  $u$  est solution faible de (5.17).
2. Si  $u \in C^1(\mathbb{R} \times ]0, +\infty[) \cap C(\mathbb{R} \times [0, +\infty[)$  est solution faible de (5.17) alors  $u$  est solution classique de (5.17).
3. Soit  $\sigma \in \mathbb{R}$ ,  $D_1 = \{(x, t) \in \mathbb{R} \times \mathbb{R}_+; x < \sigma t\}$  et  $D_2 = \{(x, t) \in \mathbb{R} \times \mathbb{R}_+; x > \sigma t\}$ . Alors si  $u \in C(\mathbb{R} \times \mathbb{R}_+)$  est telle que  $u|_{D_i} \in C^1(D_i, \mathbb{R})$ ,  $i = 1, 2$  et que (5.17) est vérifié pour tout  $(x, t) \in D_i$ ,  $i = 1, 2$ , alors  $u$  est solution faible de (5.17).

**Démonstration :**

1. Supposons que  $u$  est solution classique de (5.17), c.à.d. de :

$$\begin{cases} u_t + (f(u))_x = 0, & (x, t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x, 0) = u_0(x). \end{cases}$$

Soit  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$ . Multiplions (5.17) par  $\varphi$  et intégrons sur  $\mathbb{R} \times \mathbb{R}_+$ . On obtient :

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u_t(x, t)\varphi(x, t) dt dx + \int_{\mathbb{R}} \int_{\mathbb{R}_+} (f(u))_x(x, t)\varphi(x, t) dt dx = 0.$$

L'application du théorème de Fubini et une intégration par parties donnent alors :

$$\int_{\mathbb{R}} u(x, 0)\varphi(x, 0) dx - \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x, t)\varphi_t(x, t) dt dx - \int_{\mathbb{R}_+} \int_{\mathbb{R}} f(u)(x, t)\varphi_x(x, t) dx dt = 0,$$

(car  $\text{supp}(\varphi)$  est compact). Et on obtient donc bien la relation (5.20), grâce à la condition initiale  $u(x, 0) = u_0(x)$ .

2. Soit donc  $u$  une solution faible de (5.17), qui vérifie de plus  $u \in C^1(\mathbb{R} \times ]0, +\infty[) \cap C(\mathbb{R} \times [0, +\infty[)$ . On a donc suffisamment de régularité pour intégrer par parties dans (5.20).

Commençons par prendre  $\varphi$  à support compact dans  $\mathbb{R} \times ]0, +\infty[$ . On a donc  $\varphi(x, 0) = 0$ , et une intégration par parties dans (5.20) donne :

$$- \int_{\mathbb{R}} \int_{\mathbb{R}_+} u_t(x, t)\varphi(x, t) dt dx - \int_{\mathbb{R}_+} \int_{\mathbb{R}} (f(u))_x(x, t)\varphi(x, t) dx dt = 0.$$

On a donc :

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (u_t(x, t) + (f(u))_x(x, t)) \varphi(x, t) dt dx = 0, \forall \varphi \in C_c^1(\mathbb{R} \times ]0, +\infty[).$$

Comme  $u_t + (f(u))_x$  est continue, on en déduit que  $u_t + (f(u))_x = 0$ . En effet, on rappelle que si  $\int_{\mathbb{R}} f(x)\varphi(x) dx = 0$  pour toute fonction  $\varphi$  continue de  $\mathbb{R}$  dans  $\mathbb{R}$ , alors  $f = 0$  p.p. ; si de plus  $f$  est continue, alors  $f = 0$  partout.

On prend alors  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+)$ . Dans ce cas, une intégration par parties dans (5.20) donne

$$\int_{\mathbb{R}} u(x, 0)\varphi(x, 0) dx - \int_{\mathbb{R}} \int_{\mathbb{R}_+} (u_t(x, t) + (f(u))_x(x, t)) \varphi(x, t) dt dx - \int_{\mathbb{R}} u_0(x)\varphi(x, 0) dx = 0.$$

Mais on vient de montrer que  $u_t + (f(u))_x = 0$ . On en déduit que

$$\int_{\mathbb{R}} (u_0(x) - u(x, 0))\varphi(x, 0) dx = 0, \forall \varphi \in C_c^1(\mathbb{R}).$$

Comme  $u$  est continue, ceci entraîne  $u(x, 0) = u_0(x)$ . Donc  $u$  est solution classique de (5.17).

3. Soit  $u \in C(\mathbb{R} \times \mathbb{R}_+)$  telle que  $u|_{D_i}$  vérifie (5.17), pour tout  $(x, t) \in D_i$ . Montrons que  $u$  est solution faible. Pour cela, calculons :

$$X = \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x, t) \varphi_t(x, t) dt dx + \int_{\mathbb{R}_+} \int_{\mathbb{R}} f(u)(x, t) \varphi_x(x, t) dx dt.$$

On a donc  $X = X_1 + X_2$ , avec

$$X_1 = \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x, t) \varphi_t(x, t) dt dx \text{ et } X_2 = \int_{\mathbb{R}_+} \int_{\mathbb{R}} (f(u))(x, t) \varphi_x(x, t) dx dt.$$

Calculons  $X_1$ . Comme  $u$  n'est de classe  $C^1$  que sur chacun des domaines  $D_i$ , on n'a pas le droit d'intégrer par parties sur  $\mathbb{R} \times \mathbb{R}_+$  entier. On va donc décomposer l'intégrale sur  $D_1$  et  $D_2$ ; supposons par exemple  $\sigma < 0$ , voir figure 5.5. (Le cas  $\sigma > 0$  se traite de façon similaire). On a alors  $D_2 = \{(x, t); x \in \mathbb{R}_- \text{ et } 0 < t < \frac{x}{\sigma}\}$  et  $D_1 = \mathbb{R}_+ \times \mathbb{R}_+ \cup \{(x, t); x \in \mathbb{R}_- \text{ et } \frac{x}{\sigma} < t < +\infty\}$ .

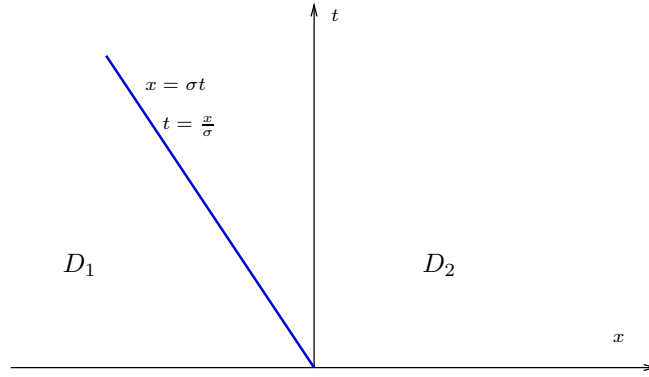


FIG. 5.5 – Les domaines  $D_1$  et  $D_2$

On a donc :

$$X_1 = \int_{\mathbb{R}_-} \int_0^{x/\sigma} u(x, t) \varphi_t(x, t) dt dx + \int_{\mathbb{R}_-} \int_{\frac{x}{\sigma}}^{+\infty} u(x, t) \varphi_t(x, t) dt dx + \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} u(x, t) \varphi_t(x, t) dt dx.$$

Comme  $u$  est de classe  $C^1$  sur chacun des domaines, on peut intégrer par parties, ce qui donne :

$$\begin{aligned} X_1 = & \int_{\mathbb{R}_-} u(x, \frac{x}{\sigma}) \varphi(x, \frac{x}{\sigma}) dx - \int_{\mathbb{R}_-} u(x, 0) \varphi(x, 0) dx - \int_{\mathbb{R}_-} \int_0^{\frac{x}{\sigma}} u_t(x, t) \varphi(x, t) dt dx \\ & + \int_{\mathbb{R}_-} u(x, \frac{x}{\sigma}) \varphi(x, \frac{x}{\sigma}) dx - \int_{\mathbb{R}_-} \int_{\frac{x}{\sigma}}^{+\infty} u_t(x, t) \varphi(x, t) dt dx \\ & + \int_{\mathbb{R}_+} (-u(x, 0) \varphi(x, 0) dx - \int_{\mathbb{R}_+} \int_{\mathbb{R}_+} u_t(x, t) \varphi(x, t) dt dx. \end{aligned} \quad (5.21)$$

En simplifiant, il vient :

$$X_1 = - \int_{\mathbb{R}} u(x, 0) \varphi(x, 0) dx - \int \int_{D_1} u_t(x, t) \varphi(x, t) dt dx - \int \int_{D_2} u_t(x, t) \varphi(x, t) dt dx.$$

On décompose de même  $X_2$  sur  $D_1 \cup D_2$ , en remarquant maintenant que  $D_1 = \{(x, t) \in \mathbb{R} \times \mathbb{R}_+; x < \sigma t\}$  et  $D_2 = \{(x, t) \in \mathbb{R} \times \mathbb{R}_+; x > \sigma t\}$  :

$$X_2 = \int_{\mathbb{R}_+} \int_{-\infty}^{\sigma t} f(u)(x, t) \varphi_x(x, t) dx dt + \int_{\mathbb{R}_+} \int_{\sigma t}^{+\infty} f(u)(x, t) \varphi_x(x, t) dx dt.$$

La fonction  $u$  est de classe  $C^1$  sur chacun des domaines, on peut là encore intégrer par parties. Comme  $\varphi$  est à support compact sur  $\mathbb{R} \times \mathbb{R}_+$ , on obtient après simplification :

$$X_2 = - \int \int_{D_1} (f(u))_x(x, t) \varphi(x, t) dx dt - \int \int_{D_2} (f(u))_x(x, t) \varphi(x, t) dx dt.$$

Comme  $u_t + (f(u))_x = 0$  sur  $D_1$  et  $D_2$ , on a donc :

$$X = X_1 + X_2 = - \int_{\mathbb{R}} u(x, 0) \varphi(x, 0) dx,$$

ce qui prouve que  $u$  est solution faible de (5.17). ■

Notons qu'il existe souvent plusieurs solutions faibles. On a donc besoin d'une notion supplémentaire pour les distinguer. C'est la notion de solution entropique, qui nous permettra d'obtenir l'unicité. Donnons tout d'abord un exemple de non-unicité de la solution faible. Pour cela on va considérer une équation modèle, appelée équation de Burgers, qui s'écrit

$$u_t + (u^2)_x = 0. \quad (5.22)$$

Pour calculer les solutions du problème de Cauchy associé à cette équation de manière analytique, on considère une donnée initiale particulière, qui s'écrit

$$u_0(x) = \begin{cases} u_g \text{ si } x < 0, \\ u_d \text{ si } x > 0, \end{cases}$$

Ces données initiales définissent un problème de Cauchy particulier, qu'on appelle problème de Riemann, que nous étudierons plus en détails par la suite.

Considérons alors le problème suivant (dit problème de Riemann, voir définition 5.28) pour l'équation de Burgers :

$$\begin{cases} u_t + (u^2)_x = 0, \\ u_0(x) = \begin{cases} u_g = -1 \text{ si } x < 0, \\ u_d = 1 \text{ si } x > 0. \end{cases} \end{cases} \quad (5.23)$$

On cherche une solution faible de la forme :

$$u(x, t) = \begin{cases} u_g \text{ si } x < \sigma t, \\ u_d \text{ si } x > \sigma t. \end{cases} \quad (5.24)$$

Notons que cette éventuelle solution est discontinue au travers de la droite d'équation  $x = \sigma t$  dans le plan  $(x, t)$ . On remplace  $u(x, t)$  par ces valeurs dans (5.20). Après calculs (voir exercice 68 page 200, ou aussi la proposition 5.29 plus loin), on s'aperçoit que  $u$  est solution faible si la condition suivante, dite condition de Rankine et Hugoniot, est vérifiée :

$$\sigma(u_d - u_g) = (f(u_d) - f(u_g)), \quad (5.25)$$

ce qui avec la condition initiale particulière choisie ici, donne  $2\sigma = 1^2 - (-1)^2 = 0$ .

Mais on peut trouver d'autres solutions faibles : en effet, on sait que sur les caractéristiques, qui ont pour équation  $x = x_0 + f'(u_0(x_0))t$ , la fonction  $u$  est constante. Comme  $f'(u) = 2u$ , les caractéristiques sont donc des droites de pente -2 si  $x_0 < 0$ , et de pente 2 si  $x_0 > 0$ . Construisons ces caractéristiques sur la figure 5.6 : Dans la zone du

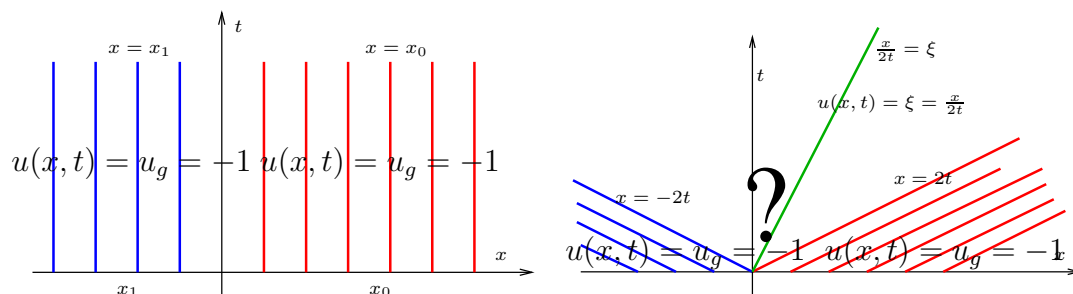


FIG. 5.6 – Problème de Riemann pour l'équation de Burgers

milieu, où l'on a représenté un point d'interrogation, on cherche  $u$  sous la forme  $u(x, t) = \varphi\left(\frac{x}{t}\right)$ , et telle que  $u$



soit continue sur  $\mathbb{R} \times \mathbb{R}_+$ . La fonction  $u$  suivante convient :

$$u(x, t) = \begin{cases} -1 & \text{si } x < -2t, \\ \frac{x}{2t} & \text{si } -2t < x < 2t, \\ 1 & \text{si } x > 2t. \end{cases} \quad (5.26)$$

Comment choisir la “bonne” solution faible, entre (5.24) et (5.26)? Comme les problèmes hyperboliques sont souvent obtenus en négligeant les termes de diffusion dans des équations paraboliques, une technique pour choisir la solution est de chercher la limite du problème de diffusion associé qui s’écrit :

$$u_t + (f(u))_x - \varepsilon u_{xx} = 0, \quad (5.27)$$

lorsque le terme de diffusion devient négligeable, c.à.d. lorsque  $\varepsilon$  tend vers 0. Soit  $u_\varepsilon$  la solution de (5.27) (on admettra l’existence et l’unicité de  $u_\varepsilon$ ). On peut montrer que  $u_\varepsilon$  tend vers  $u$  lorsque  $\varepsilon$  tend vers 0, où  $u$  est la “solution faible entropique” de (5.27), définie comme suit.

**Définition 5.21 (Solution entropique)** Soit  $u_0 \in L^\infty(\mathbb{R})$  et  $f \in C^1(\mathbb{R})$ , on dit que  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$  est solution entropique de (5.27) si pour toute fonction  $\eta \in C^1(\mathbb{R})$  convexe, appelée “entropie”, et pour toute fonction  $\phi \in C^1$  telle que  $\phi' = f'\eta'$ , appelé “flux d’entropie”, on a :

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\eta(u)\varphi_t + \phi(u)\varphi_x) dx dt + \int_{\mathbb{R}} \eta(u_0(x))\varphi(x, 0) dx \geq 0, \forall \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+). \quad (5.28)$$

**Remarque 5.22 (Condition initiale)** Noter que dans la définition 5.21, on prend une fois de plus  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+)$  de manière à bien prendre en compte la condition initiale ; ceci n’est pas toujours fait de cette manière dans les travaux plus anciens sur le sujet, mais entraîne des difficultés lorsqu’on s’intéresse à la convergence des schémas numériques.

On admettra le théorème suivant (dû à Kruskov, 1955)

**Théorème 5.23 (Kruskov)** Soient  $u_0 \in L^\infty(\mathbb{R})$  et  $f \in C^1(\mathbb{R})$  alors il existe une unique solution entropique de (5.17) au sens de la définition 5.21.

**Proposition 5.24** Si  $u$  est solution classique de (5.17), alors  $u$  est solution entropique.

**Démonstration :** Soit  $u \in C^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$ , Soit  $\eta \in C^1(\mathbb{R})$ , convexe, une entropie et  $\phi$  tel que  $\phi' = f'\eta'$ , le flux associé. Multiplions (5.17) par  $\eta'(u)$  :

$$\eta'(u)u_t + f'(u)u_x\eta'(u) = 0$$

Soit encore, puisque  $\phi' = f'\eta'$ ,

$$(\eta(u))_t + \phi'(u)u_x = 0$$

On a donc finalement :

$$(\eta(u))_t + (\phi(u))_x = 0 \quad (5.29)$$

De plus, comme  $u(x, 0) = u_0(x)$ , on a aussi :  $\eta(u(x, 0)) = \eta(u_0(x))$ . Soit  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+)$ , on multiplie (5.29) par  $\varphi$ , on intègre sur  $\mathbb{R} \times \mathbb{R}_+$  et on obtient (5.28) (avec égalité) en intégrant par parties. Dans le cas d’une solution classique, l’inégalité d’entropie est une égalité. ■

On a de même le résultat suivant :

**Proposition 5.25** Si  $u$  est solution faible entropique de (5.17), alors  $u$  est solution faible.

**Démonstration :** Il suffit de prendre  $\eta(u) = u$  et  $\eta(u) = -u$  dans (5.28) pour se convaincre du résultat. ■

On déduit de la proposition 5.24, et du théorème 5.23 de Kruskov, que si on a plusieurs solutions faibles au problème 5.17 page 186 et que l’une d’entre elles est régulière, alors cette dernière est forcément la solution entropique. Enfin, la caractérisation suivante, que l’on admettra, est souvent utilisée en pratique :

**Proposition 5.26 (Entropies de Kruskov)** Soit  $u_0 \in L^\infty(\mathbb{R})$  et  $f \in C^1(\mathbb{R})$ , alors  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$  est solution entropique de (5.27) au sens de la définition 5.21 si et seulement si pour tout  $k \in \mathbb{R}$ , alors (5.28) est vérifiée avec  $\eta$  définie par  $\eta(s) = |s - k|$ , et  $\phi$ , flux d'entropie associée, défini par :

$$\phi(u) = \max(f(u), k) - \min(f(u), k).$$

Notons que  $\eta$  n'est pas de classe  $C^1$ .

Notons que les solutions d'une équation hyperbolique non linéaire respectent les bornes de la solution initiale. Plus précisément, on a le résultat suivant, qu'on admettra :

**Proposition 5.27** Si  $u_0 \in L^\infty(\mathbb{R})$  et soit  $A$  et  $B \in \mathbb{R}$  tels que  $A \leq u_0 \leq B$  p.p.. Soit  $f \in C^1(\mathbb{R})$ , alors la solution entropique  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+)$  de (5.17) vérifie :  $A \leq u(x) \leq B$  p.p. dans  $\mathbb{R} \times \mathbb{R}_+$ .

Cette propriété est essentielle dans les phénomènes de transport, et il est souhaitable qu'elle soit préservée pour la solution approchée donnée par un schéma numérique.

Avant d'aborder l'étude des schémas numériques pour les équations hyperboliques, nous terminons par un résultat sur les solutions du problème de Riemann, dont nous nous sommes d'ailleurs servis pour montrer la non unicité des solutions faibles de (5.23).

**Définition 5.28 (Problème de Riemann)** Soient  $f \in C^1(\mathbb{R}, \mathbb{R})$ , on appelle problème de Riemann avec données  $u_g, u_d \in \mathbb{R}$ , le problème suivant :

$$\begin{cases} u_t + (f(u))_x = 0, & x \in \mathbb{R}, t > 0 \\ u(0, x) = \begin{cases} u_g & \text{si } x < 0 \\ u_d & \text{si } x > 0 \end{cases} \end{cases} \quad (5.30)$$

Lorsque la fonction  $f$  est convexe ou concave, les solutions du problème de Riemann se calculent facilement ; en effet, on peut montrer le résultat suivant (voir aussi exercice 69 page 200) :

**Proposition 5.29** Soit  $f \in C^1(\mathbb{R}, \mathbb{R})$  strictement convexe, et soient  $u_g$  et  $u_d \in \mathbb{R}$ .

1. Si  $u_g > u_d$ , on pose

$$\sigma = \frac{[f(u)]}{[u]} \text{ avec } [f(u)] = f(u_d) - f(u_g) \text{ et } [u] = u_d - u_g. \quad (5.31)$$

alors la fonction  $u$  définie par

$$\begin{cases} u(x, t) = u_g & \text{si } x < \sigma t \\ u(x, t) = u_d & \text{si } x > \sigma t \end{cases} \quad (5.32)$$

est l'unique solution entropique de (5.30). Une solution de la forme (5.32) est appelée une onde de "choc".

2. Si  $u_g < u_d$ , alors la fonction  $u$  définie par

$$\begin{cases} u(x, t) = u_g & \text{si } x < f'(u_g)t \\ u(x, t) = u_d & \text{si } x > f'(u_d)t \\ u(x, t) = \xi & \text{si } x = f'(\xi)t \text{ avec } u_g < \xi < u_d \end{cases} \quad (5.33)$$

est l'unique solution entropique de (5.30). Notons que dans ce cas, la solution entropique est continue. Une solution de la forme (5.33) est appelée une onde de "détente".

**Démonstration :** 1. Cherchons  $u$  sous la forme (5.32). Commençons par déterminer  $\sigma$  pour que  $u$  soit solution faible. On suppose, pour fixer les idées, que  $\sigma > 0$  (mais le même raisonnement marche pour  $\sigma < 0$ ). Soit  $\varphi \in C_c^\infty(\mathbb{R} \times \mathbb{R}_+, \mathbb{R})$ . On veut montrer que

$$X = X_1 + X_2 = - \int_{\mathbb{R}} u(x, 0) \varphi(x, 0) dx,$$

où  $X_1 = \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x, t) \varphi_t(x, t) dt dx$  et  $X_2 = \int_{\mathbb{R}} \int_{\mathbb{R}_+} f(u(x, t)) \varphi_x(x, t) dt dx$ .

Calculons donc  $X_1$  et  $X_2$  :

$$\begin{aligned} X_1 &= \int_{-\infty}^0 \int_0^{+\infty} u(x,t) \varphi_t(x,t) dt dx + \int_0^{+\infty} \int_0^{\frac{x}{\sigma}} u(x,t) \varphi_t(x,t) dt dx + \int_0^{+\infty} \int_{\frac{x}{\sigma}}^{+\infty} u(x,t) \varphi_t(x,t) dt dx \\ &= - \int_{-\infty}^0 u_g \varphi(x,0) dx + \int_0^{+\infty} u_d \left( \varphi(x, \frac{x}{\sigma}) - \varphi(x,0) \right) dx + \int_0^{+\infty} u_g \left( -\varphi(x, \frac{x}{\sigma}) \right) dx \\ &= - \int_{\mathbb{R}} u(x,0) \varphi(x,0) dx + \int_0^{+\infty} (u_d - u_g) \varphi(x, \frac{x}{\sigma}) dx. \end{aligned}$$

De même

$$\begin{aligned} X_2 &= \int_0^{+\infty} \int_{-\infty}^{\sigma t} f(u) \varphi_x(x,t) dx dt + \int_0^{+\infty} \left( \int_{\sigma t}^{+\infty} f(u)(x,t) \varphi_x(x,t) \right) dx dt \\ &= \int_0^{+\infty} f(u_g) \varphi(\sigma t, t) dt - \int_0^{+\infty} f(u_d) \varphi(\sigma t, t) dt. \end{aligned}$$

En posant  $[u] = u_d - u_g$  et  $[f(u)] = f(u_d) - f(u_g)$ , on obtient :

$$\begin{aligned} X + \int_{\mathbb{R}} u(x,0) \varphi(x,0) dx &= \int_0^{+\infty} [u] \varphi(x, \frac{x}{\sigma}) dx - \int_0^{+\infty} [f(u)] \varphi(\sigma t, t) dt \\ &= \int_0^{+\infty} [u] \varphi(\sigma t, t) \sigma dt - \int_0^{+\infty} [f(u)] \varphi(\sigma t, t) dt. \end{aligned}$$

On en déduit que

$$X + \int_{\mathbb{R}} u(x,0) \varphi(x,0) dx = 0 \text{ si } \sigma[u] - [f(u)] = 0,$$

ce qui est vrai si la condition suivante, dite de Rankine et Hugoniot :

$$\sigma[u] = [f(u)] \quad (5.34)$$

est vérifiée.

Voyons maintenant si  $u$  est bien solution entropique. Pour cela, on considère  $\eta \in C^1$  une “entropie”, et  $\phi \in C^1$  le flux d’entropie associé, t.q.  $\phi' = \eta' f'$ . Le même calcul que le précédent, en remplaçant  $u$  par  $\eta(u)$  et  $f(u)$  par  $\phi(u)$  donne que :

$$\begin{aligned} &\int_{\mathbb{R}} \int_{\mathbb{R}_+} \eta(u)(x,t) \varphi_t(x,t) dt dx + \int_{\mathbb{R}_+} \int_{\mathbb{R}} \phi(u)(x,t) \varphi_x(x,t) dx dt \\ &+ \int_{\mathbb{R}} \eta(u_0(x)) dx = \int_0^{+\infty} (\sigma[\eta(u)] - [\phi(u)]) \varphi(\sigma t, t) dt. \end{aligned}$$

Pour que  $u$  soit solution entropique, il faut (et il suffit) donc que

$$\sigma[\eta(u)] \geq [\phi(u)] \quad (5.35)$$

Il reste à vérifier que cette inégalité est vérifiée pour  $\sigma$  donné par (5.34), c.à.d.

$$\frac{f(u_d) - f(u_g)}{u_d - u_g} (\eta(u_d) - \eta(u_g)) \geq \phi(u_d) - \phi(u_g)$$

Ceci s’écrit encore :

$$(f(u_d) - f(u_g))(\eta(u_d) - \eta(u_g)) \leq (\phi(u_d) - \phi(u_g))(u_d - u_g).$$

Cette inégalité est vérifiée en appliquant le lemme suivant avec  $b = u_g > u_d = a$ .

**Lemme 5.30** Soient  $a, b \in \mathbb{R}$  tels que  $a < b$ , soient  $f$  et  $\eta \in C^1(\mathbb{R})$  des fonctions convexes et  $\phi \in C^1(\mathbb{R})$  telle que  $\phi' = \eta' f'$ , alors :

$$\int_a^b \phi'(s) ds (b - a) \geq \int_a^b f'(s) ds \int_a^b \eta'(s) ds$$

**Démonstration :** On a

$$\begin{aligned} \int_a^b \phi'(x) dx &= \int_a^b f'(x) \eta'(x) dx \\ &= \int_a^b f'(x) (\eta'(x) - \eta'(y)) dx + \int_a^b f'(x) \eta'(y) dx, \quad \forall y \in \mathbb{R}. \end{aligned}$$

On a donc, en intégrant par rapport à  $y$  entre  $a$  et  $b$  :

$$(b-a) \int_a^b \phi'(x) dx = \int_a^b \int_a^b f'(x) (\eta'(x) - \eta'(y)) dx dy + \int_a^b f'(x) dx \int_a^b \eta'(y) dy$$

Or

$$\int_a^b \int_a^b f'(x) [\eta'(x) - \eta'(y)] dx dy = \int_a^b \int_a^b f'(y) (\eta'(y) - \eta'(x)) dx dy$$

et donc

$$(b-a) \int_a^b \int_a^b \phi'(x) dx = \int_a^b \int_a^b (f'(x) - f'(y)) (\eta'(x) - \eta'(y)) dx dy + \left( \int_a^b f'(x) dx \right) \left( \int_a^b \eta'(y) dy \right).$$

Comme  $f'$  et  $\eta'$  sont croissantes, la première intégrale du second membre est nulle, et on a donc bien le résultat annoncé. ■

2. On vérifie facilement que la fonction  $u$  définie par (5.33) est continue sur  $\mathbb{R} \times \mathbb{R}_+^*$ , et qu'elle vérifie  $u_t + (f(u))_x = 0$  dans chacun des domaines  $D_1, D_2, D_3$  définis par

$$D_1 = \{t > 0, x < f'(u_g)t\}, D_2 = \{t > 0, f'(u_g)t < x < f'(u_d)t\} \text{ et } D_3 = \{t > 0, x > f'(u_d)t\}.$$

Donc par le point 3 de la proposition 5.20 page 188, on sait que  $u$  est solution faible (mais attention, ce n'est pas une solution classique car  $u$  n'est pas forcément  $C^1$  sur  $\mathbb{R} \times \mathbb{R}_+$  tout entier).

Soit  $\eta \in C^1(\mathbb{R}, \mathbb{R})$  une entropie (convexe) et  $\phi$  le flux d'entropie associé, comme  $u_t + (f(u))_x = 0$  dans  $D_i$  pour  $i = 1$  à  $3$ , en multipliant par  $\eta'(u)$ , on a également que  $(\eta(u))_t + (\phi(u))_x = 0$  dans  $D_i$  pour  $i = 1$  à  $3$ . Soit maintenant  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}_+)$ , on va montrer que

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\eta(u))(x, t) \varphi_t(x, t) dt dx + \int_{\mathbb{R}} \int_{\mathbb{R}} (\phi(u))(x, t) \varphi_x(x, t) dx dt + \int_{\mathbb{R}} \eta(u_0(x)) \varphi(x, 0) dx = 0$$

(dans le cas d'une solution continue, l'inégalité d'entropie est une égalité). En effet, en intégrant par parties les trois termes précédents sur  $D_1, D_2, D_3$ , comme on l'a fait dans les questions 1 et 2, comme la fonction  $u$  est continue, les traces des fonctions sur le bord des domaines s'annulent deux à deux, et il ne reste donc que la condition initiale. On montre ainsi (faire le calcul pour s'en convaincre. . .) que

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} (\eta(u))(x, t) \varphi_x(x, t) dt dx + \int_{\mathbb{R}} \int_{\mathbb{R}_+} \phi(u)(x, t) (\varphi_x(x, t)) dx dt = - \int_{\mathbb{R}} \eta(u_0(x)) \varphi(x, 0),$$

ce qui prouve que  $u$  est la solution entropique. ■

## 5.5 Schémas pour les équations non linéaires

On se donne  $u_0 \in L^\infty(\mathbb{R})$  et  $f \in C^1(\mathbb{R})$ , et on cherche à trouver une approximation de la solution entropique du problème (5.17). On utilise les mêmes notations que pour le schéma (5.15). En intégrant l'équation  $u_t + (f(u))_x = 0$  sur une maille  $K_i$ , on obtient, au temps  $t = t_n$  :

$$\int_{K_i} u_t(x, t_n) dx dt + f(u(x_{i+1/2}, t_n)) - f(u(x_{i-1/2}, t_n)) = 0.$$

En utilisant le schéma d'Euler explicite pour la discrétisation de la dérivée temporelle, et en notant  $f_{i+1/2}^n$  le flux numérique, c'est à dire l'approximation de  $f(u(x_{i+1/2}, t_n))$  on obtient le schéma numérique suivant :

$$\begin{cases} h_i \frac{u_i^{n+1} - u_i^n}{k} + f_{i+1/2}^n - f_{i-1/2}^n = 0 \\ u_i^0 = \frac{1}{h_i} \int_{K_i} u_0(x) dx. \end{cases} \quad (5.36)$$

Pour que ce schéma soit complètement défini, il reste à préciser  $f_{i+1/2}^n$  en fonction des inconnues discrètes  $u_i^n$ . Un premier choix possible est le schéma centré,

$$f_{i+1/2}^n = \frac{f(u_{i+1}^n) + f(u_i^n)}{2}$$

dont on a vu qu'il est à proscrire, puisque, dans le cas linéaire, il est instable. Rappelons que dans le cas linéaire, le choix décentré amont donne

$$\text{si } f(u) = u, f_{i+1/2}^n = f(u_i^n), \text{ et}$$

$$\text{si } f(u) = -u, f_{i+1/2}^n = f(u_{i+1}^n).$$

Dans le cadre de ce cours, on va s'intéresser aux schémas les plus simples à trois points, c.à.d. que l'équation associée à l'inconnue  $u_i^n$  fait intervenir les trois inconnues discrètes  $u_i^n, u_{i-1}^n$  et  $u_{i+1}^n$ . Le flux numérique  $g$  s'écrit sous la forme

$$f_{i+1/2}^n = g(u_i^n, u_{i+1}^n).$$

Pour obtenir un "bon" schéma, on va choisir un flux "monotone", au sens suivant :

**Définition 5.31** On dit que qu'une fonction  $g$  définie de  $\mathbb{R}^2$  dans  $\mathbb{R}$  est un flux monotone pour la discrétisation de (5.17), si

1.  $g$  est consistante par rapport à  $f$ , c.à.d.  $g(u, u) = f(u)$ ,
2.  $g$  est croissante par rapport à la première variable et décroissante par rapport à la deuxième variable,
3.  $g$  est lipschitzienne sur  $[A, B]$ , où  $A = \inf_{\mathbb{R}} u_0$  et  $B = \sup_{\mathbb{R}} u_0$ .

**Remarque 5.32 (Flux monotones et schémas monotones)** Si le schéma 5.15 est à flux monotone, et s'il vérifie la condition de CFL, on peut alors montrer que le schéma est monotone, c.à.d. qu'il s'écrit sous la forme :

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n),$$

où  $H$  est une fonction croissante de ses trois arguments.

**Cas où  $f$  est monotone** Pour illustrer le choix de  $g$ , supposons par exemple que  $f$  soit croissante. Un choix très simple consiste alors à prendre  $g(u_i^n, u_{i+1}^n) = f(u_i^n)$ . On vérifie (exercice) que dans ce cas, les trois conditions ci-dessus sont vérifiées, ce schéma est dit décentré amont. On vérifiera qu'on retrouve le schéma décentré amont exposé dans le cas linéaire. De même si  $f$  est décroissante on peut facilement vérifier que le choix  $g(u_i^n, u_{i+1}^n) = f(u_{i+1}^n)$  convient.

**Schéma à décomposition de flux** Le schéma à décomposition de flux, appelé aussi "flux splitting" en anglais, consiste comme le nom l'indique à décomposer  $f = f_1 + f_2$ , où  $f_1$  est croissante et  $f_2$  décroissante, et à prendre pour  $g$  :

$$g(u_i^n, u_{i+1}^n) = f_1(u_i^n) + f_2(u_{i+1}^n)$$

**Schéma de Lax Friedrich** Le schéma de Lax Friedrich consiste à modifier le schéma centré de manière à le rendre stable. On écrit donc :

$$g(u_i^n, u_{i+1}^n) = \frac{1}{2}(f(u_i^n) + f(u_{i+1}^n)) + D(u_i^n - u_{i+1}^n)$$

où  $D \geq 0$  est il faut avoir  $D$  suffisamment grand pour que  $g$  soit croissante par rapport à la première variable et décroissante par rapport à la seconde variable.

**Schéma de Godunov** Le schéma de Godunov<sup>1</sup> est un des schémas les plus connus pour les équations hyperboliques non linéaires. De nombreux schémas pour les systèmes ont été inspirés par ce schéma. Le flux numérique du schéma de Godunov s'écrit :

$$g(u_i^n, u_{i+1}^n) = f(w_R(u_i^n, u_{i+1}^n)) \quad (5.37)$$

où  $w_R(u_i^n, u_{i+1}^n)$  est la solution en 0 du problème de Riemann avec conditions  $u_i^n, u_{i+1}^n$ , qui s'écrit :

$$\begin{cases} u_t + (f(u))_x = 0 \\ u_0(x) = \begin{cases} u_g = u_i^n & w < 0 \\ u_d = u_{i+1}^n & w > 0 \end{cases} \end{cases}$$

On peut montrer que le flux de Godunov (5.37) vérifie les conditions de la définition 5.31.

**Schéma de Murman** Une manière de simplifier le schéma de Godunov est de remplacer la résolution du problème de Riemann linéaire. On prend alors  $g(u_i^n, u_{i+1}^n) = f(\tilde{w}_R(u_i^n, u_{i+1}^n))$  où  $\tilde{w}_R(u_i^n, u_{i+1}^n)$  est solution de

$$\begin{cases} u_t + \alpha u_x = 0 \\ u_0(x) = \begin{cases} u_i^n & x < 0 \\ u_{i+1}^n & x > 0 \end{cases} \end{cases}$$

Comme le problème est linéaire, la solution de ce problème est connue :  $u(x, t) = u_0(x - \alpha t)$ . Le schéma est donc très simple, malheureusement, le schéma de Murman n'est pas un schéma monotone (voir exercice (71), car le flux n'est pas monotone par rapport aux deux variables. De fait on peut montrer que les solutions approchées peuvent converger vers des solutions non entropiques. On peut alors envisager une procédure "correction d'entropie"...

**Théorème 5.33 (Stabilité et convergence)** Soit  $(u_i^n)_{\substack{i \in \mathbb{Z} \\ n \in \mathbb{N}}}$  donnée par le schéma

$$\begin{cases} h_i \frac{u_i^{n+1} - u_i^n}{k} + g(u_i^n, u_{i+1}^n) - g(u_{i-1}^n, u_i^n) = 0 \\ u_i^0 = \frac{1}{h_i} \int_{K_i} u_0(x) dx \end{cases}$$

On suppose que  $g$  est un flux monotone au sens de la définition 5.31. On suppose de plus que :

$$k \leq \frac{\alpha h}{2M}, \text{ et } \alpha h \leq h_i \leq h, \forall i,$$

où  $M$  est la constante de Lipschitz de  $g$  sur  $[A, B]$ , et  $A$  et  $B$  sont tels que  $A \leq u_0(x) \leq B$  p.p.. On a alors  $A \leq u_i^n \leq B$  p.p., et  $\|u_{\tau, k}\| \leq \|u_0\|_\infty$ . Sous les mêmes hypothèses, si on note  $u_{\tau, k}$  la solution approchée définie par (5.16), alors

$u_{\tau, k}$  tend vers  $u$ , solution entropique de (5.17) dans  $L^1_{loc}(\mathbb{R} \times \mathbb{R}_+)$  lorsque  $h$  (et  $k$ ) tend vers 0.

<sup>1</sup>Sergei K. Godunov est un mathématicien russe né en 1929, membre de l'Académie des Sciences russe, en poste au Sobolev Institute of Mathematics, Novosibirsk, Sibérie

## 5.6 Exercices

**Exercice 60 (Problème linéaire en dimension 1)** *Corrigé en page 205*

Calculer la solution faible du problème :

$$\begin{cases} u_t - 2u_x = 0, & x \in \mathbb{R}, t \in \mathbb{R}_+ \\ u(x, 0) = \begin{cases} 0 & \text{si } x < 0, \\ 1 & \text{sinon.} \end{cases} \end{cases} \quad (5.38)$$

1. Tracer sur un graphique la solution à  $t = 0$  et à  $t = 1$ , en fonction de  $x$ . Cette solution faible est-elle solution classique de (5.38) ?

2. Même question en remplaçant la condition initiale par  $u(x, 0) = \sin x$ .

**Exercice 61 (Problème linéaire en dimension 2)** *Suggestions en page 205, Corrigé en page 205*

Soit  $\mathbf{v} \in \mathbb{R}^2$  et soit  $u_0 \in C^1(\mathbb{R}^2, \mathbb{R})$ . On considère le problème de Cauchy suivant :

$$\begin{cases} u_t + \operatorname{div}(\mathbf{v}u) = 0, \\ u(x, 0) = u_0(x), \end{cases} \quad (5.39)$$

Calculer la solution du problème (5.39) (en fonction de  $u_0$ ) en tout point  $(x, t) \in \mathbb{R}^2 \times \mathbb{R}$ .

**Exercice 62 (Schéma de Lax–Wendroff)** *Corrigé en page 205*

Soit  $u_0 \in C^\infty(\mathbb{R}, \mathbb{R})$  et  $T > 0$ , et  $a > 0$ . On considère le problème suivant :

$$\partial_t u(x, t) + a \partial_x u(x, t) = 0, \quad x \in \mathbb{R}, t \in [0, T], \quad (5.40a)$$

$$u(x, 0) = u_0(x). \quad (5.40b)$$

Ce problème admet une et une seule solution classique, notée  $u$ . On se donne un pas de temps constant  $k$ , avec  $k = \frac{T}{N+1}$  ( $N \in \mathbb{N}$ ). On se donne également un pas d'espace constant  $h$ , et des points de discrétisation en espace,  $(x_i)_{i \in \mathbb{Z}}$  tels que  $x_{i+1} - x_i = h$  pour tout  $i$ . On pose  $t_n = nk$ , pour  $n \in \{0, \dots, N+1\}$ . On cherche une approximation de  $u(x_j, t_n)$  pour  $n \in \{0, \dots, N+1\}$  et  $i \in \mathbb{Z}$ . On pose  $\lambda = \frac{ak}{h}$ .

1. Montrer que la solution  $u$  de (5.40) satisfait :

Pour tout  $j \in \mathbb{Z}$  et pour tout  $n \in \mathbb{N}$ ,

$$u(x_j, t_{n+1}) = u(x_j, t_n) - ak \partial_x u(x_j, t_n) + \frac{1}{2} a^2 k^2 \partial_{xx}^2 u(x_j, t_n) + k^3 \varepsilon(k), \quad \text{avec } \varepsilon(k) \rightarrow 0 \text{ lorsque } k \rightarrow 0, \quad (5.41)$$

et

$$u(x_j, t_{n+1}) = u(x_{j-1}, t_n) \text{ si } \lambda = 1. \quad (5.42)$$

2. Montrer pourquoi l'égalité (5.41) suggère le schéma suivant, dit de Lax-Wendroff :

$$\begin{cases} u_j^{(n+1)} = u_j^{(n)} - \frac{1}{2} \lambda (u_{j+1}^{(n)} - u_{j-1}^{(n)}) + \frac{1}{2} \lambda^2 (u_{j+1}^{(n)} - 2u_j^{(n)} + u_{j-1}^{(n)}), & j \in \mathbb{Z}, n > 0, \\ u_j^{(0)} = u_0(x_j); & j \in \mathbb{Z}. \end{cases} \quad (5.43)$$

avec  $u_j^{(0)} = u_0(x_j)$  pour tout  $j \in \mathbb{Z}$ . Donner l'ordre de consistance du schéma (distinguer les cas  $\lambda \neq 1$  et  $\lambda = 1$ ).

3. On prend comme condition initiale  $u^0(x) = e^{ipx}$ , pour  $p \in \mathbb{Z}$  fixé (avec  $i^2 = -1$ ). Pour  $j \in \mathbb{Z}$ , calculer la valeur  $u_j^{(1)}$  donnée par le schéma (5.43) en fonction de  $u_j^{(0)}$  et en déduire le facteur d'amplification  $\xi_p$ , tel que  $u_j^{(1)} = \xi_p u_j^{(0)}$ .

Montrer que le schéma est stable au sens de Von Neumann sous condition de CFL.

4. Montrer par un contre exemple que si  $\lambda \neq 1$ , le schéma n'est pas stable pour la norme  $L^\infty$ . [On pourra par exemple prendre  $u_j^{(0)} = 0$  pour  $j < 0$ ,  $u_j^{(0)} = 1$  pour  $j \geq 0$ , et calculer  $u_0^{(1)}$  et montrer ainsi que le schéma est instable si  $\lambda < 1$ , et chercher ensuite un contre-exemple pour le cas  $\lambda \geq 1$ .]

**Exercice 63 (Stabilité du schéma amont dans le cas linéaire)** *Corrigé en page 205*

On considère le problème hyperbolique linéaire (5.9), avec  $u_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ , dont on calcule une solution approchée par le schéma volumes finis amont (5.15). Montrer que ce schéma est stable pour les normes  $L^1$ ,  $L^2$  et  $L^\infty$ , c.à.d. que la solution approchée satisfait les propriétés suivantes :

1.  $\|u_{\mathcal{T},k}(\cdot, n)\|_{L^1(\mathbb{R})} \leq \|u_0\|_{L^1(\mathbb{R})}, \forall n \in \mathbb{N}$ ,
2.  $\|u_{\mathcal{T},k}(\cdot, n)\|_{L^2(\mathbb{R})} \leq \|u_0\|_{L^2(\mathbb{R})}, \forall n \in \mathbb{N}$ ,
3.  $\|u_{\mathcal{T},k}(\cdot, n)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}, \forall n \in \mathbb{N}$ ,

où  $u_{\mathcal{T},k}$  désigne la solution approchée calculée par le schéma (voir (5.16)).

**Exercice 64 (Convergence des schémas DFDA et VFDA dans le cas linéaire)**

*Corrigé en page 206*

Soit  $u_0 \in C^2(\mathbb{R}, \mathbb{R})$  et  $T \in \mathbb{R}_+^*$ . On suppose que  $u_0$ ,  $u'_0$  et  $u''_0$  sont bornées (sur  $\mathbb{R}$ ). On considère le problème suivant :

$$u_t(x, t) + u_x(x, t) = 0, \quad x \in \mathbb{R}, \quad t \in [0, T], \quad (5.44)$$

$$u(x, 0) = u_0(x). \quad (5.45)$$

Ce problème admet une et une seule solution classique, notée  $u$ . On se donne un pas de temps,  $k$ , avec  $k = \frac{T}{N+1}$  ( $N \in \mathbb{N}$ ), et des points de discrétisation en espace,  $(x_i)_{i \in \mathbb{Z}}$ . On pose  $t_n = nk$ , pour  $n \in \{0, \dots, N+1\}$ , et  $h_{i+\frac{1}{2}} = x_{i+1} - x_i$ , pour  $i \in \mathbb{Z}$ . On note  $\bar{u}_i^n = u(t_n, x_i)$  (pour  $n \in \{0, \dots, N+1\}$  et  $i \in \mathbb{Z}$ ), et on cherche une approximation de  $\bar{u}_i^n$ .

1. Soient  $\alpha, \beta \in \mathbb{R}$ . On suppose que, pour un certain  $h \in \mathbb{R}$ ,  $\alpha h \leq h_{i+\frac{1}{2}} \leq \beta h$ , pour tout  $i \in \mathbb{Z}$ . On considère, dans cette question le schéma suivant, appelé DFDA (pour Différences Finies Décentré Amont) :

$$\frac{u_i^{n+1} - u_i^n}{k} + \frac{1}{h_{i-\frac{1}{2}}}(u_i^n - u_{i-1}^n) = 0, \quad n \in \{0, \dots, N\}, \quad i \in \mathbb{Z}, \quad (5.46)$$

$$u_i^0 = u_0(x_i), \quad i \in \mathbb{Z}. \quad (5.47)$$

- (a) (Stabilité) Montrer que  $k \leq \alpha h \Rightarrow \inf(u_0) \leq u_i^n \leq \sup(u_0), \forall n \in \{0, \dots, N+1\}, \forall i \in \mathbb{Z}$ .
- (b) (Convergence) Montrer que, si  $k \leq \alpha h$ , on a :

$$\sup_{i \in \mathbb{Z}} |u_i^n - \bar{u}_i^n| \leq CT(k+h), \quad \forall n \in \{0, \dots, N+1\},$$

où  $C$  ne dépend que de  $u_0$  et  $\beta$ .

2. On suppose maintenant que  $x_i$  est le centre de la maille  $M_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ , pour  $i \in \mathbb{Z}$ . On pose  $h_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$ . Soient  $\alpha, \beta \in \mathbb{R}$ . On suppose que, pour un certain  $h \in \mathbb{R}$ ,  $\alpha h \leq h_i \leq \beta h$ , pour tout  $i \in \mathbb{Z}$ . On considère, dans cette question le schéma suivant, appelé VFDA (pour Volumes Finis Décentré Amont) :

$$h_i \frac{u_i^{n+1} - u_i^n}{k} + (u_i^n - u_{i-1}^n) = 0, \quad n \in \{0, \dots, N\}, \quad i \in \mathbb{Z}, \quad (5.48)$$

$$u_i^0 = \frac{1}{h_i} \int_{M_i} u_0(x) dx, \quad i \in \mathbb{Z}. \quad (5.49)$$

- (a) (Stabilité) Montrer que  $k \leq \alpha h \Rightarrow \inf(u_0) \leq u_i^n \leq \sup(u_0), \forall n \in \{0, \dots, N+1\}, \forall i \in \mathbb{Z}$ .
- (b) Étudier la consistance du schéma au sens DF.



(c) (Convergence) On pose  $\bar{u}_i^n = u(t_n, x_{i+\frac{1}{2}})$ . Montrer que, si  $k \leq \alpha h$ , on a :

$$\sup_{i \in \mathbb{Z}} |u_i^n - \bar{u}_i^n| \leq C_1(k+h), \forall n \in \{0, \dots, N+1\},$$

où  $C_1$  ne dépend que de  $u_0, \beta$  et  $T$ . En déduire que :

$$\sup_{i \in \mathbb{Z}} |u_i^n - \bar{u}_i^n| \leq C_2(k+h), \forall n \in \{0, \dots, N+1\},$$

où  $C_2$  ne dépend que de  $u_0, \beta$  et  $T$ .

**Exercice 65 (Eq. lin., sol. faible, conv. des schémas VFDA et DFDA, méthode VF)**

Corrigé en page 207

Soit  $u_0 \in L^\infty(\mathbb{R}) \cap L^2(\mathbb{R})$  et  $T \in \mathbb{R}_+^*$ . On considère le problème suivant :

$$u_t(x, t) + u_x(x, t) = 0, x \in \mathbb{R}, t \in [0, T], \quad (5.50)$$

$$u(x, 0) = u_0(x). \quad (5.51)$$

Ce problème admet une et une seule solution faible, notée  $u$ . On se donne un pas de temps,  $k$ , avec  $k = \frac{T}{N+1}$  ( $N \in \mathbb{N}$ ), et on pose  $t_n = nk$ , pour  $n \in \{0, \dots, N+1\}$ ; on se donne des points de discrétisation en espace,  $(x_i)_{i \in \mathbb{Z}}$ , et on suppose que  $x_i$  est le centre de la maille  $M_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}]$ , pour  $i \in \mathbb{Z}$ . On pose  $h_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}$  et  $h_{i+\frac{1}{2}} = x_{i+1} - x_i$ . Soient  $\alpha, \beta \in \mathbb{R}$ . On suppose que, pour un certain  $h \in \mathbb{R}$ ,  $\alpha h \leq h_i \leq \beta h$ , pour tout  $i \in \mathbb{Z}$ . On considère le schéma (5.48), (5.49) (schéma "VFDA").

1. (Stabilité  $L^\infty$ ) Montrer que  $k \leq \alpha h \Rightarrow |u_i^n| \leq \|u_0\|_\infty, \forall n \in \{0, \dots, N+1\}, \forall i \in \mathbb{Z}$ .
2. Montrer que, pour tout  $n = 0, \dots, N$ , on a  $u_i^n \rightarrow 0$  lorsque  $i \rightarrow +\infty$  ou  $i \rightarrow -\infty$ .
3. (Estimation "BV faible") Soient  $\zeta > 0$ . Montrer que :

$$k \leq (1 - \zeta)\alpha h \Rightarrow \sum_{n=0, \dots, N} \sum_{i \in \mathbb{Z}} k(u_i^n - u_{i-1}^n)^2 \leq C(\zeta, u_0),$$

où  $C(\zeta, u_0)$  ne dépend que de  $\zeta$  et  $u_0$  (multiplier (5.48) par  $ku_i^n$  et sommer sur  $i$  et  $n$ .)

4. (convergence) On pose  $\mathcal{T} = (M_i)_{i \in \mathbb{Z}}$  et on définit la solution approchée sur  $[0, T] \times \mathbb{R}$ , notée  $u_{\mathcal{T}, k}$ , donnée par (5.48), (5.49), par  $u_{\mathcal{T}, k}(t, x) = u_i^n$ , si  $x \in M_i$  et  $t \in [t_n, t_{n+1}[$ .  
On admet que  $u_{\mathcal{T}, k} \rightarrow u$ , pour la topologie faible- $\star$  de  $L^\infty([0, T] \times \mathbb{R})$ , quand  $h \rightarrow 0$ , avec  $k \leq (1 - \zeta)\alpha h$  ( $\zeta$  fixé). Montrer que  $u$  est la solution faible de (5.50)-(5.51).

**Remarque 5.34 (VF, DF et convergence forte)** On peut montrer le même résultat avec (5.46) au lieu de (5.48). On peut aussi montrer (cf. la suite du cours...) que la convergence est forte dans  $L_{loc}^p([0, T] \times \mathbb{R})$ , pour tout  $p < \infty$ .

**Exercice 66 (Construction d'une solution faible)** Corrigé en page 210

1/ Construire une solution faible du problème

$$\begin{cases} u_t + (u^2)_x = 0 \\ u(x, 0) = u_0(x) = \begin{cases} 1 & \text{si } x < 0 \\ 1-x & \text{si } x \in [0, 1] \\ 0 & \text{si } x > 1 \end{cases} \end{cases}$$

2/ Même question (mais nettement plus difficile...) pour le problème

$$\begin{cases} u_t + (u^2)_x = 0 \\ u(x, 0) = u_0(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1-x & \text{si } x \in [0, 1] \\ 1 & \text{si } x > 1 \end{cases} \end{cases}$$

**Exercice 67 (Problème de Riemann)**

Soit  $f$  la fonction de  $\mathbb{R}$  dans  $\mathbb{R}$  définie par  $f(s) = s^4$ . Soit  $u_d$  et  $u_g$  des réels. Calculer la solution entropique du problème de Riemann (5.30) avec données  $u_d$  et  $u_g$  en fonction de  $u_d$  et  $u_g$ .

**Exercice 68 (Non unicité des solutions faibles)** *Corrigé en page 211*

On considère l'équation

$$\begin{cases} u_t + (u^2)_x = 0 \\ u(0, x) = \begin{cases} u_g & \text{si } x < 0 \\ u_d & \text{si } x > 0 \end{cases} \end{cases} \quad (5.52)$$

avec  $u_g < u_d$ .

1. Montrer qu'il existe  $\sigma \in \mathbb{R}$  tel que si  $\begin{cases} u(t, x) = u_g & \text{si } x < \sigma t \\ u(t, x) = u_d & \text{si } x > \sigma t \end{cases}$  alors  $u$  est solution faible de (5.52). Vérifier que  $u$  n'est pas solution entropique de (5.52).

2. Montrer que  $u$  définie par :

$$\begin{cases} u(t, x) = u_g & \text{si } x < 2u_g t \\ u(t, x) = \frac{x}{2t} & \text{si } 2u_g t \leq x \leq 2u_d t \\ u(t, x) = u_d & \text{si } x > 2u_d t \end{cases} \quad (5.53)$$

alors  $u$  est solution faible entropique de (5.52).

**Exercice 69 (Problème de Riemann)**

1. Déterminer la solution entropique de (5.30) dans le cas où  $f$  est strictement concave.
2. On se place dans le cas où  $f$  est convexe puis concave : plus précisément, on considère  $f \in C^2(\mathbb{R}, \mathbb{R})$  avec
  - (i)  $f(0) = 0$ ,  $f'(0) = f'(1) = 0$
  - (ii)  $\exists a \in ]0, 1[$ , tel que  $f$  est strictement convexe sur  $]0, a[$ ,  $f$  est strictement concave sur  $]a, 1[$ .

On supposera de plus  $u_g = 1$ ,  $u_d = 0$ .

- (a) Soit  $b$  l'unique élément  $b \in ]a, 1[$  tel que  $\frac{f(b)}{b} = f'(b)$  ; montrer que  $u$  définie par :

$$\begin{cases} u(t, x) = 1 & \text{si } x \leq 0 \\ u(t, x) = \xi & \text{si } x = f'(\xi)t, b < \xi < 1 \\ u(t, x) = 0 & \text{si } x > f'(b)t \end{cases}$$

est la solution faible entropique de (5.30) (sous les hypothèses précédentes).

- (b) Construire la solution entropique du problème de Riemann dans le cas  $f(u) = \frac{u^2}{u^2 + \frac{(1-u)^2}{4}}$  et  $u_g, u_d \in [0, 1]$ . [Complicé. On distinguera plusieurs cas.]

**Exercice 70 (Stabilité de schémas numériques)** *Corrigé en page 211*

Soient  $f \in C^1(\mathbb{R}, \mathbb{R})$  et  $u_0 \in L^\infty(\mathbb{R})$ . On considère le problème suivant :

$$u_t(x, t) + (f(u))_x(x, t) = 0, \quad x \in \mathbb{R}, \quad t \in [0, T], \quad (5.54)$$

$$u(x, 0) = u_0(x). \quad (5.55)$$

On utilise ci dessous les notations du cours. On discrétise le problème (5.54),(5.55) par l'un des schémas vu en cours ("Flux-splitting", "Godunov", "Lax-Friedrichs modifié" et "Murman"). Montrer qu'il existe  $M$  (dépendant de la fonction "flux numérique" et de  $u_0$ ) tel que  $k \leq Mh_i$ , pour tout  $i \in \mathbb{Z}$ , implique :

1.  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$  pour tout  $n \in \mathbb{N}$ .
2. (Plus difficile)  $\sum_{i \in \mathbb{Z}} |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i \in \mathbb{Z}} |u_{i+1}^n - u_i^n|$  pour tout  $n \in \mathbb{N}$ . (Cette estimation n'est intéressante que si  $\sum_{i \in \mathbb{Z}} |u_{i+1}^0 - u_i^0| < \infty$ , ce qui n'est pas toujours vrai pour  $u_0 \in L^\infty(\mathbb{R})$ . Cela est vrai si  $u_0$  est une fonction à "variation bornée".)

**Exercice 71 (Schéma de Murman)** *Corrigé en page 212*

Soient  $f \in C^1(\mathbb{R}, \mathbb{R})$  et  $u_0 \in L^\infty(\mathbb{R})$ . On suppose que  $A \leq u_0 \leq B$ , p.p. sur  $\mathbb{R}$ . On s'intéresse au problème suivant :

$$u_t(x, t) + (f(u))_x(x, t) = 0, \quad x \in \mathbb{R}, \quad t \in \mathbb{R}_+, \quad (5.56)$$

$$u(x, 0) = u_0(x), \quad x \in \mathbb{R}. \quad (5.57)$$

Pour discrétiser le problème (5.56)-(5.57), on se donne un pas d'espace  $h > 0$  et un pas de temps  $k > 0$ . On pose  $M_i = ]ih, ih + h[$  et on note  $u_i^n$  l'approximation recherchée de la solution exacte dans la maille  $M_i$  à l'instant  $nk$ . On considère le schéma de Murmann :

$$h \frac{u_i^{n+1} - u_i^n}{k} + (f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n) = 0, \quad n \in \mathbb{N}, \quad i \in \mathbb{Z}, \quad (5.58)$$

$$u_i^0 = \frac{1}{h} \int_{M_i} u_0(x) dx, \quad i \in \mathbb{Z}, \quad (5.59)$$

avec  $f_{i+\frac{1}{2}}^n = g(u_i^n, u_{i+1}^n)$  et  $g \in C(\mathbb{R} \times \mathbb{R}, \mathbb{R})$  définie par :

$g(a, a) = f(a)$  et, pour  $a \neq b$ ,

$$g(a, b) = \begin{cases} f(a) & \text{si } \frac{f(b)-f(a)}{b-a} \geq 0, \\ f(b) & \text{si } \frac{f(b)-f(a)}{b-a} < 0. \end{cases}$$

1. (Stabilité) Montrer qu'il existe  $M$ , ne dépendant que de  $f$ ,  $A$  et  $B$  (on donnera la valeur de  $M$  en fonction de  $f$ ,  $A$  et  $B$ ) t.q. pour  $k \leq Mh$  on ait :

(a) (Stabilité  $L^\infty$ )  $A \leq u_i^n \leq B$ , pour tout  $n \in \mathbb{N}$  et tout  $i \in \mathbb{Z}$ ,

(b) (Stabilité BV)  $\sum_{i \in \mathbb{Z}} |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i \in \mathbb{Z}} |u_{i+1}^n - u_i^n|$  pour tout  $n \in \mathbb{N}$ . (Cette estimation n'est intéressante que si  $\sum_{i \in \mathbb{Z}} |u_{i+1}^0 - u_i^0| < \infty$ , ce qui n'est pas toujours vrai pour  $u_0 \in L^\infty(\mathbb{R})$ . Cela est vrai si  $u_0$  est une fonction à "variation bornée".)

2. On prend, dans cette question,  $f(s) = s^2$ .

(a) (Non monotonie) Montrer que si  $A < 0$  et  $B > 0$ , la fonction  $g$  n'est pas "croissante par rapport à son premier argument et décroissante par rapport à son deuxième argument" sur  $[A, B]^2$ .

(b) (Exemple de non convergence) Donner un exemple de non convergence du schéma. Plus précisément, donner  $u_0$  t.q., pour tout  $h > 0$  et tout  $k > 0$ , on ait  $u_i^n = u_i^0$  pour tout  $i \in \mathbb{Z}$  et pour tout  $n \in \mathbb{N}$  (la solution discrète est donc "stationnaire") et pourtant  $u(\cdot, T)$  ( $u$  est la solution exacte de (5.56)-(5.57)) est différent de  $u_0$  pour tout  $T > 0$  (la solution exacte n'est donc pas stationnaire).

3. (Schéma "ordre 2", question plus difficile) Pour avoir un schéma "plus précis", on pose maintenant  $p_i^n = \min \text{mod}(\frac{u_{i+1}^n - u_{i-1}^n}{2h}, 2\frac{u_{i+1}^n - u_i^n}{h}, 2\frac{u_i^n - u_{i-1}^n}{h})$  et on remplace, dans le schéma précédent,  $f_{i+\frac{1}{2}}^n = g(u_i^n, u_{i+1}^n)$  par  $f_{i+\frac{1}{2}}^n = g(u_i^n + (h/2)p_i^n, u_{i+1}^n - (h/2)p_{i+1}^n)$ . Reprendre les 2 questions précédentes (c'est à dire : "Stabilité  $L^\infty$ ", "Stabilité BV", "non monotonie" et "Exemple de non convergence").

### Exercice 72 (Flux monotones)

Soient  $f \in C^1(\mathbb{R}, \mathbb{R})$  et  $u_0 \in L^\infty(\mathbb{R})$  ; on considère l'équation hyperbolique non linéaire (5.17) qu'on rappelle :

$$\begin{cases} u_t + (f(u))_x = 0, & (x, t) \in \mathbb{R} \times \mathbb{R}_+, \\ u(x, 0) = u_0(x). \end{cases} \quad (5.60)$$

On se donne un maillage  $(]x_{i-1/2}, x_{i+1/2}[)_{i \in \mathbb{Z}}$  de  $\mathbb{R}$  et  $k > 0$  et, pour  $i \in \mathbb{Z}$ , on définit une condition initiale approchée :  $u_i^0 = \frac{1}{h_i} \int_{M_i} u_0(x) dx$ , avec  $h_i = x_{i+1/2} - x_{i-1/2}$ .

Pour calculer la solution entropique de l'équation (5.17), on considère un schéma de type volumes finis explicite à trois points, défini par un flux numérique  $g$ , fonction de deux variables.

1. Ecrire le schéma numérique (i.e. donner l'expression de  $u_i^{n+1}$  en fonction des  $(u_i^n)_{i \in \mathbb{Z}}$ ).

2. On suppose dans cette question que le flux  $g$  est monotone et lipschitzien en ses deux variables, c.à.d. qu'il existe  $M \geq 0$  tel que pour tout  $(x, y, z) \in \mathbb{R}^3$ ,  $|g(x, z) - g(y, z)| \leq M|x - y|$  et  $|g(x, y) - g(x, z)| \leq M|y - z|$ . Montrer que le schéma numérique de la question précédente peut s'écrire sous la forme

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n)$$

où  $H$  est une fonction croissante de ses trois arguments si  $k$  satisfait une condition de type  $k \leq Ch_i$  pour tout  $i$ , où  $C$  est une constante à déterminer.

3. Montrer que si la fonction  $g$  est croissante par rapport à son premier argument et décroissant par rapport au second, et si  $a, b \in \mathbb{R}$  sont tels que  $a \leq b$ , alors  $g(a, b) \leq g(\xi, \xi)$  pour tout  $\xi \in [a, b]$ .

4. En déduire que si le flux  $g$  est monotone, alors il vérifie la propriété suivante :

$$\forall (a, b) \in \mathbb{R}^2, \begin{cases} g(a, b) \leq \min_{s \in [a, b]} f(s) \text{ si } a \leq b \\ g(a, b) \geq \max_{s \in [a, b]} f(s) \text{ si } a \geq b. \end{cases}$$

5. Soit  $g$  un flux monotone qui est tel que pour tous  $a, b \in \mathbb{R}$ , il existe  $u_{a,b}$  dans l'intervalle d'extrémités  $a$  et  $b$ , tel que  $g(a, b) = f(u_{a,b})$ . Montrer que

$$\forall (a, b) \in \mathbb{R}^2, \begin{cases} g(a, b) = \min_{s \in [a, b]} f(s) \text{ si } a \leq b \\ g(a, b) = \max_{s \in [b, a]} f(s) \text{ si } a \geq b. \end{cases}$$

### Exercice 73 (Schémas pour les problèmes hyperboliques)

Soient  $f \in C^2(\mathbb{R}, \mathbb{R})$ ,  $T > 0$  et  $u_0 \in L^\infty(\mathbb{R}) \cap BV(\mathbb{R})$ ; on cherche une approximation de la solution de l'équation hyperbolique avec condition initiale :

$$u_t(x, t) + (f(u))_x(x, t) = 0, \quad x \in \mathbb{R}, \quad t \in [0, T], \quad (5.61)$$

$$u(x, 0) = u_0(x). \quad (5.62)$$

On note  $h$  (resp.  $k = \frac{1}{N+1}$ ) le pas (constant, pour simplifier) de la discrétisation en espace (resp. en temps), et  $u_i^n$  la valeur approchée recherchée de  $u$  au temps  $nk$  dans la maille  $M_i = [(i - \frac{1}{2})h, (i + \frac{1}{2})h]$ , pour  $n \in \{0, \dots, N+1\}$  et  $i \in \mathbb{Z}$ . On considère le schéma obtenu par une discrétisation par volumes finis explicite à trois points :

$$\frac{u_i^{n+1} - u_i^n}{k} + \frac{1}{h}(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n) = 0, \quad n \in \{0, \dots, N+1\}, \quad i \in \mathbb{Z}, \quad (5.63)$$

$$u_i^0 = \frac{1}{h} \int_{M_i} u_0(x) dx, \quad (5.64)$$

avec  $f_{i+\frac{1}{2}}^n = g(u_i^n, u_{i+1}^n)$ , où  $g \in C^1(\mathbb{R}, \mathbb{R})$ .

1. Montrer que le schéma (5.63),(5.64) possède la propriété de "consistance des flux" ssi  $g$  est telle que :

$$g(s, s) = f(s), \quad \forall s \in \mathbb{R}. \quad (5.65)$$

2. Montrer que le schéma, vu comme un schéma de différences finies, est, avec la condition (5.65), d'ordre 1 (c.à.d. que l'erreur de consistance est majorée par  $C(h+k)$ , où  $C$  ne dépend que de  $f$  et de la solution exacte, que l'on suppose régulière). Montrer que si le pas d'espace est non constant, la condition (5.65) est (en général) insuffisante pour assurer que le schéma (5.63),(5.64) (convenablement modifié) est consistant au sens des différences finies, et que le schéma est alors d'ordre 0.

3. On étudie, dans cette question, le schéma de Godunov, c.à.d. qu'on prend :

$$g(u_g, u_d) = f(u_{u_g, u_d}(0, t)),$$

où  $u_{u_g, u_d}$  est la solution du problème de Riemann :

$$u_t(x, t) + (f(u))_x(x, t) = 0, \quad (x, t) \in \mathbb{R} \times \mathbb{R}_+ \quad (5.66)$$

$$u(x, 0) = u_g \text{ si } x < 0, \quad (5.67)$$

$$u(x, 0) = u_d \text{ si } x > 0. \quad (5.68)$$

(a) Montrer que le schéma (5.63), (5.64) peut s'écrire :

$$u_i^{n+1} = u_i^n + C_i(u_{i+1}^n - u_i^n) + D_i(u_{i-1}^n - u_i^n),$$

$$\text{avec : } C_i = \frac{k}{h} \frac{f(u_i^n) - g(u_i^n, u_{i+1}^n)}{u_{i+1}^n - u_i^n} \geq 0 \text{ et } D_i = \frac{k}{h} \frac{g(u_{i-1}^n, u_i^n) - f(u_i^n)}{u_{i-1}^n - u_i^n} \geq 0.$$

(b) On pose  $A = \|u_0\|_\infty$ ,  $M = \sup_{s \in [-A, A]} |f'(s)|$ , et  $h$  le pas (constant) d'espace. On suppose que  $k$  et  $h$  vérifient la condition de CFL :

$$k \leq \frac{h}{2M}.$$

On note  $u^n$  la fonction définie par :  $u^n(x) = (u_i^n)$  si  $x \in M_i$  ; montrer que :

$$\text{– Stabilité } L^\infty : \|u^{n+1}\|_\infty \leq \|u^n\|_\infty (\leq \dots \leq \|u^0\|_\infty), \forall n \in \{0, \dots, N+1\}. \quad (E1)$$

$$\text{– Stabilité } BV : \|u^{n+1}\|_{BV} \leq \|u^n\|_{BV} (\leq \dots \leq \|u^0\|_{BV}), \forall n \in \{0, \dots, N+1\}. \quad (E2)$$

On rappelle que, comme  $u_n$  est une fonction constante par morceaux, on a :

$$\|u^n\|_{BV} = \sum_{i \in \mathbb{Z}} |u_{i+1}^n - u_i^n|.$$

(c) Remarque : on peut montrer (ce n'est pas facile) que si on a la condition "CFL" le schéma de Godunov converge.

4. On suppose maintenant  $f(u) = au$ ,  $a \in \mathbb{R}$ , et on prend  $g(\lambda, \mu) = \frac{\lambda + \mu}{2}$  (schéma centré).

Montrer que pour tous  $k, h > 0$ , les conditions (E1) et (E2) sont fausses, c.à.d. qu'il existe  $u_0 \in L^\infty \cap BV$  t.q.  $\|u^1\|_\infty \not\leq \|u_0\|_\infty$ , et  $\|u^1\|_{BV} \not\leq \|u_0\|_{BV}$ .

5. On étudie maintenant un schéma de type "MUSCL", i.e. On prend dans le schéma (5.63)  $f_{i+\frac{1}{2}}^n = f(u_i^n + \frac{h}{2} p_i^n)$ , où :

$$p_i^n = \begin{cases} \frac{\varepsilon_i^n}{2h} \min(|u_{i+1}^n - u_{i-1}^n|, 4|u_{i+1}^n - u_i^n|, 4|u_i^n - u_{i-1}^n|), & \text{où } \varepsilon_i^n = \text{sign}(u_{i+1}^n - u_{i-1}^n) \\ & \text{si } \text{sign}(u_{i+1}^n - u_{i-1}^n) = \text{sign}(u_{i+1}^n - u_i^n) = \text{sign}(u_i^n - u_{i-1}^n) \\ 0 & \text{sinon.} \end{cases}$$

(a) Montrer que  $\frac{1}{h}(f_{i+\frac{1}{2}}^n - f_{i-\frac{1}{2}}^n)$  est une approximation d'ordre 2 de  $(f(u))_x(x_i, t_n)$  aux points où  $u \in C^2$  et  $u_x \neq 0$ .

(b) Montrer que sous une condition de type  $k \leq Ch$ , où  $C$  ne dépend que de  $u_0$  et  $f$ , les conditions de stabilité (E1) et (E2) sont vérifiées.

### Exercice 74 (Éléments finis pour une équation hyperbolique)

Soit  $f \in C^1(\mathbb{R}, \mathbb{R})$ ,  $u_0 \in C(\mathbb{R})$  t.q.  $u_0$  bornée ; on considère la loi de conservation scalaire suivante :

$$\frac{\partial u}{\partial t}(x, t) + \frac{\partial}{\partial x}(f(u))(x, t) = 0, \quad x \in \mathbb{R}, \quad t \in \mathbb{R}_+, \quad (5.69)$$

avec la condition initiale :

$$u(x, 0) = u_0(x). \quad (5.70)$$

On se donne un pas de discrétisation en temps constant  $k$ , on note  $t_n = nk$  pour  $n \in \mathbb{N}$ , et on cherche à approcher  $u(\cdot, t_n)$ . On note  $u^{(n)}$  la solution approchée recherchée.

1. Montrer qu'une discrétisation par le schéma d'Euler explicite en temps amène au schéma en temps suivant :

$$\frac{1}{k}(u^{(n+1)} - u^{(n)}) + \frac{\partial}{\partial x}(f(u^{(n)}))(x) = 0, \quad x \in \mathbb{R}, \quad n \in \mathbb{N}^*, \quad (5.71)$$

$$u^0(x) = u_0(x). \quad (5.72)$$

On cherche à discrétiser (5.71) par une méthode d'éléments finis. On se donne pour cela une famille de points  $(x_i)_{i \in \mathbb{Z}} \subset \mathbb{R}$ , avec  $x_i < x_{i+1}$ .

2. On introduit les fonctions de forme  $P_1$ , notées  $\Phi_i, i \in \mathbb{Z}$ , des éléments finis associés au maillage donné par la famille de points  $(x_i)_{i \in \mathbb{Z}}$ ; on effectue un développement de Galerkin de  $u^{(n)}$  sur ces fonctions de forme dans (5.71) et (5.72); on multiplie l'équation ainsi obtenue par chaque fonction de forme, et on approche le terme  $f(\sum_{j \in \mathbb{Z}} u_j^{(n)} \Phi_j)$  par  $\sum_{j \in \mathbb{Z}} f(u_j^{(n)}) \Phi_j$ , et on intègre sur  $\mathbb{R}$ . Montrer qu'on obtient ainsi un système d'équations de la forme :

$$\sum_{j \in \mathbb{Z}} a_{i,j} \frac{u_j^{(n+1)} - u_j^{(n)}}{k} + \sum_{j \in \mathbb{Z}} b_{i,j} f(u_j^{(n)}) = 0, \quad i \in \mathbb{Z}, \quad n \in \mathbb{N}^*. \quad (5.73)$$

$$u_i^0 = u_0(x_i) \quad i \in \mathbb{Z}. \quad (5.74)$$

(les  $a_{i,j}$  et  $b_{i,j}$  sont à déterminer).

3. On effectue une "condensation de la matrice de masse", c.à.d. qu'on remplace les  $a_{i,j}$  dans (5.73) par  $\tilde{a}_{i,j}$  avec  $\tilde{a}_{i,j} = 0$  si  $i \neq j$  et  $\tilde{a}_{i,i} = \sum_{j \in \mathbb{Z}} a_{i,j}$ . Montrer que le schéma ainsi obtenu est identique à un schéma volumes finis sur le maillage  $(K_i)_{i \in \mathbb{Z}}$  où  $K_i = ]x_{i-1/2}, x_{i+1/2}[$ ,  $x_{i+1/2} = (x_i + x_{i+1})/2$ , avec approximation centrée du flux.

4. Montrer que ce schéma est instable, dans un (ou plusieurs) sens à préciser.

5. On remplace le flux numérique centré  $F_{i+1/2}$  du schéma volumes finis obtenu à la question 3 par  $G_{i+1/2} = F_{i+1/2} + D_{i+1/2}(u_i^{(n)} - u_{i+1}^{(n)})$ . Montrer que l'approximation du flux reste consistante et que si les  $D_{i+1/2}$  sont bien choisis, le nouveau schéma est stable sous une condition de CFL à préciser.

On considère maintenant la même équation de conservation, mais sur  $\mathbb{R}^2$  (avec  $f \in C^1(\mathbb{R}, \mathbb{R}^2)$ ,  $u_0 \in C(\mathbb{R}^2)$ , bornée).

$$u_t(x, t) + \operatorname{div}(f(u))(x, t) = 0, \quad x \in \mathbb{R}^2, \quad t \in \mathbb{R}_+, \quad (5.75)$$

$$u(x, 0) = u_0(x). \quad (5.76)$$

Soit  $\mathcal{T}$  un maillage en triangles de  $\mathbb{R}^2$ , admissible pour une discrétisation par éléments finis  $P_1$ . Soit  $\mathcal{S}$  l'ensemble des noeuds de ce maillage et  $(\Phi_j)_{j \in \mathcal{S}}$  la famille des fonctions de forme éléments finis bilinéaires  $P_1$ . En conservant la même discrétisation en temps, on cherche une approximation de  $u(\cdot, t_n)$  dans l'espace engendré par les fonctions  $\Phi_j$ .

6. Montrer qu'en suivant la même démarche qu'aux questions 2 et 3, on aboutit au schéma :

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} \int_{\mathbb{R}^2} \Phi_i(x) dx - \sum_{j \in \mathcal{S}} f(u_j^{(n)}) \cdot \int_{\mathbb{R}^2} \Phi_j(x) \nabla \Phi_i(x) dx = 0, \quad n \in \mathbb{N}^* \quad (5.77)$$

7. Montrer que ce schéma peut encore s'écrire :

$$\frac{u_i^{(n+1)} - u_i^{(n)}}{k} \int_{\mathbb{R}^2} \Phi_i(x) dx + \sum_{j \in \mathcal{S}} E_{i,j} = 0, \quad (5.78)$$

avec

$$E_{i,j} = \frac{1}{2} (f(u_i^{(n)}) + f(u_j^{(n)})) \cdot \int_{\mathbb{R}^2} (\Phi_i(x) \nabla \Phi_j(x) - \Phi_j(x) \nabla \Phi_i(x)) dx.$$

Montrer que ce schéma est instable.

8. Dans le schéma (5.78), on remplace  $E_{i,j}$  par

$$\tilde{E}_{i,j}^n = E_{i,j}^n + D_{i,j}(u_i^n - u_j^n),$$

où  $D_{i,j} = D_{j,i}$  (pour que le schéma reste conservatif). Montrer que pour un choix judicieux de  $D_{i,j}$ , le schéma ainsi obtenu est à flux monotone et stable sous condition de CFL.

## 5.7 Suggestions pour les exercices

### Exercice 61 page 197

Chercher les solutions sous la forme  $u(\mathbf{x}, t) = u_0(\mathbf{x} - \mathbf{v}t)$ .

## 5.8 Corrigés des exercices

### Exercice 60 page 197

1. En appliquant les résultats de la section 5.2 page 180, la solution faible du problème s'écrit  $u(x, t) = u_0(x + 2t)$ , pour  $x \in \mathbb{R}$ , et  $t \in \mathbb{R}_+$ , c.à.d.

$$\begin{cases} u(x, t) = \begin{cases} 0, & \text{si } x < -2t, \\ 1 & \text{si } x > -2t. \end{cases} \end{cases} \quad (5.79)$$

La représentation graphique de la solution à  $t = 0$  et à  $t = 1$ , en fonction de  $x$  est donnée en Figure 5.7. Cette

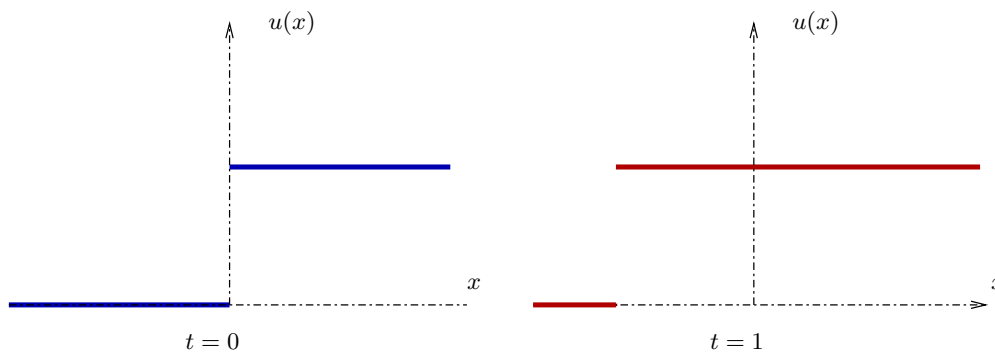


FIG. 5.7 – Représentation graphique de la solution

solution faible n'est pas solution classique de (5.79) car elle n'est pas continue, donc ses dérivées en temps et espace ne sont pas définies partout.

2. Dans le cas où  $u_0(x) = \sin x$ , la solution faible du problème s'écrit  $u(x, t) = \sin(x + 2t)$ , pour  $x \in \mathbb{R}$ , et  $t \in \mathbb{R}_+$ , et cette solution est régulière, donc solution classique.

### Exercice 61 page 197

Pour  $(x, t) \in \mathbb{R}^2 \times \mathbb{R}$ , on pose  $u(x, t) = u_0(x - \mathbf{v}t)$ . Comme  $u_0 \in C^1(\mathbb{R}, \mathbb{R})$ , on a  $u \in C^1(\mathbb{R}^2 \times \mathbb{R}_+, \mathbb{R})$ ; on peut donc calculer les dérivées partielles de  $u$  par rapport à au temps  $t$ , qu'on notera  $\partial_t u$  et par rapport aux deux variables d'espace  $x_1$  et  $x_2$ , qu'on notera  $\partial_1 u$  et  $\partial_2 u$ . On a :  $\partial_t u(x, t) = \nabla u_0(x - \mathbf{v}t) \cdot \mathbf{v}$ . Or  $\text{div}(\mathbf{v}u) = \mathbf{v} \cdot \nabla u$  car  $\mathbf{v}$  est constant, et  $\nabla u = \nabla u_0$ . On en déduit que  $u_t(x, t) + \text{div}(\mathbf{v}u)(x, t) = 0$ , et donc  $u$  est solution (classique) de (5.39).

### Exercice 62 page 197

Corrigé en cours d'élaboration

### Exercice 63 page 198 (Stabilité du schéma amont dans le cas linéaire)

On considère le problème hyperbolique linéaire (5.9), avec  $u_0 \in L^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ , dont on calcule une solution approchée par le schéma volumes finis amont (5.15). Montrer que ce schéma est stable pour les normes  $L^2$  et  $L^\infty$ , c.à.d. que la solution approchée satisfait les propriétés suivantes :

1.  $\|u_{\mathcal{T},k}(\cdot, n)\|_{L^2(\mathbb{R})} \leq \|u_0\|_{L^2(\mathbb{R})}, \forall n \in \mathbb{N}$ ,
2.  $\|u_{\mathcal{T},k}(\cdot, n)\|_{L^\infty(\mathbb{R})} \leq \|u_0\|_{L^\infty(\mathbb{R})}, \forall n \in \mathbb{N}$ ,

où  $u_{\mathcal{T},k}$  désigne la solution approchée calculée par le schéma (voir (5.16)).

Le schéma (5.15) s'écrit encore :

$$h_i(u_i^{n+1} - u_i^n) = k(u_i^n - u_{i-1}^n).$$

Multiplions par  $u_i^{n+1}$ . On obtient :

$$\frac{1}{2}h_i(u_i^{n+1} - u_i^n)^2 + \frac{1}{2}h_i(u_i^{n+1})^2 - \frac{1}{2}h_i(u_i^n)^2 + k(u_i^{n+1} - u_i^n)(u_i^n - u_{i-1}^n) + ku_i^n(u_i^n - u_{i-1}^n) = 0.$$

### Exercice 64 page 198

1.a) Le schéma numérique s'écrit :

$$u_i^{n+1} = \left(1 - \frac{k}{h_{i-\frac{1}{2}}}\right)u_i^n + \frac{k}{h_{i-\frac{1}{2}}}u_{i-1}^n \quad (5.80)$$

Comme  $k \leq \alpha h \leq h_{i-\frac{1}{2}}$  on a  $\frac{k}{h_{i-\frac{1}{2}}} \in [0, 1]$  On a donc

$$\min(u_i^n, u_{i-1}^n) \leq u_i^{n+1} \leq \max(u_i^n, u_{i-1}^n)$$

d'où on déduit que

$$\min_j(u_j^n) \leq u_i^{n+1} \leq \max_j(u_j^n), \quad \forall i \in \mathbb{Z},$$

puis, par récurrence sur  $n$ , que

$$\inf u_0 \leq u_i^n \leq \sup u_0 \quad \forall i \in \mathbb{Z}, \forall n \in \mathbb{N}.$$

1.b) Par définition de l'erreur de consistance, on a :

$$\begin{aligned} \frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} &= u_t(x_i, tn) + R_i^n, \text{ où } |R_i^n| \leq \|u_{tt}\|_\infty k, \\ \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_{i-\frac{1}{2}}} &= u_x(x_i, tn) + S_i^n, |S_i^n| \leq \|u_{xx}\|_\infty h. \beta \end{aligned}$$

En posant  $e_i^n = \bar{u}_i^n - u_i^n$ , on a donc

$$\frac{e_i^{n+1} - e_i^n}{k} + \frac{1}{h_{i-\frac{1}{2}}}(e_i^n - e_{i-1}^n) = R_i^n + S_i^n \leq C(u_0, \beta)(h + k),$$

avec  $C(u_0, \beta) = \|u_0''\|_\infty \max(\beta, 1)$ , car  $(u(t, x) = u_0(x, t))$  et donc  $\|u_{tt}\|_\infty = \|u_{xx}\|_\infty = \|u_0''\|_\infty$ . On pose  $C(u_0, \beta) = C$ , on obtient alors

$$e_i^{n+1} = \left(1 - \frac{k}{h_{i-\frac{1}{2}}}\right) e_i^n + \frac{k}{h_{i-\frac{1}{2}}} e_{i-1}^n + Ck(h + k)$$

donc  $\sup_i |e_i^{n+1}| \leq \sup_j |e_j^n| + Ck(h + k)$ . Par récurrence sur  $n$ , on en déduit

$$\sup_i |e_i^n| \leq Ckn(h + k) \text{ et donc } \sup_i |e_i^n| \leq CT(h + k) \text{ si } 0 \leq n \leq N + 1, \text{ où } (N + 1)k = T.$$

2.a) On a  $\inf u_0 \leq u_i^0 \leq \sup u_0$  puis, par récurrence :

$$u_i^{n+1} = \left(1 - \frac{k}{h_i}\right)u_i^n + \frac{k}{h_i}u_{i-1}^n.$$

Comme  $k \leq \alpha h \leq h_i$  on en déduit comme en 1) a) que :

$$\inf(u_0) \leq u_i^h \leq \sup(u_0).$$

2.b) Consistance

$$\frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} = u_t(x_i, t_n) + R_i^n, |R_i^n| \leq \|u_{tt}\|_\infty k$$



mais

$$\frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_i} = \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_{i-\frac{1}{2}}} \frac{h_{i-\frac{1}{2}}}{h_i} = [u_x(x_i, t_n) + S_i^n] \frac{h_{i-\frac{1}{2}}}{h}, \text{ avec } |S_i^n| \leq \|u_{xx}\|_\infty \beta h,$$

donc

$$\begin{aligned} \frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} + \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_i} &= (u_t + u_x)(x_i, t_n) + R_i^n + S_i^n \frac{h_{i-\frac{1}{2}}}{h_i} + T_i^n, \\ &= R_i^n + S_i^n \frac{h_{i-\frac{1}{2}}}{h_i} + T_i^n, \end{aligned}$$

avec :

$$\begin{aligned} |R_i^n| &\leq \|u_{tt}\|_\infty k, \\ \left| \frac{h_{i-\frac{1}{2}}}{h_i} \right| |S_i^n| &\leq \|u_{xx}\|_\infty \frac{\beta h}{\alpha h} \beta h = \frac{\beta^2}{\alpha} \|u_{xx}\|_\infty h, \\ T_i^n &= u_x(x_i, t_n) \frac{h_{i-\frac{1}{2}} - h_i}{h_i} = u_x(x_i, t_n) \frac{h_{i-1} - h_i}{2h_i}. \end{aligned}$$

En prenant par exemple un pas tel que  $h_i = h$  si  $i$  est pair et  $h_i = h/2$  si  $i$  est impair, on voit  $T_i^n$  ne tend pas vers 0 lorsque  $h$  tend vers 0 ; le schéma apparait donc comme non consistant au sens des différences finies.

c) Convergence. On a

$$\begin{aligned} \frac{\bar{u}_i^{n+1} - \bar{u}_i^n}{k} &= u_t(x_{i+\frac{1}{2}}, t_n) + R_i^n, |R_i^n| \leq \|u_{tt}\|_\infty k \\ \frac{\bar{u}_i^n - \bar{u}_{i-1}^n}{h_i} &= u_x(x_{i+\frac{1}{2}}, t_n) + S_i^n, |S_i^n| \leq \|u_{xx}\|_\infty \beta h \end{aligned}$$

donc, avec  $f_i^n = \bar{u}_i^n - u_i^n$

$$f_i^{n+1} = \left(1 - \frac{k}{h_i}\right) f_i^n + f_{i-1}^n \left(\frac{k}{h_i}\right) + k(S_i^n + R_i^n)$$

On a donc :

$$\begin{aligned} \sup_i |f_i^{n+1}| &\leq \sup_i |f_i^n| + k \|u_0''\|_\infty (k + \beta h) \\ &\leq \sup_i |f_i^n| + k C_1 (k + h) \quad C_1 = \|u_0''\|_\infty \max(\beta, 1), \end{aligned}$$

et par récurrence sur  $n$

$$\sup_i |f_i^n| \leq C_1 n k (k + h) + \|u_0'\|_\infty h \beta$$

car  $\sup_i |f_i^0| \leq \|u_0'\|_\infty h \beta$ . D'où on déduit que

$$\begin{aligned} \sup_i |f_i^n| &\leq C_1 T (k + h) + \|u_0'\|_\infty h \beta \\ &\leq C_2 (k + h), \quad 0 \leq n \leq N + 1. \end{aligned}$$

avec  $C_2 = C_1 T + \beta \|u_0'\|_\infty$ . Il reste à remarquer que  $|\bar{u}_i^n - \bar{u}_i^n| \leq \|u_0'\|_\infty \beta h$  pour avoir

$$\sup_i |\bar{u}_i^n - u_i^n| \leq C_3 (h + k) \quad \text{avec}$$

$$C_3 = C_2 + \beta \|u_0'\|_\infty = \|u_0''\|_\infty \max(\beta, 1) T + 2\beta \|u_0'\|_\infty.$$

### Exercice 65 page 199

1. On remarque d'abord que  $|u_i^0| \in [-\|u_0\|_\infty, \|u_0\|_\infty]$ . On a vu à la question 2) a) de l'exercice 64 que  $u_i^{n+1} \in [u_i^n, u_{i-1}^n]$  [ou  $[u_{i-1}^n, u_i^n]$ ]. On en déduit par une récurrence sur  $n$  que  $u_i^n \in [-\|u_0\|_\infty, \|u_0\|_\infty] \quad \forall i, \forall n \geq 0$ .

2. On va utiliser le fait que  $u_0 \in L^2$  et montrer la propriété par récurrence sur  $n$ . Pour  $n = 0$ , on a :

$$|u_i^0|^2 \leq \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} (u_0(x))^2 dx \frac{1}{h_i} \rightarrow 0 \text{ lorsque } i \rightarrow \pm\infty \quad (5.81)$$

En effet, comme  $u_0 1_{[x, x+\eta]} \rightarrow 0$  pp,  $u_0 1_{[x, x+\eta]} \leq u_0 \in L^2$  donc  $\int_x^{x+\eta} |u_0|^2 dx \rightarrow 0$  lorsque  $x \rightarrow +\infty$  par convergence dominée, pour tout  $\eta > 0$ . De plus,  $h \geq \alpha h (\Rightarrow \frac{1}{h_i} \leq \frac{1}{\alpha h})$ , d'où on déduit que (5.81) est vérifiée. On conclut ensuite par une récurrence immédiate sur  $n$ , que :

$$|u_i^{n+1}| \leq \max(|u_i^n| |u_{i-1}^n|) \rightarrow 0 \text{ quand } i \rightarrow \pm\infty. \quad (5.82)$$

2. On veut montrer que  $\sum_{n=0}^N \sum_{i \in \mathbb{Z}} k(u_i^n - u_{i-1}^n)^2 \leq C(\zeta, u_0)$ . On multiplie le schéma par  $ku_i^n$ , on obtient :

$$h_i(u_i^{n+1} - u_i^n)u_i^n + (u_i^n - u_{i-1}^n)ku_i^n = 0,$$

ce qu'on peut réécrire :

$$h_i \left[ -\frac{(u_i^{n+1} - u_i^n)^2}{2} + \frac{(u_i^{n+1})^2}{2} - \frac{(u_i^n)^2}{2} \right] + k \left[ \frac{(u_i^n - u_{i-1}^n)^2}{2} + \frac{(u_i^n)^2}{2} - \frac{(u_{i-1}^n)^2}{2} \right] = 0.$$

Comme  $|u_i^{n+1} - u_i^n| = \frac{k}{h_i} |u_i^n - u_{i-1}^n|$ , ceci s'écrit aussi :

$$k(1 - \frac{k}{h_i})(u_i^n - u_{i-1}^n)^2 + h_i(u_i^{n+1})^2 - h_i(u_i^n)^2 + k(u_i^n)^2 - k(u_{i-1}^n)^2 = 0,$$

et comme  $\frac{k}{h} \leq 1 - \zeta$ , on a donc  $1 - \frac{k}{h_i} \geq \zeta$  et  $\zeta(u_i^n - u_{i-1}^n)^2 + h_i(u_i^{n+1})^2 - h_i(u_i^n)^2 + (u_i^n)^2 - (u_{i-1}^n)^2 \leq 0$  en sommant pour  $i \in \{-M, \dots, M\}$ , et  $h \in \{0, \dots, N\}$ , on obtient alors :

$$\zeta \sum_{i=-M}^M \sum_{n=0}^M (u_i^n - u_{i-1}^n)^2 + \alpha h \sum_{n=0}^N (u_M^n)^2 - \beta h \sum_{n=0}^N (u_{-M}^n)^2 \leq \sum_{i=-M}^M (u_i^0)^2.$$

En remarquant que

$$k \sum_{i=-M}^M (u_i^0)^2 \leq \sum_{i=-M}^M h_i (u_i^0)^2 \leq \|u_0\|_2^2$$

(voir (5.81)) et que  $u_{-M}^n \rightarrow 0$  qd  $M \rightarrow \infty$  (voir (5.82)), on en déduit

$$\zeta k \sum_{i=-\infty}^{\infty} \sum_{n=0}^N (u_i^n - u_{i-1}^n)^2 \leq \|u_0\|_2^2,$$

donc  $C = \frac{\|u_0\|_2^2}{\zeta}$  convient.

3) (Convergence) Pour montrer la convergence, on va passer à la limite sur le schéma numérique. On aura pour cela besoin du lemme suivant :

**Lemme 5.35** Soit  $(u_n)_{n \in \mathbb{N}}$  une suite bornée dans  $L^\infty(\mathbb{R})$ . Si  $u_n \rightarrow u$  dans  $L^\infty(\mathbb{R})$  pour la topologie faible \* lorsque  $n \rightarrow +\infty$ , (c.à.d

$$\int_{\mathbb{R}} u_n(x) \varphi(x) dx \xrightarrow{n \rightarrow +\infty} \int_{\mathbb{R}} u(x) \varphi(x) dx, \quad \forall \varphi \in L^1(\mathbb{R}),$$

et  $v_n \rightarrow v$  dans  $L^1$  lorsque  $n \rightarrow +\infty$ , alors

$$\int u_n(x) v_n(x) dx \xrightarrow{n \rightarrow +\infty} \int u(x) v(x) dx.$$

**Démonstration :**

$$\begin{aligned} \left| \int u_n(x) v_n(x) dx - \int u(x) v(x) dx \right| &\leq \|u_n\|_\infty \|v_n - v\|_1 + \left| \int u_n(x) v(x) dx - \int u(x) v(x) dx \right| \\ &\leq C \|u_n - v\|_1 + \left| \int u_n(x) v_n(x) dx - \int u(x) v(x) dx \right| \xrightarrow{n \rightarrow +\infty} 0, \end{aligned}$$

car  $(u_n)_n$  est bornée dans  $L^\infty$ . ■

On multiplie le schéma numérique par  $k\varphi_i^n$ ,  $\varphi \in C_c^\infty(\mathbb{R} \times [0, T])$  et  $\varphi_i^n = \varphi(x, t_n)$ , et en somme sur  $i$  et  $n$  (toutes les sommes sont finies, car  $\varphi$  est à support compact); on obtient :

$$\sum_{i \in \mathbb{Z}} \sum_{n \in \mathbb{N}} \frac{u_i^{n+1} - u_i^n}{k} k h_i \varphi_i^n + \sum_{i \in \mathbb{Z}} \sum_{n=1}^N (u_i^n - u_{i-1}^n) k \varphi_i^n = 0.$$

Comme  $\varphi_i^n = 0$  si  $n \geq N+1$ , on a :

$$\sum_{i \in \mathbb{Z}} \sum_{n=1}^N h_i u_i^n (\varphi_i^{n-1} - \varphi_i^n) - \sum_i u_i^0 \varphi_i^0 h_i + \sum_{i \in \mathbb{Z}} \sum_n (\varphi_i^n - \varphi_{i+1}^n) u_i^n k = 0.$$

Or :

- $T_1 = \sum_{i \in \mathbb{Z}} u_i^0 \varphi_i^0 h_i = \sum_i \int_{x_{i-\frac{1}{2}}}^{x_{i+\frac{1}{2}}} u_0(x) \varphi_0(x_i) dx \xrightarrow{h \rightarrow 0} \int u_0 \varphi dx.$  (avec  $\varphi_0 = \varphi(\cdot, 0)$ )

car  $\sum_i \varphi_0(x_i) 1_{]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[} \rightarrow \varphi(\cdot, 0)$  dans  $L^1$  quand  $h \rightarrow 0$ .

- $T_2 = \sum_{i \in \mathbb{Z}} \sum_{n=1}^N h_i u_i^n \frac{\varphi_i^{n-1} - \varphi_i^n}{k} k = - \int_{\mathbb{R}_+} \int_{\mathbb{R}} u_{\mathcal{T},k} \psi_{\mathcal{T},k} dx dt.$  Soit

$$\psi_{\mathcal{T},k}(x, t) = \sum_{i \in \mathbb{Z}} \sum_{n=1}^N \frac{\varphi_i^{n-1} - \varphi_i^n}{k} 1_{]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[} 1_{]nk, (n+1)k[}.$$

En effet, pour  $x \in \mathbb{R}$  et  $t > 0$ ,  $|\frac{\varphi_i^{n-1} - \varphi_i^n}{k} - \varphi_t(x, t)| \leq k \|\varphi_{tt}\|_\infty$  si  $(x, t) \in ]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[ \times ]nk, (n+1)k[$ , pour  $n \geq 1$ . On a donc  $\psi_{\mathcal{T},k} \rightarrow \varphi_t$  p.p. sur  $\mathbb{R} \times ]0, T[$ . De plus, et  $|\psi_{\mathcal{T},k}| \leq \|\varphi_t\|_\infty 1_K$  si  $\beta h \leq 1$ , où  $K = [-a-1, a+1] \times [0, T]$ , et  $a$  est tel que  $\varphi = 0$  sur  $([-a, a] \times [0, T])^c$ . Donc, par convergence dominée,  $\psi_{\mathcal{T},k} \rightarrow -\varphi_t$  dans  $L^1(\mathbb{R} \times ]0, T])$  lorsque  $k \rightarrow 0$ . Comme  $u_{\mathcal{T},k}$  converge vers  $u$  dans  $L^1$  faible \*, on en déduit par le lemme 5.35 que :

$$T_2 = - \int_{\mathbb{R}_+} \int_{\mathbb{R}} u_{\mathcal{T},k}(x, t) \psi_{\mathcal{T},k}(x, t) dx dt \xrightarrow{h, k \rightarrow 0} - \int_{\mathbb{R}_+} \int_{\mathbb{R}} u(x, t) \varphi_t(x, t) dx dt.$$

- $T_3 = \sum_{i \in \mathbb{Z}} \sum_{n \in \mathbb{N}} \frac{\varphi_i^n - \varphi_{i+1}^n}{h_i} u_i^n k h_i = \sum_{i \in \mathbb{Z}} \sum_{n \in \mathbb{N}} k \varphi_i^n (u_i^n - u_{i-1}^n)$   
 $= \sum_{i \in \mathbb{Z}} \sum_{n \in \mathbb{N}} k \varphi_{i-1}^n (u_i^n - u_{i-1}^n) + \sum_{i \in \mathbb{Z}} \sum_{n \in \mathbb{N}} k (\varphi_i^n - \varphi_{i-\frac{1}{2}}^n) (u_i^n - u_{i-1}^n) = T_4 + T_5$ , avec :

- \*  $T_4 = \sum_i \sum_n k h_i \frac{\varphi_{i-\frac{1}{2}}^n - \varphi_{i+\frac{1}{2}}^n}{h} u_i^n = \int \int u_{\mathcal{T}k}(x) \chi_{\mathcal{T}k}(x) dx$  où

$$\chi_{\mathcal{T}k} = \frac{\varphi_{i-\frac{1}{2}}^n - \varphi_{i+\frac{1}{2}}^n}{h_i} \text{ sur } ]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[ \times ]t_n, t_{n+1}[;$$

et donc  $\chi_{\mathcal{T}k} \rightarrow -\varphi_x$  dans  $L^1(\mathbb{R} \times ]0, 1])$  et  $T_4 \rightarrow - \int \int u(x) \varphi_x(x) dx dt$  lorsque  $h \rightarrow 0$ ,

- \*  $T_5 \leq \sum_{i=M_1}^{M_2} \sum_{n=0}^N k \beta h \|\varphi_x\|_\infty (u_i - u_{i-1}^h) \leq \beta k h \|\varphi_x\|_\infty \sum_{n=0}^N \sum_{i=M_1}^{M_2} (u_i^n - u_{i-1}^n)$  si  $\beta h \leq 1$ , où  $M_1$  et  $M_2$  sont tels que  $i \notin \{M_1, \dots, M_2\} \Rightarrow ]x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}[ \subset [-a, a]^c$ , et  $\varphi = 0$  sur  $([-a, a] \times [0, T])^c$ . On a donc :

$$\begin{aligned} T_5 &\leq \beta k h \|\varphi_x\|_\infty \left( \sum_{i=M_1}^{M_2} \sum_{n=0}^N (u_i^n - u_{i-1}^n)^2 \right)^{1/2} \left( \sum_{n=0}^N \sum_{i=M_1}^{M_2} 1 \right)^{1/2} \\ &\leq \beta k h \|\varphi_x\|_\infty \frac{\sqrt{c}}{\sqrt{k}} \left( \sum_{n=0}^N \sum_{i=M_1}^{M_2} 1 \right)^{1/2}, \\ &\leq \beta \sqrt{k} h \|\varphi_x\|_\infty \sqrt{c} \sqrt{N+1} \sqrt{M_2 - M_1} \quad (M_2 - M_1) \alpha h \leq 2a. \\ &\leq \beta \sqrt{k} h \|\varphi_x\|_\infty \sqrt{c} \frac{\sqrt{T}}{\sqrt{k}} \frac{\sqrt{2a}}{\sqrt{\alpha h}} = \beta \|\psi_x\|_\infty \sqrt{c} \frac{\sqrt{T}}{\sqrt{\alpha}} \sqrt{h} \\ &\rightarrow 0 \text{ quand } h \rightarrow 0. \end{aligned}$$

On en déduit que  $T_3 \rightarrow -\int \int u(x)\varphi_x(x)dx$  quand  $h \rightarrow 0$ .  
Comme  $T_1 + T_2 + T_3 = 0$ , on a donc

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x,t)\varphi_t(x;t)dxdt + \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(x,t)\varphi_x(x,t)dxdt + \int_{\mathbb{R}} u_0(x)\varphi(x,.)dx = 0$$

et donc  $u$  est solution faible de (5.50)-(5.51).

### Exercice 66 page 199

1/ Dans le premier cas, la solution est facile à construire par la méthode des caractéristiques, pour tout  $t < 1/2$ . En effet, les droites caractéristiques sont d'équation :  $x(t) = 2u_0(x_0)t + x_0$ , c'est-à-dire

$$x(t) = \begin{cases} 2t + x_0, & \text{si } x_0 < 0, \\ 2(1 - x_0)t + x_0, & \text{si } x_0 \in ]0, 1[, \\ 0 & \text{si } x_0 > 1. \end{cases}$$

Les droites caractéristiques se rencontrent à partir de  $t = 1/2$ , il y alors apparition d'un choc, dont la vitesse est donnée par la relation de Rankine-Hugoniot :

$$\sigma(u_g - u_d) = (u_g^2 - u_d^2), \text{ et donc } \sigma = u_g + u_d = 1.$$

La solution entropique est donc :

$$u(x,t) = \begin{cases} 1 & \text{si } t < \frac{1}{2} \text{ et } x < 2t \text{ ou si } t > \frac{1}{2} \text{ et } x < t + \frac{1}{2}, \\ \frac{x-1}{2t-1} & \\ 0 & \text{si } t < \frac{1}{2} \text{ et } x > 1 \text{ ou si } t > \frac{1}{2} \text{ et } x > t + \frac{1}{2}. \end{cases}$$

2/ On pourra montrer que la fonction définie par les formules suivantes est la solution pour  $t < \frac{1}{2}$  (c'est-à-dire avant que les droites caractéristiques ne se rencontrent, la solution contient deux zones de détente).

$$\begin{aligned} u(x,t) &= 0, \text{ si } x < 0, t < \frac{1}{2}, \\ u(x,t) &= \frac{x}{2t}, \text{ si } 0 < x < 2t, t < \frac{1}{2}, \\ u(x,t) &= \frac{1-x}{1-2t}, \text{ si } 2t < x < 1, t < \frac{1}{2}, \\ u(x,t) &= \frac{x-1}{2t}, \text{ si } 1 < x < 1+2t, t < \frac{1}{2}, \\ u(x,t) &= 1, \text{ si } 1+2t < x, t < \frac{1}{2}. \end{aligned}$$

En  $t = \frac{1}{2}$ , on pourra vérifier qu'un choc apparaît en  $x = 1$  et se propage à la vitesse 1. On obtient alors pour  $t > \frac{1}{2}$  la solution suivante :

$$\begin{aligned} u(x,t) &= 0, \text{ si } x < 0, t > \frac{1}{2}, \\ u(x,t) &= \frac{x}{2t}, \text{ si } 0 < x < \frac{1}{2} + t, t > \frac{1}{2}, \\ u(x,t) &= \frac{x-1}{2t}, \text{ si } \frac{1}{2} + t < x < 1+2t, t > \frac{1}{2}, \\ u(x,t) &= 1, \text{ si } 1+2t < x, t > \frac{1}{2}. \end{aligned}$$

Remarquons que, bien que la solution initiale soit discontinue, la solution entropique est continue pour  $t \in ]0, 1/2[$ .

**Exercice 68 page 200**

1. La question 1 découle du point 1 de la proposition 5.29 page 192 (il faut que  $\sigma$  satisfasse la condition de Rankine–Hugoniot).
2. La question 2 découle du point 2 de la proposition 5.29 page 192.

**Exercice 70 page 200**

Les quatre schémas s'écrivent sous la forme :

$$u_i^{n+1} = u_i^n - \frac{k}{h_i} (g(u_i^n, u_{i+1}^n) - g(u_i^n, u_i^n)) + \frac{k}{h_i} (g(u_{i-1}^n, u_i^n) - g(u_i^n, u_i^n))$$

soit encore

$$u_i^{n+1} = u_i^n + C_i^n (u_{i+1}^n - u_i^n) + D_i^n (u_{i-1}^n - u_i^n),$$

avec

$$C_i^n = \frac{k}{h_i} \frac{g(u_i^n, u_i^n) - g(u_i^n, u_{i+1}^n)}{u_{i+1}^n - u_i^n} \quad \text{si } u_i^n \neq u_{i+1}^n \text{ (0 sinon)}$$

$$D_i^n = \frac{k}{h_i} \frac{g(u_{i-1}^n, u_i^n) - g(u_i^n, u_i^n)}{u_{i-1}^n - u_i^n} \quad \text{si } u_i^n \neq u_{i+1}^n \text{ (0 sinon)}$$

On suppose que  $A \leq u_0 \leq B$  p.p. et on remarque qu'il existe  $L \in \mathbb{R}_+$  tel que :

$$\left. \begin{array}{l} |g(a, b) - g(a, c)| \leq L|b - c|, \\ |g(b, a) - g(c, a)| \leq L|b - c| \end{array} \right\} \quad \forall a, b, c \in [A, B]$$

(On laisse le lecteur vérifier qu'un tel  $L$  existe pour les 4 schémas considérés).

1) Dans le cas des 3 premiers schémas (FS, Godunov et LFM), la fonction  $g$  est croissante par rapport au 1er argument et décroissante par rapport au 2ème argument. Donc si  $u_i^n \in [A, B], \forall i$  (pour  $n$  fixé), on a  $C_i^n \geq 0$   $D_i^n \geq 0$ . En prenant  $2k \leq Lh_i \quad \forall i$  on a aussi :  $C_i^n, D_i^n \leq \frac{1}{2}$  et donc  $u_i^{n+1}$  est une combinaison convexe de  $u_{i-1}^n, u_i^n, u_{i+1}^n$  donc  $u_i^{n+1} \in [A, B] \quad \forall i$  (et aussi  $\|u^{n+1}\|_\infty \leq \|u^n\|_\infty$ ). Par récurrence sur  $n$  on en déduit :

$$u_i^n \in [A, B] \quad \forall i, \forall n \text{ si } k \leq uh_i \forall i \text{ avec } M = \frac{L}{2}$$

Dans le dernier cas (Murman), on a

$$g(a, b) = f(a) \text{ si } \frac{f(b) - f(a)}{b - a} \geq 0 \quad (a \neq b), \quad g(a, b) = f(b) \text{ si } \frac{f(b) - f(a)}{b - a} < 0 \quad (a \neq b) \text{ et } g(a, a) = f(a).$$

Si  $\frac{f(u_{i+1}^n) - f(u_i^n)}{u_{i+1}^n - u_i^n} \geq 0$ , on a :  $g(u_i^n, u_{i+1}^n) = f(u_i^n)$ , donc  $C_i^n = 0$ .

Si  $\frac{f(u_{i+1}^n) - f(u_i^n)}{u_{i+1}^n - u_i^n} < 0$ , on a :  $g(u_i^n, u_{i+1}^n) = f(u_{i+1}^n)$ ,  $C_i^n = \frac{-f(u_{i+1}^n) + f(u_i^n)}{u_{i+1}^n - u_i^n} > 0$ , et  $C_i^n \leq \frac{1}{2}$  si  $k \leq Mh_i$  avec

$$M = \frac{L}{2} \quad (L \text{ est ici la constante de Lipschitz de } f).$$

Le même calcul vaut pour  $D_i^n$  et on conclut comme précédemment car

$$u^{n+1} = (1 - C_i^n - D_i^n)u^n + C_i^n u_{i+1}^n + D_i^n u_{i-1}^n$$

2) On reprend la formule de 1) et la même limitation sur  $k$  (pour les 4 schémas). On a :

$$u_{i+1}^{n+1} = u_{i+1}^n + C_{i+1}^n (u_{i+2}^n - u_{i+1}^n) + D_{i+1}^n (u_i^n - u_{i+1}^n)$$

$$u_i^{n+1} = u_i^n + C_i^n (u_{i+1}^n - u_i^n) + D_i^n (u_{i-1}^n - u_i^n)$$

et donc, en soustrayant membre à membre :

$$u_{i+1}^{n+1} - u_i^{n+1} = (u_{i+1}^n - u_i^n) \underbrace{(1 - C_{i+1}^n - D_{i+1}^n)}_{\geq 0} + \underbrace{C_{i+1}^n}_{\geq 0} (u_{i+2}^n - u_{i+1}^n) + \underbrace{D_i^n}_{\geq 0} (u_i^n - u_{i-1}^n)$$

Par inégalité triangulaire, on a donc :

$$|u_{i+1}^{n+1} - u_i^{n+1}| \leq |u_{i+1}^n - u_i^n|(1 - C_i^n - D_{i+1}^n) + C_{i+1}^n |u_{i+2}^n - u_{i+1}^n| + D_i^n |u_i^n - u_{i-1}^n|$$

Sommons alors entre  $i = -P$  à  $P$  :

$$\begin{aligned} \sum_{i=-P}^P |u_{i+1}^{n+1} - u_i^{n+1}| &\leq \sum_{i=-P}^P |u_{i+1}^n - u_i^n| - \sum_{i=-P}^P C_i^n |u_{i+1}^n - u_i^n| + \sum_{i=-P}^P C_{i+1}^n |u_{i+2}^n - u_{i+1}^n| \\ &- \sum_{i=-P}^P D_{i+1}^n |u_{i+1}^n - u_i^n| + \sum_{i=-P}^P D_i^n |u_i^n - u_{i-1}^n|. \end{aligned}$$

En regroupant :

$$\sum_{i=-P}^P |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i=-P}^P |u_{i+1}^n - u_i^n| + C_{P+1}^n |u_{P+2}^n - u_{P+1}^n| + D_{-P}^n |u_{-P}^n - u_{-P-1}^n|.$$

Or  $C_{P+1}^n \in [0, 1]$  et  $D_{-P}^n \in [0, 1]$  donc

$$\sum_{i=-P}^P |u_{i+1}^{n+1} - u_i^{n+1}| \leq \sum_{i=-P-1}^{P+1} |u_{i+1}^n - u_i^n| \leq \sum_{i=-\infty}^{+\infty} |u_{i+1}^n - u_i^n|.$$

Il ne reste plus qu'à faire tendre  $P$  vers  $+\infty$  pour obtenir le résultat.

### Exercice 71 page 200

1) Cette question a été complètement traitée dans l'exercice 70.

Les estimations sont vérifiées avec  $M = \frac{L}{2}$ , où  $L$  est la constante de Lipschitz de  $f$  sur  $[A, B]$ .

2) Remarquons que si  $f(s) = s^2$  alors  $\frac{f(b) - f(a)}{b - a} = b + a$ .

a) Soit  $\bar{b} \in ]0, B[$ ,  $\bar{a} \in ]A, 0[$  tel que  $\bar{b} + \bar{a} > 0$ , (par exemple :  $\bar{a} = -\frac{\epsilon}{2}$ ,  $\bar{b} = \epsilon$  avec  $0 < \epsilon < \min(-A, B)$ ). Soit  $\alpha \in ]0, \bar{a} + \bar{b}[$ . Pour  $a \in [\bar{a} - \alpha, \bar{a} + \alpha]$ , on a  $\bar{b} + a > 0$ , et donc  $g(a, \bar{b}) = f(a) = a^2$ , ce qui prouve que sur l'ensemble  $[\bar{a} - \alpha, \bar{a} + \alpha] \times \{\bar{b}\}$ , la fonction  $g$  est décroissante par rapport à  $a$ .

b) Soit  $u_0$  définie par :  $u_0 = \begin{cases} -1 & \text{sur } \mathbb{R}_- \\ +1 & \text{sur } \mathbb{R}_+ \end{cases}$  de sorte que  $u_i^0 = \begin{cases} +1 & \text{si } i \geq 0 \\ -1 & \text{si } i < 0 \end{cases}$

Comme  $f(u_i^0) = +1 \quad \forall i$  on a  $u_i^1 = u_i^0 \quad \forall i$  et donc  $u_i^n = u_i^0$  pour tout  $i$  et pour tout  $n$ . Par une récurrence facile, la solution approchée est donc stationnaire. La solution exacte n'est pas stationnaire (voir proposition 5.29, cas où  $f$  est strictement convexe et  $u_g < u_d$ ).

### Exercice 72 page 201 (Flux monotones)

1. Le schéma s'écrit :  $u_i^{n+1} = u_i^n - \frac{h_i}{k}(g(u_i^n, u_{i+1}^n) - g(u_i^n, u_{i-1}^n))$

2.

$$\begin{aligned} u_{i+1}^n &= u_i^n + C_i^n (u_{i+1}^n - u_i^n) + D_i^n (u_{i-1}^n - u_i^n) \\ &= (1 - C_i^n - D_i^n)u_i^n + C_i^n u_{i+1}^n + D_i^n u_{i-1}^n \end{aligned}$$

$$\text{avec } C_i^n = \frac{h_i}{k} \frac{g(u_i^n, u_{i+1}^n) - g(u_i^n, u_i^n)}{u_{i+1}^n - u_i^n} \text{ si } u_{i+1}^n \neq u_i^n \text{ (et 0 sinon)}$$

$$\text{et } D_i^n = \frac{h_i}{k} \frac{g(u_i^n, u_{i+1}^n) - g(u_i^n, u_i^n)}{u_{i+1}^n - u_i^n} \text{ si } u_{i-1}^n \neq u_i^n \text{ (et 0 sinon)}$$

Remarquons que  $C_i^n \geq 0$  et  $D_i^n \geq 0$  car  $g$  est monotone. On en déduit que  $H$  définie par

$$H(u_{i-1}^n, u_i^n, u_{i+1}^n) = (1 - C_i^n - D_i^n)u_i^n + C_i^n u_{i+1}^n + D_i^n u_{i-1}^n$$

est une fonction croissante de ses arguments si  $1 - C_i^n - D_i^n \geq 0$ , ce qui est vérifié si  $k \leq \frac{h_i}{2M}$  pour tout  $i \in \mathbb{Z}$ .

3. Comme  $a \leq \xi$   $g(a, b) \leq g(\xi, b)$ , et comme  $\xi \leq b$ ,  $g(\xi, b) \leq g(\xi, \xi)$ .

4. D'après la question précédente, si  $a \leq b$ , on a bien  $g(a, b) \leq \min\{g(\xi, \xi), \xi \in [a, b]\}$ , et comme  $g(\xi, \xi) = f(\xi)$ , on a le résultat souhaité.

Si  $a \geq b$ , alors on vérifie facilement que :  $g(a, b) \geq g(\xi, \xi)$  pour tout  $\xi \in [b, a]$ , ce qui prouve le résultat.

5. Comme  $g(a, b) = f(u_{a,b})$ , on a  $\min_{s \in [a,b]} f(s) \leq g(a, b)$  si  $a \leq b$  et  $g(a, b) \leq \max_{s \in [b,a]} f(s)$  si  $a \geq b$ . On a donc égalité dans les inégalités de la question 3.

2. Montrer que sous une condition à préciser, le schéma peut s'écrire sous la forme

$$u_i^{n+1} = H(u_{i-1}^n, u_i^n, u_{i+1}^n)$$

où  $H$  est une fonction croissante de ses trois arguments.

# Bibliographie

- [1] BREZIS, H. (1983), *Analyse Fonctionnelle : Théorie et Applications* (Masson, Paris).
- [2] P.G. Ciarlet, Introduction à l'analyse numérique et à l'optimisation, Masson 1982.
- [3] CIARLET, P.G. (1978), *The Finite Element Method for Elliptic Problems* (North-Holland, Amsterdam).
- [4] CIARLET, P.G. (1991), *Basic error estimates for elliptic problems* in : *Handbook of Numerical Analysis II* (North-Holland, Amsterdam) 17-352.
- [5] R. Eymard, T. Gallouët and R. Herbin, Finite Volume Methods, *Handbook of Numerical Analysis*, Vol. VII, pp. 713-1020. Edited by P.G. Ciarlet and J.L. Lions (North Holland). Version en ligne : <http://www.cmi.univ-mrs.fr/~herbin/PUBLI/bookevol.pdf>
- [6] T. Gallouët and R. Herbin, Mesures, Intégration, Probabilités. <http://www.cmi.univ-mrs.fr/gallouet/licence.d/mes-int-pro.pdf>
- [7] GODLEWSKI E. and P. A. RAVIART (1991), *Hyperbolic systems of conservation laws*, Ellipses.
- [8] GODLEWSKI E. and P.A. RAVIART (1996), *Numerical approximation of hyperbolic systems of conservation laws*, *Applied Mathematical Sciences* **118** (Springer, New York).
- [9] GODUNOV S. (1976), *Résolution numérique des problèmes multidimensionnels de la dynamique des gaz* (Editions de Moscou).
- [10] R. Herbin, Analyse numérique. <http://www.cmi.univ-mrs.fr/~herbin/PUBLI/anamat.pdf>
- [11] KRÖNER D. (1997) *Numerical schemes for conservation laws in two dimensions*, Wiley-Teubner Series Advances in Numerical Mathematics. (John Wiley and Sons, Ltd., Chichester ; B. G. Teubner, Stuttgart).
- [12] LEVEQUE, R. J. (1990), *Numerical methods for conservation laws* (Birkhauser verlag).
- [13] QUARTERONI A., SACCO R., AND SALERI F. Numerical mathematics. Springer, 2000.
- [14] J. RAPPAZ AND M. PICASSO Introduction à l'analyse numérique. Presses Polytechniques et Universitaires Romandes, Lausanne, 1998.
- [15] P.A. RAVIART AND JM THOMAS. Introduction à l'analyse numérique des équations aux dérivées partielles.