Université H. Poincaré, Nancy I

Master 2 IMOI

Ingénierie Mathématique et Outils Informatiques

Option Calcul Scientifique

Méthodes numériques pour la dynamique des fluides

Notes de cours/J.-F. Scheid

Année 2011-2012

Table des matières

1 Introduction et classification des EDP **3**
- 1.1 Introduction . . . 3
- 1.2 EDP linéaires du second ordre . . . 3
- 1.3 EDP du premier ordre . . . 4
- 1.4 Problème bien posé . . . 5
- 1.5 Classification des EDP du second ordre . . . 5

2 EDP elliptiques linéaires **9**
- 2.1 Introduction - Propriétés des solutions . . . 9
  - 2.1.1 Quelques rappels sur l'existence, l'unicité et la régularité des solutions . . . 9
  - 2.1.2 Principes du maximum . . . 10
- 2.2 Différences finies pour le cas 1D . . . 13
  - 2.2.1 Erreur de consistance . . . 14
  - 2.2.2 Matrices monotones . . . 15
  - 2.2.3 Stabilité . . . 17
  - 2.2.4 Convergence . . . 18
  - 2.2.5 Autres conditions limites . . . 19
- 2.3 Différences finies pour le cas 2D . . . 20
  - 2.3.1 Un schéma à 5 points pour le Laplacien . . . 21
  - 2.3.2 Un autre schéma à 5 points . . . 24
  - 2.3.3 Un schéma à 9 points pour fonction harmonique . . . 25
  - 2.3.4 Cas d'un domaine non-rectangulaire . . . 26
  - 2.3.5 Opérateur sous forme de divergence . . . 31
  - 2.3.6 Autres conditions limites . . . 32
- 2.4 Evaluation pratique de l'ordre de convergence d'une méthode . . . 34
- 2.5 Méthode de Richardson . . . 34
- 2.6 Conditionnement . . . 35

3 EDP paraboliques linéaires **37**
- 3.1 Introduction . . . 37
  - 3.1.1 Existence et unicité des solutions . . . 37
  - 3.1.2 Principes du maximum . . . 38
- 3.2 Equation de la chaleur en dimension 1 d'espace . . . 38
  - 3.2.1 Schéma d'Euler explicite . . . 39
  - 3.2.2 Schéma d'Euler implicite . . . 42
  - 3.2.3 Schéma de Crank-Nicholson (1947) . . . 46
  - 3.2.4 θ-schéma pour l'équation de la chaleur . . . 48
  - 3.2.5 Autres schémas . . . 48
- 3.3 Cas de la dimension 2 d'espace . . . 49
  - 3.3.1 θ-schéma pour l'équation de la chaleur en 2D . . . 49
  - 3.3.2 Directions alternées . . . 50

4 EDP hyperboliques linéaires **55**
4.1 Equation des ondes . . . 55
4.1.1 Existence, unicité et propriétés des solutions . . . 55
4.1.2 Un θ-schéma centré pour l'équation des ondes (1D) . . . 57
4.2 Equation de transport . . . 59
4.2.1 Introduction . . . 59
4.2.2 Schéma centré . . . 61
4.2.3 Schéma de Lax . . . 62
4.2.4 Schéma décentré (upwind) . . . 66
4.2.5 Schéma de Lax-Wendroff . . . 69
4.2.6 Comparaison des différents schémas . . . 73
4.2.7 Quelques schémas pour le 2D . . . 75

5 EDP hyperboliques non-linéaires - Lois de conservation **79**
5.1 Introduction . . . 79
5.2 Solutions classiques . . . 80
5.3 Solutions faibles . . . 81
5.4 Relations de Rankine-Hugoniot . . . 82
5.5 Solutions d'entropie . . . 85
5.6 Problème de Riemann . . . 90
5.6.1 Cas où f est strictement convexe . . . 90
5.6.2 Cas général . . . 91
5.7 Schémas d'approximations aux Différences Finies . . . 94
5.7.1 Introduction et généralités . . . 94
5.7.2 Schéma de Godounov . . . 95
5.7.3 Schéma d'Engquist-Osher . . . 97

6 Equations de Stokes **99**
6.1 Introduction . . . 99
6.2 Adimensionalisation . . . 100
6.3 Réductions des équations . . . 100
6.4 Discrétisation des équations de Stokes par *Différences Finies* . . . 101
6.4.1 Introduction . . . 101
6.4.2 Schéma MAC pour le problème de Stokes . . . 102
6.4.3 Forme matricielle du schéma de MAC . . . 103
6.5 Résolution du système discrétisé de Stokes . . . 105
6.6 Conditions de Dirichlet non-homogènes . . . 106
6.7 Traitement pratique des conditions limites . . . 108
6.8 Exemples . . . 109

7 Equations de Navier-Stokes **113**
7.1 Introduction . . . 113
7.2 Semi-discrétisation en temps . . . 113
7.3 Discrétisation totale . . . 114
7.4 Forme matricielle du schéma semi-implicite . . . 116
7.5 Résolution du système discrétisé de Navier-Stokes . . . 118
7.6 Conditions de Dirichlet non-homogènes - Traitement des conditions limites . . . 118
7.7 Exemples . . . 118

Références **121**

Chapitre 1

Introduction et classification des EDP

1.1 Introduction

On s'intéresse aux équations aux dérivéees partielles sous la forme générale

$$F(\mathbf{x}, u, Du, \cdots, D^\alpha u) = 0, \tag{1.1}$$

où u est une fonction inconnue des N variables regroupées dans le vecteur $\mathbf{x} = (x_1, \cdots, x_N)$; α est un multi-indice $\alpha = (\alpha_1, \cdots, \alpha_N)$ et $D^\alpha = \dfrac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_N^{\alpha_N}}$ avec $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_N$.

L'équation (1.1) est du **premier ordre** si $|\alpha| = 1$ et du **deuxième ordre** si $|\alpha| = 2$.

L'équation est dite :

- **linéaire** si F est linéaire en $u, Du, \cdots, D^\alpha u$ (u et v solutions $\Rightarrow \lambda u + \beta v$ solution).
- **semi-linéaire** si F est linéaire en $Du, \cdots, D^\alpha u$.
- **quasi-linéaire** si F est linéaire en $D^\alpha u$.
- **non-linéaire** si F n'est pas linéaire en au moins une dérivée.

1.2 EDP linéaires du second ordre

Donnons pour commencer, quelques exemples fondamentaux d'EDP linéaires du second ordre qu'on étudiera dans les chapitres suivants.

Equation de Laplace.
En dimension 2, on cherche $u = u(x, y)$ vérifiant

$$u_{xx} + u_{yy} = f(x, y), \quad \text{pour } (x, y) \in (0, 1) \times (0, 1), \tag{1.2}$$

avec les conditions limites

$$\begin{array}{ll} u(0, y) = \phi_0(y), & u(1, y) = \phi_1(y), \\ u(x, 0) = \psi_0(x), & u(x, 1) = \psi_1(x), \end{array} \tag{1.3}$$

les fonctions $f, \phi_0, \phi_1, \psi_0, \psi_1$ étant données.

Plus généralement en dimension n, on cherche $u = u(x_1, \cdots, x_n)$ telle que

$$\left\{ \begin{array}{rcl} -\Delta u & = & f, \quad \text{dans un ouvert } \Omega \subset \mathbb{R}^n \\ u & = & g, \quad \text{sur } \partial\Omega \end{array} \right.$$

où Δ désigne l'opérateur de Laplace et g est une fonction donnée sur le bord. Cette équation modélise par exemple (de façon très simplifiée...) le déplacement d'une membrane soumise à une force extérieure f avec un déplacement g imposé sur le bord $\partial\Omega$. L'équation avec $f \equiv 0$ est appelée *équation de Poisson*

et correspond à la recherche des fonctions harmoniques. L'équation de Poisson peut être obtenue de la façon suivante. En l'absence de force extérieure ($f \equiv 0$) on peut écrire la loi de conservation du flux sur tout le bord fermé $\partial\Sigma$ d'un sous-domaine quelconque Σ, c'est-à-dire

$$\int_{\partial\Sigma} \nabla u \cdot \mathbf{n}\, d\sigma = 0,$$

où $\mathbf{n}$ désigne la normale extérieure au domaine Σ. La formule de Green donne alors

$$\int_{\Sigma} \Delta u\, d\mathbf{x} = 0.$$

Ceci étant vrai pour tout sous-domaine Σ (et u suffisamment régulière), on en déduit $\Delta u = 0$ dans Ω.

Equation de la chaleur.
Pour un domaine $\Omega \subset \mathbb{R}^n$ et $T > 0$, on cherche une fonction $u = u(x_1, \cdots, x_n, t)$ dépendante du temps t telle que

$$\left\{ \begin{array}{rcll} \dfrac{\partial u}{\partial t} - \Delta u & = & f, & \text{dans } \Omega \times (0,T) \\ u & = & g, & \text{sur } \partial\Omega \times (0,T) \\ u(\mathbf{x}, t=0) & = & u_0(\mathbf{x}), & \mathbf{x} \in \Omega. \end{array} \right.$$

La fonction u représente par exemple la température en un point $\mathbf{x}$ du domaine Ω et à l'instant $t \in [0,T]$.

Equation des ondes.
On cherche $u = u(x,y)$ (y représente le temps) vérifiant

$$u_{xx} - u_{yy} = f(x,y), \quad \text{pour } (x,y) \in (0,1) \times (0,T), \tag{1.4}$$

avec les conditions limites

$$\begin{array}{l} u(0,y) = \phi_0(y), \quad u(1,y) = \phi_1(y), \\ u(x,0) = \psi_0(x), \quad u_y(x,0) = \psi_1(x), \end{array} \tag{1.5}$$

ou bien plus généralement, pour un domaine $\Omega \subset \mathbb{R}^n$ on cherche $u = u(x_1, \cdots, x_n, t)$ telle que

$$\left\{ \begin{array}{rcll} \dfrac{\partial^2 u}{\partial t^2} - \Delta u & = & f, & \text{dans } \Omega \times (0,T) \\ u & = & g, & \text{sur } \partial\Omega \times (0,T) \\ u(\mathbf{x}, t=0) & = & u_0(\mathbf{x}), & \\ \dfrac{\partial u}{\partial t}(\mathbf{x}, t=0) & = & v_0(\mathbf{x}), & \mathbf{x} \in \Omega. \end{array} \right.$$

1.3 EDP du premier ordre

On s'intéressera également aux EDP du premier ordre à partir de quelques exemples fondamentaux.

Equation de transport.
En dimension 1 d'espace, on cherche $u = u(x,t)$ avec $x \in \mathbb{R}$ et $t \in \mathbb{R}^+$, vérifiant

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0. \tag{1.6}$$

La fonction u représente par exemple une quantité en x et à l'instant t transportée par une vitesse $c \in \mathbb{R}$.

Equation de Bürgers.
En dimension 1 d'espace, on cherche $u = u(x,t)$ avec $x \in \mathbb{R}$ et $t \in \mathbb{R}^+$, vérifiant

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(\frac{u^2}{2}\right) = 0. \tag{1.7}$$

Il s'agit d'un modèle simplifié de la dynamique des gaz. Cette équation sert aussi de modèle du *bang sonique* : le bruit engendré par un avion supersonique, loin de l'avion (près du sol), se concentre dans certaines zones où la pression est gouvernée par l'équation de Bürgers.

Lois de conservation.

Les deux équations précédentes sont des cas particuliers de *lois de conservation* qui s'écrivent plus généralement (toujours en dimension 1 d'espace)

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x}\left(f(u)\right) = 0. \tag{1.8}$$

Ces équations sont quasi-linéaires car linéaire par rapport aux dérivées d'ordre le plus élévé (u_t et u_x).

1.4 Problème bien posé

> *On dit qu'un problème est bien posé au sens d'Hadamard si sa solution existe, est unique et dépend continûment des données.*

Les conditions aux limites jouent un rôle essentiel quant au caractère bien posé ou non d'un problème. Par exemple, l'équation de Laplace avec une condition de Neumann sur le bord ($\partial u/\partial \mathbf{n} = 0$ sur $\partial\Omega$) n'est pas bien posé en général (il faut des conditions de compatibilité sur la fonction f). De même l'équation de Laplace (1.2) avec les conditions aux limites (1.5) n'est pas bien posé (pas de continuité par rapport aux données), ni l'équation des ondes (1.4) avec les conditions aux limites (1.3)

1.5 Classification des EDP du second ordre

Comme on le verra plus tard, les solutions d'EDP du second ordre ont des propriétés différentes selon le type d'équations. Dans un domaine $Q \subset \mathbb{R}^N$, on considère pour $u = u(\mathbf{x})$, l'équation linéaire aux dérivées partielles du second ordre de la forme

$$\sum_{i,j=1}^{N} a_{ij}(\mathbf{x})\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i=1}^{N} b_i(\mathbf{x})\frac{\partial u}{\partial x_i} + c(\mathbf{x})u = f, \tag{1.9}$$

avec $\mathbf{x} = (x_1, \cdots, x_N) \in Q$. Les coefficients a_{ij} sont réels et la matrice A formée des coefficients a_{ij} est supposée symétrique([1]). La fonction f est donnée.

Pour un point $\mathbf{x}_0$ fixé dans Q, on désigne par $N_+ = N_+(\mathbf{x}_0)$ le nombre de valeurs propres de $A(\mathbf{x}_0)$ strictement positives, par $N_- = N_-(\mathbf{x}_0)$ le nombre de celles strictement négatives et par $N_0 = N_0(\mathbf{x}_0)$ le nombre de celles nulles ($N_+ + N_- + N_0 = N$).

L'équation (1.9) est dite

- **elliptique** en $\mathbf{x}_0$ si $N_+ = N$ ou $N_- = N$,
- **parabolique** en $\mathbf{x}_0$ si $N_0 > 0$,
- **hyperbolique** en $\mathbf{x}_0$ si ($N_+ = N-1$ et $N_- = 1$) ou bien ($N_- = N-1$ et $N_+ = 1$),

L'équation (1.9) est dite elliptique, parabolique ou hyperbolique en $\mathcal{O} \subset Q$, si elle jouit de la propriété en tout point $\mathbf{x} \in \mathcal{O}$.

---

1. L'hypothèse de symétrie de la matrice A est *raisonnable* au sens où l'on peut écrire $\sum_{i,j} a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} = \sum_{i,j} a'_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} + \sum_{i,j} a''_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}$, avec $a'_{ij} = (a_{ij}+a_{ji})/2$ et $a''_{ij} = (a_{ij}-a_{ji})/2$. Or $\sum_{i,j} a''_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} = 0$ car si $u \in C^2(Q)$, on a $\frac{\partial^2 u}{\partial x_i \partial x_j} = \frac{\partial^2 u}{\partial x_j \partial x_i}$. Par conséquent, on obtient $\sum_{i,j} a_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j} = \sum_{i,j} a'_{ij}\frac{\partial^2 u}{\partial x_i \partial x_j}$ et la matrice formée des coefficients a'_{ij} est *symétrique*.

Dans le cas où N est pair, on en déduit que l'équation (1.9) est elliptique, parabolique ou hyperbolique si on a respectivement, $\det(A) > 0$, $\det(A) = 0$ ou $\det(A) < 0$.

Donnons à présent une interprétation géométrique de la classification précédente. On introduit la fonction $F : \mathbb{R}^N \to \mathbb{R}$ définie par $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x})$ où $(\cdot, \cdot)$ désigne le produit scalaire dans $\mathbb{R}^N$. L'équation $x_{N+1} = F(x_1, ..., x_N)$ est l'équation d'une ellipse, d'une parabole ou d'une hyperbole en dimension $N+1$, selon que l'équation (1.9) est respectivement elliptique, parabolique ou hyperbolique. En effet, la matrice A étant diagonalisable, on a la décomposition $A = Q^T DQ$ où D est la matrice diagonale formée des valeurs propres λ_i $(i = 1, ..., N)$ de A et Q est la matrice orthogonale dont les colonnes sont formées des vecteurs propres. Ainsi $F(\mathbf{x}) = (A\mathbf{x}, \mathbf{x}) = (DQ\mathbf{x}, Q\mathbf{x}) = \sum_{i=1}^{N} \lambda_i z_i^2$ avec $z_i = (Q\mathbf{x})_i$ et en fonction du signe des valeurs propres, l'équation $x_{N+1} = F(x_1, ..., x_N)$ est celle d'une ellipse, d'une parabole ou d'une hyperbole.

Exemples.

Considérons le cas $N = 2$. Les *courbes de niveaux* de F sont des ellipses, des hyperboles ou des droites, selon que l'équation (1.9) est respectivement elliptique, hyperbolique ou parabolique . Si la matrice A est définie positive $(\det(A) > 0)$ alors l'équation (1.9) est elliptique et les courbes de niveaux de F sont des ellipses. Par exemple avec la matrice $A = \begin{pmatrix} 4 & 2 \\ 2 & 3 \end{pmatrix}$ dont les valeurs propres sont $\lambda_1 \simeq 1.4384472$ et $\lambda_2 \simeq 5.561552$, on obtient les ellipses de la figure 1.1. Les axes principaux sont déterminés par les vecteurs propres de A.

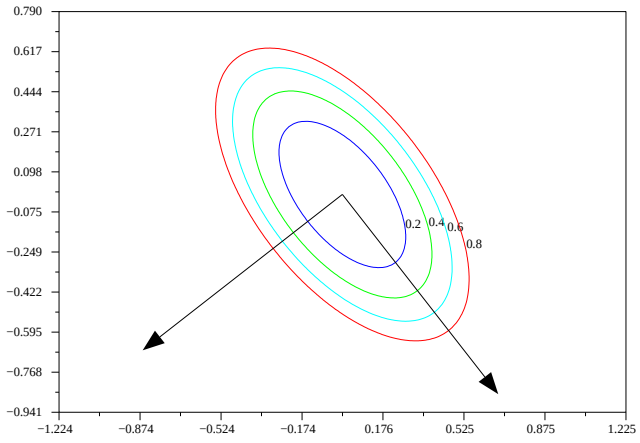


FIGURE 1.1 – Courbes de niveaux $(A\mathbf{x}, \mathbf{x}) = cste$ avec $\det(A) > 0$ $(N_+ = N = 2)$.

La figure 1.2 montre un exemple *hyperbolique* lorsque la matrice A a toutes ses valeurs propres positives sauf une strictement négative $(\det(A) < 0)$. On a choisi la matrice $A = \begin{pmatrix} 2 & 3 \\ 3 & 1 \end{pmatrix}$ dont les valeurs propres sont $\lambda_1 \simeq -1.5413813$ et $\lambda_2 \simeq 4.5413813$.

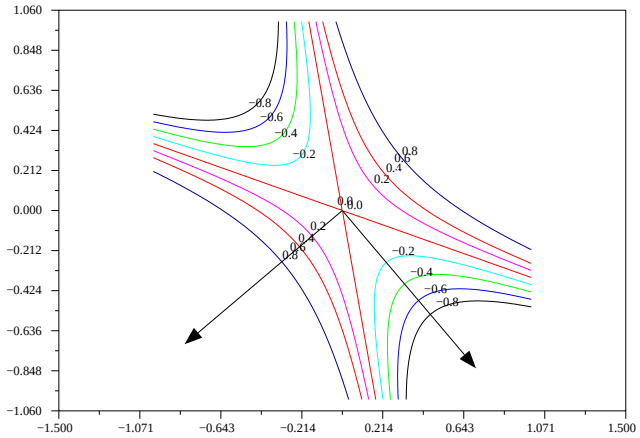


FIGURE 1.2 – Courbes de niveaux $(A\mathbf{x}, \mathbf{x}) = cste$ avec $\det(A) < 0$ ($N_+ = N - 1 = 1$ et $N_- = 1$).

La figure 1.3 montre un exemple *parabolique* lorsque la matrice A est singulière ($\det(A) = 0$). On a choisi la matrice $A = \begin{pmatrix} 2 & 1 \\ 1 & 0.5 \end{pmatrix}$ dont les valeurs propres sont $\lambda_1 = 0$ et $\lambda_2 = 2.5$.

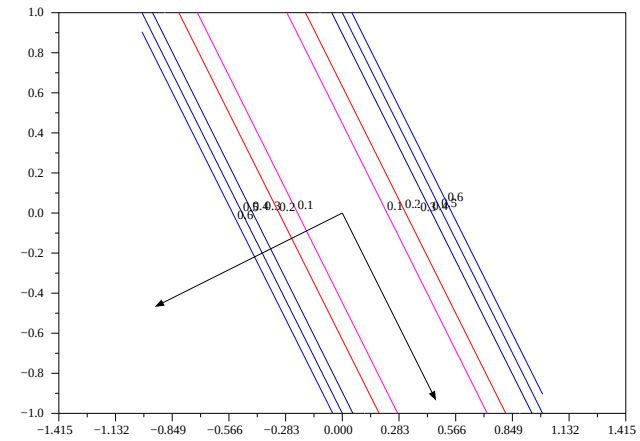


FIGURE 1.3 – Courbes de niveaux $(A\mathbf{x}, \mathbf{x}) = cste$ avec $\det(A) = 0$ ($N_0 = 1$).

Remarque : En général, une équation peut avoir différents types dans différentes parties du domaine Q. Par exemple l'équation de Tricomi

$$yu_{xx} + u_{yy} = f$$

est *elliptique* pour $y > 0$, *parabolique* pour $y = 0$ et *hyperbolique* pour $y < 0$.

Chapitre 2

EDP elliptiques linéaires

2.1 Introduction - Propriétés des solutions

Pour un domaine (ouvert connexe) $\Omega \subset \mathbb{R}^n$ borné et de frontière $\partial\Omega$ *régulière*, on considère le problème aux limites suivant pour une fonction u :

$$\mathcal{L}u := -\mathrm{div}\,(A\nabla u) + c\,u = f \quad \text{dans } \Omega \tag{2.1}$$

$$u = 0 \quad \text{sur } \partial\Omega \tag{2.2}$$

où A est une matrice de taille $n \times n$ avec $A = A(\mathbf{x}) = (a_{ij}(\mathbf{x}))_{1\le i,j\le n}$ pour $\mathbf{x} \in \Omega$. On suppose dans tout ce chapitre (sauf précision contraire) que les coefficients $a_{ij} \in C^1(\overline{\Omega})$. Les fonctions f et c sont données et on suppose également dans tout ce chapitre, que la fonction $c \in C(\overline{\Omega})$ est **positive ou nulle** i.e. $c(\mathbf{x}) \ge 0,\ \forall \mathbf{x} \in \Omega$. L'opérateur $\mathcal{L}$ apparaît ici sous forme divergentielle et on doit lire

$$\mathrm{div}\,(A\nabla u) = \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right).$$

On dit que l'équation (2.1) est (uniformément) **elliptique** ou bien que l'opérateur $\mathcal{L}$ est elliptique si

$$\exists\, \alpha > 0 \ \text{ tel que } \ (A(\mathbf{x})\xi,\ \xi)_{\mathbb{R}^n} = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x})\xi_i\xi_j \ \ge\ \alpha|\xi|^2, \quad \forall \xi \in \mathbb{R}^n,\ \forall \mathbf{x} \in \Omega. \tag{2.3}$$

On supposera désormais que la condition d'ellipticité (2.3) est vérifiée.

Remarque. Si A vérifie la condition (2.3) alors l'équation (2.1) est elliptique au sens de la classification faite au chapitre 1 (cf. section 1.5). En effet, soit λ une valeur propre de A et $\mathbf{u} \neq 0$ un vecteur propre associé. On a d'une part $(A\mathbf{u},\, \mathbf{u}) \ge \alpha|\mathbf{u}|^2$ et d'autre part $(A\mathbf{u},\, \mathbf{u}) = \lambda(\mathbf{u},\, \mathbf{u}) = \lambda|\mathbf{u}|^2$. On en déduit que $\lambda \ge \alpha > 0$ donc toutes les valeurs propres de A sont strictement positives. La condition d'ellipticité (2.3) implique en fait que la matrice A est définie positive.

2.1.1 Quelques rappels sur l'existence, l'unicité et la régularité des solutions

Pour les quelques résultats d'existence, d'unicité et de principe du maximum rappelés ci-après, on pourra consulter par exemple [1], [3], [5], [4].

Commençons par le cas de la dimension 1 avec $\Omega = (0,1)$. Dans ce cas, le problème s'écrit

$$-\left(au'\right)' + c\,u = f \quad \text{dans } (0,1) \tag{2.4}$$

$$u(0) = u(1) = 0. \tag{2.5}$$

- Si $f,\ c \in C([0,1])$ avec $c \ge 0$ et si $a \in C^1([0,1])$ avec $a(x) \ge \alpha > 0$ pour tout $x \in (0,1)$([1]), alors le problème (2.4)-(2.5) admet une unique solution *classique* $u \in C^2([0,1])$.

1. Cette condition ne traduit rien d'autre que la condition d'ellipticité (2.3) dans le cas 1D.

Considérons à présent le cas de la dimension quelconque avec $\Omega \subset \mathbb{R}^n$ un domaine borné et *régulier*. On rappelle tout d'abord un résultat d'existence de solution *faible*([1]).

- Si $f \in L^2(\Omega)$, $c \in C(\overline{\Omega})$ avec $c \geq 0$ et si les coefficients $a_{ij} \in C(\overline{\Omega})$ vérifient (2.3), alors le problème (2.1)-(2.2) admet une unique solution *faible* $u \in H_0^1(\Omega)$ (Lax-Milgram) et $\|u\|_{H^1} \leq C\|f\|_{L^2}$ où C est une constante indépendante de u et f.

  Si de plus, $a_{ij} \in C^1(\overline{\Omega})$ alors $u \in H^2(\Omega)$([2]) et u vérifie les équations (2.1)-(2.2) au sens *presque partout*.

Donnons enfin un résultat de régularité dans les espaces de Hölder. Pour $0 < \alpha < 1$, on définit les espaces

$$C^{0,\alpha}(\overline{\Omega}) = \{u \in C^0(\overline{\Omega}),\ \sup_{\mathbf{x}\neq\mathbf{y}} \frac{|u(\mathbf{x}) - u(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|^\alpha} < \infty\}$$

et pour $m \in \mathbb{N}$,

$$C^{m,\alpha}(\overline{\Omega}) = \{u \in C^m(\overline{\Omega}),\ D^\beta u \in C^{0,\alpha}(\overline{\Omega}) \text{ avec } |\beta| = m\}.$$

- Si $a_{ij} \in C^{k+1,\alpha}(\overline{\Omega})$ vérifient (2.3) et f, $c \in C^{k,\alpha}(\overline{\Omega})$ avec $c \geq 0$, alors il existe une unique solution $u \in C^{k+2,\alpha}(\overline{\Omega})$.

Ces résultats indiquent un effet "régularisant" d'un opérateur elliptique, par rapport aux données. Par exemple pour l'opérateur Laplacien ($A = I_d$), si on prend un second membre f aléatoire (entre -100 et 100) sur le pavé unité $\Omega = (0,1)\times(0,1)$ on obtient une solution "plus régulière", comme l'illustre la figure 2.1. Si à présent on prend cette solution comme second membre de l'équation de Laplace, on obtient une solution encore plus régulière (cf. Fig. 2.1).

Remarques sur la régularité des solutions.

L'étude de la convergence d'approximation de solution doit tenir compte de la régularité de la solution exacte. Si la solution est faible (L^2, $H^1, \cdots$), il faut regarder la convergence dans L^2, $H^1, \cdots$; si la solution est régulière ($C^2, \cdots$), alors on peut regarder la convergence dans $L^2, H^1, H^2, L^\infty, C^1, C^2, \cdots$.

Par ailleurs, même pour des problèmes elliptiques simples, la solution n'est pas toujours régulière ... jusqu'au bord. Par exemple, considérons le problème

$$-\Delta u = 1 \quad \text{dans } \Omega = (0,1)\times(0,1) \tag{2.6}$$
$$u = 0 \quad \text{sur } \partial\Omega \tag{2.7}$$

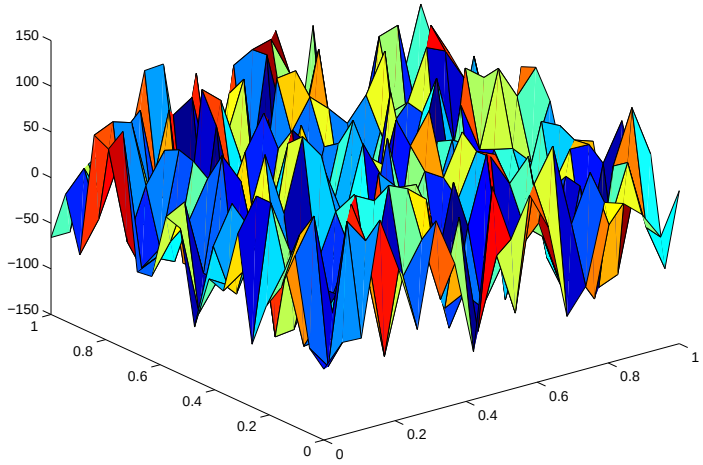
Ce problème n'a pas de solution dans $C^2(\overline{\Omega})$. En effet, si une telle solution existait, on aurait $u(x,0) = 0$ pour tout $x \in [0,1]$ et donc $\dfrac{\partial^2 u}{\partial x^2}(0,0) = 0$. De même, on aurait $u(0,y) = 0$ pour tout $y \in [0,1]$ et donc $\dfrac{\partial^2 u}{\partial y^2}(0,0) = 0$. Ainsi, on obtiendrait $\Delta u(0,0) = 0$, ce qui contredirait (2.6) en faisant tendre (x,y) vers $(0,0)$.

2.1.2 Principes du maximum

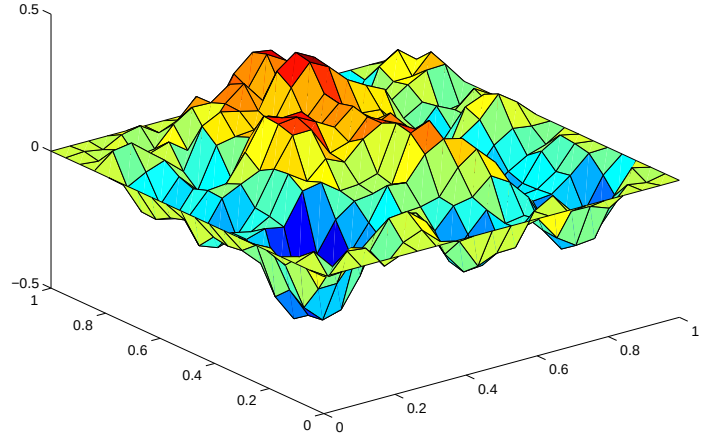
On donne à présent des résultats de principe du maximum pour des solutions classiques.

On suppose que $a_{ij} \in C^1(\overline{\Omega})$ vérifient (2.3) et $c \in C(\overline{\Omega})$ avec $c \geq 0$. Soit $u \in C^2(\Omega) \cap C^0(\overline{\Omega})$ vérifiant $\mathcal{L}u = f$ dans Ω.

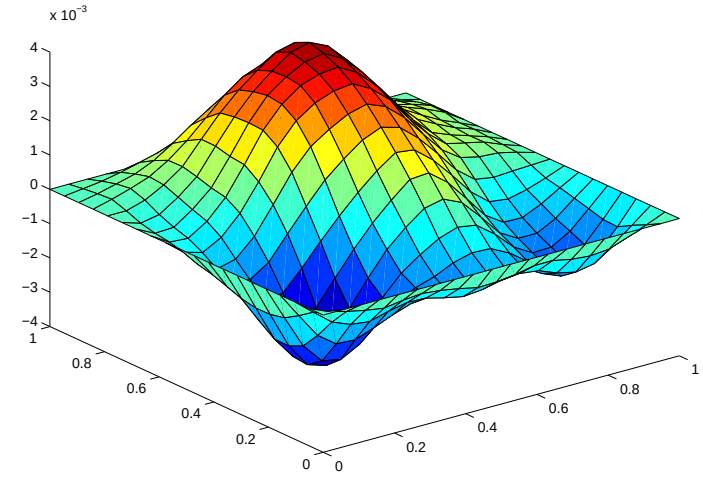
1. u est solution faible de (2.1),(2.2) si $u \in H_0^1(\Omega)$ et vérifie $\int_\Omega A\nabla u \cdot \nabla v + \int_\Omega c\,uv = \int_\Omega fv, \quad \forall v \in H_0^1(\Omega)$.
2. On a supposé Ω régulier ; en général, on n'a pas la régularité $u \in H^2(\Omega)$ si Ω présente des coins rentrants par exemple.



Terme source initial f.



Solution de -$\Delta u_1 = f$.



Solution de -$\Delta u_2 = u_1$.

FIGURE 2.1 – Effet “régularisant” du Laplacien.

i) • On prend $c \equiv 0$. Si $f \leq 0$ (resp. $f \geq 0$) alors u atteint son maximum (resp. minimum) sur $\partial\Omega$.

ii) • Si $c \equiv 0$ et $f \equiv 0$ alors $\inf_{\partial\Omega} u \leq u \leq \sup_{\partial\Omega} u$. Ceci montre qu'avec $c \equiv 0$ et $f \equiv 0$, si on choisit en particulier $u_{|\partial\Omega} = 0$ alors on obtient $u \equiv 0$.

iii) • Si $f \geq 0$ et $u_{|\partial\Omega} \geq 0$ alors $u \geq 0$ dans Ω.

iv) • (Principe de Hopf) Soit $f \leq 0$ et soit $\mathbf{x}_0 \in \partial\Omega$ tel que $u(\mathbf{x}_0) > u(\mathbf{x})$, $\forall \mathbf{x} \in \Omega$. On suppose que u est dérivable en $\mathbf{x}_0$. Si $c \equiv 0$ ou bien si $u(\mathbf{x}_0) = 0$, alors

$$\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}_0) > 0,$$

où $\mathbf{n}$ désigne la normale extérieure à $\partial\Omega$.

Démonstration (directe) de iii) dans le cas $c > 0$.

Soit $\mathbf{x_0} \in \overline{\Omega}$ tel que $u(\mathbf{x_0}) = \min_{\overline{\Omega}} u$, i.e. le point de $\overline{\Omega}$ où le minimum de u est atteint.

- Si $\mathbf{x_0} \in \partial\Omega$ alors $u(\mathbf{x}) \geq u(\mathbf{x_0}) \geq 0$ pour tout $\mathbf{x} \in \Omega$ et la démonstration est terminée.

- Si $\mathbf{x_0} \in \Omega$ alors $\nabla u(\mathbf{x_0}) = 0$ et $\dfrac{\partial^2 u}{\partial x_i^2}(\mathbf{x_0}) \geq 0$ pour $i = 1, \cdots, n$. On va alors montrer que

$$\sum_{i,j} a_{ij}(\mathbf{x_0}) \frac{\partial^2 u}{\partial x_i x_j}(\mathbf{x_0}) \geq 0. \tag{2.8}$$

On peut tout d'abord supposé, sans perte de généralité, que la matrice A est symétrique (cf. note de bas de page de la Section 1.5, Chap. 1). La matrice A est donc diagonalisable par

$$D = CAC^T, \tag{2.9}$$

avec $D = \mathrm{diag}(\lambda_1, \cdots, \lambda_n)$ où les λ_i sont les valeurs propres (positives) de A et $C = C(\mathbf{x_0})$ est la matrice orthogonale ($C^{-1} = C^T$) formée des vecteurs propres associés. On considère alors la fonction $\tilde{u}(\mathbf{y}) = u(\mathbf{x})$ avec $\mathbf{y} = C\mathbf{x}$. La fonction $\tilde{u}$ est définie sur l'ouvert $\mathcal{O} = \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{y} = C\mathbf{x},\ \mathbf{x} \in \Omega\}$ et atteint son minimum au point intérieur $\mathbf{y_0} = C\mathbf{x_0} \in \mathcal{O}$([1]). Par conséquent, on a, pour $j = 1, \cdots, n$,

$$\frac{\partial^2 \tilde{u}}{\partial y_j^2}(\mathbf{y_0}) \geq 0, \quad \text{avec } \mathbf{y_0} = C\mathbf{x_0}. \tag{2.10}$$

On note (c_{ij}) et (d_{ij}) les coefficients de C et D. En utilisant le fait que $\dfrac{\partial y_i}{\partial x_j} = c_{ij}$ et la décomposition (2.9), on montre que

$$\mathcal{L}u = \sum_{i,j} a_{ij} \frac{\partial^2 u}{\partial x_i x_j} = \sum_{i,j} d_{ij} \frac{\partial^2 \tilde{u}}{\partial y_i y_j} = \sum_i \lambda_i \frac{\partial^2 \tilde{u}}{\partial y_i^2}. \tag{2.11}$$

On conclut à (2.8) par (2.11), (2.10) et le fait que toutes les valeurs propres λ_i de A soient positives (condition d'ellipticité (2.3)).

On termine alors la démonstration en écrivant que $-\mathrm{div}\,(A\nabla u)\,(\mathbf{x_0}) + c(\mathbf{x_0})u(\mathbf{x_0}) = f(\mathbf{x_0})$; puisque $-\mathrm{div}\,(A\nabla u)\,(\mathbf{x_0}) = -\sum_{i,j} a_{ij}(\mathbf{x_0}) \dfrac{\partial^2 u}{\partial x_i x_j}(\mathbf{x_0}) \leq 0$ et $f(\mathbf{x_0}) \geq 0$, on en déduit que $c(\mathbf{x_0})u(\mathbf{x_0}) \geq f(\mathbf{x_0}) \geq 0$ et donc que $u(\mathbf{x_0}) \geq 0$ (on a supposé $c > 0$), ce qui montre que $u \geq 0$ dans Ω. □

1. $\mathcal{O} = f^{-1}(\Omega)$ est un ouvert comme image réciproque de l'ouvert Ω par l'application continue $f : \mathbf{y} \mapsto C^{-1}\mathbf{y}$.

2.2 Différences finies pour le cas 1D

On choisit pour simplifier $\Omega = (0,1)$. Le problème consiste à trouver une fonction $u = u(x)$ qui vérifie

$$(P)\ \left\{ \begin{array}{rcl} Lu := -(au')' + cu & = & f \quad \text{dans } (0,1) \\ u(0) = u(1) & = & 0. \end{array} \right.$$

On se donne les fonctions $f,\ c \in C([0,1])$ avec $c \geq 0$ et $a \in C^1([0,1])$ avec $a(x) \geq \alpha > 0$ pour tout $x \in [0,1]$, de sorte que (P) admet une unique solution $u \in C^2([0,1])$.

Principe des Différences Finies (DF).

On discrétise le segment $[0,1]$ et on approche les dérivées aux points de discrétisation, par des opérateurs aux différences. Plus précisément, on se donne un entier N à partir duquel on définit le pas de discrétisation $h = 1/(N+1)$; on introduit les $N+2$ points $x_i = ih$ pour $i = 0, \cdots, N+1$ qui forment alors une subdivision régulière de l'intervalle $[0,1]$. On définit Ω_h le *maillage intérieur* à $(0,1)$ par

$$\Omega_h = \{x_i,\ 1 \leq i \leq N\} \tag{2.12}$$

et $\overline{\Omega}_h$ le *maillage complet* de $[0,1]$ par

$$\overline{\Omega}_h = \{x_i,\ 0 \leq i \leq N+1\}. \tag{2.13}$$

On désigne par V_h (resp. $\overline{V}_h$) l'ensemble des fonctions (*discrètes*) définies sur le maillage Ω_h (resp. $\overline{\Omega}_h$), qu'on identifie à $\mathbb{R}^N$ (resp. $\mathbb{R}^{N+2}$) c'est-à-dire qu'un élément v_h de V_h (resp. $\overline{V}_h$) est identifié à un vecteur $\mathbf{v}_h$ de $\mathbb{R}^N$ (resp. $\mathbb{R}^{N+2}$) dont les composantes sont $(v_h(x_1), \cdots, v_h(x_N))$ (resp. $(v_h(x_0), \cdots, v_h(x_{N+1}))$). Les composantes de $\mathbf{v}_h$ seront désormais notées $v_i = v_h(x_i)$.

On définit l'opérateur aux différences centrées $D_h : \overline{V}_{h/2} \mapsto V_{h/2}$ par

$$D_h v(x) = \frac{v(x+h/2) - v(x-h/2)}{h}, \quad x \in \Omega_h \tag{2.14}$$

pour une fonction $v \in \overline{V}_{h/2}$. Comme on le verra plus tard, $D_h v$ est une approximation de la dérivée de v aux points du maillage. Par application itérée de l'opérateur aux différences D_h, on obtient

$$D_h\left(a\, D_h u_h\right)(x) = \frac{1}{h^2}\left[a(x+h/2)u_h(x+h) - \left(a(x+h/2) + a(x-h/2)\right)u_h(x) + a(x-h/2)u_h(x-h)\right], \tag{2.15}$$

qui est défini pour $u_h \in \overline{V}_h$ et $x \in \Omega_h$. On remarquera que <u>seules interviennent les valeurs de la fonction u_h aux points x_i</u>.

On considère alors le problème approché qui consiste à trouver une fonction $u_h \in \overline{V}_h$ telle que

$$(P_h)\ \left\{ \begin{array}{rcl} L_h u_h := -D_h\left(a\, D_h u_h\right) + c\, u_h & = & f \quad \text{dans } \Omega_h \\ u_h(0) = u_h(1) & = & 0. \end{array} \right.$$

Les valeurs des fonctions c et f aux pointx x_i seront notées c_i et f_i.

Cas particulier $a \equiv 1$.

Le problème approché (P_h) s'écrit dans ce cas, de la façon suivante. Trouver les u_i tels que

$$\left\{ \begin{array}{rcl} -\dfrac{(u_{i+1} - 2u_i + u_{i-1})}{h^2} + c_i u_i & = & f_i \quad \text{pour } i = 1, \cdots, N \\ u_0 = u_{N+1} & = & 0. \end{array} \right. \tag{2.16}$$

On peut écrire le système précédent sous forme matricielle. Les inconnues sont regroupées dans le vecteur $\mathbf{u}_h = (u_1, \cdots, u_N)^\top \in \mathbb{R}^N$ qui vérifie

$$A_h \mathbf{u}_h = \mathbf{b}, \tag{2.17}$$

où la matrice A_h de taille $N \times N$ est donnée par

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix} + C_h \text{ avec } C_h = \mathrm{diag}(c_1, \cdots, c_N) \tag{2.18}$$

et le second membre $\mathbf{b} = (f_1, \cdots, f_N)^\top \in \mathbb{R}^N$.

Cas général $a \not\equiv 1$.

Le problème approché (P_h) peut s'écrire de la façon suivante. Trouver les u_i tels que

$$\left\{ \begin{array}{rcl} -\frac{1}{h^2}\left[a_{i+1/2}u_{i+1} - \left(a_{i+1/2} + a_{i-1/2}\right)u_i + a_{i-1/2}u_{i-1}\right] + c_i u_i &=& f_i \quad \text{pour } i = 1, \cdots, N \\ u_0 = u_{N+1} &=& 0, \end{array} \right. \tag{2.19}$$

avec $a_{i+1/2} = a(x_i + h/2)$ et $a_{i-1/2} = a(x_i - h/2)$. La matrice A_h du système linéaire (2.17) correspondant est donnée dans le cas général par

$$A_h = \frac{1}{h^2}\begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{N-2} & \alpha_{N-1} & \beta_{N-1} \\ 0 & & & \beta_{N-1} & \alpha_N \end{pmatrix} + C_h \tag{2.20}$$

avec $\alpha_i = a_{i+1/2} + a_{i-1/2}$, $\beta_i = -a_{i+1/2}$ et $C_h = \mathrm{diag}(c_1, \cdots, c_N)$. Cette matrice est symétrique et définie positive (voir Proposition 2.7), de sorte que le système linéaire (2.17) (et donc le problème approché) admet une unique solution $\mathbf{u}_h$.

2.2.1 Erreur de consistance

L'erreur de consistance correspond à l'erreur commise lorsqu'on remplace l'opérateur différentiel $L : v \mapsto -(a\,v')' + cv$ par l'opérateur aux différences $L_h : v \mapsto -D_h(a\,D_h v) + cv$. L'erreur de consistance de (P_h) par rapport à (P) est définie par

$$R_h v(x) = Lv(x) - L_h v(x), \tag{2.21}$$

pour $v \in C^2([0,1])$ et $x \in \Omega_h$ (i.e. pour $x = x_i$, $i = 1, \cdots, N$). Le problème (P_h) est dit consistant par rapport à (P) si pour chaque v, on a $R_h v = Lv - L_h v \to 0$, quand $h \to 0$. La convergence précédente doit être précisée. On s'intéresse ici à la norme discrète du maximum, définie pour des fonctions $v \in V_h$ par

$$\|v\|_{h,\infty} = \max_{1 \le i \le N} |v(x_i)|. \tag{2.22}$$

On remarquera que dans la norme discrète, on ne prend pas en compte les valeurs de la fonction aux points extrêmes $x_0 = 0$ et $x_{N+1} = 1$. Mais, comme on le verra, ceci n'est pas génant pour évaluer la différence entre les solutions exacte et approchée car toutes deux coïncident et sont nulles en 0 et 1 (Dirichlet homogène). Enfin, on rappelle qu'on identifie une fonction $v \in V_h$ avec un vecteur $\mathbf{v} \in \mathbb{R}^N$ et on a $\|v\|_{h,\infty} = \|\mathbf{v}\|_\infty := \max_{1 \le i \le N} |v_i|$ où les v_i sont les composantes de $\mathbf{v}$ (avec $v_i = v(x_i)$).

Proposition 2.1 *Le problème (P_h) est consistant par rapport à (P) et on a*

i) Pour $a \equiv 1$, si $v \in C^4([0,1])$ alors on a

$$\|R_h v\|_{h,\infty} = \|Lv - L_h v\|_{h,\infty} \le \frac{1}{12}\|v^{(4)}\|_\infty \, h^2. \tag{2.23}$$

ii) Si $a \in C^3([0,1])$ et $v \in C^4([0,1])$ alors il existe une constante $C = C(\|a\|_{C^3}, \|v\|_{C^4}) > 0$ indépendante de h telle que

$$\|R_h v\|_{h,\infty} = \|Lv - L_h v\|_{h,\infty} \le C\, h^2. \tag{2.24}$$

Démonstration. La preuve est donnée dans le cas $a \equiv 1$ (traiter le cas général $a \in C^3$ en exercice). Remarquons tout d'abord que

$$D_h\left(D_h v\right) = \frac{v(x+h) - 2v(x) + v(x-h)}{h^2}.$$

Les développements de Taylor-MacLaurin de $v \in C^4([0,1])$ donnent

$$\begin{aligned} v(x+h) &= v(x) + hv'(x) + \frac{h^2}{2}v''(x) + \frac{h^3}{3!}v^{(3)}(x) + \frac{h^4}{4!}v^{(4)}(x+\theta_1 h) \quad \text{avec } 0 < \theta_1 < 1. \\ v(x-h) &= v(x) - hv'(x) + \frac{h^2}{2}v''(x) - \frac{h^3}{3!}v^{(3)}(x) + \frac{h^4}{4!}v^{(4)}(x+\theta_2 h) \quad \text{avec } 0 < \theta_2 < 1. \end{aligned}$$

Des deux équations précédentes, on obtient

$$v(x+h) - 2v(x) + v(x-h) = h^2 v''(x) + \frac{h^4}{4!}\left(v^{(4)}(x+\theta_1 h) + v^{(4)}(x+\theta_2 h)\right)$$

d'où l'on déduit (2.23). □

2.2.2 Matrices monotones

Avant de s'intéresser à la notion de stabilité, on rappelle quelques résultats utiles par la suite, sur les matrices monotones et les M-matrices.

Définition 2.1 *Une matrice $A \in \mathbb{R}^{N\times N}$ est dite monotone si A est inversible et si $A^{-1} \geq 0$ i.e. tous les coefficients de A^{-1} sont positifs ou nuls.*

La terminologie est en fait justifiée par le résultat suivant.

Proposition 2.2 *Une matrice A est monotone si et seulement si $(A\mathbf{x} \geq 0 \Rightarrow \mathbf{x} \geq 0)$.*

Démonstration.

- Supposons tout d'abord que A soit une matrice monotone et considérons $\mathbf{x} \in \mathbb{R}^N$ tel que $A\mathbf{x} \geq 0$. Montrons que $\mathbf{x} \geq 0$. La matrice A étant inversible, on écrit $\mathbf{x} = A^{-1}A\mathbf{x}$. Comme $A^{-1} \geq 0$ et que $A\mathbf{x} \geq 0$, on en déduit que $\mathbf{x} \geq 0$.

- Supposons maintenant que la propriété $(A\mathbf{x} \geq 0 \Rightarrow \mathbf{x} \geq 0)$ soit vraie. Montrons d'abord que A est inversible. Soit $\mathbf{x} \in \mathbb{R}^N$ tel que $A\mathbf{x} = 0$. On déduit de l'hypothèse que $\mathbf{x} \geq 0$. De plus, on a $A(-\mathbf{x}) = 0$ d'où l'on déduit également que $-\mathbf{x} \geq 0$. On a ainsi établi que $\mathbf{x} = 0$. Il reste à montrer que $A^{-1} \geq 0$. Soit alors $\mathbf{y} \in \mathbb{R}^N$ tel que $\mathbf{y} \geq 0$ et on pose $\mathbf{x} = A^{-1}\mathbf{y}$ et par conséquent $\mathbf{y} = A\mathbf{x} \geq 0$. L'hypothèse implique que $\mathbf{x} \geq 0$ et donc que $A^{-1}\mathbf{y} \geq 0$. Par des choix convenables de $\mathbf{y}$ ($\mathbf{y} = (1,0,\cdots,0)^\top, \mathbf{y} = (0,1,0,\cdots,0)^\top$, $\cdots, \mathbf{y} = (0,\cdots,0,1)^\top$), on en déduit que $A^{-1} \geq 0$. □

Définition 2.2 *Une matrice $A \in \mathbb{R}^{N\times N}$ est une M-matrice si elle est monotone et si $a_{ij} \leq 0$ pour $i \neq j$.*

Voici quelques critères pratiques de détermination de M-matrices.

Proposition 2.3 *Soit une matrice $A \in \mathbb{R}^{N\times N}$ telle que*

i) $a_{ij} \leq 0$ pour tous $i \neq j$.

ii) $\displaystyle\sum_{1\leq j\leq N} a_{ij} > 0$ pour $i = 1, \cdots, N$.

Alors A est une M-matrice.

Démonstration.

On a clairement $a_{ii} > 0$, pour tout i. On introduit alors la matrice diagonale $D = \mathrm{diag}\,(a_{ii})$ qui est inversible avec $D^{-1} = \mathrm{diag}\,(1/a_{ii})$ et par conséquent $D^{-1} \geq 0$. On décompose alors A sous la forme $A = D(I - M)$ où $I \in \mathbb{R}^{N\times N}$ est la matrice identité et $M = (m_{ij})$ avec $m_{ij} = -\dfrac{a_{ij}}{a_{ii}}$ pour $i \neq j$ et $m_{ii} = 0$. Par l'hypothèse *i)*, on a $M \geq 0$. On va montrer que $I - M$ est inversible puis monotone. Pour cela, on va utiliser un résultat bien connu sur le rayon spectral de matrice (cf. [2]).

Lemme 2.1 *Soit $M \in \mathbb{R}^{N\times N}$ une matrice dont on note $\lambda_i(M)$, $i = 1, \cdots, N$ les valeurs propres et $\rho(M) = \max_i |\lambda_i(M)|$ le rayon spectral([1]).*

i) Pour toute norme matricielle $\|\cdot\|$ subordonnée([2]),on a $\rho(M) \leq \|M\|$.

ii) Si $\rho(M) < 1$ alors $I - M$ est inversible et $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$.

En utilisant le point *i)* de ce lemme, on a $\rho(M) \leq \|M\|_\infty = \sup_{1\leq i\leq N} \sum_{j=1}^{N} |m_{ij}|$. Or, d'après les hypothèses *i)* et *ii)* de la Proposition, il vient

$$\sum_{j=1}^{N} |m_{ij}| = \sum_{j\neq i} \left|-\frac{a_{ij}}{a_{ii}}\right| = -\frac{1}{a_{ii}} \sum_{j\neq i} a_{ij} < 1,$$

ce qui implique d'après le point *ii)* du lemme, que la matrice $I-M$ est inversible et $(I - M)^{-1} = \sum_{k=0}^{\infty} M^k$. Comme $M \geq 0$, on a aussi $M^k \geq 0$ et donc $(I - M)^{-1} \geq 0$. Ainsi, $A = D(I - M)$ est inversible et $A^{-1} = (I - M)^{-1} D^{-1} \geq 0$. Les coefficients a_{ij} étant tous négatifs ou nuls par hypothèse, la matrice A est bien une M-matrice. □

On peut "relacher" la contrainte de positivité stricte de la somme des coefficients et obtenir le résultat suivant.

Proposition 2.4 *Soit une matrice $A \in \mathbb{R}^{N\times N}$ telle que*

i) $a_{ij} \leq 0$ pour tous $i \neq j$.

ii) $\sum_{1\leq j\leq N} a_{ij} \geq 0$ pour $i = 1, \cdots, N$.

iii) A est inversible.

Alors A est une M-matrice.

Démonstration.

- Montrons d'abord que $A + \varepsilon I$ est monotone, quelque soit $\varepsilon > 0$. On pose $A + \varepsilon I = \left(a_{ij}^{\varepsilon}\right)_{i,j}$ avec $a_{ij}^{\varepsilon} = a_{ij} + \varepsilon\delta_{ij}$. Pour $\varepsilon > 0$, on a $a_{ij}^{\varepsilon} \leq 0$ pour $i \neq j$ et $\sum_j a_{ij}^{\varepsilon} = \sum_j a_{ij} + \varepsilon \geq \varepsilon > 0$. Par conséquent, d'après la proposition 2.3, la matrice $A + \varepsilon I$ est monotone.

- Montrons à présent que A est monotone par passage à la limite $\varepsilon \to 0$. La matrice A est inversible donc on peut écrire $A + \varepsilon I = A\left(I + \varepsilon A^{-1}\right)$. De plus, on a $\rho(\varepsilon A) = \varepsilon\rho(A) < 1$ pour ε suffisamment petit donc $I + \varepsilon A^{-1}$ est inversible et $\left(I + \varepsilon A^{-1}\right)^{-1} = \sum_{k=0}^{\infty} \left(-\varepsilon A^{-1}\right)^k$. Par ailleurs, $A + \varepsilon I$ est inversible avec

$$(A + \varepsilon I)^{-1} = \left(I + \varepsilon A^{-1}\right)^{-1} A^{-1} = \left(\sum_{k=0}^{\infty} (-\varepsilon)^k (A^{-1})^k\right) A^{-1} = \left(I + \sum_{k=1}^{\infty} (-\varepsilon)^k (A^{-1})^k\right) A^{-1}.$$

Ainsi, $(A + \varepsilon I)^{-1} - A^{-1} = \sum_{k=1}^{\infty} (-\varepsilon)^k (A^{-1})^k\ A^{-1}$ et donc

$$\left\|(A + \varepsilon I)^{-1} - A^{-1}\right\| \leq \left(\sum_{k=1}^{\infty} \varepsilon^k \|A^{-1}\|^k\right) \|A^{-1}\| = \frac{\varepsilon\|A^{-1}\|}{1 - \varepsilon\|A^{-1}\|} \|A^{-1}\| \to 0, \text{ quand } \varepsilon \to 0,$$

1. Si M est symétrique, on a $\|M\|_2 := \sup_{\mathbf{x}\neq 0} \frac{\|M\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \rho(M)$.

2. $\|M\| = \sup_{\mathbf{x}\neq 0} \frac{\|M\mathbf{x}\|}{\|\mathbf{x}\|}$

pour n'importe quelle norme matricielle $\|\cdot\|$ subordonnée. Par conséquent chaque coefficient de $(A+\varepsilon I)^{-1}$ qui est positif ou nul, tend vers le coefficient correspondant de A^{-1} qui est donc également positif ou nul. Ainsi $A^{-1} \geq 0$ et A est une M-matrice. □

2.2.3 Stabilité

Soient u la solution de (P) (avec $Lu = f$ dans Ω) et u_h la solution de (P_h) (avec $L_h u_h = f$ dans Ω_h). On a $Lu - L_h u_h = 0$ dans Ω_h et donc $0 = Lu - L_h u + L_h(u - u_h) = R_h u + L_h(u - u_h)$ où $R_h u$ est l'erreur de consistance (encore appelée résidu). On obtient ainsi

$$R_h u = L_h(u_h - u) \quad \text{dans } \Omega_h. \tag{2.25}$$

On rappelle (cf. Proposition 2.1) que $\|R_h u\|_{h,\infty} \leq Ch^2$ si la solution u est suffisamment régulière. La question que l'on se pose à présent est de savoir si la différence $u - u_h$ entre les solutions exacte et approchée est petite, lorsque le résidu $R_h u$ est "petit" (avec h "petit") ? C'est la notion de stabilité.

Plus précisement, on dit que le problème approché (P_h) est stable si pour toute fonction f, la solution correspondante u_h de (P_h) avec f, est bornée par la donnée f. Les normes utilisées sont liées à la régularité des solutions c'est-à-dire aux espaces dans lesquels on cherche la solution et dans lesquels les fonctions sont données. La notion de stabilité dépend donc des normes utilisées et comme on le verra plus tard, un problème peut être stable pour une norme et ne pas l'être pour une autre norme.

On va montrer que le problème (P_h) est stable pour la norme discrète du maximum. Commençons d'abord par un principe du maximum discret.

Proposition 2.5 (Principe du maximum discret)
Si $f \geq 0$, alors la solution u_h du problème (P_h) vérifie $u_h \geq 0$ dans $\overline{\Omega}_h$.

Démonstration. On donne une démonstration directe dans le cas où $a \equiv 1$ et $c > 0$. On retrouvera ce résultat plus tard, en montrant que la matrice du système linéaire est monotone. La solution u_h vérifie

$$-D_h(D_h u_h)(x) + c(x)u_h(x) = -\frac{u_h(x-h) - 2u_h(x) + u_h(x+h)}{h^2} + c(x)u_h(x) = f(x),$$

pour $x \in \Omega_h$, c'est-à-dire aux points intérieurs $x = x_i$, $i = 1, ..., N$ de l'intervalle $(0,1)$. Soit x_{i_0} le point de $\overline{\Omega}_h$ où le minimum de $u_h \in \overline{V}_h$ est atteint i.e. $u_h(x_{i_0}) \leq u_h(x_i)$ pour tout $i = 0, \cdots, N+1$. Si $i_0 \in \{1, \cdots, N\}$, alors $-D_h\left(D_h u_h\right)(x_{i_0}) \leq 0$ et donc $c(x_{i_0})u_h(x_{i_0}) \geq f(x_{i_0}) \geq 0$, d'où $u_h(x_{i_0}) \geq 0$. Si $i_0 = 1$ ou $i_0 = N$, alors $u_h(x_{i_0}) = 0$. Dans tous les cas, on a $0 \leq u_h(x_i)$, $\forall i$, d'où la conclusion. □

Comme nous allons le voir, ce résultat est en fait une conséquence directe d'une propriété plus générale de monotonie de la matrice A_h du problème approché. Les résultats sur les matrices monotones vont surtout nous servir à établir la stabilité du problème (P_h). On va examiner tout d'abord le cas homogène où $a \equiv 1$ puis le cas hétérogène où a est quelconque ($a \in C^3([0,1])$).

Stabilité dans le cas homogène $a \equiv 1$.

On suppose pour simplifier que $c \equiv 0$. La stabilité s'établit à partir des propriétés de la matrice $A_h \in \mathbb{R}^{N \times N}$ du système linéaire (2.17). On rappelle (avec $c \equiv 0$) que

$$A_h = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix}. \tag{2.26}$$

Proposition 2.6 *La matrice A_h (avec $c \equiv 0$) vérifie les propriétés suivantes.*

i) *A_h est une M-matrice, symétrique définie positive.*

ii) (STABILITÉ) $\|A_h^{-1}\|_\infty = \max_i \sum_j |b_{ij}| \leq \frac{1}{8}$ *où les b_{ij} sont les coefficients de la matrice A_h^{-1}.*

Remarques :

- D'après la propriété *i)* précédente, la matrice A_h est monotone, ce qui permet de retrouver le principe du maximum discret établi à la Proposition 2.5.
- La propriété *ii)* correspond bien à la notion de stabilité introduite au début du paragraphe. En effet, la solution u_h du problème approché est identifiée au vecteur $\mathbf{u}_h$ qui vérifie $A_h\mathbf{u}_h = \mathbf{b}$. Par conséquent, on a $\|u_h\|_{h,\infty} = \|\mathbf{u}_h\|_\infty = \|A_h^{-1}\mathbf{b}\|_\infty \leq \|A_h^{-1}\|_\infty \|\mathbf{b}\|_\infty \leq \frac{1}{8}\|f\|_\infty$.

Démonstration de la Proposition. i) Il suffit de vérifier que $(A_h\mathbf{x}, \mathbf{x}) = \dfrac{1}{h^2}\left(x_1^2 + x_N^2 + \displaystyle\sum_{i=2}^{N}(x_i - x_{i-1})^2\right)$ pour tout $\mathbf{x} = (x_1, \cdots, x_N)^\top$ et d'utiliser la Proposition 2.4 sur les M-matrices.

ii) On a $\|A_h^{-1}\|_\infty = \max_i \sum_j |b_{ij}| = \max_i \sum_j b_{ij}$ car $A_h^{-1} \geq 0$ et par conséquent $\|A_h^{-1}\|_\infty = \|A_h^{-1}\mathbf{e}\|_\infty$ avec $\mathbf{e} = (1, 1, \cdots, 1)^\top \in \mathbb{R}^N$. On considère alors le problème suivant

$$\begin{aligned} Lu := -u'' &= 1 \quad \text{dans } (0,1) \\ u(0) = u(1) &= 0, \end{aligned}$$

qui admet pour solution $u(x) = \dfrac{1}{2}x(1-x)$. On considère alors la solution u_h du problème approché associé au problème précédent, c'est-à-dire $u_h \in \overline{V}_h$ telle que $L_h u_h := -D_h(D_h u_h) = 1$ dans Ω_h et $u_h(0) = u_h(1) = 0$. On a $R_h u = L_h u - Lu = 0$ car d'après la Proposition 2.1, il existe $C > 0$ telle que $\|R_h u\|_{h,\infty} \leq C\|u^{(4)}\|_\infty h^2$ et la fonction u est telle que $u^{(4)} \equiv 0$ car u est un polynôme de degré 2. Ainsi, le résidu étant nul, on obtient d'après (2.25), $L_h u = L_h u_h = 1$ dans Ω_h et par conséquent $u_h(x_i) = u(x_i)$ pour tout $i = 0, \cdots, N+1$, par unicité de la solution du problème approché. Finalement, on remarque que matriciellement, la solution approchée vérifie le système linéaire $A_h\mathbf{u}_h = \mathbf{e}$ et par conséquent, on obtient $\|A_h^{-1}\mathbf{e}\|_\infty = \|\mathbf{u}_h\|_\infty = \|u_h\|_{h,\infty} = \|u\|_{h,\infty} \leq \sup_{x\in[0,1]} |u(x)| = \dfrac{1}{8}$. □

Stabilité dans le cas hétérogène.

On suppose seulement que $a \in C^3([0,1])$ mais toujours (pour simplifier) que $c \equiv 0$. La matrice A_h est donnée cette fois par (2.20) (avec $C_h = 0$). En suivant la même démarche que dans le cas homogène précédent, on peut montrer les propriétés suivantes sur cette matrice.

Proposition 2.7

i) *A_h est une M-matrice, symétrique définie positive.*

ii) (STABILITÉ) *Si h est suffisamment petit, alors il existe une constante $C > 0$ indépendante de h telle que $\|A_h^{-1}\|_\infty \leq C$.*

2.2.4 Convergence

La convergence est établie dans le cas homogène $a \equiv 1$ et avec $c \equiv 0$ (pour simplifier).

Théorème 2.1 *Soient $f \in C^2([0,1])$, $a \equiv 1$ et $c \equiv 0$. On note u la solution de (P) et u_h la solution de (P_h). Alors il existe une constante $C > 0$ indépendante de h telle que*

$$\|u - u_h\|_{h,\infty} \leq C\|f^{(2)}\|_\infty h^2.$$

Remarque. La norme discrète $\|\cdot\|_{h,\infty}$ ne porte que sur les points intérieurs x_i, $i = 1, \cdots, N$. Mais les fonctions u et u_h coïncident et sont nulles en $x_0 = 0$ et $x_{N+1} = 1$.

Démonstration. On a $Lu = f$ dans Ω, $L_h u_h = f$ dans Ω_h et $R_h u = Lu - L_h u = L_h(u_h - u)$ (cf. (2.25)). Matriciellement, on peut écrire $A_h \mathbf{u}_h = \mathbf{b}$ et $A_h(\mathbf{u} - \mathbf{u}_h) = \mathbf{R}_h u$ avec $\mathbf{u} = (u(x_1), \cdots, u(x_N))^\top$, $\mathbf{R}_h u = (R_h u(x_1), \cdots, R_h u(x_N))^\top$. On en déduit que $\mathbf{u} - \mathbf{u}_h = A_h^{-1} \mathbf{R}_h u$. Puisque $f \in C^2([0,1])$, alors $u \in C^4([0,1])$ et on a

$$\|u - u_h\|_{h,\infty} = \|\mathbf{u} - \mathbf{u}_h\|_\infty \leq \|A_h^{-1}\|_\infty \|R_h u\|_{h,\infty} \leq C\|u^{(4)}\|_\infty h^2 = C\|f^{(2)}\|_\infty h^2,$$

d'après les Propositions 2.1 et 2.6. □

2.2.5 Autres conditions limites

On peut choisir d'autres types de conditions limites pour le problème (P).

Conditions de Dirichlet non-homogènes.

Plus généralement, des conditions de Dirichlet non-homogènes

$$u(0) = g_0, \quad u(1) = g_1,$$

peuvent être imposées avec g_0, g_1 données. Il faut alors modifier le second membre $\mathbf{b}$ du système linéaire $A_h \mathbf{u}_h = \mathbf{b}$. Par exemple dans le cas $a \equiv 1$, la matrice A_h de taille $N \times N$ est toujours donnée par (2.18) mais le second membre devient

$$\mathbf{b} = \left(f_1 + \frac{g_0}{h^2}, f_2, \cdots, f_{N-1}, f_N + \frac{g_1}{h^2}\right)^\top. \tag{2.27}$$

Conditions de Neumann.

On peut imposer aussi une condition de Neumann en choisissant par exemple

$$a(1)u'(1) = g_1 \tag{2.28}$$

avec g_1 donnée, au lieu de $u(1) = 0$. En revanche, on conserve (pour simplifier) la condition $u(0) = 0$. Il y a maintenant $N+1$ inconnues $\mathbf{u}_h = (u_1, \cdots, u_N, u_{N+1}) \in \mathbb{R}^{N+1}$. Pour les points *intérieurs* $x_1, \cdots, x_N$, on utilise la même discrétisation que précédemment. Dans le cas où $c \equiv 0$ (pour simplifier...), ceci fournit les N équations (cf. (2.19)) :

$$-\frac{1}{h^2}\left[a_{i+1/2}u_{i+1} - \left(a_{i+1/2} + a_{i-1/2}\right)u_i + a_{i-1/2}u_{i-1}\right] = f_i \quad \text{pour } i = 1, \cdots, N. \tag{2.29}$$

Mais, ayant une inconnue en plus (u_{N+1}), il nous faut une équation supplémentaire. Cette dernière est obtenue par discrétisation de la condition limite (2.28). **Il faut porter une attention particulière à cette discrétisation si l'on ne veut pas perdre l'ordre de convergence en** h^2. Une façon de faire est d'utiliser des différences finies centrées qui font intervenir des points "virtuels". Par développement de Taylor, on a

$$v(x_{N+1}) = \frac{v(x_{N+1} - h/2) + v(x_{N+1} + h/2)}{2} - \frac{h^2}{16}\left(v''(\theta_1) + v''(\theta_2)\right),$$

pour $\theta_1 \in (x_{N+1} - h/2, x_{N+1})$ et $\theta_2 \in (x_{N+1}, x_{N+1} + h/2)$. En prenant $v = au'$, on obtient

$$a(x_{N+1})u'(x_{N+1}) = g_1 = \frac{a(x_{N+1} - h/2)u'(x_{N+1} - h/2) + a(x_{N+1} + h/2)u'(x_{N+1} + h/2)}{2} + \mathcal{O}(h^2).$$

Or $u'(x_{N+1} \pm h/2) = D_h u(x_{N+1} \pm h/2) + \mathcal{O}(h^2)$. L'approximation considérée est alors la suivante (qui conserve l'ordre h^2).

$$2g_1 = a(x_{N+1} - h/2)D_h u_h(x_{N+1} - h/2) + a(x_{N+1} + h/2)D_h u_h(x_{N+1} + h/2)$$

soit

$$2g_1 = a_{N+1/2}\frac{u_{N+1}-u_N}{h} + a_{N+3/2}\frac{u_{N+2}-u_{N+1}}{h}. \tag{2.30}$$

L'équation aux différences finies $L_h u_h = f$, s'écrit au noeud x_{N+1} :

$$-\frac{1}{h^2}\left(a_{N+3/2}u_{N+2} - (a_{N+3/2}+a_{N+1/2})u_{N+1} + a_{N+1/2}u_N\right) = f_{N+1}.$$

On substitue alors la relation (2.30) donnant $a_{N+3/2}(u_{N+2}-u_{N+1})$ dans l'équation précédente pour obtenir

$$-\frac{1}{h^2}\left(2g_1 h - a_{N+1/2}(u_{N+1}-u_N) - a_{N+1/2}u_{N+1} + a_{N+1/2}u_N\right) = f_{N+1},$$

ce qui donne

$$-\frac{1}{h^2}\left(a_{N+1/2}u_N - a_{N+1/2}u_{N+1}\right) = \frac{f_{N+1}}{2} + \frac{g_1}{h}. \tag{2.31}$$

On peut à présent, donner le système linéaire $A_h\mathbf{u}_h = \mathbf{b}$ correspondant aux équations (2.29) et (2.31). La matrice A_h est de taille $(N+1)\times(N+1)$ et vaut

$$A_h = \frac{1}{h^2}\begin{pmatrix} \alpha_1 & \beta_1 & & & 0 \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{N-1} & \alpha_N & \beta_N \\ 0 & & & \beta_N & -\beta_N \end{pmatrix}, \tag{2.32}$$

avec $\alpha_i = a_{i+1/2} + a_{i-1/2}$ et $\beta_i = -a_{i+1/2}$. Le second membre est donné par

$$\mathbf{b} = \left(f_1, \cdots, f_N, \frac{f_{N+1}}{2} + \frac{g_1}{h}\right)^\top \in \mathbb{R}^{N+1}. \tag{2.33}$$

Remarque. On a introduit un point "fantôme" $x_{N+3/2} = x_{N+1} + h/2$ et le flux au' en ce point est calculé par extrapolation linéaire du flux aux points $x_{N+1/2}$ et x_{N+1}. Cette technique parfois appelée *mirror imaging* permet de conserver un ordre global en h^2, ce qui n'est pas le cas si on prend par exemple (ce qu'on aurait tendance à faire...) des différences finies non-centrées pour discrétiser les flux au bord. Ainsi l'approximation

$$a(x_{N+1})u'(x_{N+1}) = g_1 \simeq a(x_{N+1})D_h^- u(x_{N+1})$$

avec l'opérateur *décentré* $D_h^- v(x) = \dfrac{v(x)-v(x-h)}{h}$, ne fournit qu'un ordre de convergence en h car $v'(x) = D_h^- v(x) + \mathcal{O}(h)$.

2.3 Différences finies pour le cas 2D

Pour un ouvert borné Ω de $\mathbb{R}^2$, on considère le problème suivant : trouver une fonction $u = u(x,y)$ qui vérifie

$$(P)\ \begin{cases} Lu := -\mathrm{div}\,(A(\mathbf{x})\nabla u) + c\,u &= f \quad \text{dans } \Omega \\ u|_{\partial\Omega} &= 0 \quad \text{sur } \partial\Omega. \end{cases}$$

On suppose que la fonction $c = c(x,y)$ est positive ou nulle et que l'équation est elliptique c'est-à-dire qu'il existe $\alpha > 0$ tel que $(A(\mathbf{x})\xi,\ \xi)_{\mathbb{R}^n} \geq \alpha|\xi|^2, \quad \forall \xi \in \mathbb{R}^n,\ \forall \mathbf{x} \in \Omega$. On suppose enfin que (P) admet une unique solution $u \in C^2(\Omega)\cap C(\overline{\Omega})$.

2.3.1 Un schéma à 5 points pour le Laplacien

Commençons par le cas du Laplacien, c'est-à-dire que l'on prend

$$L = -\Delta + c\, I_d \tag{2.34}$$

et plaçons nous pour commencer dans un domaine rectangulaire en choisissant

$$\Omega = (0, a) \times (0, b). \tag{2.35}$$

Soient deux entiers N et M qui permettent de définir les paramètres de discrétisation $h_x = a/(N+1)$ et $h_y = b/(M+1)$. On notera $h = \max(h_x, h_y)$. Enfin, on introduit les points $P_{ij} = (x_i, y_j)$ de $\overline{\Omega}$ avec

$$x_i = i\, h_x, \quad i = 0, \cdots, N+1 \tag{2.36}$$

$$y_j = j\, h_y, \quad j = 0, \cdots, M+1. \tag{2.37}$$

Ces points appelés aussi *noeuds* définissent un *maillage* du domaine $\overline{\Omega}$. Plus précisement, on définit Ω_h le *maillage intérieur* à Ω par

$$\Omega_h = \{P_{ij} = (x_i, y_j),\ i = 1, \cdots, N,\ j = 1, \cdots, M\} \tag{2.38}$$

et $\overline{\Omega}_h$ le *maillage complet* de $\overline{\Omega}$ par

$$\overline{\Omega}_h = \{P_{ij} = (x_i, y_j),\ i = 0, \cdots, N+1,\ j = 0, \cdots, M+1\}. \tag{2.39}$$

On désignera par Γ_h les points du maillage complet $\overline{\Omega}_h$ qui sont sur le bord $\partial\Omega$, c'est-à-dire

$$\Gamma_h = \{P_{ij} = (x_i, y_j),\ \text{avec } i = 0 \text{ ou } N+1, \text{ou } j = 0 \text{ ou } M+1\}. \tag{2.40}$$

Comme dans le cas monodimensionnel, on désignera par V_h (resp. $\overline{V}_h$) l'ensemble des fonctions (*discrètes*) définies sur le maillage Ω_h (resp. $\overline{\Omega}_h$). On identifiera un élément v_h de V_h (resp. $\overline{V}_h$) à un vecteur $\mathbf{v}_h$ de $\mathbb{R}^{N\times M}$ (resp. $\mathbb{R}^{(N+2)\times(M+2)}$).

Le problème approché s'écrit alors : trouver $u_h \in \overline{V}_h$ telle que

$$(P_h) \left\{ \begin{array}{rcl} -\Delta_h u_h + c\, u_h &=& f \quad \text{dans } \Omega_h \\ u_h &=& 0 \quad \text{sur } \Gamma_h. \end{array} \right.$$

L'opérateur discret Δ_h est donné par un schéma à 5 points. Pour un point intérieur P du maillage Ω_h et une fonction $v \in \overline{V}_h$, on a

$$\Delta_h v(P) = \frac{1}{h_x^2}\left[v(W) - 2v(P) + v(E)\right] + \frac{1}{h_y^2}\left[v(S) - 2v(P) + v(N)\right], \tag{2.41}$$

où les points E, W, N, S sont les *voisins* de P, c'est-à-dire les 4 points du maillage $\overline{\Omega}_h$ les plus proches de P comme indiqué sur la figure 2.2.

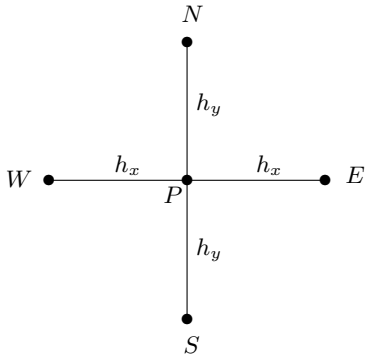


FIGURE 2.2 – Schéma à 5 points pour le Laplacien.

Le système linéaire correspondant au problème (P_h) s'écrit

$$A_h \mathbf{u}_h = \mathbf{b}, \tag{2.42}$$

où $\mathbf{u}_h$ est le vecteur de $\mathbb{R}^{N\times M}$ composé des $N \times M$ inconnues $\mathbf{u}_h = (u_h(P_{11}), \cdots, u_h(P_{NM}))^\top$. Le système linéaire dépend de la façon dont on numérote les noeuds du maillage. On choisit de numéroter les noeuds dans l'ordre **naturel** suivant : **de la gauche vers la droite et de bas en haut** (voir la figure 2.3).

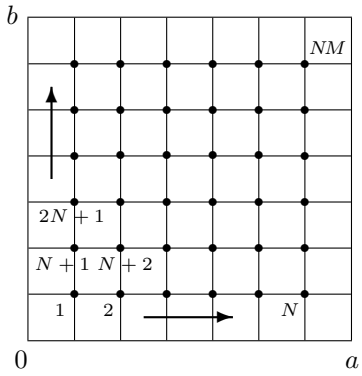


FIGURE 2.3 – Numérotation des noeuds dans l'ordre "naturel".

De cette façon, le vecteur $\mathbf{u}_h$ est donné par

$$\mathbf{u}_h = (u_{11}, u_{21}, \cdots, u_{N1},\ u_{12}, \cdots, u_{N2},\ \cdots,\ u_{1M}, \cdots, u_{NM})^\top, \tag{2.43}$$

où $u_{ij} = u_h(P_{ij})$. De même, le second membre du système linéaire (2.42) est donné par

$$\mathbf{b} = (f(P_{11}), \cdots, f(P_{NM}))^\top \in \mathbb{R}^{N\times M}. \tag{2.44}$$

La matrice A_h est une matrice de taille $NM \times NM$, tridiagonale par blocs. Donnons enfin la forme de la matrice A_h dans le cas (pour simplifier) d'un maillage uniforme où $h_x = h_y = h$. On a

$$A_h = \frac{1}{h^2}\begin{pmatrix} D_N & -I_N & & & 0 \\ -I_N & D_N & -I_N & & \\ & \ddots & \ddots & \ddots & \\ & & -I_N & D_N & -I_N \\ 0 & & & -I_N & D_N \end{pmatrix} + \operatorname{diag}\big(c(P_{11}), c(P_{21}), \cdots, c(P_{NM})\big), \tag{2.45}$$

où I_N désigne la matrice identité de taille $N \times N$ et D_N est la matrice carrée de taille $N \times N$ donnée par

$$D_N = \begin{pmatrix} 4 & -1 & & & 0 \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ 0 & & & -1 & 4 \end{pmatrix}. \tag{2.46}$$

Remarque. Si on considère de façon plus générale la condition aux limites de type Dirichlet non-homogène $u|_{\partial\Omega} = g$ sur le bord $\partial\Omega$, la matrice A_h du système linéaire (2.42) reste inchangée, en revanche le second membre devient

$$\mathbf{b} = (f(P_{11}), \cdots, f(P_{NM}))^\top + \frac{1}{h^2}\,(G(P_{11}), \cdots, G(P_{NM}))^\top, \tag{2.47}$$

avec $G(P) = \displaystyle\sum_{\substack{Q\in\mathcal{V}(P)\\ Q\in\partial\Omega}} g(Q)$ où $\mathcal{V}(P)$ est l'ensemble des noeuds du maillage $\overline{\Omega}_h$ qui sont les voisins de P([1]).

1. Par convention lorsque l'ensemble $\mathcal{V}(P) \cap \partial\Omega$ est vide, il n'y a pas de somme et le terme supplémentaire $G(P)$ est considéré comme nul.

Consistance du schéma à 5 points.

Pour une fonction v suffisamment régulière ($v \in C^4(\Omega)$), on obtient par développement de Taylor au point $P = (x_P, y_P) \in \Omega_h$,

$$\begin{aligned}\frac{v(W) - 2v(P) + v(E)}{h_x^2} &= \frac{\partial^2 v}{\partial x^2}(P) + \frac{h_x^2}{24}\left[\frac{\partial^4 v}{\partial x^4}(\theta_1, y_P) + \frac{\partial^4 v}{\partial x^4}(\theta_2, y_P)\right],\\ \frac{v(S) - 2v(P) + v(N)}{h_y^2} &= \frac{\partial^2 v}{\partial y^2}(P) + \frac{h_y^2}{24}\left[\frac{\partial^4 v}{\partial y^4}(x_P, \theta_3) + \frac{\partial^4 v}{\partial y^4}(x_P, \theta_4)\right],\end{aligned}$$

avec $x_P < \theta_1 < x_E$, $x_W < \theta_2 < x_P$ et $y_P < \theta_3 < y_N$, $y_S < \theta_4 < y_P$. On obtient ainsi,

$$\Delta_h v(P) = \Delta v(P) + R_h v(P) \text{ avec } \|R_h v\|_{h,\infty} \leq \frac{h^2}{6} M_4(v) \text{ où } M_4(v) = \max\left(\left\|\partial_x^4 v\right\|_\infty, \left\|\partial_y^4 v\right\|_\infty\right). \quad (2.48)$$

L'erreur de consistance est donc en $\mathcal{O}(h^2)$.

Stabilité.

On ne traite ici que le cas $c \equiv 0$. La matrice A_h est symétrique définie positive et c'est une M-matrice (cf. Proposition 2.4). On souhaite évaluer la norme matricielle $\|A_h^{-1}\|_\infty$. Si on note (b_{ij}) les coefficients de la matrice inverse A_h^{-1}, on a $\|A_h^{-1}\|_\infty = \max_i \sum_j |b_{ij}| = \max_i \sum_j b_{ij}$ car $A_h^{-1} \geq 0$. On obtient ainsi $\|A_h^{-1}\|_\infty = \max_i (A_h^{-1}\mathbf{e})_i = \|A_h^{-1}\mathbf{e}\|_\infty$ avec $\mathbf{e} = (1, \cdots, 1)^\top \in \mathbb{R}^{NM}$. Soit alors $\mathbf{w}_h$ la solution du système $A_h \mathbf{w}_h = \mathbf{e}$. On veut donc estimer $\|A_h^{-1}\|_\infty = \|\mathbf{w}_h\|_\infty$. L'idée est de trouver un vecteur $\varphi_h \geq 0$ tel que

$$A_h(\pm \mathbf{w}_h + \varphi_h) \geq 0. \quad (2.49)$$

La positivité de A_h^{-1} implique alors

$$\pm \mathbf{w}_h + \varphi_h \geq 0,$$

et puisque $\varphi_h \geq 0$, on obtient $|(\mathbf{w}_h)_i| \leq (\varphi_h)_i$ pour tout i, c'est-à-dire

$$\|\mathbf{w}_h\|_\infty \leq \|\varphi_h\|_\infty.$$

Si φ_h est uniformément borné par rapport à h alors on obtiendra le résultat de stabilité annoncé.

Introduisons la fonction $\psi(x, y) = \frac{1}{4}\left(x(a - x) + y(b - y)\right)$ qui vérifie

$$\begin{aligned}-\Delta\psi &= 1 \quad \text{dans } \Omega = (0, a) \times (0, b)\\ \psi &\geq 0 \quad \text{dans } \overline{\Omega}.\end{aligned}$$

On considère alors $\varphi_h = (\psi(P_{11}), \cdots, \psi(P_{NM}))$ pour $P_i \in \Omega_h$. On a

$$A_h\left(\pm \mathbf{w}_h + \varphi_h\right) = \pm \underbrace{A\mathbf{w}_h}_{\mathbf{e}} + A_h \varphi_h. \quad (2.50)$$

De plus,

$$(A_h \varphi_h)_i = -\Delta_h \psi(P_i) + \frac{1}{h^2} \sum_{\substack{Q \in \mathcal{V}(P_i)\\ Q \in \partial\Omega}} \psi(Q) \quad (2.51)$$

où $\mathcal{V}(P_i)$ désigne l'ensemble des voisins de P_i. D'après l'analyse de consistance du schéma à 5 points pour le Laplacien, on rappelle que

$$\Delta_h v(P) = \Delta v(P) + R_h v(P)$$

avec $\|R_h v\|_{h,\infty} = \max_{P\in\Omega_h} |R_h v(P)| \leq \dfrac{h^2}{6} M_4(v)$ où $M_4(v) = \max\left(\|\partial_x^4 v\|_\infty, \|\partial_y^4 v\|_\infty\right)$. On obtient donc

$$R_h \psi = 0,$$

puisque ψ est un polynôme du deuxième degré. Par conséquent, on a

$$\Delta\psi(P) = \Delta_h\psi(P) \quad \forall P \in \Omega_h,$$

et donc

$$-\Delta_h\psi(P) = 1 \quad \forall P \in \Omega_h.$$

La relation (2.51) devient donc

$$(A_h\varphi_h)_i = 1 + \frac{1}{h^2} \sum_{\substack{Q\in\mathcal{V}(P_i)\\ Q\in\partial\Omega}} \underbrace{\psi(Q)}_{\geq 0}. \tag{2.52}$$

En combinant (2.50) et (2.52), on obtient

$$\left(A_h\left(\pm\mathbf{w}_h + \varphi_h\right)\right)_i \geq (\pm 1) + 1 \geq 0.$$

L'objectif (2.49) est ainsi atteint. On a donc

$$\|A_h^{-1}\|_\infty = \|\mathbf{w}_h\|_\infty \leq \|\varphi_h\|_\infty.$$

Or $\|\varphi_h\|_\infty = \dfrac{1}{4}\left((a/2)^2 + (b/2)^2\right)$ et donc

$$\boxed{\|A_h^{-1}\|_\infty \leq \frac{1}{16}(a^2+b^2).} \tag{2.53}$$

Convergence.

La convergence est établie à partir de la consistance et de la stabilité, (avec $c \equiv 0$).

Théorème 2.2 *On suppose que $A \equiv I_d$ et $c \equiv 0$. On note u la solution de (P) supposée régulière ($u \in C^4(\Omega)$) et u_h la solution de (P_h). Il existe une constante $C > 0$ indépendante de h et u telle que*

$$\|u - u_h\|_{h,\infty} \leq C\, M_4(u)\, h^2, \tag{2.54}$$

où $M_4(u) = \max\left(\left\|\partial_x^4 u\right\|_\infty, \left\|\partial_y^4 u\right\|_\infty\right)$.

Démonstration. La démonstration est identique à celle du Théorème 2.1 dans le cas 1D. On obtient $\|u - u_h\|_{h,\infty} \leq \|A_h^{-1}\|_\infty \|R_h u\|_{h,\infty}$; en utilisant alors la consistance (cf. (2.48)) et la stabilité (cf. (2.53)), on obtient l'estimation voulue avec la constante $C = \left(a^2+b^2\right)/96$. □

2.3.2 Un autre schéma à 5 points

Pour certaines géométries, il peut être préférable d'utiliser un autre schéma à 5 points, "en croix". L'opérateur discret $\Delta_h^\times$ de ce schéma est donné par

$$\Delta_h^\times v(P) = \frac{1}{2h^2}\left[v(NE) + v(NW) + v(SW) + v(SE) - 4v(P)\right], \tag{2.55}$$

où P est un point intérieur du maillage Ω_h et les points NE, NW, SW, SE sont les 4 points du maillage $\overline{\Omega}_h$, les plus proches de P *dans les directions diagonales* (voir la figure 2.4).

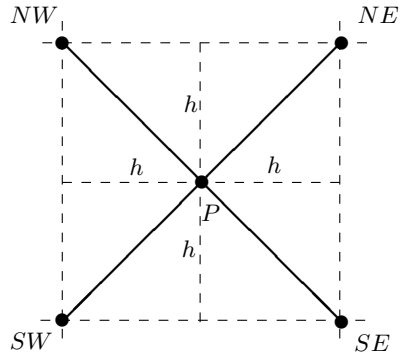


FIGURE 2.4 – Autre schéma à 5 points pour le Laplacien.

On montre alors que l'ordre de consistance de ce schéma est encore en $\mathcal{O}(h^2)$. Si $v \in C^4(\Omega)$, on a

$$\Delta_h^{\times} v(P) - \Delta v(P) = \frac{h^2}{12}\left[\frac{\partial^4 v}{\partial x^4}(Q_1) + 6\frac{\partial^4 v}{\partial x^2 \partial y^2}(Q_2) + \frac{\partial^4 v}{\partial y^4}(Q_3)\right], \tag{2.56}$$

où $|Q_i - P| \leq h$.

2.3.3 Un schéma à 9 points pour fonction harmonique

A partir des deux schémas à 5 points précédents, on peut construire un schéma à 9 points (cf. Fig. 2.5), très précis pour les fonction harmoniques. On choisit un maillage uniforme ($h_x = h_y = h$). L'opérateur discret $\Delta_h^{(9)}$ de ce schéma est donné par

$$\Delta_h^{(9)} = \frac{2}{3}\Delta_h + \frac{1}{3}\Delta_h^{\times}. \tag{2.57}$$

De façon explicite, pour un point intérieur P du maillage Ω_h, on a

$$\Delta_h^{(9)} v(P) = \frac{1}{6h^2}\left[v(NE) + v(NW) + v(SW) + v(SE) + 4\big(v(E) + v(N) + v(W) + v(S)\big) - 20v(P)\right].$$

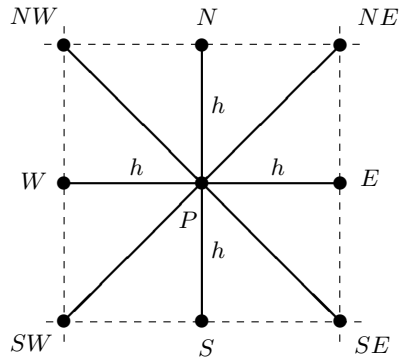


FIGURE 2.5 – Un schéma à 9 points pour les fonctions harmoniques.

Si v est une fonction suffisamment régulière, on montre que

$$\Delta_h^{(9)} v(P) - \Delta v(P) = \frac{h^2}{2}\Delta^2 v(P) + \frac{2}{6!}h^4\left[\Delta^3 v(P) + 2\frac{\partial^4(\Delta v)}{\partial x^2 \partial y^2}(P)\right] + \mathcal{O}(h^6).$$

Si u vérifie $-\Delta u = f$ avec f harmonique i.e $\Delta f = 0$, alors $\Delta^2 u = 0$ et on obtient

$$\Delta_h^{(9)} u(P) - \Delta u(P) = \mathcal{O}(h^4).$$

Si maintenant $f \equiv 0$ c'est-à-dire si u est harmonique ($\Delta u = 0$) alors on voit que

$$\Delta_h^{(9)} u(P) - \Delta u(P) = \mathcal{O}(h^6),$$

ce qui rend ce schéma à 9 points très précis pour les fonctions harmoniques.

2.3.4 Cas d'un domaine non-rectangulaire

On va voir qu'il faut être prudent pour discrétiser l'opérateur près du bord d'un domaine non-rectangulaire. Nous décrivons une première méthode qui semble naturelle, mais qui conduit malheureusement à une perte de l'ordre de consistance. Nous verrons ensuite une autre méthode par interpolation, qui fournit quant à elle un meilleur ordre de consistance. Pour fixer les idées, on considère le cas du laplacien ($A \equiv I_d$) et on suppose que $c \equiv 0$. On choisit un pas de discrétisation uniforme avec $h = h_x = h_y$ et on note

$$\mathcal{S}_h = \{(ih, jh),\ i, j \in \mathbb{Z}\},$$

un maillage uniforme de $\mathbb{R}^2$ tout entier. On définit alors le *maillage intérieur*

$$\Omega_h = \mathcal{S}_h \cap \Omega \tag{2.58}$$

et on note

$$\overline{\Omega}_h = \mathcal{S}_h \cap \overline{\Omega}. \tag{2.59}$$

Ces définitions coïncident avec celles données dans le cadre d'un domaine rectangulaire. Enfin, on désigne par *voisins* d'un point P du maillage $\mathcal{S}_h$, les points du maillage $\mathcal{S}_h$ ayant même abscisse ou même ordonnée que P et situés à une distance h de P.

Première (mauvaise) méthode.

Pour un point P du maillage Ω_h qui n'est pas près du bord (c'est-à-dire dont les 4 voisins W, N, S, E situés à une distance h de P sont dans $\overline{\Omega}_h$), on utilise le schéma à 5 points standard (cf. section 2.3.1).

Pour un point P du maillage Ω_h qui est près du bord (c'est-à-dire dont au moins un voisin n'est pas dans $\overline{\Omega}_h$), on procède de la façon suivante. Supposons pour fixer les idées, que W, N, S soient dans $\overline{\Omega}_h$ mais que $E \notin \overline{\Omega}_h$ (voir la figure 2.6). On considère alors le point A d'intersection du bord $\partial\Omega$ avec le segment PE et on note θh la distance de A à P ($0 < \theta < 1$).

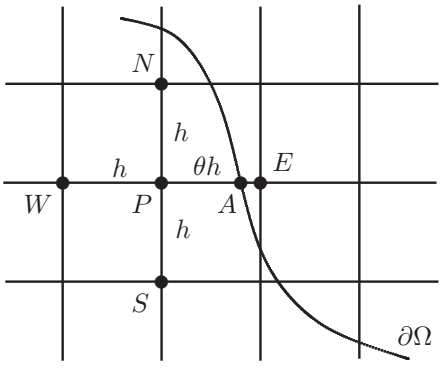


FIGURE 2.6 – Cas d'un domaine non-rectangulaire.

Par développement de Taylor, on a

$$u(A) = u(P) + \theta h \frac{\partial u}{\partial x}(P) + \frac{(\theta h)^2}{2}\frac{\partial^2 u}{\partial x^2}(P) + \frac{(\theta h)^3}{6}\frac{\partial^3 u}{\partial x^3}(P) + \mathcal{O}(h^4), \tag{2.60}$$

$$u(W) = u(P) - h\frac{\partial u}{\partial x}(P) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(P) - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(P) + \mathcal{O}(h^4). \tag{2.61}$$

En éliminant $(\partial u/\partial x)(P)$ dans les deux équations précédentes, on obtient

$$\frac{\partial^2 u}{\partial x^2}(P) = \frac{2}{h^2}\left(\frac{u(A)}{\theta(1+\theta)} + \frac{u(W)}{(1+\theta)} - \frac{u(P)}{\theta}\right) + (1-\theta)\frac{h}{3}\frac{\partial^3 u}{\partial x^3}(P) + \mathcal{O}(h^2). \tag{2.62}$$

Si on considère l'approximation

$$\frac{\partial^2 u}{\partial x^2}(P) \simeq \frac{2}{h^2}\left(\frac{u(A)}{\theta(1+\theta)} + \frac{u(W)}{(1+\theta)} - \frac{u(P)}{\theta}\right),$$

pour un point P près du bord, on voit qu'on obtient un ordre de consistance en $\mathcal{O}(h)$, là où on avait un ordre $\mathcal{O}(h^2)$ avec le schéma standard à 5 points (pour un point P loin du bord). Cette méthode est donc à éviter.

Deuxième (bonne) méthode.

On utilise les formules standards (schéma à 5 points de la section 2.3.1) pour les points P entourés par des points intérieurs. Pour les autres points proches du bord, on procède par interpolation linéaire. Reprenons la situation décrite à la figure 2.6. On détermine une valeur $\tilde{u}(P)$ par interpolation linéaire de u entre W et A, selon la direction WE (voir la figure 2.7).

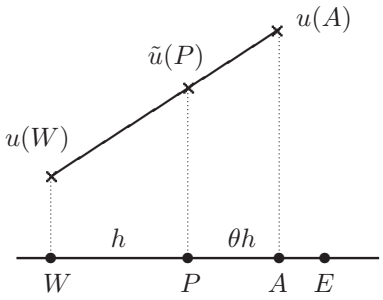


FIGURE 2.7 – Interpolation linéaire en un point P près du bord.

Plus précisement, on a $\dfrac{u(A)-u(W)}{(1+\theta)h} = \dfrac{\tilde{u}(P)-u(W)}{h}$, ce qui donne

$$\tilde{u}(P) = \frac{u(A)+\theta u(W)}{1+\theta}. \tag{2.63}$$

Par ailleurs, en utilisant les développements de Taylor (2.60) et (2.61), on obtient

$$u(P) = \tilde{u}(P) - \theta\frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(P) + \mathcal{O}(h^3). \tag{2.64}$$

Si on considère l'approximation $u(P) \simeq \tilde{u}(P) = \dfrac{u(A)+\theta u(W)}{1+\theta}$ pour un point P près du bord, on voit qu'on obtient un ordre de consistance en $\mathcal{O}(h^2)$. On conserve ainsi l'ordre du schéma standard à 5 points.

Etude de la convergence de la deuxième méthode.

On rappelle que par *voisins* d'un point P du maillage, on désigne les points du maillage $\mathcal{S}_h$ ayant même abscisse ou même ordonnée que P et situés à une distance h de P. L'ensemble des voisins de P est noté $\mathcal{V}(P)$.

Soient Ω_h^{int} l'ensemble des points du maillage dont les 4 voisins sont dans $\overline{\Omega}_h$ et Ω_h^b l'ensemble des points du maillage $\overline{\Omega}_h$ dont au moins un voisin n'appartient pas à $\overline{\Omega}_h$ (voir la figure 2.8).

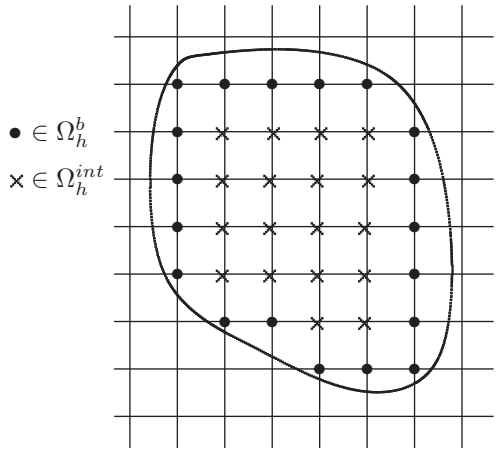


FIGURE 2.8 – La distinction des noeuds du maillage.

Pour un point $P \in \Omega_h^{int}$, on utilise le schéma standard à 5 points, c'est-à-dire

$$-\Delta_h u(P) := -\frac{1}{h^2}\Big(u(W) + u(E) + u(N) + u(S) - 4u(P)\Big) = f(P)$$

Pour un point $P \in \Omega_h^b$, il existe au moins un point $P' \in \partial\Omega$ ayant même abscisse ou même ordonnée que P et situé à une distance $h' = \theta h$ de P avec $0 < \theta < 1$. On note alors P'' le point de $\overline{\Omega}$ ayant même abscisse ou même ordonnée que P', situé à une distance $h'' \leq h$ de P mais de l'autre côté de P. Dans l'exemple de la figure 2.9 suivante, on a $P' = A$ et $P'' = W$ avec $h'' = h$.

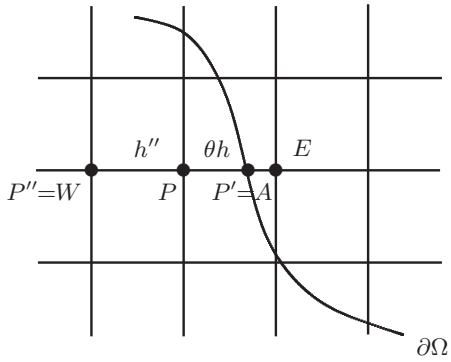


FIGURE 2.9 – Exemple d'interpolation dans le cas d'un domaine non-rectangulaire.

Ainsi, en utilisant la procédure d'interpolation de la deuxième méthode, on obtient

$$u(P) - \frac{h''}{\theta h + h''}\, u(P') - \frac{\theta h}{\theta h + h''}\, u(P'') = 0 \tag{2.65}$$

et compte tenu de la condition limite du problème (P) qui donne $u(P') = 0$ (ou en tout cas, une valeur connue si on impose une condition de Dirichlet non-homogène), on obtient

$$u(P) - \frac{\theta h}{\theta h + h''} u(P'') = 0. \tag{2.66}$$

En résumé, on a l'approximation suivante, pour un point P quelconque du maillage

$$L_h u(P) = \begin{cases} -\Delta_h u(P), & \text{si } P \in \Omega_h^{int} \\ u(P) - \dfrac{\theta h}{\theta h + h''}\, u(P''), & \text{si } P \in \Omega_h^b \end{cases}$$

et le problème approché s'écrit : trouver $u_h \in V_h$ telle que

$$L_h u_h(P) = \tilde{f}(P) \text{ pour } P \in \Omega_h, \text{ avec } \tilde{f}(P) = \begin{cases} f(P) & \text{si } P \in \Omega_h^{int} \\ 0 & \text{si } P \in \Omega_h^b. \end{cases} \tag{2.67}$$

Système linéaire.

Soit n le nombre de noeuds du maillage Ω_h. On numérote les noeuds de gauche à droite et de bas en haut (cf. Section 2.3.1). L'utilisation du schéma à 5 points pour les points intérieurs et de la discrétisation ci-dessus pour la condition limite, conduit au système linéaire $A_h \mathbf{u}_h = \mathbf{b}_h$. La matrice A_h de taille $n \times n$, est donnée (avec $c \equiv 0$) par

$$A_h = \frac{1}{h^2} \begin{array}{c} \begin{matrix} S & & W & P & E & & N \end{matrix} \\ \left(\begin{array}{ccccccc} \cdot & -1 & \cdots & -1 & 4 & -1 & \cdots & -1 & \cdots \\ \cdots & \cdots & \cdots & 1 & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & -\frac{h'}{h'+h''} & 1 & \cdots & \cdots & \cdots & \cdots \end{array} \right) \\ \begin{matrix} & & & P'' & P & & & & \end{matrix} \end{array} \begin{array}{l} P \in \Omega_h^{int} \quad (S, W, E, N \notin \partial\Omega) \\ P \in \Omega_h^b, \ P'' \in \partial\Omega \\ P \in \Omega_h^b, \ P'' \notin \partial\Omega, \ P'' \in \Omega_h \end{array} \tag{2.68}$$

et le second membre $\mathbf{b}_h = (b_h(P_1), \cdots, b_h(P_n))^\top \in \mathbb{R}^n$. Dans le cas d'une condition de Dirichlet générale $u|_{\partial\Omega} = g$, les composantes de $\mathbf{b}_h$ sont données par

$$b_h(P) = \begin{cases} f(P) + \dfrac{1}{h^2} \displaystyle\sum_{\substack{Q \in \mathcal{V}(P) \\ Q \in \partial\Omega}} g(Q), & \text{si } P \in \Omega_h^{int} \\ \displaystyle\sum_{P' \in \partial\Omega} \frac{h''}{h'+h''}\, g(P') + \sum_{P'' \in \partial\Omega} \frac{h'}{h'+h''}\, g(P''), & \text{si } P \in \Omega_h^b. \end{cases} \tag{2.69}$$

Consistance.

• En étudiant la consistance du schéma à 5 points, on a obtenu (cf. (2.48)) pour $u \in C^4(\Omega)$ et $P \in \Omega_h^{int}$,

$$|R_h^1 u(P)| = |L_h u(P) - Lu(P)| \leq \frac{M_4(u)}{6} h^2, \tag{2.70}$$

où $M_4 = \max\left(\|\partial_x^4 u\|_\infty, \ \|\partial_y^4 u\|_\infty \right)$

• Considérons à présent le cas d'un point $P \in \Omega_h^b$. On note $R_h^2 u(P) = u(P) - \dfrac{\theta h}{\theta h + h''} u(P'')$. Par développement de Taylor, on obtient (avec $h' = \theta h$ et en supposant que P' et P ont même ordonnée)

$$\begin{aligned} 0 = u(P') &= u(P) + h' \frac{\partial u}{\partial x}(P) + \frac{h'^2}{2} \frac{\partial^2 u}{\partial x^2}(P) + \frac{h'^3}{6} \frac{\partial^3 u}{\partial x^3}(Q_1) \\ u(P'') &= u(P) - h'' \frac{\partial u}{\partial x}(P) + \frac{h''^2}{2} \frac{\partial^2 u}{\partial x^2}(P) - \frac{h''^3}{6} \frac{\partial^3 u}{\partial x^3}(Q_2), \end{aligned}$$

ce qui permet d'établir que

$$R_h^2 u(P) = -\frac{h'h''}{2} \frac{\partial^2 u}{\partial x^2}(P) - \frac{h'h''}{6(h'+h'')} \left(h'^2 \partial_x^3 u(Q_1) - h''^2 \partial_x^3 u(Q_2) \right). \tag{2.71}$$

Ainsi, on obtient l'estimation

$$|R_h^2 u(P)| \leq \frac{M_2}{2} h^2 \quad \text{pour } P \in \Omega_h^b, \tag{2.72}$$

avec $M_2 = \max\left(\|\partial_x^2 u\|_\infty,\ \|\partial_y^2 u\|_\infty\right)$.

Convergence.

On va établir le résultat suivant.

Proposition 2.8 *Soient $\Omega \subset \mathbb{R}^2$ un domaine borné, régulier et R le rayon du plus petit cercle contenant Ω. On note u la solution de (P) (avec $A \equiv I_d$ et $c \equiv 0$), supposée régulière ($u \in C^4(\Omega)$) et u_h la solution du problème approché (P_h) correspondant. On a alors*

$$\|u - u_h\|_{h,\infty} \leq \left(\frac{M_4(u)}{24} R^2 + M_2(u)\right) h^2, \tag{2.73}$$

où $M_k(u) = \max\left(\|\partial_x^k u\|_\infty,\ \|\partial_y^k u\|_\infty\right)$ pour $k = 2, 4$.

Démonstration. Remarquons tout d'abord que la matrice A_h est une M-matrice. En effet, pour les lignes i qui correspondent à des points $P \in \Omega_h^b$, on vérifie facilement que

- $\displaystyle\sum_j a_{ij} = \begin{cases} -\dfrac{h'}{h'+h''} + 1 = \dfrac{h''}{h'+h''} > 0, & \text{si } P'' \notin \partial\Omega \\ 1, & \text{si } P'' \in \partial\Omega \end{cases}$
- $a_{ij} \leq 0$ pour tout $j \neq i$.

Il suffit alors d'appliquer la Proposition 2.3.

On note l'erreur $w_h = u_h - u \in \overline{V}_h$ et $\mathbf{w}_h = \mathbf{u}_h - \mathbf{u}$ le vecteur de $\mathbb{R}^n$ correspondant. L'idée est de trouver un vecteur φ_h positif tel que $A_h(\pm\mathbf{w}_h + \varphi_h) \geq 0$ et $\varphi_h \to 0$ quand $h \to 0$. Par la monotonie de la matrice A_h, on en déduira que $\pm\mathbf{w}_h + \varphi_h \geq 0$ et donc que $|w_h(P_i)| \leq \varphi_h(P_i)$ pour tout $P_i \in \Omega_h$, soit encore $\|w_h\|_{h,\infty} \leq \|\varphi_h\|_\infty$. Une estimation de φ_h en fonction de h permettra alors de conclure à l'estimation (2.73).

Considérons le cercle $\mathcal{C}$ de centre (x_0, y_0) et de rayon R, circonscrit au domaine Ω et soit la fonction ϕ définie dans $\mathcal{C}$ par

$$\phi(x, y) = \frac{1}{4}\left(R^2 - (x - x_0)^2 - (y - y_0)^2\right). \tag{2.74}$$

On a $\phi \geq 0$ dans $\overline{\Omega}$. On note alors $\mathbf{\Phi} = (\phi(P_1), \cdots, \phi(P_n))^\top$, $\mathbf{e} = (1, \cdots, 1)^\top \in \mathbb{R}^n$ et on définit

$$\varphi_h = \frac{M_4(u) h^2}{6} \mathbf{\Phi} + M_2(u) h^2 \mathbf{e}, \tag{2.75}$$

avec $M_k(u) = \max\left(\|\partial_x^k u\|_\infty,\ \|\partial_y^k u\|_\infty\right)$ pour $k = 2, 4$. Puisque $A_h \mathbf{u}_h = \mathbf{b}_h$, on a

$$A_h(\pm\mathbf{w}_h + \varphi_h) = \pm(\mathbf{b}_h - A_h\mathbf{u}) + \frac{M_4(u) h^2}{6} A_h \mathbf{\Phi} + M_2(u) h^2 A_h \mathbf{e}. \tag{2.76}$$

Etudions maintenant le signe de chacune des composantes du vecteur précédent. Chaque composante est liée à un point P différent. Il faut distinguer les cas selon que P est un point intérieur ou un point du bord.

- Soit $P \in \Omega_h^{int}$.

On a $\mathbf{b}_h(P) = b_h(P) = f(P)$ et $A_h\mathbf{u}(P) = -\Delta_h u(P) + \dfrac{1}{h^2} \displaystyle\sum_{\substack{Q \in \mathcal{V}(P) \\ Q \in \partial\Omega}} \overbrace{u(Q)}^{=0} = \underbrace{-\Delta u(P)}_{=f(P)} + R_h^1 u(P)$. On obtient ainsi

$$A_h\mathbf{u}(P) - b_h(P) = R_h^1 u(P) \quad \text{avec } |R_h^1 u(P)| \leq \frac{M_4(u)}{6} h^2. \tag{2.77}$$

Par ailleurs, $A_h\mathbf{\Phi}(P) = -\Delta_h\phi(P) + \dfrac{1}{h^2}\sum_{\substack{Q\in\mathcal{V}(P)\\ Q\in\partial\Omega}} \overbrace{\phi(Q)}^{\geq 0} \geq -\Delta\phi(P) + R_h^1\phi(P)$. Or, on a $-\Delta\phi \equiv 1$ et $R_h^1\phi(P) \equiv 0$ car $|R_h^1\phi(P)| \leq M_4(\phi)h^2/6$ avec $M_4(\phi) = \max\left(\|\partial_x^4\phi\|_\infty, \|\partial_y^4\phi\|_\infty\right)$ et il est clair que $M_4(\phi) = 0$ puisque ϕ est un polynôme de degré 2. Par conséquent, on a

$$A_h\mathbf{\Phi}(P) \geq 1. \tag{2.78}$$

Enfin,

$$A_h\mathbf{e}(P) = \sum_j a_{ij} \geq 0. \tag{2.79}$$

Ainsi d'après (2.76)-(2.79), on obtient pour tout $P \in \Omega_h^{int}$,

$$A_h\left(\pm\mathbf{w}_h + \varphi_h\right)(P) \geq -|R_h^1 u(P)| + \frac{M_4(u)}{6}h^2 \geq 0. \tag{2.80}$$

• Soit $P \in \Omega_h^b$.

On a $\mathbf{b}_h(P) = 0$ (condition de Dirichlet homogène) et $A_h\mathbf{u}(P) = R_h^2 u(P)$ avec $|R_h^2 u(P)| \leq \dfrac{M_2(u)}{2}h^2$ d'après (2.72).
Par ailleurs, $A_h\mathbf{\Phi}(P) = R_h^2\phi(P)$. Or, on a

$$\partial_x^2\phi(P) = -1/2 \quad \text{et } \partial_x^3\phi(P) = \partial_y^3\phi(P) = 0,$$

donc $R_h^2\phi(P) = \dfrac{h'h''}{4} > 0$ d'après (2.71) et par conséquent

$$A_h\mathbf{\Phi}(P) > 0 \quad \text{ pour tout } P \in \Omega_h^b.$$

Enfin, on vérifie que $A_h\mathbf{e}(P) = \begin{cases} 1 - \dfrac{h'}{h'+h''} & \text{si} \quad P'' \notin \partial\Omega \\ 1 & \text{si} \quad P'' \in \partial\Omega \end{cases}$
Or, on a $1 - h'/(h'+h'') = h''/(h'+h'') > 1/2$ si $P'' \notin \partial\Omega$ et donc

$$A_h\mathbf{e}(P) > 1/2 \quad \text{pour tout } P \in \Omega_h^b.$$

Ainsi d'après (2.76), on obtient pour tout $P \in \Omega_h^b$,

$$A_h\left(\pm\mathbf{w}_h + \varphi_h\right)(P) \geq -|R_h^2 u(P)| + \frac{M_2(u)}{2}h^2 \geq 0. \tag{2.81}$$

On a ainsi montrer que $A_h\left(\pm\mathbf{w}_h + \varphi_h\right) \geq 0$ et par conséquent $\|w_h\|_\infty \leq \|\varphi_h\|_\infty$. On vérifie par ailleurs que $\|\mathbf{\Phi}\|_\infty = R^2/4$, ce qui donne l'estimation cherchée en utilisant (2.75). □

2.3.5 Opérateur sous forme de divergence

On considère à présent le cas d'un opérateur général (en 2D) sous la forme

$$Lu := -\mathrm{div}\left(A(\mathbf{x})\nabla u\right). \tag{2.82}$$

On suppose que le domaine est rectangulaire et les notations sont celles de la section 2.3.1. La discrétisation de L se fait en introduisant les opérateurs aux différences (centrées) suivants. Pour $\mathbf{v} = (v_1, v_2)$, on définit

$$\mathrm{div}_h\ \mathbf{v} = D_{h,x}v_1 + D_{h,y}v_2 \tag{2.83}$$

avec

$$\begin{aligned} D_{h,x}v &= \frac{1}{h_x}\left[v(x+h_x/2,y)-v(x-h_x/2,y)\right] \\ D_{h,y}v &= \frac{1}{h_y}\left[v(x,y+h_y/2)-v(x,y-h_y/2)\right] \end{aligned}$$

et

$$\nabla_h v = \left(D_{h,x}v, D_{h,y}v\right)^\top . \tag{2.84}$$

L'opérateur aux différences L_h associé à L est alors défini par

$$L_h u := -\mathrm{div}_h\left(A\nabla_h u\right). \tag{2.85}$$

Si on note $(a_{ij})_{1\le i,j\le 2}$ les coefficients de la matrice A, l'opérateur L_h s'écrit alors en un point P du maillage Ω_h

$$L_h u(P) = D_{h,x}\big(a_{11}D_{h,x}u + a_{12}D_{h,y}u\big) + D_{h,y}\big(a_{21}D_{h,x}u + a_{22}D_{h,y}u\big) \tag{2.86}$$

et on vérifie sans peine que

$$L_h u(P) = \sum_{Q\in\{P,S,W,E,N\}} \alpha(h_x,h_y,Q)\ u(Q) \tag{2.87}$$

où les coefficients α dépendent des valeurs des a_{ij} aux points milieux des segments $[W,P]$, $[P,E]$, $[S,P]$ et $[P,N]$. En revanche, l'expression (2.87) montre que L_h ne fait intervenir que les valeurs de u aux noeuds du maillage $\overline{\Omega}_h$.

2.3.6 Autres conditions limites

• **Condition de Dirichlet non-homogène.** On impose $u_{|_{\partial\Omega}} = g$. Par rapport aux conditions de Dirichlet nulles ($u_{|_{\partial\Omega}} = 0$), le nombre d'inconnues reste le même et la matrice A_h est inchangée. Seul le second membre est modifié (cf. (2.47) pour un domaine rectangle ou (2.69) pour un domaine non-rectangle).

• **Condition de Neumann pour le Laplacien sur un carré.** On impose $\dfrac{\partial u}{\partial \mathbf{n}} := \nabla u\cdot\mathbf{n} = g$ sur $\partial\Omega$ où $\mathbf{n}$ est la normale extérieure unitaire au bord $\partial\Omega$.

On procède comme dans le cas 1D, en utilisant une technique de *mirror imaging*. Soit P_0 un noeud du maillage appartenant au bord $\partial\Omega$. On a

$$g(P_0) = \frac{\partial u}{\partial \mathbf{n}}(P_0) = \frac{u(P_1)-u(P_3)}{2h} + \mathcal{O}(h^2), \tag{2.88}$$

où P_1 et P_3 sont les deux voisins de P_0 de part et d'autre du bord $\partial\Omega$, avec $P_1\notin\Omega$ et $P_3\in\Omega$ (cf. Fig. 2.10). D'après (2.88), on écrit pour l'approximation u_h :

$$u_h(P_1) = 2hg(P_0) + u_h(P_3) \tag{2.89}$$

et on substitue l'expression précédente de $u_h(P_1)$ dans l'expression de $\Delta_h u_h(P_0)$. Pour le schéma à 5 points, cela donne

$$\begin{aligned} -\Delta_h u_h(P_0) &= \frac{1}{h^2}\left[4u_h(P_0)-u_h(P_1)-u_h(P_2)-u_h(P_3)-u_h(P_4)\right] \\ &= \frac{1}{h^2}\left[4u_h(P_0)-(2hg(P_0)+u_h(P_3))-u_h(P_2)-u_h(P_3)-u_h(P_4)\right] \end{aligned}$$

et le schéma approché en P_0 s'écrit

$$L_h u_h(P_0) := \frac{1}{h^2}\left[4u_h(P_0)-u_h(P_2)-2u_h(P_3)-u_h(P_4)\right] = \frac{2}{h}g(P_0)+f(P_0). \tag{2.90}$$

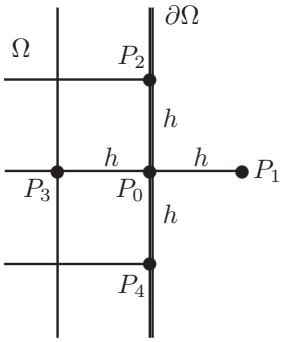


FIGURE 2.10 – Condition de Neumann pour un carré.

D'après (2.88), on voit qu'on conserve de cette façon un ordre de consistance en $\mathcal{O}(h^2)$.

Remarques. Si on écrit l'approximation $u_h(P_0) = u_h(P_3) + hg(P_0)$ qui vient de l'utilisation de différences finies *décentrées* : $g(P_0) = \dfrac{\partial u}{\partial \mathbf{n}}(P_0) = \dfrac{u(P_0) - u(P_3)}{h} + \mathcal{O}(h)$, on obtient seulement un ordre de consistance en $\mathcal{O}(h)$ sur le bord...
Pour le schéma à 9 points, on obtient un ordre en $\mathcal{O}(h^4)$ en tout point intérieur et en $\mathcal{O}(h^2)$ en tout point du bord avec le schéma (2.90).

• **Condition de Neumann pour le Laplacien sur un domaine non-rectangulaire.** Soit un noeud P du maillage $\overline{\Omega}_h$ tel qu'au moins un de ses voisins parmi P_1, P_2, P_3, P_4, ne soit pas dans $\overline{\Omega}$. Pour fixer les idées, on suppose qu'on a la situation décrite à la figure 2.11 c'est-à-dire que $P_1 \notin \overline{\Omega}$ et $P_2 \in \Omega$.

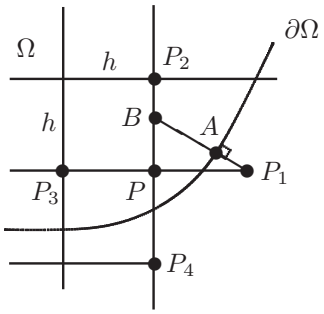


FIGURE 2.11 – Condition de Neumann pour un domaine non-rectangulaire.

On écrit en P, le schéma standard à 5 points.

$$-\Delta_h u(P) = -\frac{1}{h^2}\left[4u(P) - u(P_1) - u(P_2) - u(P_3) - u(P_4)\right] = f(P). \tag{2.91}$$

On élimine alors la valeur $u(P_1)$ de la façon suivante. Soient A le point de $\partial\Omega$, projeté orthogonal de P_1 sur le bord $\partial\Omega$ et B l'intersection des segments $[PP_2]$ et $[AP_1]$. En écrivant la condition de Neumann en A, on obtient ($\mathbf{n}$ est colinéaire à $\overrightarrow{AP_1}$) :

$$g(A) = \frac{\partial u}{\partial \mathbf{n}}(A) \simeq \frac{u(P_1) - u(B)}{P_1 B}. \tag{2.92}$$

Puis, on détermine une valeur approchée de $u(B)$ par interpolation linéaire de u entre P et P_2. Ceci fournit avec (2.92), une expression (une approximation) de $u(P_1)$ en fonction de $u(P)$ et $u(P_2)$. On substitue alors cette expression de $u(P_1)$, dans l'équation (2.91) pour obtenir l'expression de l'opérateur aux différences $L_h u$ au noeud P (dans l'exemple de la figure 2.11 on élimine aussi de la même façon $u(P_4)$). Cette méthode est peu précise (et c'est la limite des D.F...).

2.4 Evaluation pratique de l'ordre de convergence d'une méthode

Pour déterminer de façon pratique l'ordre de convergence p d'une méthode, on calcule l'erreur $u - u_h$ en fonction de différents pas de discrétisation h. Pour calculer l'erreur $u - u_h$, on se donne généralement une fonction u avec laquelle on détermine le second membre $f := Lu$, puis on résout le problème approché $L_h u_h = f$ pour obtenir u_h.

Si on représente l'erreur en fonction du pas h, on obtient en échelle logarithmique, des points répartis à peu près le long d'une droite (droite de régression, par exemple) et dont la pente fournit l'ordre p. Par exemple, supposons que $erreur := \|u - u_h\| \simeq \alpha h^p$ pour une certaine norme $\|\cdot\|$. On a

$$\log(erreur) \simeq p \log(h) + \log(\alpha),$$

ce qui correspond bien à une droite de pente p, en échelle logarithmique.

2.5 Méthode de Richardson

C'est une technique (générale) qui permet d'améliorer l'ordre de convergence d'une méthode, à partir de solutions calculées sur des maillages différents.

Supposons que l'on dispose d'une méthode d'ordre $\mathcal{O}(h^p)$ pour calculer une approximation d'une solution exacte u. On considère alors les solutions approchées u_{h_1} et u_{h_2} obtenues par cette méthode avec des maillages de pas uniformes h_1 et h_2 respectivement. On suppose que l'on a

$$\begin{aligned} u - u_{h_1} &= \alpha h_1^p + \mathcal{O}(h_1^q) \quad \text{avec } q > p \\ u - u_{h_2} &= \alpha h_2^p + \mathcal{O}(h_2^q). \end{aligned}$$

On multiplie la première équation par h_2^p et la seconde par h_1^p, puis on forme la différence. On obtient ainsi

$$(h_2^p - h_1^p)u - h_2^p u_{h_1} + h_1^p u_{h_2} = \mathcal{O}(h_1^q h_2^p) + \mathcal{O}(h_1^p h_2^q)$$

et par conséquent

$$u = \frac{h_2^p u_{h_1} - h_1^p u_{h_2}}{h_2^p - h_1^p} + \mathcal{O}\left(\frac{h_1^q h_2^p + h_1^p h_2^q}{h_2^p - h_1^p}\right).$$

Si on prend

$$h_1 = 2h_2,$$

alors on a

$$u = \frac{u_{h_1} - 2^p u_{h_2}}{1 - 2^p} + \mathcal{O}(h_2^q).$$

Ainsi, en calculant

$$u_h = \frac{u_{h_1} - 2^p u_{h_2}}{1 - 2^p}, \tag{2.93}$$

on obtient une approximation de u en $\mathcal{O}(h_2^q)$ au lieu de $\mathcal{O}(h_2^p)$ (avec $q > p$).

Coût de calcul pour Richardson. Supposons pour fixer les idées, que l'on résolve un laplacien sur le carré unité et que l'on résout le système linéaire avec une méthode de Gauss (sans stratégie de pivot). Le nombre d'opérations pour la résolution du système linéaire est $\sim$ (ordre de la matrice) $\times$ (largeur de bande)^2. Dans le cas de la matrice tridiagonale bloc (2.45), la complexité est en $\mathcal{O}(N^4)$ (en $\mathcal{O}(N^6)$ dans le cas d'une matrice pleine) où $h = 1/(N+1)$ (N est le nombre de noeuds dans une direction horizontale ou verticale). Les calculs de u_{h_1} et u_{h_2} avec les pas de discrétisation respectifs h_1 et $h_2 = h_1/2$, nécessitent $\mathcal{O}(N^4)$ opérations (avec $h_1 = 1/(N+1)$). Par conséquent le calcul de u_h par (2.93) requiert $\mathcal{O}(N^4)$ opérations pour une précision en $\mathcal{O}(h_2^q)$.

Pour calculer directement une solution $u_{\tilde{h}}$ avec la même précision, il faut prendre un pas de discrétisation $\tilde{h}$ tel que $\tilde{h}^p = h_2^q$, soit $\tilde{h} = h_2^{q/p}$ ce qui correspond à $\tilde{N} = N^{q/p}$. La complexité pour le calcul de $u_{\tilde{h}}$ est donc en $\mathcal{O}(\tilde{N}^4) = \mathcal{O}(N^{4q/p})$ avec $q/p > 1$. Par exemple, avec $p = 2$ et $q = 3$, le calcul de u_h nécessite $\mathcal{O}(N^4)$ opérations, alors qu'il en faut $\mathcal{O}(N^6)$ pour $u_{\tilde{h}}$.

2.6 Conditionnement

Les matrices A_h ne sont pas bien conditionnées. En effet, le conditionnement de A_h défini par

$$K(A_h) = \|A_h\| \, \|A_h^{-1}\|,$$

pour une norme matricielle $\|\cdot\|$ donnée, est en général grand lorsque h est petit. Par exemple, en 2D avec le Laplacien sur le carré unité, on a $\|A_h\|_\infty = \dfrac{8}{h^2}$ et $\|A_h^{-1}\|_\infty \leq \dfrac{1}{8}$, de sorte que $K_\infty(A_h) \leq \dfrac{1}{h^2}$ et la borne "explose" quand h tend vers 0. On peut montrer (cf. exercice) que dans le cas 1D, le conditionnement (spectral) admet le développement

$$K_2(A_h) := \|A_h\|_2 \|A_h^{-1}\|_2 = \frac{4(b-a)^2}{\pi^2 h^2} + \mathcal{O}(h^2), \tag{2.94}$$

pour $\Omega = (a, b)$. On a donc $K_2(A_h) \to +\infty$ quand $h \to 0$!!! La matrice A_h étant symétrique, le conditionnement spectral de A_h est plus petit que n'importe quel conditionnement en norme subordonnée. Dans ce cas (en 1D), tout conditionnement de A_h associé à une norme matricielle subordonnée tend vers $+\infty$ quand h tend vers 0.

Chapitre 3

EDP paraboliques linéaires

3.1 Introduction

Pour un domaine (ouvert connexe) $\Omega \subset \mathbb{R}^n$ borné, de frontière $\partial\Omega$ *régulière* et pour un temps $T > 0$ fixé, on considère le problème aux limites suivant : trouver une fonction $u = u(\mathbf{x}, t)$ avec $\mathbf{x} \in \Omega$ et $t \in (0, T)$, telle que

$$\begin{aligned} \frac{\partial u}{\partial t} - Lu &= f \quad \text{dans } \Omega \times (0,T) && (3.1)\\ u &= 0 \quad \text{sur } \partial\Omega \times (0,T) && (3.2)\\ u(\mathbf{x}, 0) &= u_0(\mathbf{x}) \quad \text{dans } \Omega. && (3.3) \end{aligned}$$

La condition (3.2) est la *condition limite* sur le bord du domaine Ω (ici Dirichlet homogène). La condition (3.2) est la *condition initiale* à l'instant $t = 0$. L'opérateur L est défini par

$$\begin{aligned} Lu &:= \operatorname{div}(A\nabla u) - \mathbf{b}\cdot\nabla u - c\,u && (3.4)\\ &= \sum_{i,j=1}^{n} \frac{\partial}{\partial x_i}\left(a_{ij}\frac{\partial u}{\partial x_j}\right) - \sum_{i=1}^{n} b_i \frac{\partial u}{\partial x_i} - c\,u && (3.5) \end{aligned}$$

avec $A = A(\mathbf{x}, t) = (a_{ij}(\mathbf{x}, t)) \in \mathbb{R}^{n,n}$, $\mathbf{b} = \mathbf{b}(\mathbf{x}, t) = (b_i(\mathbf{x}, t)) \in \mathbb{R}^n$ et $c = c(\mathbf{x}, t)$. On suppose dans tout ce chapitre (sauf précision contraire) que les coefficients $a_{ij} \in C^1(\overline{\Omega} \times [0, T])$. Les fonctions $\mathbf{b}$, c et $f = f(\mathbf{x}, t)$ sont données dans $C(\overline{\Omega} \times [0, T])$.

On dit que l'équation (3.1) est uniformément **parabolique** si l'opérateur $-L$ est uniformément elliptique en la variable d'espace $\mathbf{x}$, c'est-à-dire si la condition suivante est satisfaite :

$$\exists\, \alpha > 0 \ \text{ tel que } \ (A(\mathbf{x},t))\xi,\ \xi)_{\mathbb{R}^n} = \sum_{i,j=1}^{n} a_{ij}(\mathbf{x},t)\xi_i\xi_j \ \geq\ \alpha|\xi|^2, \quad \forall \xi \in \mathbb{R}^n,\ \forall (\mathbf{x},t) \in \Omega \times (0,T). \quad (3.6)$$

On supposera désormais que la condition (3.6) est vérifiée.

Voici à présent quelques résultats d'existence, d'unicité et de principe du maximum relatifs aux équations paraboliques (on renvoit par exemple aux ouvrages [1], [3], [5], [4]).

3.1.1 Existence et unicité des solutions

Commençons par le cas de l'équation de la chaleur avec $L = \Delta$.

- Si $f \in L^2(\Omega \times (0,T))$ et $u_0 \in H_0^1(\Omega)$ alors le problème (3.1)-(3.3) avec $L = \Delta$ admet une unique solution

$$\begin{gathered} u \in L^2(0,T; H^2(\Omega) \cap H_0^1(\Omega)) \cap C([0,T]; H_0^1(\Omega)) \\ \frac{\partial u}{\partial t} \in L^2(0,T; L^2(\Omega)) \end{gathered} \quad (3.7)$$

  vérifiant les équations (3.1), (3.2) et (3.3) au sens *presque partout*.

- Si $f \in C^\infty(\overline{\Omega}\times[0,T])$ et $u_0 \in L^2(\Omega)$ alors il existe une unique solution
$$u \in C^\infty(\overline{\Omega}\times[\varepsilon,T]) \quad \forall\, \varepsilon > 0. \tag{3.8}$$
Si de plus, $u_0 \in C^\infty(\overline{\Omega})$ et (f,u_0) vérifie certaines relations de compatibilité([1]) sur $\partial\Omega$ alors $u \in C^\infty(\overline{\Omega}\times[0,T])$.

D'après (3.8), on voit que l'équation de la chaleur a un effet fortement régularisant sur la donnée initiale u_0. En effet, la solution u est C^∞ en x pour chaque $t>0$, même si la donnée initiale u_0 est discontinue.

Ces résultats se généralisent au cas d'un opérateur L à coefficients *réguliers* :

- Si $a_{ij}, b_i, c \in C^\infty(\overline{\Omega}\times[0,T])$ et si les coefficients a_{ij} vérifient la condition d'ellipticité (3.6) alors tous les résultats précédents restent vrais. On remarquera qu'il n'y a pas d'hypothèse sur le signe de la fonction c, contrairement au cas elliptique (cf. Chap. 2).

3.1.2 Principes du maximum

Soit $u \in C^2(\Omega\times(0,T]) \cap C^0(\overline{\Omega}\times[0,T])$ vérifiant $u_t - Lu = f$.

- On prend $c \equiv 0$. Si $f \le 0$ (resp. $f \ge 0$) alors u atteint son maximum (resp. minimum) sur $\Sigma = \partial\Omega\times[0,T] \cup \Omega\times\{t=0\}$. En particulier, si u est solution de (3.1)-(3.3) avec $c\equiv 0$, $f\ge 0$ et $u_0 \ge 0$, alors $u \ge 0$ dans $\overline{\Omega}\times[0,T]$.
- (Principe de Hopf) Soit $f \le 0$ et soient $\mathbf{x}_0 \in \partial\Omega$, $t_0 \in (0,T)$ tels que $u(\mathbf{x},t_0) < u(\mathbf{x}_0,t_0)$ pour tout $\mathbf{x}\in\Omega$. On suppose que u est dérivable en $\mathbf{x}_0$. Si $c\equiv 0$ ou bien si $u(\mathbf{x}_0,t_0)=0$ alors
$$\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}_0,t_0) > 0,$$
où $\mathbf{n}$ désigne la normale extérieure à $\partial\Omega$.

Pour l'équation de la chaleur avec $f\equiv 0$ et $c\equiv 0$, si $u_0 \ge 0$ avec $u_0 \not\equiv 0$ alors la solution u de (3.1)-(3.3) vérifie $u(\mathbf{x},t) > 0$, $\forall \mathbf{x}\in\Omega$, $\forall t>0$. Autrement dit, l'effet d'une petite perturbation initiale est ressenti immédiatement partout : la chaleur se propage avec une vitesse infinie.

3.2 Equation de la chaleur en dimension 1 d'espace

On considère l'équation en 1D d'espace avec $\Omega=(0,1)$: trouver $u=u(x,t)$ pour $(x,t)\in(0,1)\times(0,T)$ telle que

$$(P)\quad \begin{cases} \dfrac{\partial u}{\partial t} - \gamma \dfrac{\partial^2 u}{\partial x^2} &= f & \text{dans } Q_T := (0,1)\times(0,T)\\ u(0,t) &= \alpha, & t\in(0,T)\\ u(1,t) &= \eta, & \\ u(x,0) &= u_0(x), & x\in(0,1) \end{cases}$$

avec $\gamma>0$, α et η sont des constantes données. On suppose que le problème (P) admet une unique solution $u\in C^2(\overline{Q}_T)$. On notera qu'avec des conditions aux limites de Dirichlet homogènes c'est-à-dire avec $\alpha=\eta=0$, on a l'estimation suivante (avec $u_0\in H_0^1(0,1)$ et $f\in L^2(Q_T)$) :

$$\|u(t)\|_{H^1(0,1)} \le C\left(\|u_0\|_{H^1(0,1)} + \|f\|_{L^2(Q_T)}\right) \quad \text{pour tout } t\in(0,T). \tag{3.9}$$

1. Les relations de compatibilité sont des conditions nécessaires pour que $u\in C^\infty(\overline{\Omega}\times[0,T])$. Elles s'obtiennent en écrivant que toutes les dérivées de u par rapport à t, sont nulles sur $\partial\Omega\times[0,T]$ i.e. $u=\partial_t u = \cdots = \partial_t^j u = \cdots = 0$ sur $\partial\Omega\times[0,T]$. On dérive de façon itérée l'équation (3.1) par rapport à t et on utilise les relations précédentes. Par exemple, l'équation $u_t = \Delta u + f$ dans $\overline{\Omega}\times[0,T]$, fournit sur $\partial\Omega\times\{t=0\}$ la relation $-\Delta u_0 = f(\mathbf{x},0)$ pour $\mathbf{x}\in\partial\Omega$. Puis l'équation $u_{tt} = \Delta u_t + f_t = \Delta^2 u + \Delta f + f_t$ dans $\overline{\Omega}\times[0,T]$, fournit sur $\partial\Omega\times\{t=0\}$ la relation $-\Delta^2 u_0 = \Delta f(\mathbf{x},0) + f_t(\mathbf{x},0)$ pour $\mathbf{x}\in\partial\Omega$, et ainsi de suite ... On peut prendre par exemple, les relations suivantes
$$u_0 = \Delta u_0 = \cdots = \Delta^j u_0 = \cdots = 0 \text{ sur } \partial\Omega,$$
$$f = \Delta f = \cdots = \Delta^j f = \cdots = 0 \text{ sur } \partial\Omega\times(0,T).$$
On remarquera enfin, que l'hypothèse $u_0\in C^\infty(\overline{\Omega})$ avec $u_0=0$ sur $\partial\Omega$ ne suffit pas pour avoir $u\in C^\infty(\overline{\Omega}\times[0,T])$...

3.2.1 Schéma d'Euler explicite

Pour deux entiers M et N donnés, on discrétise de façon uniforme les intervalles d'espace $\overline{\Omega} = [0,1]$ et de temps $[0,T]$ en introduisant les points

$$x_j = jh, \quad j = 0, \cdots, M+1 \tag{3.10}$$
$$t^n = n\Delta t, \quad n = 0, \cdots, N \tag{3.11}$$

où h est le pas de discrétisation en espace donné par $h = 1/(M+1)$ et Δt est le pas de discrétisation en temps avec $\Delta t = T/N$. On introduit enfin les points

$$P_j^n = (x_j, t^n) \tag{3.12}$$

pour $j = 0, \cdots, M+1$ et $n = 0, \cdots, N$, qui définissent un maillage du domaine spatio-temporel $\overline{Q}_T$ et qu'on appellera *noeuds* du maillage.

On cherche alors une approximation $u_j^n \simeq u(x_j, t^n)$ de la solution exacte aux noeuds P_j^n, en discrétisant la dérivée en espace par un schéma *centré* au temps t^n et la dérivée en temps par un schéma *décentré progressif*, ce qui donne

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \gamma \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} + f_j^n \tag{3.13}$$

où l'on a noté $f_j^n = f(x_j, t^n)$.

Le problème approché consiste alors à trouver u_j^n pour $j = 0, \cdots, M+1$, $n = 0, \cdots, N$ telle que

$$u_j^{n+1} = \beta u_{j+1}^n + (1-2\beta)u_j^n + \beta u_{j-1}^n + \Delta t f_j^n, \qquad j = 1, \cdots, M \tag{3.14}$$
$$u_0^n = \alpha \tag{3.15}$$
$$u_{M+1}^n = \eta \tag{3.16}$$
$$u_j^0 = u_0(x_j) \tag{3.17}$$

avec

$$\beta = \gamma \frac{\Delta t}{h^2}. \tag{3.18}$$

Le schéma (3.14)-(3.17) est *explicite* dans la mesure où l'on calcule u_j^{n+1} directement à partir de u_j^n. La figure 3.1 indique les valeurs nécessaires au calcul de u_j^n en un noeud P_j^n donné. Ceci définit le *cône de dépendance numérique* du schéma d'Euler explicite.

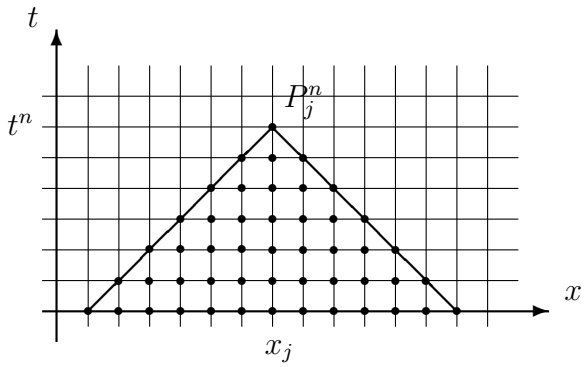


FIGURE 3.1 – Cône de dépendance numérique du schéma d'Euler explicite.

On peut écrire le problème approché sous forme matricielle en regroupant les inconnues dans un vecteur $\mathbf{u}^n = (u_1^n, \cdots, u_M^n)^\top$. Le problème approché s'écrit alors

$$\mathbf{u}^{n+1} = (I_d - \beta A)\,\mathbf{u}^n + \Delta t\,\mathbf{f}^n + \beta\,\mathbf{b} \tag{3.19}$$

avec $\mathbf{f}^n = (f_1^n, \cdots, f_M^n)^\top$, $\mathbf{b} = (\alpha, 0, \cdots, 0, \eta)^\top \in \mathbb{R}^M$ et

$$A = \begin{pmatrix} 2 & -1 & & & 0 \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ 0 & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{M,M}. \tag{3.20}$$

Consistance.

On note $L_{h,\Delta t}$ l'opérateur aux différences associé au problème approché :

$$L_{h,\Delta t} v(x,t) = \frac{v(x,t+\Delta t) - v(x,t)}{\Delta t} - \gamma \frac{v(x+h,t) - 2v(x,t) + v(x-h,t)}{h^2}. \tag{3.21}$$

En posant $L = \partial/\partial t - \partial^2/\partial x^2$ l'opérateur de la chaleur, on obtient pour une fonction v régulière ($v \in C^{4,2}$),

$$\left(L_{h,\Delta t} v - Lv\right)(x,t) = \frac{\Delta t}{2} \partial_t^2 v(x,\bar{t}) + \gamma \frac{h^2}{12} \partial_x^4 v(\bar{x},t) = \mathcal{O}(\Delta t + h^2). \tag{3.22}$$

Stabilité.

Il y a plusieurs notions de stabilité.

⋆ **Critère de Von Neumann - Fourier.**

Dans ce critère, on ne prend pas en compte les effets de bords de la discrétisation (conditions limites) et on analyse seulement l'équation (3.14). Ceci revient à considérer le problème (P) non plus dans un intervalle borné mais dans $\mathbb{R}$ tout entier et à ignorer les conditions limites. On prend également $f \equiv 0$. Dans ce cas, la solution exacte de (P) est bornée([1]). On veut alors retrouver cette propriété sur les approximations u_j^n. On cherche une solution sous la forme particulière suivante, avec $\xi \in \mathbb{R}$:

$$u_j^n = \xi^n \mathrm{e}^{ik\pi jh} \quad \text{pour } k \text{ fixé.} \tag{3.23}$$

Puisque $|u_j^n| = |\xi|^n$, on impose, pour tout mode k, la condition

$$|\xi| \leq 1, \tag{3.24}$$

afin que la solution approchée soit bornée, pour tout n. C'est la condition de stabilité correspondant au critère de Von Neumann. Le paramètre ξ s'appelle *facteur d'amplification* associé au mode k.

En reportant cette expression dans l'équation (3.14) (avec $f \equiv 0$), on obtient

$$\xi^{n+1} \mathrm{e}^{ik\pi jh} = \beta \xi^n \mathrm{e}^{ik\pi(j+1)h} + (1-2\beta)\xi^n \mathrm{e}^{ik\pi(j+1)h} + \beta \xi^n \mathrm{e}^{ik\pi(j-1)h},$$

ce qui donne

$$\begin{aligned} \xi &= \beta \mathrm{e}^{ik\pi h} + (1-2\beta) + \beta \mathrm{e}^{-ik\pi h} \\ &= 2\beta \cos(k\pi h) + 1 - 2\beta \\ &= 1 + 2\beta(\cos(k\pi h) - 1). \end{aligned}$$

Or $\cos(x) - 1 = -2\sin^2(x/2)$ et par conséquent on obtient la relation

$$\xi = 1 - 4\beta \sin^2\left(\frac{k\pi h}{2}\right). \tag{3.25}$$

1. L'équation $u_t - u_{xx} = 0$ posée pour $x \in \mathbb{R}$ et $t > 0$ avec $u(\cdot, 0) = u_0$, a pour solution $u(x,t) = E * u_0(x,t) = \int_{\mathbb{R}} E(x-y,t) u_0(y)\, dy$ avec $E(x,t) = (4\pi t)^{-1/2} \exp(-|x|^2/4t)$. La solution u est bornée uniformément en x et en t et on a même $\|u(t)\|_{L^\infty(\mathbb{R})} \to 0$ quand $t \to +\infty$.

Avec la relation (3.25), la condition de stabilité (3.24) devient $-1 \leq 1 - 4\beta \sin^2(k\pi h/2) \leq 1$ pour tout k, soit $4\beta \sin^2(k\pi h/2) \leq 2$. Pour que cette dernière inégalité soit vérifiée quelque soit k, on impose $4\beta \leq 2$ c'est-à-dire

$$\beta \leq \frac{1}{2} \quad \text{i.e.} \quad \gamma \frac{\Delta t}{h^2} \leq \frac{1}{2}. \tag{3.26}$$

La condition (3.26) est une condition de stabilité (selon le critère de Von Neumann-Fourier) qui relie le pas de temps au pas d'espace.

⋆ Stabilité L^∞ et L^2 du schéma explicite.

Introduisons les normes L^∞ et L^2 discrètes suivantes. Pour $\mathbf{v} = (v_1, \cdots, v_M) \in \mathbb{R}^M$, on note

$$\|\mathbf{v}\|_\infty := \max_{1\leq i\leq M} |v_i|, \tag{3.27}$$

$$\|\mathbf{v}\|_{2,h} := \sqrt{h}\|\mathbf{v}\|_2 = \sqrt{h}\left(\sum_{i=1}^{M} v_i^2\right)^{1/2}. \tag{3.28}$$

La norme $\|\cdot\|_{2,h}$ est une norme *admissible* au sens où l'on a $\|\mathbf{v}\|_{2,h} \to \|v\|_{L^2(0,1)}$ quand $h \to 0$, pour toute fonction $v \in L^2(0,1)$ continue sur $(0,1)$, avec $\mathbf{v} = (v(x_1), \cdots, v(x_M))$. Ceci n'est pas le cas de $\|\mathbf{v}\|_2$ en général, ce qui explique la présence du $\sqrt{h}$ dans la définition de la norme $\|\cdot\|_{2,h}$ ...

1. Principe du maximum discret : Soit $f \geq 0$, $u_0 \geq 0$ et $\alpha, \eta \geq 0$. Si $\beta \leq \frac{1}{2}$ alors $\mathbf{u}^n \geq 0$, $\forall n$.
2. Dans le cas de conditions aux limites homogènes ($\alpha = \eta = 0$), si $\beta \leq \frac{1}{2}$ alors on a les estimations suivantes, pour tout $n = 0, \cdots, N$,

$$\|\mathbf{u}^n\|_\infty \leq \|\mathbf{u}^0\|_\infty + T\|f\|_{L^\infty(Q_T)} \tag{3.29}$$

$$\|\mathbf{u}^n\|_{2,h} \leq \|\mathbf{u}^0\|_{2,h} + T\max_n \|\mathbf{f}^n\|_{2,h} \tag{3.30}$$

Démonstration de 2. Avec $\alpha = \eta = 0$, le système linéaire devient

$$\mathbf{u}^{n+1} = \mathcal{M}\mathbf{u}^n + \Delta t \mathbf{f}^n \text{ avec } \mathcal{M} = I_d - \beta A. \tag{3.31}$$

De cette relation, on tire directement

$$\|\mathbf{u}^{n+1}\|_\infty \leq \|\mathcal{M}\|_\infty \|\mathbf{u}^n\|_\infty + \Delta t \|\mathbf{f}^n\|_\infty. \tag{3.32}$$

Or, puisque $\beta \leq 1/2$, on en déduit que

$$\|\mathcal{M}\|_\infty = \max_{1\leq i\leq M} \sum_{j=1}^{M} |m_{ij}| = |1 - 2\beta| + 2\beta = 1.$$

En sommant sur n, on obtient alors l'estimation (3.29).

De même, avec la norme euclidienne $\|\mathbf{v}\|_2 = \left(\sum_{i=1}^{M} v_i^2\right)^{1/2}$, on a

$$\|\mathbf{u}^{n+1}\|_2 \leq \|\mathcal{M}\|_2 \|\mathbf{u}^n\|_2 + \Delta t \|\mathbf{f}^n\|_2. \tag{3.33}$$

La matrice $\mathcal{M}$ étant symétrique, on a $\|\mathcal{M}\|_2 = \rho(\mathcal{M}) = \max_k |\lambda_k|$ où λ_k désigne les valeurs propres de $\mathcal{M} = I_d - \beta A$. On a (cf. exercice),

$$\lambda_k = 1 - 4\beta \sin^2\left(\frac{k\pi}{2(M+1)}\right) \quad \text{pour } k = 1, \cdots, M.$$

Par conséquent, on voit que λ_k est exactement le facteur d'amplification ξ associé au mode k du critère de Von Neumann. Ainsi, avec $\beta \leq 1/2$, on a $|\lambda_k| \leq 1$ et donc $\|\mathcal{M}\|_2 \leq 1$. En sommant sur n et en passant à la norme $\|\cdot\|_{2,h}$, on obtient l'estimation (3.30). □

Convergence.

On pose $e_j^n = u(jh, n\Delta t) - u_j^n$. En supposant la solution exacte u suffisamment régulière, l'erreur de consistance donne, pour $j = 1, \cdots, M$

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} = \gamma \frac{e_{j+1}^n - 2e_j^n + e_{j-1}^n}{h^2} + \mathcal{O}(\Delta t + h^2)$$

avec $e_0^n = \alpha - \alpha = 0$ et $e_{M+1}^n = \eta - \eta = 0$. Matriciellement en posant $\mathbf{e}^n = (e_1^n, \cdots, e_M^n)$, on obtient

$$\mathbf{e}^{n+1} = \mathcal{M}\mathbf{e}^n + \Delta t\, \mathbf{O}(\Delta t + h^2). \tag{3.34}$$

Pour $\beta \leq 1/2$, on montre que (cf. démonstration stabilité L^∞), pour tout $n = 0, \cdots, N$,

$$\|\mathbf{e}^n\|_\infty \leq \|\mathbf{e}^0\|_\infty + T\, \mathcal{O}(\Delta t + h^2).$$

Puisque $\mathbf{e}^0 = \mathbf{0}$, on a

$$\|\mathbf{e}^n\|_\infty = \mathcal{O}(\Delta t + h^2) \quad \text{si} \quad \beta \leq 1/2. \tag{3.35}$$

Conclusion. Le schéma d'Euler explicite est simple (pas de système linéaire à résoudre) mais il y a une condition de stabilité à respecter, ce qui limite le pas de temps pour un pas h donné.

3.2.2 Schéma d'Euler implicite

Ce schéma s'écrit

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \gamma \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} + f_j^{n+1} \tag{3.36}$$

En notant $\beta = \gamma \Delta t / h^2$, le problème approché consiste alors à trouver u_j^n pour $j = 0, \cdots, M+1$ et $n = 0, \cdots, N$ telle que

$$-\beta u_{j+1}^{n+1} + (1 + 2\beta) u_j^{n+1} - \beta u_{j-1}^{n+1} = u_j^n + \Delta t\, f_j^{n+1}, \quad \text{pour } j = 1, \cdots, M, \tag{3.37}$$

avec les conditions limites $u_0^n = \alpha$, $u_{M+1}^n = \eta$ et la donnée initiale $u_j^0 = u_0(x_j)$.

En regroupant les inconnues dans les vecteurs $\mathbf{u}^n = (u_1^n, \cdots, u_M^n)^\top$, le système précédent s'écrit sous forme matricielle

$$\mathcal{M}\mathbf{u}^{n+1} = \mathbf{u}^n + \Delta t\, \mathbf{f}^n + \beta\, (\alpha, 0, \cdots, 0, \eta)^\top \tag{3.38}$$

La matrice $\mathcal{M} = I_d + \beta A$ est à diagonale strictement dominante donc définie positive ; de plus elle est indépendante du temps. On effectue ainsi une décomposition de Choleski ($CC^\top$) de la matrice $\mathcal{M}$, une fois pour toute.

Consistance.

Pour une fonction v régulière ($v \in C^{4,2}$), on a

$$(L_{h,\Delta t} v - Lv)\,(x,t) = \mathcal{O}(\Delta t + h^2). \tag{3.39}$$

où L est l'opérateur de la chaleur et $L_{h,\Delta t}$ est l'opérateur aux différences associé au problème approché :

$$L_{h,\Delta t} v(x,t) = \frac{v(x,t) - v(x,t-\Delta t)}{\Delta t} - \gamma \frac{v(x+h,t) - 2v(x,t) + v(x-h,t)}{h^2}. \tag{3.40}$$

Stabilité.

⋆ **Critère de Von Neumann - Fourier.**

Avec $f \equiv 0$, on cherche une solution de la relation (3.37) sous la forme particulière suivante

$$u_j^n = \xi^n \mathrm{e}^{ik\pi jh} \quad \text{pour } k \text{ fixé.} \tag{3.41}$$

Le paramètre ξ est le facteur d'amplification associé au mode k et la condition de stabilité du critère de Von Neumann s'écrit $|\xi| \leq 1$ pour tout mode k.

En reportant cette expression dans l'équation (3.37) (avec $f \equiv 0$), on obtient

$$-\beta\xi \mathrm{e}^{ik\pi h} + (1+2\beta)\xi - \beta\xi \mathrm{e}^{-ik\pi h} = 1,$$

ce qui donne

$$\xi\,(1 + 2\beta(1 - \cos(k\pi h)) = 1,$$

soit encore

$$\xi = \frac{1}{1 + 4\beta \sin^2(k\pi h/2)}. \tag{3.42}$$

Pour tout k, on a clairement $0 < \xi \leq 1$ et par conséquent le schéma implicite est *inconditionellement stable* au sens du critère de Von Neumann.

⋆ **Stabilité L^∞ et L^2 du schéma implicite.**

1. Principe du maximum discret : Si $f \geq 0$, $u_0 \geq 0$ et $\alpha, \eta \geq 0$ alors $\mathbf{u}^n \geq 0$, $\forall n$ (pas de condition sur β).
2. Dans le cas de conditions aux limites homogènes ($\alpha = \eta = 0$), on a les estimations suivantes, pour tout $n = 0, \cdots, N$ (sans condition sur β),

$$\|\mathbf{u}^n\|_\infty \leq \|\mathbf{u}^0\|_\infty + T\|f\|_{L^\infty(Q_T)} \tag{3.43}$$
$$\|\mathbf{u}^n\|_{2,h} \leq \|\mathbf{u}^0\|_{2,h} + T \max_n \|\mathbf{f}^n\|_{2,h}. \tag{3.44}$$

Démonstration de 1. Le résultat provient de la forme matricielle (3.38) et du fait que $I_d + \beta A$ est une M-matrice.

Démonstration de 2.

i) Stabilité L^∞ : Avec $\alpha = \eta = 0$, on a

$$(1+2\beta)|u_j^{n+1}| \leq \beta\left(|u_{j+1}^{n+1}| + |u_{j-1}^{n+1}|\right) + |u_j^n| + \Delta t\,|f_j^{n+1}|, \text{ pour } j = 1, \cdots, M$$

On en déduit, pour $j = 1, \cdots, M$

$$(1+2\beta)|u_j^{n+1}| \leq 2\beta\|\mathbf{u}^{n+1}\|_\infty + \|\mathbf{u}^n\|_\infty + \Delta t\,\|f\|_{L^\infty(Q_T)}$$

d'où

$$(1+2\beta)\|\mathbf{u}^{n+1}\|_\infty \leq 2\beta\|\mathbf{u}^{n+1}\|_\infty + \|\mathbf{u}^n\|_\infty + \Delta t\,\|f\|_{L^\infty(Q_T)}$$

ce qui donne

$$\|\mathbf{u}^{n+1}\|_\infty \leq \|\mathbf{u}^n\|_\infty + \Delta t\,\|f\|_{L^\infty(Q_T)} \tag{3.45}$$

et on en déduit l'estimation (3.43).

ii) Stabilité L^2 : Avec $\alpha = \eta = 0$, le système linéaire devient

$$\mathbf{u}^{n+1} = \mathcal{M}^{-1}\left(\mathbf{u}^n + \Delta t\, \mathbf{f}^n\right) \text{ avec } \mathcal{M} = I_d - \beta A. \tag{3.46}$$

On en déduit que

$$\|\mathbf{u}^{n+1}\|_2 \leq \|\mathcal{M}^{-1}\|_2 \left(\|\mathbf{u}^n\|_2 + \Delta t \|\mathbf{f}^n\|_2\right). \tag{3.47}$$

Or, $\|\mathcal{M}^{-1}\|_2 = \dfrac{1}{\min_k |\mu_k|}$ où μ_k sont les valeurs propres de $\mathcal{M} = I_d + \beta A$. On a donc $\mu_k = 1 + \beta\lambda_k$ où $\lambda_k = 4\sin^2(k\pi/(2(M+1))$ sont les valeurs propres de A. On obtient donc

$$\mu_k = 1 + 4\beta \sin^2(k\pi/(2(M+1)) \geq 1, \text{ pour } k = 1, \cdots, M. \tag{3.48}$$

Par conséquent $\|\mathcal{M}^{-1}\|_2 \leq 1$. On en déduit alors facilement l'estimation (3.44). □

Remarque. On notera que $1/\mu_k = \xi$ est le facteur d'amplification du critère de Von Neumann. Ce critère est en fait directement lié à la stabilité L^2.

Convergence.

On pose $e_j^n = u(jh, n\Delta t) - u_j^n$. En supposant la solution exacte u suffisamment régulière, on a, pour $j = 1, \cdots, M$

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} = \gamma \frac{e_{j+1}^{n+1} - 2e_j^{n+1} + e_{j-1}^{n+1}}{h^2} + \mathcal{O}(\Delta t + h^2)$$

avec $e_0^{n+1} = 0$ et $e_{M+1}^{n+1} = 0$. Matriciellement en posant $\mathbf{e}^n = (e_1^n, \cdots, e_M^n)$, on obtient

$$\mathbf{e}^{n+1} = \mathcal{M}^{-1}\mathbf{e}^n + \Delta t\, \mathbf{O}(\Delta t + h^2). \tag{3.49}$$

Ainsi on obtient, pour tout $n = 0, \cdots, N$,

$$\|\mathbf{e}^n\|_\infty = \mathcal{O}(\Delta t + h^2), \tag{3.50}$$

sans condition sur le pas de temps et d'espace.

Conclusion. Le schéma d'Euler implicite nécessite la résolution d'un système linéaire (contrairement au schéma explicite) mais il est inconditionnellement stable. Il n'y a pas de restriction sur les pas de temps et d'espace. On peut prendre ainsi des pas de temps assez grands. La figure 3.2 montre le comportement des schémas d'Euler explicite et implicite pour différentes valeurs du paramètre $\beta = \gamma\Delta t/h^2$.

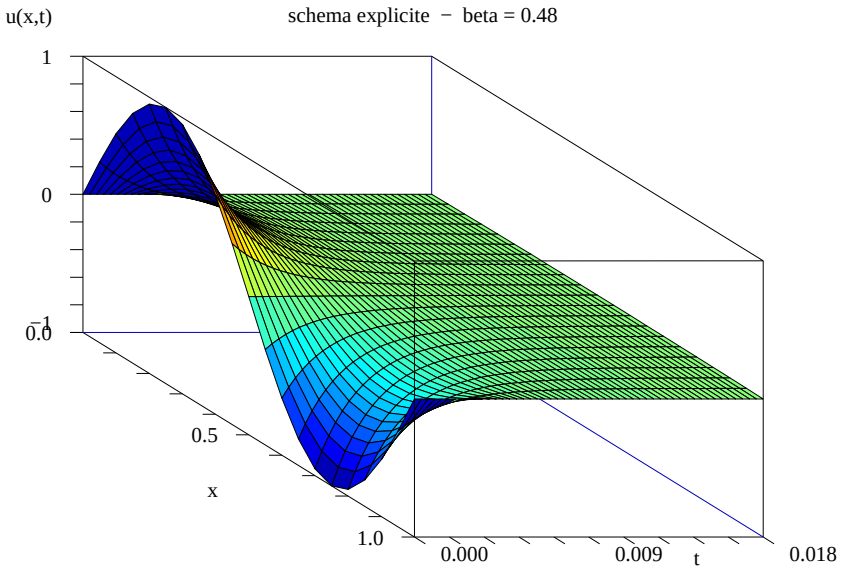


Schéma d'Euler explicite avec $\beta = 0.48$.

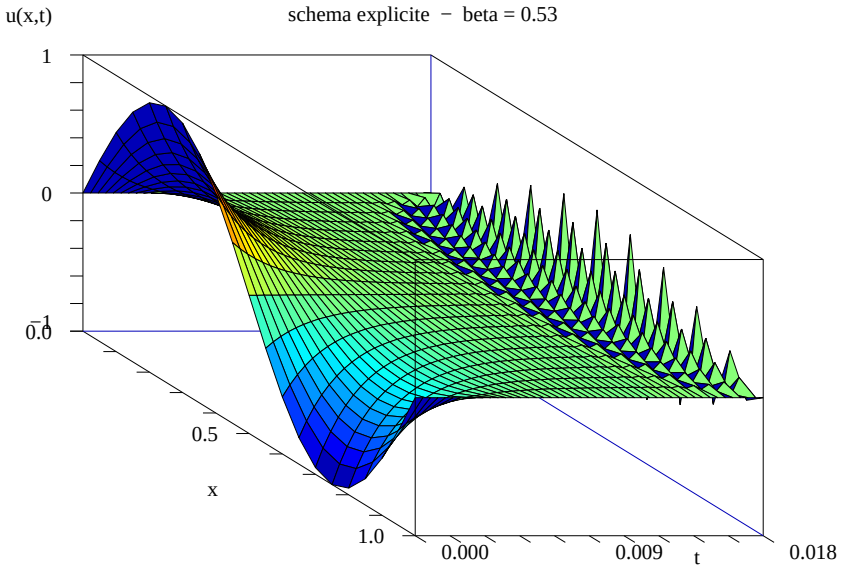


Schéma d'Euler explicite avec $\beta = 0.53$.

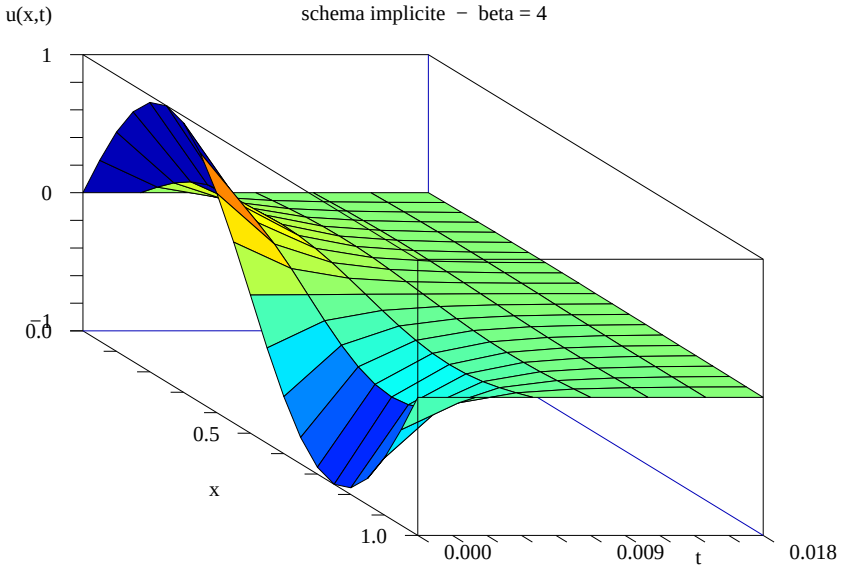


Schéma d'Euler implicite avec $\beta = 4$.

FIGURE 3.2 – Schémas d'Euler pour l'équation de la chaleur.

3.2.3 Schéma de Crank-Nicholson (1947)

Ce schéma s'écrit

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\frac{\gamma}{2}\left(\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{h^2}+\frac{u_{j+1}^{n}-2u_j^{n}+u_{j-1}^{n}}{h^2}\right)=\frac{1}{2}\left(f_j^n+f_j^{n+1}\right) \tag{3.51}$$

En posant $\beta=\gamma\Delta t/h^2$, le problème approché consiste alors à trouver u_j^n pour $j=0,\cdots,M+1$ et $n=0,\cdots,N$ telle que

$$-\frac{\beta}{2}u_{j+1}^{n+1}+(1+\beta)u_j^{n+1}-\frac{\beta}{2}u_{j-1}^{n+1}=\frac{\beta}{2}u_{j+1}^{n}+(1-\beta)u_j^n+\frac{\beta}{2}u_{j-1}^n+\frac{\Delta t}{2}\left(f_j^n+f_j^{n+1}\right), \tag{3.52}$$

pour $j=1,\cdots,M$ et avec les conditions aux limites $u_0^n=\alpha$, $u_{M+1}^n=\eta$ et la condition initiale $u_j^0=u_0(x_j)$.

En regroupant les inconnues dans les vecteurs $\mathbf{u}^n=(u_1^n,\cdots,u_M^n)^\top$, le système précédent s'écrit sous forme matricielle

$$\left(I_d+\frac{\beta}{2}A\right)\mathbf{u}^{n+1}=\left(I_d-\frac{\beta}{2}A\right)\mathbf{u}^n+\frac{\Delta t}{2}\left(\mathbf{f}^n+\mathbf{f}^{n+1}\right)+\beta\left(\alpha,0,\cdots,0,\eta\right)^\top. \tag{3.53}$$

La matrice $I_d+\beta/2\,A$ est symétrique définie positive.

Consistance.

Pour une fonction v régulière ($v\in C^{4,2}$), on a

$$\begin{aligned}\frac{\partial v}{\partial t}(x,t+\Delta t/2)&=\frac{v(x,t+\Delta t)-v(x,t)}{\Delta t}+\mathcal{O}(\Delta t^2)\\ \frac{\partial^2 v}{\partial x^2}(x,t+\Delta t/2)&=\frac{1}{2}\left(\frac{\partial^2 v}{\partial x^2}(x,t)+\frac{\partial^2 v}{\partial x^2}(x,t+\Delta t)\right)+\mathcal{O}(\Delta t^2)\end{aligned}$$

On en déduit que

$$\left(L_{h,\Delta t}v-Lv\right)(x,t+\Delta t/2)=\mathcal{O}(\Delta t^2+h^2). \tag{3.54}$$

où L est l'opérateur de la chaleur et $L_{h,\Delta t}$ est l'opérateur aux différences du schéma de Crank-Nicholson. Le schéma de Crank-Nicholson est d'ordre 2 en espace et en temps. Il s'agit donc d'un schéma précis en temps.

Stabilité.

⋆ **Critère de Von Neumann - Fourier.**

On montre que le facteur d'amplification ξ associé à un mode k est donné par

$$\xi=\frac{1-2\beta\sin^2\left(k\pi h/2\right)}{1+2\beta\sin^2\left(k\pi h/2\right)} \tag{3.55}$$

et par conséquent on a $|\xi|\leq 1$. Le schéma de Crank-Nicholson est *inconditionnellement stable au sens de Von Neumann.*

⋆ **Stabilité L^∞ et L^2 du schéma de Crank-Nicholson.**

1. Principe du maximum discret : Soit $f\geq 0$, $u_0\geq 0$ et $\alpha,\eta\geq 0$. Si $\beta\leq 1$ alors $\mathbf{u}^n\geq 0$, $\forall n$.
2. Dans le cas de conditions aux limites homogènes ($\alpha=\eta=0$), on a les estimations suivantes, pour tout $n=0,\cdots,N$,

$$\begin{aligned}&\text{Si } \beta\leq 1 \text{ alors} \quad \|\mathbf{u}^n\|_\infty\leq\|\mathbf{u}^0\|_\infty+T\|f\|_{L^\infty(Q_T)}\\ &\text{Sans condition sur } \beta, \quad \|\mathbf{u}^n\|_{2,h}\leq\|\mathbf{u}^0\|_{2,h}+T\max_n\|\mathbf{f}^n\|_{2,h}.\end{aligned}$$

Démonstration de 1. Le résultat provient de la forme matricielle (3.53),du fait que $I_d + \beta/2\, A$ est une M-matrice et que sous la condition $\beta \leq 1$, la matrice $I_d - \beta/2\, A$ est positive.

Démonstration de 2.

i) Stabilité L^∞ : Avec $\alpha = \eta = 0$, on a

$$(1+\beta)u_j^{n+1} = \frac{\beta}{2}\left(u_{j+1}^{n+1} + u_{j-1}^{n+1}\right) + \frac{\beta}{2}\left(u_{j+1}^{n} + u_{j-1}^{n}\right) + (1-\beta)u_j^n + \frac{\Delta t}{2}\left(f_j^n + f_j^{n+1}\right),\ j = 1, \cdots, M$$

On en déduit

$$(1+\beta)\|\mathbf{u}^{n+1}\|_\infty \leq \beta\|\mathbf{u}^{n+1}\|_\infty + (\beta + |1-\beta|)\|\mathbf{u}^n\|_\infty + \Delta t\, \|f\|_{L^\infty(Q_T)}. \tag{3.56}$$

En choisissant la condition $\beta \leq 1$, on a $\beta + |1-\beta| = 1$ et l'inégalité (3.56) donne

$$\|\mathbf{u}^{n+1}\|_\infty \leq \|\mathbf{u}^n\|_\infty + \Delta t\, \|f\|_{L^\infty(Q_T)} \tag{3.57}$$

ce qui fournit l'estimation (2).

ii) Stabilité L^2 : Commençons par des résultats préliminaires.

Lemme 3.1 *(a) Soient M_1 et M_2 deux matrices qui commutent et M_1 inversible. Alors M_1^{-1} et M_2 commutent.*

(b) Soient M_1 et M_2 deux matrices symétriques qui commutent. Alors le produit M_1M_2 est symétrique.

(c) Soit M une matrice carrée symétrique semi-définie positive. Pour $\alpha \geq 0$, la matrice $I_d + \alpha M$ est inversible et la matrice $B = (I_d + \alpha M)^{-1}(I_d - \alpha M)$ est symétrique avec $\rho(B) \leq 1$.

Démonstration du lemme.

(a) Par hypothèse $M_1M_2 = M_2M_1$. Donc $M_2 = M_1^{-1}M_2M_1$ et $M_2M_1^{-1} = M_1^{-1}M_2$.

(b) $(M_1M_2)^\top = M_2^\top M_1^\top = M_2M_1 = M_1M_2$.

(c) Montrons d'abord que $I_d + \alpha M$ est inversible. Soit x tel que $(I_d + \alpha M)x = 0$. On a $Mx = -\dfrac{1}{\alpha}x$ pour $\alpha > 0$ (si $\alpha = 0$ alors $x = 0$). Par conséquent $(Mx, x) = -\dfrac{1}{\alpha}\|x\|^2 \geq 0$ car M est semi-définie positive. On en déduit $x = 0$ car $\alpha \geq 0$.

Etablissons à présent la propriété sur la matrice B. Les matrices $I_d + \alpha M$ et $I_d - \alpha M$ commutent donc d'après *(a)* et *(b)*, B est symétrique. En notant $C = I_d + \alpha M$, on a $B = C^{-1}(2I_d - C) = 2C^{-1} - I_d$. Les valeurs propres de B sont $\lambda_k = \dfrac{2}{1+\alpha\mu_k} - 1$ où μ_k sont les v.p. de M. La matrice M étant semi-définie positive, on a $\mu_k \geq 0$. Par ailleurs, $\rho(B) = \max\limits_k |\lambda_k|$ et $|\lambda_k| = \dfrac{|1-\alpha\mu_k|}{1+\alpha\mu_k} \leq 1$, $\forall k$, donc $\rho(B) \leq 1$. □

Etudions à présent la stabilité L^2 du schéma de Crank-Nicholson. Avec $\alpha = \eta = 0$, le système linéaire s'écrit

$$\mathbf{u}^{n+1} = B\mathbf{u}^n + \frac{\Delta t}{2}\left(I_d + \frac{\beta}{2}A\right)^{-1}\left(\mathbf{f}^n + \mathbf{f}^{n+1}\right) \text{ avec } B = \left(I_d + \frac{\beta}{2}A\right)^{-1}\left(I_d - \frac{\beta}{2}A\right). \tag{3.58}$$

On en déduit que

$$\|\mathbf{u}^{n+1}\|_2 \leq \|B\|_2\|\mathbf{u}^n\|_2 + \frac{\Delta t}{2}\|(I_d + \frac{\beta}{2}A)^{-1}\|_2\left(\|\mathbf{f}^n\|_2 + \|\mathbf{f}^{n+1}\|_2\right). \tag{3.59}$$

D'après la propriété *(c)* du Lemme 3.1, B est symétrique et $\|B\|_2 = \rho(B) \leq 1$.

Notons maintenant $C = (I_d + \frac{\beta}{2}A)^{-1}$. On a $\|C\|_2 = \rho(C)$ car C est symétrique et les v.p. de C sont données par $\gamma_k = \dfrac{1}{1+\frac{\beta}{2}\mu_k}$ où μ_k sont les v.p. de A. Puisque $\mu_k > 0$, on a $\gamma_k < 1$, $\forall k$ et par conséquent $\|C\|_2 \leq 1$. On obtient ainsi à partir de (3.59),

$$\|\mathbf{u}^{n+1}\|_2 \leq \|\mathbf{u}^n\|_2 + \Delta t \max_n \|\mathbf{f}^n\|_2. \tag{3.60}$$

On conclut en passant à la norme $\|\cdot\|_{2,h}$ et en sommant sur n. □

Remarque. Les valeurs propres de B sont données par

$$\lambda_k = \frac{1 - 2\beta \sin^2(k\pi/(2(M+1)))}{1 + 2\beta \sin^2(k\pi/(2(M+1)))} = \xi$$

où ξ est le facteur d'amplification du critère de Von Neumann.

Conclusion. Le schéma de Crank-Nicholson est un schéma précis en temps (ordre 2) qui est inconditionnellement L^2-stable mais qui est L^∞-stable sous la condition $\beta \leq 1$.

3.2.4 θ-schéma pour l'équation de la chaleur

Il s'agit d'une généralisation des schémas précédents. Avec $0 \leq \theta \leq 1$, ce schéma s'écrit

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \gamma\left(\theta \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} + (1-\theta)\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}\right) = \theta f_j^n + (1-\theta) f_j^{n+1}. \tag{3.61}$$

Sous forme matricielle, le système correspondant s'écrit

$$(I_d + \theta\beta A)\,\mathbf{u}^{n+1} = (I_d - (1-\theta)\beta A)\,\mathbf{u}^n + \Delta t\left(\theta\mathbf{f}^n + (1-\theta)\mathbf{f}^{n+1}\right) + \beta\,(\alpha, 0, \cdots, 0, \eta)^\top. \tag{3.62}$$

Cas particuliers.
- $\theta = 0$ → Euler explicite
- $\theta = 1$ → Euler implicite
- $\theta = 1/2$ → Crank-Nicholson

Consistance.

L'erreur de consistance est en $\mathcal{O}(\Delta t + h^2)$ si $\theta \neq 1/2$ et en $\mathcal{O}(\Delta t^2 + h^2)$ si $\theta = 1/2$.

Stabilité.

i) Pour $0 \leq \theta \leq 1/2$, le schéma est L^2-stable (et stable au sens de Von Neumann) si $\beta \leq \dfrac{1}{2-4\theta}$.

ii) Pour $1/2 \leq \theta \leq 1$, le schéma est inconditionnellement L^2-stable.

iii) Pour $0 \leq \theta \leq 1$, le schéma est L^∞-stable si $\beta \leq \dfrac{1}{2(1-\theta)}$.

3.2.5 Autres schémas

⋆ Richardson

Il s'agit d'un schéma centré en temps. Ce schéma s'écrit

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} - \gamma\left(\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2}\right) = f_j^n. \tag{3.63}$$

Ce schéma est toujours instable (exercice) !

⋆ **Dufort-Frankel** (1953)

Ce schéma s'écrit

$$\frac{u_j^{n+1}-u_j^{n-1}}{2\Delta t}-\frac{\gamma}{h^2}\left(u_{j+1}^n-u_j^{n+1}-u_j^{n-1}+u_{j-1}^n\right)=f_j^n. \tag{3.64}$$

Il s'agit d'un schéma explicite. L'erreur de consistance est en $\mathcal{O}\left(\Delta t^2+h^2+\left(\frac{\Delta t}{h}\right)^2\right)$. Par conséquent, il faut que $\Delta t/h\to 0$. Pour avoir un schéma d'ordre 2 en h, il faut prendre $\Delta t=\mathcal{O}(h^2)$. Ce schéma est inconditionnellement L^2-stable.

⋆ **Schéma rétrograde**

Il s'agit d'un schéma implicite qui s'écrit

$$\frac{1}{\Delta t}\left(\frac{3}{2}u_j^{n+1}-2u_j^n+\frac{1}{2}u_j^{n-1}\right)-\frac{\gamma}{h^2}\left(u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}\right)=f_j^{n+1}. \tag{3.65}$$

L'erreur de consistance est en $\mathcal{O}\left(\Delta t^2+h^2\right)$ et le schéma est inconditionnellement L^2-stable. Ce schéma est plus stable que celui de Crank-Nicholson au sens suivant : quand $\beta=\gamma\Delta t/h^2\to+\infty$, le facteur d'amplification ξ du schéma de Crank-Nicholson tend vers -1 alors que celui du schéma rétrograde tend vers 0.

3.3 Cas de la dimension 2 d'espace

3.3.1 θ-schéma pour l'équation de la chaleur en 2D

On cherche une fonction $u:\Omega\times(0,T)\to\mathbb{R}$ avec $\Omega=(a,b)\times(c,d)$ telle que

$$(P)\begin{cases} u_t-\gamma\Delta u=f(x,y,t) & \text{pour } (x,y,t)\in\Omega\times(0,T),\\ u=0 & \text{sur } \partial\Omega,\\ u(x,y,0)=u_0(x,y) & \text{pour } (x,y)\in\Omega.\end{cases}$$

Pour cela, on considère une discrétisation de $\overline{\Omega}$ en introduisant les points $P_{i,j}=(x_i,y_j)$ avec $x_i=a+i\,h_x$, $i=0,\cdots,m+1$ et $y_j=c+j\,h_y$, $j=0,\cdots,l+1$. Les pas de discrétisation en espace dans les directions x et y sont donnés respectivement par $h_x=(b-a)/(m+1)$ et $h_y=(d-c)/(l+1)$.

On cherche alors une approximation $\{u_{i,j}^n\}$ de la solution exacte u de (P) au point $P_{i,j}$ et à l'instant $t^n=n\Delta t$, c'est-à-dire $u_{i,j}^n\simeq u(x_i,y_j,t^n)$. Pour déterminer $u_{i,j}^n$, on utilise le θ-schéma suivant

$$\frac{u_{i,j}^{n+1}-u_{i,j}^n}{\Delta t}-\gamma\left(\theta\Delta_{h_x,h_y}u_{i,j}^{n+1}+(1-\theta)\Delta_{h_x,h_y}u_{i,j}^n\right)=\theta f_{i,j}^{n+1}+(1-\theta)f_{i,j}^n$$

où $0\le\theta\le 1$ et Δ_{h_x,h_y} désigne le schéma à 5 points du Laplacien. Comme pour le cas 1D, on a les cas particuliers suivants.

$$\begin{array}{ll}\theta=0 & \to \text{Euler explicite}\\ \theta=1 & \to \text{Euler implicite}\\ \theta=1/2 & \to \text{Crank-Nicholson}\end{array}$$

On regroupe les inconnues dans le vecteur $\mathbf{u}^n=(u_{11}^n,u_{21}^n,\cdots,u_{m1}^n,\ u_{12}^n,\cdots,u_{m2}^n,\ \cdots,u_{1l}^n,\cdots,u_{ml}^n)^\top\in\mathbb{R}^{ml}$, c'est-à-dire qu'on ordonne les noeuds du maillage de la gauche vers la droite et de bas en haut. Sous forme matricielle, le θ-schéma s'écrit de la façon suivante

$$(I_d+\theta\Delta t\,M)\,\mathbf{u}^{n+1}=(I_d-(1-\theta)\Delta t\,M)\,\mathbf{u}^n+\Delta t\left(\theta\mathbf{f}^{n+1}+(1-\theta)\mathbf{f}^n\right) \tag{3.66}$$

où la matrice carrée M d'ordre ml est une matrice tridiagonale par blocs donnée par

$$M = \begin{pmatrix} D_m & E_m & & 0 \\ E_m & \ddots & \ddots & \\ & \ddots & \ddots & E_m \\ 0 & & E_m & D_m \end{pmatrix}. \tag{3.67}$$

Les matrices carrées D_m et E_m d'ordre m sont données par

$$D_m = \begin{pmatrix} d_1 & d_2 & & 0 \\ d_2 & \ddots & \ddots & \\ & \ddots & \ddots & d_2 \\ 0 & & d_2 & d_1 \end{pmatrix}, \quad E_m = -\frac{\gamma}{h_y^2} I_m \tag{3.68}$$

avec $d_1 = 2\gamma\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right)$ et $d_2 = -\frac{\gamma}{h_x^2}$. La matrice I_m représente la matrice identité d'ordre m.

Stabilité.

On pose $\beta = \gamma\Delta t\left(\frac{1}{h_x^2} + \frac{1}{h_y^2}\right)$.

i) Pour $0 \leq \theta < 1/2$, le schéma est L^2-stable si $\beta \leq \dfrac{1}{2-4\theta}$.

ii) Pour $1/2 \leq \theta \leq 1$, le schéma est inconditionnellement L^2-stable.

iii) Pour $0 \leq \theta \leq 1$, le schéma est L^∞-stable si $\beta \leq \dfrac{1}{2(1-\theta)}$.

La résolution du système (3.66) est assez coûteuse, la matrice M n'étant pas tridiagonale. On aimerait garder des systèmes linéaires tridiagonaux à résoudre. C'est ce qui motive l'emploi de la méthode suivante dite des directions alternées.

3.3.2 Directions alternées

Cette méthode est aussi appelée *méthode des pas fractionnaires*. L'idée de cette méthode est d'introduire des calculs intermédiaires - pour passer de $\mathbf{u}^n$ à $\mathbf{u}^{n+1}$ - qui ne nécessitent la résolution que de problèmes monodimensionnels. Prenons l'exemple suivant.

$$\begin{array}{ll} \dfrac{\partial u}{\partial t} + Lu = 0 & \text{dans } \Omega \times (0,T) \\ u = 0 & \text{sur } \partial\Omega \times (0,T) \\ u(0) = u_0 & \text{dans } \Omega \end{array}$$

où l'opérateur L est indépendant du temps et se décompose sous la forme

$$L = L_1 + L_2. \tag{3.69}$$

On note A_1 et A_2 les matrices associées aux opérateurs approchés de L_1 et L_2. Le schéma aux directions alternées s'écrit

$$\frac{2}{\Delta t}\left(\mathbf{u}^{n+1/2} - \mathbf{u}^n\right) + A_1\mathbf{u}^{n+1/2} + A_2\mathbf{u}^n = 0 \tag{3.70}$$

$$\frac{2}{\Delta t}\left(\mathbf{u}^{n+1} - \mathbf{u}^{n+1/2}\right) + A_1\mathbf{u}^{n+1/2} + A_2\mathbf{u}^{n+1} = 0 \tag{3.71}$$

Le schéma précédent se réécrit sous la forme

$$\begin{array}{rcl}\left(I_d+\dfrac{\Delta t}{2}A_1\right)\mathbf{u}^{n+1/2} &=& \left(I_d-\dfrac{\Delta t}{2}A_2\right)\mathbf{u}^n\\ \left(I_d+\dfrac{\Delta t}{2}A_2\right)\mathbf{u}^{n+1} &=& \left(I_d-\dfrac{\Delta t}{2}A_1\right)\mathbf{u}^{n+1/2}.\end{array} \tag{3.72}$$

Par exemple, avec $L_1=-\partial^2/\partial x^2$ et $L_2=-\partial^2/\partial y^2$, on se ramène à résoudre deux systèmes tridiagonaux correspondants chacun à un problème monodimensionnel.

Consistance.

En ajoutant (3.70) à (3.71), on obtient

$$\frac{1}{\Delta t}\left(\mathbf{u}^{n+1}-\mathbf{u}^n\right)+A_1\mathbf{u}^{n+1/2}+A_2\left(\frac{\mathbf{u}^{n+1}+\mathbf{u}^n}{2}\right)=0 \tag{3.73}$$

et en retranchant (3.70) à (3.71), on a

$$\mathbf{u}^{n+1}-2\mathbf{u}^{n+1/2}+\mathbf{u}^n+\frac{\Delta t}{2}A_2\left(\mathbf{u}^{n+1}-\mathbf{u}^n\right)=0. \tag{3.74}$$

En éliminant $\mathbf{u}^{n+1/2}$, on obtient

$$\frac{1}{\Delta t}\left(\mathbf{u}^{n+1}-\mathbf{u}^n\right)+(A_1+A_2)\left(\frac{\mathbf{u}^{n+1}+\mathbf{u}^n}{2}\right)+\frac{\Delta t^2}{4}A_1A_2\left(\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}\right)=0. \tag{3.75}$$

Cette dernière relation montre que le schéma est d'ordre $\mathcal{O}(\Delta t^2)$ par rapport à la discrétisation en temps.

Stabilité.

En éliminant $\mathbf{u}^{n+1/2}$ dans les relations (3.72), le système donnant $\mathbf{u}^{n+1}$ directement en fonction de $\mathbf{u}^n$ s'écrit

$$\mathbf{u}^{n+1}=\left(I_d+\frac{\Delta t}{2}A_2\right)^{-1}\left(I_d-\frac{\Delta t}{2}A_1\right)\left(I_d+\frac{\Delta t}{2}A_1\right)^{-1}\left(I_d-\frac{\Delta t}{2}A_2\right)\mathbf{u}^n \tag{3.76}$$

Si les matrices A_1 et A_2 commutent([1]) alors d'après le Lemme 3.1 *(a)*, on peut écrire

$$\mathbf{u}^{n+1}=B_1B_2\mathbf{u}^n \tag{3.77}$$

avec

$$B_1=\left(I_d-\frac{\Delta t}{2}A_1\right)\left(I_d+\frac{\Delta t}{2}A_1\right)^{-1}\text{ et }B_2=\left(I_d-\frac{\Delta t}{2}A_2\right)\left(I_d+\frac{\Delta t}{2}A_2\right)^{-1}. \tag{3.78}$$

Si les matrices A_1 et A_2 sont symétriques, alors d'après le Lemme 3.1 *(c)*, les matrices B_1 et B_2 sont symétriques. Si de plus, les matrices A_1 et A_2 sont semi-définies positives alors on a

$$\|B_1\|_2=\rho(B_1)\le 1\text{ et }\|B_2\|_2=\rho(B_2)\le 1, \tag{3.79}$$

ce qui établit la stabilité inconditionnelle au sens L^2.

Remarque 1. Cas de matrices qui ne commutent pas.

Si on considère l'opérateur $Lu=-\mathrm{div}\,(A(x,y)\nabla u)$ avec $A=\begin{pmatrix}a_{11}&0\\0&a_{22}\end{pmatrix}$ et la décomposition $L_1u=-\dfrac{\partial}{\partial x}(a_{11}(x,y)\dfrac{\partial u}{\partial x})$ et $L_2u=-\dfrac{\partial}{\partial y}(a_{22}(x,y)\dfrac{\partial u}{\partial y})$, alors les matrices associées A_1 et A_2 ne commutent pas

1. C'est le cas avec les opérateurs $L_1=-\partial^2/\partial x^2$ et $L_2=-\partial^2/\partial y^2$ avec une condition de Dirichlet sur le bord.

en général. Elles sont en revanche symétriques et semi-définies positives. Dans ce cas, le schéma aux directions alternées (3.70),(3.71) est L^2-stable sous la condition qu'il existe une constante $C > 0$ indépendante de Δt et h telle que

$$\Delta t \|A_2\|_2 \leq C. \tag{3.80}$$

En effet, si on pose $\mathbf{v}^{n+1} = (I_d + \frac{\Delta t}{2} A_2)\mathbf{u}^{n+1}$, alors (3.72) entraine

$$\mathbf{v}^{n+1} = B_1 B_2 \mathbf{v}^n, \tag{3.81}$$

sans que l'on est besoin de savoir que A_1 et A_2 commutent. Les matrices B_1 et B_2 sont symétriques et $\rho(B_1), \rho(B_2) \leq 1$, ce qui implique

$$\|\mathbf{v}^{n+1}\|_2 \leq \|\mathbf{v}^n\|_2. \tag{3.82}$$

On a donc une stabilité inconditionnelle pour la norme $\|\mathbf{u}\|_{A_2} = \|(I_d + \frac{\Delta t}{2} A_2)\mathbf{u}\|_2$. Par ailleurs,

$$\begin{aligned} \|\mathbf{u}\|_2 &= \|(I_d + \frac{\Delta t}{2} A_2)^{-1}(I_d + \frac{\Delta t}{2} A_2)\mathbf{u}\|_2 \\ &\leq \underbrace{\|(I_d + \frac{\Delta t}{2} A_2)^{-1}\|_2}_{\leq 1} \|\mathbf{u}\|_{A_2} \\ &\leq \|\mathbf{u}\|_{A_2}. \end{aligned}$$

De plus,

$$\|\mathbf{u}\|_{A_2} \leq \|I_d + \frac{\Delta t}{2} A_2\|_2 \|\mathbf{u}\|_2 \leq (1 + \frac{\Delta t}{2}\|A_2\|_2)\|\mathbf{u}\|_2 \leq (1 + C)\|\mathbf{u}\|_2,$$

sous la condition $\Delta t \|A_2\|_2 \leq C$. Par conséquent les normes $\|\cdot\|_{A_2}$ et $\|\cdot\|_2$ sont équivalentes avec des constantes d'équivalence indépendantes de Δt et h. On déduit alors de (3.82) que le schéma est L^2-stable.

Remarque 2. Avec l'exemple précédent, on a $\|A_2\|_2 = \mathcal{O}(1/h^2)$. La condition de stabilité s'écrit alors $\Delta t/h^2 \leq C$. Dans la pratique, cette condition n'est pas restrictive car pour Δt et h fixés on peut toujours trouver C ...

Coût de calcul de la méthode des Directions Alternées.

Supposons qu'on veuille résoudre l'équation de la chaleur sur un carré avec la méthode d'Euler implicite. La résolution par une méthode de Gauss (sans stratégie de pivot) du système linéaire associé, nécessite pour une itération en temps, un nombre d'opérations de l'ordre de (*ordre de la matrice*) $\times$ (*largeur de bande*)^2. Dans le cas de la matrice tridiagonale bloc (3.67), la complexité pour une itération est en $\mathcal{O}(N^4)$ (avec un pas uniforme). Avec la méthode des Directions Alternées (3.72), les matrices sont tridiagonales (largeur de bande = 3) et par conséquent le nombre d'opérations pour une itération en temps est en $\mathcal{O}(N^2)$.

Directions alternées pour les problèmes stationnaires.

On peut utiliser le principe des directions alternées pour résoudre des problèmes stationnaires elliptiques. Considérons par exemple le problème

$$\begin{array}{ll} Lu(\mathbf{x}) = f(\mathbf{x}) & \text{pour } \mathbf{x} \in \Omega \\ u = 0 & \text{sur } \partial\Omega \end{array} \tag{3.83}$$

où l'opérateur elliptique L et la fonction f sont indépendants du temps. On suppose que L se décompose en $L = L_1 + L_2$. La solution de (3.83) peut être vue comme l'état stationnaire atteint par la solution du problème d'évolution

$$\begin{array}{ll} \frac{\partial u}{\partial t}(\mathbf{x}, t) + Lu(\mathbf{x}, t) = f(\mathbf{x}) & \text{dans } \Omega \times (0, T) \\ u = 0 & \text{sur } \partial\Omega \times (0, T) \\ u(0) = u_0 & \text{dans } \Omega \end{array} \tag{3.84}$$

On peut alors résoudre (3.84) avec $t \to +\infty$ pour obtenir une solution approchée de (3.83). Pour résoudre (3.84), on peut utiliser le schéma aux directions alternées (3.70),(3.71) avec les seconds membres remplacés par $\mathbf{f}$:

$$\begin{aligned} \frac{2}{\Delta t}\left(\mathbf{u}^{n+1/2} - \mathbf{u}^n\right) + A_1\mathbf{u}^{n+1/2} + A_2\mathbf{u}^n &= \mathbf{f} \\ \frac{2}{\Delta t}\left(\mathbf{u}^{n+1} - \mathbf{u}^{n+1/2}\right) + A_1\mathbf{u}^{n+1/2} + A_2\mathbf{u}^{n+1} &= \mathbf{f}. \end{aligned} \tag{3.85}$$

On peut choisir $\Delta t \simeq \dfrac{h}{\sqrt{2}\pi}$ pour qu'à chaque itération, l'erreur entre $\mathbf{u}^n$ et la solution exacte u du problème stationnaire (3.83) soit minimale (cf. [9], Chap.11, §35). On montre que le nombre p d'itérations nécessaire pour diminuer l'erreur initiale d'un nombre de fois k donné est $p \simeq \dfrac{N}{\pi\sqrt{2}} \ln k = \mathcal{O}(N \ln k)$, ce qui fait un nombre total d'opérations en $\mathcal{O}(N^3 \ln k)$ (au lieu de $\mathcal{O}(N^4)$, si on résout directement le système avec une matrice tridiagonale par blocs). Dans la pratique, on fait autant d'itérations nécessaires dans (3.85) pour que la différence entre deux solutions $\mathbf{u}^{n+1}$ et $\mathbf{u}^n$ soit suffisamment petite.

Chapitre 4

EDP hyperboliques linéaires

4.1 Equation des ondes

Pour un domaine (ouvert connexe) $\Omega \subset \mathbb{R}^n$ borné, de frontière $\partial\Omega$ *régulière* et pour un temps $T > 0$ fixé, on considère le problème aux limites suivant : trouver une fonction $u = u(\mathbf{x}, t)$ avec $\mathbf{x} \in \Omega$ et $t \in (0, T)$, telle que

$$\begin{aligned} \frac{\partial^2 u}{\partial t^2} - c^2 \Delta u &= f \quad \text{dans } \Omega \times (0,T) && (4.1)\\ u &= 0 \quad \text{sur } \partial\Omega \times (0,T) && (4.2)\\ u(\mathbf{x}, 0) &= u_0(\mathbf{x}) \quad \text{dans } \Omega && (4.3)\\ \frac{\partial u}{\partial t}(\mathbf{x}, 0) &= v_0(\mathbf{x}) \quad \text{dans } \Omega. && (4.4) \end{aligned}$$

A l'instant $t = 0$, on impose la valeur de u ainsi que la vitesse initiale $\partial u/\partial t$ par la donnée de deux fonctions u_0 et v_0. Le paramètre $c \in \mathbb{R}$ représente la célérité des ondes dans le milieu considéré. L'équation des ondes est le prototype d'équation hyperbolique linéaire. Cette équation est bien hyperbolique (dans $\mathbb{R}^n \times \mathbb{R}$) au sens de la classification faite au Chapitre 1.

4.1.1 Existence, unicité et propriétés des solutions

Commençons par un résultat d'existence et d'unicité :

Si $f \in C([0,T], L^2(\Omega))$, $u_0 \in H^2(\Omega) \cap H_0^1(\Omega)$ et $v_0 \in H_0^1(\Omega)$ alors il existe une unique solution de (4.1)-(4.4) telle que

$$u \in C([0,T], H^2(\Omega) \cap H_0^1(\Omega)) \ \cap \ C^1([0,T], H_0^1(\Omega)) \ \cap \ C^2([0,T], L^2(\Omega)) \tag{4.5}$$

vérifiant les équations (4.1)-(4.4) *presque partout* dans Ω et pour tout $t \in [0,T]$. De plus,

- il existe une constante $C > 0$ telle que

$$\|\frac{\partial u}{\partial t}(t)\|^2_{L^2(\Omega)} + \|\nabla u(t)\|^2_{L^2(\Omega)} \leq C \left(\|v_0\|^2_{L^2(\Omega)} + \|\nabla u_0\|^2_{L^2(\Omega)} + \int_0^T \|f\|^2_{L^2(\Omega)} \right), \tag{4.6}$$

  pour tout $t \in [0,T]$.

- si $f \equiv 0$ on a

$$\|\frac{\partial u}{\partial t}(t)\|^2_{L^2(\Omega)} + c^2\|\nabla u(t)\|^2_{L^2(\Omega)} = \|v_0\|^2_{L^2(\Omega)} + c^2\|\nabla u_0\|^2_{L^2(\Omega)}, \tag{4.7}$$

  pour tout $t \in [0,T]$.

La relation (4.7) est une loi de conservation qui traduit la conservation d'une énergie au cours du temps.

Contrairement à l'équation de la chaleur, l'équation des ondes n'a aucun effet régularisant sur les données initiales. Par exemple, avec $\Omega = \mathbb{R}$ et $f \equiv 0$, la solution du problème

$$\begin{array}{rcll} u_{tt} - c^2 u_{xx} & = & f(x,t) & \text{dans } \mathbb{R} \times (0,T) \\ u(x,0) & = & u_0(x) & \text{dans } \mathbb{R} \\ u_t(x,0) & = & v_0(x) & \text{dans } \mathbb{R}, \end{array} \tag{4.8}$$

est donnée par

$$u(x,t) = \frac{1}{2}\left[u_0(x+ct) + u_0(x-ct)\right] + \frac{1}{2}\int_{x-ct}^{x+ct} v_0(s)\, ds. \tag{4.9}$$

Avec $v_0 = 0$, on voit que la solution n'est pas plus régulière que u_0. Si $u_0 \in C^\infty$ sauf en x_0 alors u est C^∞ sur $\mathbb{R}\times\mathbb{R}$ sauf sur les droites $x+ct = x_0$ et $x-ct = x_0$. Ces droites sont appelées les caractéristiques issues du point $(x_0, t=0)$. Les singularités se propagent le long des caractéristiques. Par ailleurs, si u_0 est continue en x_0 alors la solution est constante le long des caractéristiques.

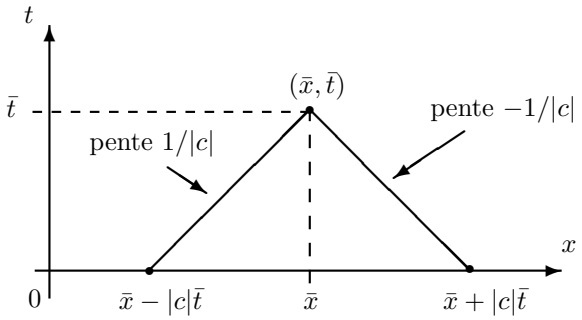


FIGURE 4.1 – Caractéristiques de l'équation des ondes 1D passant par $(\bar{x},\bar{t})$ et intervalle de dépendance correspondant.

La formule (4.9) montre que $u(\bar{x},\bar{t})$ dépend uniquement de u_0 et v_0 dans l'intervalle $[\bar{x}-|c|\bar{t},\ \bar{x}+|c|\bar{t}]$ qui définit le domaine de dépendance du point $(\bar{x},\bar{t})$ (voir Figure 4.1). On a vu que l'équation de la chaleur se propage avec une vitesse infinie. Pour l'équation des ondes, la vitesse de propagation est finie et vaut c : un signal localisé en x_0 se fait sentir en $\bar{x}$ à partir du temps $\bar{t} \geq |\bar{x}-x_0|/|c|$ (voir Figure 4.2).

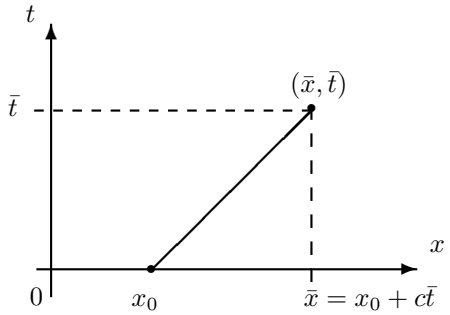


FIGURE 4.2 – Propagation d'un signal localisé en x_0 ($c>0$).

En dimension supérieure, le domaine de dépendance est $D = \{\mathbf{x}\in\mathbb{R}^n,\ |\mathbf{x}-\bar{\mathbf{x}}| \leq |c|\bar{t}\,\} \times \{t=0\}$. Lorsque $n>1$ est *impair*, la valeur $u(\bar{x},\bar{t})$ dépend uniquement des valeurs de u_0 et v_0 sur la sphère $\{\mathbf{x}\in\mathbb{R}^n,\ |\mathbf{x}-\bar{\mathbf{x}}| = |c|\bar{t}\,\}$ (*Principe d'Huygens*).

Principe du maximum. Donnons quelques formes très particulières du principe du maximum pour l'équation des ondes. Avec $\underline{f \equiv 0}$:

- si $\Omega = \mathbb{R}$ alors $(u_0 \geq 0,\ v_0 \geq 0) \Rightarrow u \geq 0$.
- si $\Omega = \mathbb{R}^2$ alors $(u_0 = 0,\ v_0 \geq 0) \Rightarrow u \geq 0$.

Mais :

- si $\Omega =]0,1[$ alors $(u_0 \geq 0,\ v_0 = 0) \not\Rightarrow u \geq 0$.
- si $\Omega = \mathbb{R}^2$ alors $(u_0 \geq 0,\ v_0 = 0) \not\Rightarrow u \geq 0$.

Pour l'équation des ondes, il n'y a donc pas de principe général du maximum.

4.1.2 Un θ-schéma centré pour l'équation des ondes (1D)

On considère l'équation des ondes en dimension 1 d'espace.

$$\begin{aligned} u_{tt} - c^2 u_{xx} &= f(x,t), \quad \text{pour } (x,t) \in (0,1)\times(0,T) && (4.10)\\ u(0,t) &= u(1,t) = 0, \quad t \in (0,T) && (4.11)\\ u(x,0) &= u_0(x), && (4.12)\\ u_t(x,0) &= v_0(x), \quad \text{pour } x \in (0,1). && (4.13) \end{aligned}$$

On discrétise $[0,1]\times[0,T]$ en introduisant les points $x_j = j\Delta x$ pour $j = 0,\cdots,m+1$ et les instants $t^n = n\Delta t$. On cherche alors $u_j^n \simeq u(x_j,t^n)$ pour $j = 1,\cdots,m$ (on a $u_0^n = u_{m+1}^n = 0$ compte tenu des conditions limites (4.11)).

Le θ-schéma centré en temps et en espace s'écrit, avec $0 \leq \theta \leq 1$ et pour $n \geq 1$:

$$\frac{u_j^{n+1} - 2u_j^n + u_j^{n-1}}{\Delta t^2} - \frac{c^2}{\Delta x^2}\left(\frac{\theta}{2}\delta_x^2 u_j^{n+1} + (1-\theta)\delta_x^2 u_j^n + \frac{\theta}{2}\delta_x^2 u_j^{n-1}\right) = \frac{\theta}{2}\left(f_j^{n+1} + f_j^{n-1}\right) + (1-\theta) f_j^n, \quad (4.14)$$

où $\delta_x v_j = v_{j+1} - 2v_j - v_{j-1}$.

Les données initiales u_0 et v_0 permettent de définir (u_j^0) et (u_j^1). On choisit

$$u_j^0 = u_0(x_j) \ \text{ pour } j = 1,\cdots,m. \quad (4.15)$$

Pour déterminer u_j^1 on procède de la façon suivante. On a utilisé une différence centrée pour la dérivée (seconde) en temps dans (4.10) ; on espère ainsi un ordre de consistance en $\mathcal{O}(\Delta t^2)$ et on souhaite conserver cet ordre en approchant la condition initiale (4.13) : on écrit

$$\frac{u(x_j,t^1) - u(x_j,t^0)}{\Delta t} = v_0(x_j) + \frac{\Delta t}{2} u_{tt}(x_j,t^0) + \mathcal{O}(\Delta t^2) \quad (4.16)$$

En utilisant (4.10), on obtient

$$\begin{aligned} \frac{u(x_j,t^1) - u(x_j,t^0)}{\Delta t} &= v_0(x_j) + \frac{\Delta t}{2}\left(c^2 u_{xx}(x_j,t^0) + f(x_j,t^0)\right) + \mathcal{O}(\Delta t^2)\\ &= v_0(x_j) + \frac{\Delta t}{2}\left(c^2\,\frac{\delta_x^2 u(x_j,t^0)}{\Delta x^2} + \mathcal{O}(\Delta x^2) + f(x_j,t^0)\right) + \mathcal{O}(\Delta t^2). \end{aligned}$$

L'équation pour déterminer (u_j^1) s'écrit donc (pour $j = 1,\cdots,m$) :

$$\frac{u_j^1 - u_j^0}{\Delta t} = v_0(x_j) + \frac{\Delta t}{2}\left(f_j^0 + \frac{c^2}{\Delta x^2}\delta_x^2 u_j^0\right). \quad (4.17)$$

Cette approximation de (4.13) est d'ordre 2 en Δt et Δx.

Remarques. Pour $\theta = 0$, le schéma est explicite, sinon il est implicite et il faut résoudre un système linéaire à chaque pas de temps.

En notant $\mathbf{u}^n = (u_1^n, \cdots, u_m^n)$, le θ-schéma (4.14) s'écrit sous forme matricielle, pour $n \geq 1$:

$$\frac{\mathbf{u}^{n+1} - 2\mathbf{u}^n + \mathbf{u}^{n-1}}{\Delta t^2} + \frac{c^2}{\Delta x^2} A \left(\frac{\theta}{2}\mathbf{u}^{n+1} + (1-\theta)\mathbf{u}^n + \frac{\theta}{2}\mathbf{u}^{n-1} \right) = \frac{\theta}{2}\left(\mathbf{f}^{n+1} + \mathbf{f}^{n-1}\right) + (1-\theta)\mathbf{f}^n, \quad (4.18)$$

où A est la matrice tridiagonale (3.20) de taille $m \times m$, formée des 2 et des -1.

Consistance.

L'erreur globale de consistance (i.e. pour (4.14) et (4.17)) est en $\mathcal{O}(\Delta t^2 + \Delta x^2)$.

Stabilité.

On pose

$$\lambda = c\frac{\Delta t}{\Delta x}. \quad (4.19)$$

On étudie la stabilité du θ-schéma pour $f \equiv 0$.

⋆ **Critère de Von Neumann - Fourier.**

On cherche la solution du θ-schéma sous la forme $u_j^n = \xi^n \mathrm{e}^{ik\pi x_j}$ $(n > 1)$. En injectant cette expression dans (4.14), on obtient

$$\xi^2 - 2\xi + 1 = \lambda^2 \left(\frac{\theta}{2}\xi^2 + (1-\theta)\xi + \frac{\theta}{2} \right) (2(\underbrace{\cos k\pi\Delta x - 1}_{=-2\sin^2(k\pi\Delta x/2)})),$$

soit encore

$$\xi^2 - 2\left(\frac{1 - 2(1-\theta)\lambda^2 \sin^2(k\pi\Delta x/2)}{1 + 2\theta\lambda^2 \sin^2(k\pi\Delta x/2)} \right) \xi + 1 = 0.$$

Le produit des racines ξ_1 et ξ_2 vaut 1. Par conséquent, on a $|\xi_1|,\ |\xi_2| \leq 1$ si et seulement si les racines sont complexes conjuguées i.e. $\xi_1 = \bar{\xi_2}$ (dans ce cas $|\xi_1| = |\xi_2| = 1$) c'est-à-dire si le discriminant est négatif :

$$\left(\frac{1 - 2(1-\theta)\lambda^2 \sin^2(k\pi\Delta x/2)}{1 + 2\theta\lambda^2 \sin^2(k\pi\Delta x/2)} \right)^2 - 1 \leq 0$$

Cette condition est équivalente à :

$$(1 - 2\theta)\lambda^2 \sin^2(k\pi\Delta x/2) \leq 1. \quad (4.20)$$

Ainsi, on en déduit les conclusions suivantes sur la stabilité du θ-schéma.

- Pour $\theta \geq 1/2$, le θ-schéma est inconditionnellement stable (au sens de Von Neumann).
- Pour $\theta < 1/2$, le θ-schéma est stable (au sens de Von Neumann) si la condition suivante dite *CFL* (Courant-Friedrichs-Levy) est satisfaite :

$$|\lambda| = |c|\frac{\Delta t}{\Delta x} \leq (1 - 2\theta)^{-1/2}. \quad (4.21)$$

⋆ **Stabilité L^2.**

Si $f \equiv 0$, on montre [1] que

$$E^{n+1} = E^n \quad (4.22)$$

avec

$$E^{n+1} = \| \frac{\mathbf{u}^{n+1} - \mathbf{u}^n}{\Delta t} \|_{2,\Delta x}^2 + c^2 (M\mathbf{u}^n, \mathbf{u}^{n+1})_{\Delta x} + \frac{\theta c^2}{2} \left(M(\mathbf{u}^{n+1} - \mathbf{u}^n), \mathbf{u}^{n+1} - \mathbf{u}^n \right)_{\Delta x} \quad (4.23)$$

où $M = A/\Delta x^2$. La norme $\| \cdot \|_{2,\Delta x}$ est la norme L^2-discrète définie par (3.28) et le produit scalaire correspondant vaut $(\mathbf{v}, \mathbf{w})_{\Delta x} = \Delta x \sum_j v_j w_j$. A partir de cette relation de conservation, on montre les résultats de stabilité suivants :

1. Il faut multiplier l'équation (4.18) par $\mathbf{u}^{n+1} - \mathbf{u}^{n-1}$ et utiliser le fait que $(M\mathbf{u}, \mathbf{v}) = (\mathbf{u}, M\mathbf{v})$

- Si $\theta > 1/2$ ou bien si $\theta < 1/2$ avec la condition CFL : $|\lambda| < (1-2\theta)^{-1/2}$, alors il existe une constante $C > 0$ indépendante de Δt et Δx telle que

$$\|\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}\|_{2,\Delta x} \leq C, \quad (M\mathbf{u}^n, \mathbf{u}^n)_{\Delta x} \leq C.$$

- Si $\theta = 1/2$, il existe une constante $C > 0$ indépendante de Δt et Δx telle que

$$\|\frac{\mathbf{u}^{n+1}-\mathbf{u}^n}{\Delta t}\|_{2,\Delta x} \leq C, \quad \left(M(\frac{\mathbf{u}^{n+1}+\mathbf{u}^n}{2}), \frac{\mathbf{u}^{n+1}+\mathbf{u}^n}{2}\right)_{\Delta x} \leq C.$$

La relation (4.22) est l'analogue discret de la loi de conservation continue (4.7). Cette loi de conservation discrète indique qu'il n'y a pas de diffusion numérique (dissipation) du θ-schéma.

Dispersion.

Considérons le schéma explicite avec $\theta = 0$. On choisit $f = 0$, $v_0 = 0$ et $u_0(x) = \sin(k\pi x)$. La solution exacte est donnée par $u(x,t) = \cos(k\pi ct)\sin(k\pi x)$ et la solution numérique par $u_j^n = \xi_{k,n}\sin(k\pi x_j)$. Si on compare $\xi_{k,n}$ et $\cos(k\pi cn\Delta t)$, on constate une dispersion c'est-à-dire un déphasage d'amplitude au cours du temps, qui est d'autant plus grande que n est grand (voir Figure 4.3).

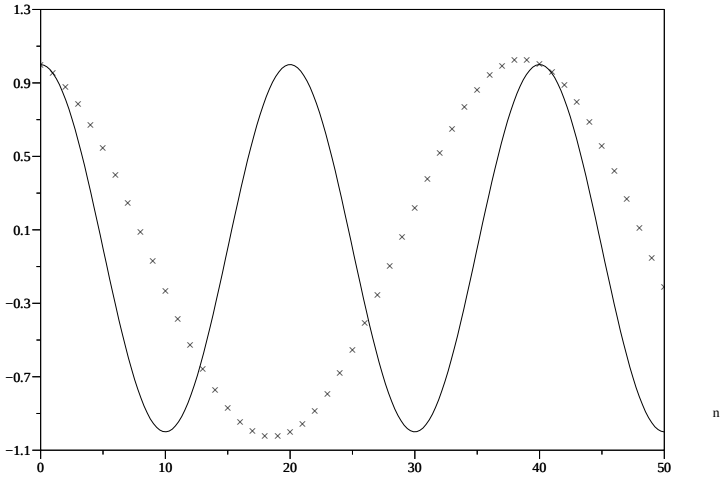


FIGURE 4.3 – Dispersion du θ-schéma ($\theta = 0$) pour l'équation des ondes ($\times$: $\xi_{k,n}$; $-$: $\cos(k\pi cn\Delta t)$).

4.2 Equation de transport

4.2.1 Introduction

Pour l'équation des ondes en 1D, on pose $\mathbf{v}(x,t) = (u_t(x,t),\ u_x(x,t))^\top$ et on a $\mathbf{v}_t(x,t) = \begin{pmatrix} 0 & c \\ 1 & 0 \end{pmatrix} \mathbf{v}_x(x,t)$. On se ramène ainsi à un système d'équations du premier ordre. Il s'agit de la généralisation aux systèmes, de l'équation de transport en dimension 1 d'espace. Ce problème de transport consiste à chercher $u : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ vérifiant

$$(P) \begin{cases} u_t(x,t) + cu_x(x,t) &= f(x,t), \quad x \in \mathbb{R}, t > 0 \\ u(x,0) &= u_0(x). \end{cases}$$

On suppose pour fixer les idées que le paramètre c est positive.

Cas $f \equiv 0$

⋆ Caractéristiques et solution de l'équation de transport.

On introduit les trajectoires $t \mapsto X(t)$ telles que $X'(t) = c$. Ces trajectoires sont des demi-droites et sont les *caractéristiques* de l'équation de transport. Toute solution *régulière* u de (P) est constante le long des caractéristiques. En effet, on a

$$\frac{d}{dt}u(X(t),t) = u_x(X(t),t)X'(t) + u_t(X(t),t) = (cu_x + u_t)(X(t),t) = 0. \tag{4.24}$$

Les caractéristiques permettent de trouver la solution exacte de (P). Soit (x,t) quelconque. On introduit X la caractéristique passant par ce point (cf. Figure 4.4) :

$$\begin{array}{rcl} \dfrac{dX}{ds}(s) & = & c, \quad s \geq 0 \\ X(t) & = & x. \end{array} \tag{4.25}$$

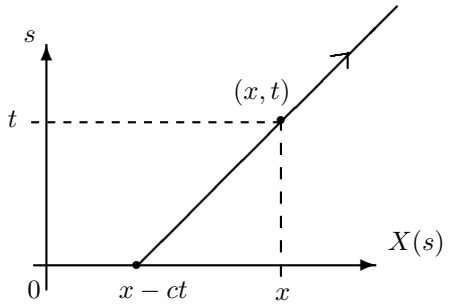


FIGURE 4.4 – Caractéristique de l'équation de transport ($c > 0$), passant par un point (x,t) donné.

En intégrant (4.25), on a clairement $X(s) = x + c(s-t)$ pour tout $s \geq 0$. En particulier, $X(0) = x - ct$ et comme u (régulière) est constante le long de X, on en déduit $u(X(t),t) = u(X(0),0)$, ce qui donne la solution

$$u(x,t) = u_0(x - ct). \tag{4.26}$$

La solution exacte de (P) est donc obtenue en transportant la donnée initiale u_0 avec une vitesse $c > 0$ (cf. Figure 4.5).

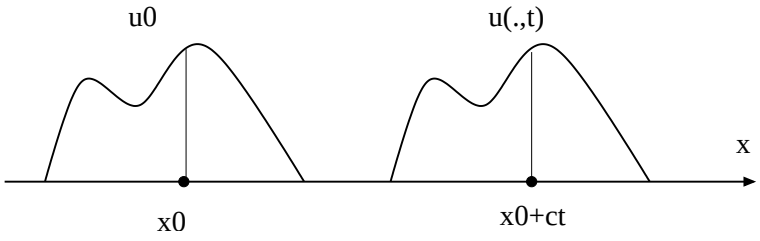


FIGURE 4.5 – Solution de l'équation de transport ($c > 0$).

Si la donnée initiale u_0 est discontinue en x_0, alors la discontinuité se propage le long des caractéristiques issues de x_0. Dans ce cas, la notion de solution qui intervient est celle de *solution faible* (voir le chapitre suivant sur les lois de conservation).

⋆ Domaine borné et conditions aux limites.

Si on pose le problème (P) dans un intervalle d'espace $[a,b]$ au lieu de $\mathbb{R}$ tout entier, il faut imposer une valeur à u au *point rentrant*[1] c'est-à-dire en $x = a$ si $c > 0$ ou en $x = b$ si $c < 0$.

1. et seulement au point rentrant !

Cas $f \not\equiv 0$

On peut encore utiliser les caractéristiques pour résoudre (P). Fixons (x,t) quelconque et considérons X la caractéristique passant par ce point i.e. vérifiant (4.25). Au lieu de (4.24), on a maintenant

$$\frac{d}{ds}u(X(s),s) = f(X(s),s).$$

En intégrant la relation précédente entre 0 et t, on obtient

$$u(X(t),t) = u(X(0),0) + \int_0^t f(X(s),s)\,ds,$$

soit encore

$$u(x,t) = u_0(x-ct) + \int_0^t f(x+c(s-t),s)\,ds. \tag{4.27}$$

4.2.2 Schéma centré

On discrétise $\mathbb{R}\times\mathbb{R}^+$ en introduisant les points $x_j = j\Delta x$ pour $j\in\mathbb{Z}$ et les instants $t^n = n\Delta t$. On cherche alors une approximation $u_j^n \simeq u(x_j,t^n)$. Le schéma centré s'écrit

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = f_j^n, \qquad \text{pour } n\geq 1 \tag{4.28}$$

$$u_j^0 = u_0(x_j), \tag{4.29}$$

avec $f_j^n = f(x_j,t^n)$. Il s'agit d'un schéma explicite qui peut s'écrire

$$u_j^{n+1} = u_j^n + \frac{\lambda}{2}(u_{j-1}^n - u_{j+1}^n) + \Delta t f_j^n \tag{4.30}$$

avec

$$\lambda = c\frac{\Delta t}{\Delta x}. \tag{4.31}$$

$\star$ Cône de dépendance numérique.

Le cône de dépendance numérique d'un point $M=(x_j,t^n)$ doit contenir la caractéristique passant par M (cf. Figure 4.6).

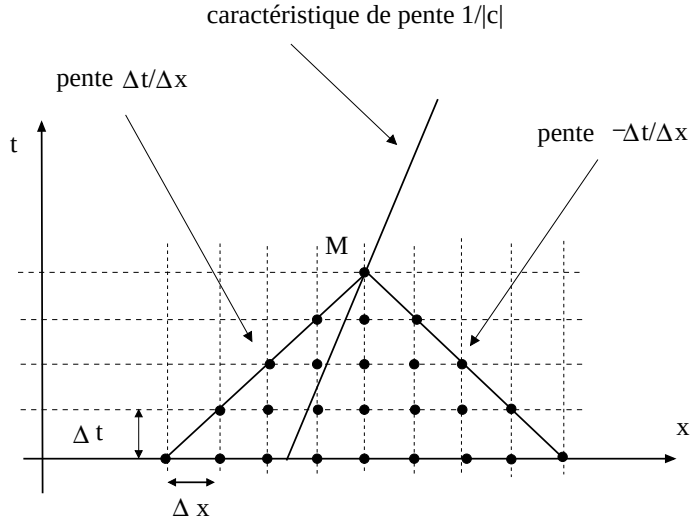


FIGURE 4.6 – Cône de dépendance numérique du schéma centré pour l'équation de transport.

Cette condition nécessaire se traduit par $-\dfrac{\Delta t}{\Delta x} \leq \dfrac{1}{|c|} \leq \dfrac{\Delta t}{\Delta x}$, c'est-à-dire

$$|\lambda| = |c|\frac{\Delta t}{\Delta x} \leq 1. \tag{4.32}$$

⋆ **Analyse de Von Neumann/Fourier.**

On prend $f \equiv 0$ et on choisit la donnée initiale $u_0(x) = \xi_0 \mathrm{e}^{ikx}$. Le paramètre k représente le nombre d'onde et $k = 2\pi/l$ où l désigne la longueur d'onde du mode retenu. On a $u_j^0 = \xi_0 \mathrm{e}^{ikj\Delta x}$ et on cherche la solution du schéma sous la forme

$$u_j^n = \xi_n \mathrm{e}^{ikj\Delta x}. \tag{4.33}$$

En injectant cette expression dans la relation (4.28) du schéma centré, on obtient

$$\xi_{n+1} = \left[1 + \frac{\lambda}{2}(\mathrm{e}^{-ik\Delta x} - \mathrm{e}^{ik\Delta x})\right] \xi_n$$

d'où

$$\xi_{n+1} = (1 - i\lambda \sin(k\Delta x))\, \xi_n. \tag{4.34}$$

Le coefficient d'amplification est donné par

$$g(\lambda, k) = 1 - i\lambda \sin(k\Delta x) \tag{4.35}$$

et on a

$$|g(\lambda, k)|^2 = 1 + \lambda^2 \sin^2(k\Delta x)$$

de sorte que (pour $k\Delta x \notin \pi\mathbb{Z}$), on a $|g(\lambda, k)| > 1$ quelque soit λ. Le schéma centré est donc toujours instable au sens de Von Neumann/Fourier.

4.2.3 Schéma de Lax

On remplace u_j^n dans le schéma centré précédent par la valeur moyenne $\tilde{u_j^n} = (u_{j-1}^n + u_{j+1}^n)/2$, ce qui donne :

$$\frac{u_j^{n+1} - \tilde{u_j^n}}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = f_j^n. \tag{4.36}$$

Le schéma de Lax peut s'écrire sous la forme

$$\begin{aligned} u_j^{n+1} &= \alpha_1 u_{j-1}^n + \alpha_2 u_{j+1}^n + \Delta t f_j^n, \qquad n \geq 1 && (4.37)\\ u_j^0 &= u_0(x_j), && (4.38) \end{aligned}$$

avec

$$\begin{gathered} \lambda = c\frac{\Delta t}{\Delta x} \\ \alpha_1 = \frac{1}{2}(1+\lambda), \quad \alpha_2 = \frac{1}{2}(1-\lambda). \end{gathered} \tag{4.39}$$

⋆ **Cône de dépendance numérique.**

Le cône de dépendance numérique du schéma de Lax coïncide avec celui du schéma centré (cf. Figure 4.6). Par conséquent on obtient la condition CFL nécessaire

$$|\lambda| \leq 1. \tag{4.40}$$

De plus, si $f \equiv 0$ et si $|\lambda| = 1$, alors on a $u_j^n = u(x_j, t^n)$ pour tout j et n. Dans ce cas, la solution approchée coïncide exactement avec la solution exacte aux points de discrétisation.

⋆ **Analyse de von Neumann/Fourier.**

On prend $f \equiv 0$ et on choisit la donnée initiale $u_j^0 = \xi_0 \mathrm{e}^{ikj\Delta x}$ (le nombre d'onde k est relié à la longueur d'onde l par $k = 2\pi/l$). On cherche la solution du schéma de Lax sous la forme

$$u_j^n = \xi_n \mathrm{e}^{ikj\Delta x}. \tag{4.41}$$

En injectant cette expression dans la relation (4.37) du schéma de Lax, on obtient

$$\begin{aligned} \xi_{n+1} &= \left[\alpha_1 \mathrm{e}^{-ik\Delta x} + \alpha_2 \mathrm{e}^{ik\Delta x}\right) \xi_n \\ &= \left[\frac{1}{2}(\mathrm{e}^{-ik\Delta x} + \mathrm{e}^{ik\Delta x}) + \frac{\lambda}{2}(\mathrm{e}^{-ik\Delta x} - \mathrm{e}^{ik\Delta x})\right] \xi_n \end{aligned}$$

d'où

$$\xi_{n+1} = (\cos(k\Delta x) - i\lambda \sin(k\Delta x))\, \xi_n. \tag{4.42}$$

Le coefficient d'amplification du schéma de Lax est donné par

$$g_L(\lambda, k) = \cos(k\Delta x) - i\lambda \sin(k\Delta x) \tag{4.43}$$

et on a

$$\begin{aligned} |g_L(\lambda, k)|^2 &= \cos^2(k\Delta x) + \lambda^2 \sin^2(k\Delta x) \\ &= 1 + (\lambda^2 - 1)\sin^2(k\Delta x). \end{aligned} \tag{4.44}$$

- Si $|\lambda| > 1$ alors $|g_L(\lambda, k)| > 1$ (pour $k\Delta x \notin \pi\mathbb{Z}$) et le schéma de Lax est <u>instable</u>.
- Si $|\lambda| \leq 1$ alors $|g_L(\lambda, k)| \leq 1$ et le schéma de Lax est <u>stable</u> au sens de Von Neumann/Fourier.

⋆ **Principe du maximum discret.**

Proposition 4.1 *On suppose que* $|\lambda| \leq 1$.

1. *Si* $f \equiv 0$ *et* $u_* \leq u_j^0 \leq u^*$, $\forall j$ *alors* $u_* \leq u_j^n \leq u^*$, $\forall j$, $\forall n$.
2. *(Positivité du schéma de Lax) Si* $f \geq 0$ *et* $u_0 \geq 0$ *alors* $u_j^n \geq 0$, $\forall j$, $\forall n$.

Démonstration de 1. Montrons par récurrence sur n que $u_j^n \leq u^*$. Pour $n = 0$, c'est l'hypothèse. Supposons que $u_j^n \leq u^*$, $\forall j$ et montrons cette propriété pour $\{u_j^{n+1}\}_j$. Si $|\lambda| \leq 1$ alors $\alpha_1 \geq 0$, $\alpha_2 \geq 0$ et $\alpha_1 + \alpha_2 = 1$. La relation (4.37) donne alors pour tout j

$$u_j^{n+1} \leq \alpha_1 u^* + \alpha_2 u^* = u^*.$$

On montre de la même façon que $u_j^n \geq u_*$. □

⋆ **Stabilité L^p.**

On introduit l'espace $l^p_{\Delta x}$ pour $p \in [1, +\infty]$, qui est l'analogue discret de l'espace $L^p(\mathbb{R})$:

- Pour $1 \leq p < +\infty$, on définit l'espace

$$l^p_{\Delta x} = \left\{\mathbf{v} = \{v_j\}_{j\in\mathbb{Z}} \text{ telle que } \Delta x \sum_{j\in\mathbb{Z}} |v_j|^p < +\infty\right\} \tag{4.45}$$

  et pour $\mathbf{v} \in l^p_{\Delta x}$, on note la norme $\|\mathbf{v}\|_{p,\Delta x} = \left(\Delta x \sum_{j\in\mathbb{Z}} |v_j|^p\right)^{1/p}$.

- Pour $p = +\infty$, on définit l'espace

$$l^\infty_{\Delta x} = \left\{\mathbf{v} = \{v_j\}_{j\in\mathbb{Z}} \text{ telle que } \sup_{j\in\mathbb{Z}} |v_j| < +\infty\right\} \tag{4.46}$$

  et pour $\mathbf{v} \in l^\infty_{\Delta x}$, on note la norme $\|\mathbf{v}\|_{\infty,\Delta x} = \sup_{j\in\mathbb{Z}} |v_j|$.

On prend (pour simplifier) $\underline{f \equiv 0}$ et on note $\mathbf{u} = \{u_j^n\}_{j\in\mathbb{Z}}$ la famille définie par le schéma de Lax.

Proposition 4.2 *Soit $\mathbf{u}^0 \in l^p_{\Delta x}$ pour $p \in [1, +\infty]$. Si $|\lambda| \leq 1$ alors $\mathbf{u}^n \in l^p_{\Delta x}$ et*

$$\|\mathbf{u}^n\|_{p,\Delta x} \leq \|\mathbf{u}^0\|_{p,\Delta x}, \text{ pour tout } n \in \mathbb{N}. \tag{4.47}$$

L'égalité a lieu si $|\lambda| = 1$.

Cette proposition découle d'un résultat plus général.

Lemme 4.1 *Soit ϕ une fonction convexe et positive sur $\mathbb{R}$ et soit $\{u_j^0\}_j$ telle que $\{\phi(u_j^0)\}_j \in l^1_{\Delta x}$. Si $|\lambda| \leq 1$ alors $\{\phi(u_j^n)\}_j \in l^1_{\Delta x}$ et*

$$\sum_{j\in\mathbb{Z}} \phi(u_j^n) \leq \sum_{j\in\mathbb{Z}} \phi(u_j^0), \tag{4.48}$$

pour tout $n \in \mathbb{N}$. L'égalité a lieu si $|\lambda| = 1$.

Démonstration du Lemme. Par récurrence sur n. Supposons que l'inégalité (4.48) soit vraie pour $\mathbf{u}^n$ et montrons la pour $\mathbf{u}^{n+1}$. D'après (4.37) et la convexité de ϕ, on a

$$\phi(u_j^{n+1}) = \phi(\alpha_1 u_{j-1}^n + \alpha_2 u_{j+1}^n) \leq \alpha_1 \phi(u_{j-1}^n) + \alpha_2 \phi(u_{j+1}^n)$$

car sous l'hypothèse $|\lambda| \leq 1$, on a $\alpha_1 \geq 0$, $\alpha_2 \geq 0$ et $\alpha_1 + \alpha_2 = 1$. Soit $k \in \mathbb{N}$ fixé. On somme sur $j = -k, \cdots, k$ et compte tenu de la positivité de ϕ, on obtient

$$\begin{aligned} \sum_{j=-k}^{k} \phi(u_j^{n+1}) &\leq \alpha_1 \sum_{j=-k}^{k} \phi(u_{j-1}^n) + \alpha_2 \sum_{j=-k}^{k} \phi(u_{j+1}^n) \\ &\leq \alpha_1 \sum_{j\in\mathbb{Z}} \phi(u_j^n) + \alpha_2 \sum_{j\in\mathbb{Z}} \phi(u_j^n) = \sum_{j\in\mathbb{Z}} \phi(u_j^n). \end{aligned}$$

En faisant tendre $k \to +\infty$, on voit que $\{\phi(u_j^{n+1})\}_j \in l^1_{\Delta x}$ et que l'inégalité (4.48) est vérifiée par $\mathbf{u}^{n+1}$.

Enfin, l'égalité a lieu si $|\lambda| = 1$ car alors $u_j^{n+1} = u_{j\pm1}^n$ et dans ce cas la solution approchée coïncide avec la solution exacte aux points de discrétisation i.e. $u_j^n = u(x_j, t^n)$. □

Démonstration de la Proposition. Pour $p \in [1, +\infty[$, la proposition se démontre en appliquant le lemme avec $\phi(s) = |s|^p$. Pour $p = +\infty$, il suffit d'adapter directement la démonstration du lemme. □

⋆ Consistance.

On note $L = \partial/\partial t + c\partial/\partial x$ l'opérateur de l'équation de transport et $L_{\Delta x,\Delta t}$ l'opérateur approché du schéma de Lax défini par

$$L_{\Delta x,\Delta t} v(x,t) = \frac{v(x,t+\Delta t) - \big(v(x-\Delta x,t) + v(x+\Delta x,t)\big)/2}{\Delta t} + c\,\frac{v(x+\Delta x,t) - v(x-\Delta x,t)}{2\Delta x}.$$

On montre que l'erreur de consistance du schéma de Lax vérifie

$$(L - L_{\Delta x,\Delta t})v(x,t) = \mathcal{O}(\Delta t + \frac{\Delta x}{\lambda}). \tag{4.49}$$

Pour λ fixé, le schéma de Lax est consistant à l'ordre 1 en temps et en espace. En fait, le terme intervenant avec le facteur $\Delta x/\lambda$ est $(\Delta x^2/\Delta t)\, u_{xx}$, ce qui correspond à un terme de diffusion. Ce terme de diffusion (numérique) est d'autant plus grand que λ est petit. On n'a donc pas intérêt à prendre λ trop petit.

⋆ Diffusion du schéma de Lax.

Si $f \equiv 0$ et si $u_0(x) = \mathrm{e}^{ikx}$ alors la solution de l'équation de transport avec u_0 comme donnée initiale, est donnée par $u(x,t) = u_0(x - ct) = \mathrm{e}^{ikx}\mathrm{e}^{-ikct}$ et par conséquent $|u(x,t)| = 1$, pour tout (x,t).

Examinons, comment se comporte la solution correspondante du schéma de Lax. On a $|u_j^n| = |g_L(\lambda,k)|^n |u_j^0|$. Si $|g_L(\lambda,k)| < 1$ alors $\max_j |u_j^n| \to 0$ quand $n \to +\infty$. Il y a donc diffusion (ou amortissement) du k-ième mode. Le module $|g_l(\lambda,k)|$ du facteur d'amplification mesure cette diffusion. Plus $|g_L(\lambda,k)|$ est petit, plus la diffusion du k-ième mode est grande. On a (cf. (4.44))

$$|g_L(\lambda,k)| = \left(1 + (\lambda^2 - 1)\sin^2(k\Delta x)\right)^{1/2}.$$

La Figure 4.7 montre le module $|g_L|$ en fonction de $k\Delta x$ et pour différentes valeurs de λ.

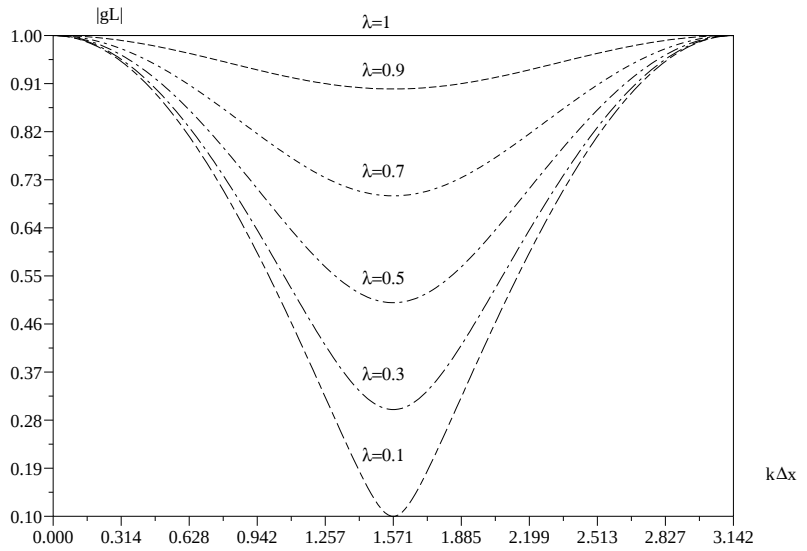


FIGURE 4.7 – Facteur d'amplification $|g_L|$ du schéma de Lax (en fonction de $k\Delta x$).

Plus λ est petit, plus le schéma de Lax diffuse. On a donc intérêt à prendre $|\lambda| \simeq 1$ (et toujours $|\lambda| \leq 1$ pour assurer la stabilité du schéma...).

⋆ Dispersion du schéma de Lax.

Reprenons la situation précédente avec $f \equiv 0$ et $u_0(x) = \mathrm{e}^{ikx}$. On a alors

$$u(x_j, t^n) = \mathrm{e}^{ik(x_j - cn\Delta t)}. \tag{4.50}$$

Par ailleurs, en posant $\varphi_k = -\arg(g_L(\lambda,k))$ c'est-à-dire $g_L(\lambda,k) = |g_L(\lambda,k)|\mathrm{e}^{-i\varphi_k}$, la solution du schéma de Lax, est donnée par

$$u_j^n = (g_L(\lambda,k))^n\, u_j^0 = |g_L(\lambda,k)|^n \mathrm{e}^{i(kx_j - n\varphi_k)}. \tag{4.51}$$

Si on compare les arguments de (4.50) et(4.51), on voit que c'est la quantité $kc\Delta t - \varphi_k$ qui intervient. La quantité $q_L(\lambda,k) = kc\Delta t - \varphi_k$ correspond au déphasage introduit à chaque pas de temps entre la solution exacte et la solution du schéma de Lax, lorsqu'on propage le k-ième mode. Cette quantité s'appelle le *facteur de dispersion* du k-ième mode. Plus $q_L(\lambda,k)$ est grand, plus la dispersion du k-ième mode est grande.

On a $\arg(z) = \arctan(\Im m(z)/\Re e(z))$, d'où $\varphi_k = -\arctan\left(-\lambda \sin(k\Delta x)/\cos(k\Delta x)\right)$, soit encore

$$\varphi_k = \arctan\left(\lambda \tan(k\Delta x)\right).$$

Le facteur de dispersion du schéma de Lax est donc donné par

$$q_L(\lambda,k) = kc\Delta t - \varphi_k = \lambda k\Delta x - \arctan(\lambda \tan(k\Delta x)). \tag{4.52}$$

La Figure 4.8 montre le facteur de dispersion q_L en fonction de $k\Delta x$ et pour différentes valeurs de λ.

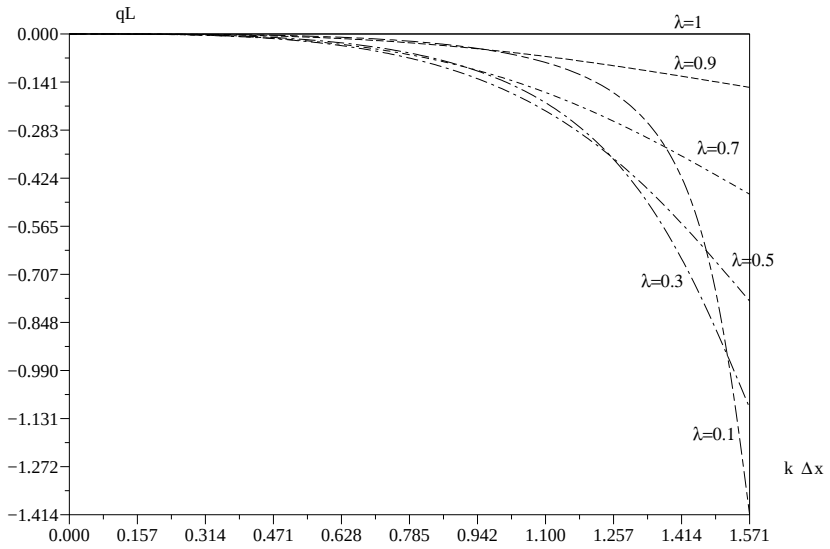


FIGURE 4.8 – Facteur de dispersion q_L du schéma de Lax (en fonction de $k\Delta x$).

Ici encore, on voit qu'on a intérêt à prendre $|\lambda| \simeq 1$ pour que le schéma de Lax disperse peu.

4.2.4 Schéma décentré (upwind)

On veut suivre au mieux la propagation le long des caractéristiques en décentrant l'approximation de la dérivée en espace. Ce décentrement tient compte de la direction de propagation. On note

$$D_x^+ v_j = \frac{v_{j+1} - v_j}{\Delta x}, \quad D_x^- v_j = \frac{v_j - v_{j-1}}{\Delta x}.$$

Si $c > 0$, on approche la dérivée en espace par D_x^- et si $c < 0$, on utilise D_x^+ (cf. Figure 4.9). Plus précisement, le schéma décentré s'écrit

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\, D_x u_j^n = f_j^n, \tag{4.53}$$

avec

$$D_x = \begin{cases} D_x^+, & \text{si } c < 0 \\ D_x^-, & \text{si } c > 0. \end{cases} \tag{4.54}$$

La relation (4.53) peut encore s'écrire

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c^+ D_x^- u_j^n + c^- D_x^+ u_j^n = f_j^n, \tag{4.55}$$

avec

$$c^+ = \max(c, 0), \quad c^- = \min(c, 0).$$

Le schéma décentré s'écrit alors de façon générale

$$\begin{array}{rcl} u_j^{n+1} & = & (1 - |\lambda|) u_j^n + \lambda^+ u_{j-1}^n - \lambda^- u_{j+1}^n + \Delta t f_j^n, \quad n \geq 1 \\ u_j^0 & = & u_0(x_j), \end{array} \tag{4.56}$$

avec

$$\lambda = c\Delta t/\Delta x, \quad \lambda^+ = \max(\lambda, 0), \quad \lambda^- = \min(\lambda, 0).$$

⋆ **Cône de dépendance numérique.**

La condition que le cône de dépendance numérique doit contenir la caractéristique passant par M (cf. Figure 4.9) entraine, en considérant les deux cas $c > 0$ et $c < 0$, la condition $|\lambda| \leq 1$.

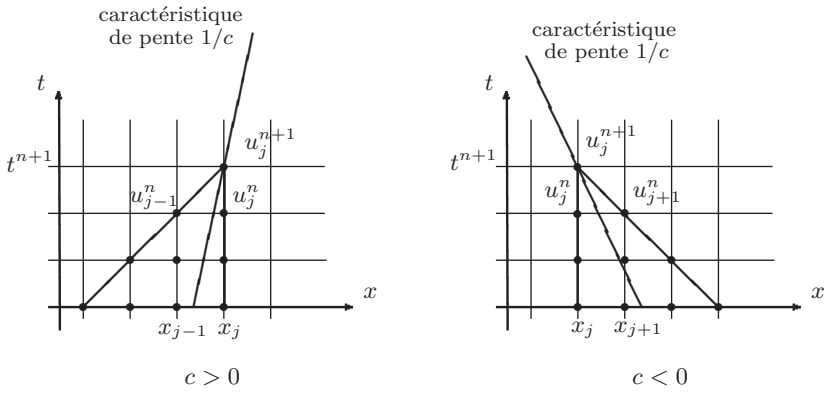


FIGURE 4.9 – Cônes de dépendance numérique du schéma décentré en fonction du signe de la vitesse c.

De plus, avec $f \equiv 0$ et $|\lambda| = 1$, on a $u_j^n = u(x_j, t^n)$ pour tout j et n. Dans ce cas, la solution approchée coïncide exactement avec la solution exacte aux points de discrétisation.

⋆ **Analyse de von Neumann/Fourier.**

On suppose que $c > 0$ et on prend $f \equiv 0$. Avec la donnée initiale $u_j^0 = \xi_0 \mathrm{e}^{ikj\Delta x}$, on cherche la solution du schéma décentré sous la forme

$$u_j^n = \xi_n \mathrm{e}^{ikj\Delta x}. \tag{4.57}$$

D'après la relation (4.56) du schéma décentré, on a

$$\xi_{n+1} = \big[1 - \lambda(1 - \cos k\Delta x) - i\lambda \sin k\Delta x\big]\xi_n. \tag{4.58}$$

Le coefficient d'amplification du schéma décentré est donné par

$$g_D(\lambda, k) = 1 - \lambda(1 - \cos k\Delta x) - i\lambda \sin k\Delta x \tag{4.59}$$

et on a

$$|g_D(\lambda, k)|^2 = 1 - 2\lambda(1 - \lambda)(1 - \cos k\Delta x). \tag{4.60}$$

- Si $\lambda > 1$ alors $|g_D(\lambda, k)| > 1$ (pour $k\Delta x \notin \pi\mathbb{Z}$) et le schéma décentré est <u>instable</u>.
- Si $\lambda \leq 1$ alors $|g_D(\lambda, k)| \leq 1$ et le schéma décentré est <u>stable</u> au sens de Von Neumann/Fourier.

Dans le cas où $c < 0$, on trouve la même condition de stabilité avec $-\lambda$.

⋆ **Principe du maximum discret.**

Sous la condition $|\lambda| \leq 1$, le schéma décentré vérifie le principe du maximum de la Proposition 4.1. En particulier le schéma décentré est positif. La démonstration est tout à fait identique à celle du schéma de Lax.

⋆ **Stabilité L^p.**

On prend (pour simplifier) <u>$f \equiv 0$</u> et on note $\mathbf{u}^n = \{u_j^n\}_{j\in\mathbb{Z}}$ la famille définie par le schéma décentré. Sous la condition $|\lambda| \leq 1$ (condition CFL), le schéma décentré est l^p-stable c'est-à-dire que les $\mathbf{u}^n$ du schéma décentré vérifient la Proposition 4.2 (l^p-stabilité pour le schéma de Lax). La démonstration de ce résultat est d'ailleurs identique à celle de la Proposition 4.2.

⋆ **Consistance.**

On montre facilement que l'erreur de consistance est en $\mathcal{O}(\Delta t + \Delta x)$. Le schéma décentré est donc consistant à l'ordre 1 en temps et en espace.

⋆ **Diffusion du schéma décentré.**

Le module du facteur d'amplification est donné par (cf. (4.60))

$$|g_D(\lambda,k)| = \Big(1-2\lambda(1-\lambda)(1-\cos k\Delta x)\Big)^{1/2}.$$

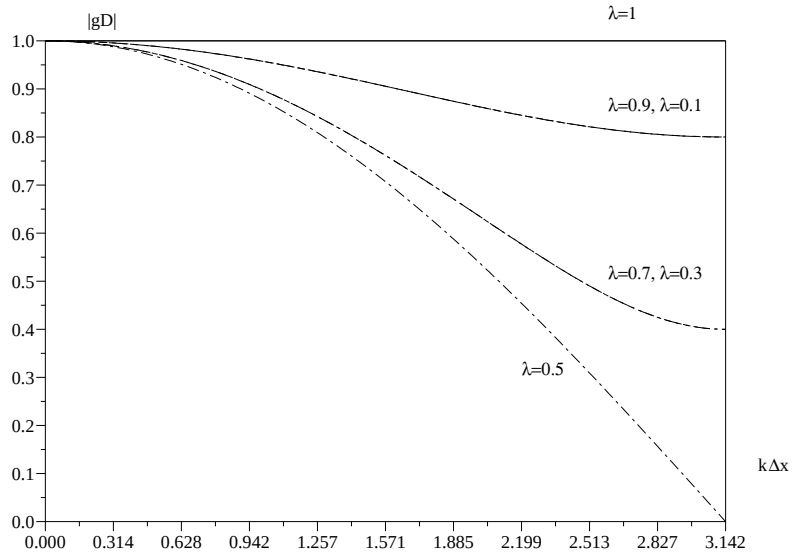


FIGURE 4.10 – Facteur d'amplification $|g_D|$ du schéma de décentré (en fonction de $k\Delta x$).

La diffusion est maximale pour $\lambda = 0.5$ (à $k\Delta x$ fixé). Et il n'y a évidemment pas de diffusion avec $\lambda = 1$.

⋆ **Dispersion du schéma décentré.**

On note $q_D(\lambda,k) = k\Delta t - \varphi_k$ avec $\varphi_k = -\arg(g_D(\lambda,k))$, le facteur de dispersion du schéma décentré. Ce facteur vaut

$$q_D(\lambda,k) = \lambda k\Delta x - \arctan\left(\frac{\lambda\sin k\Delta x}{1-\lambda(1-\cos k\Delta x)}\right). \tag{4.61}$$

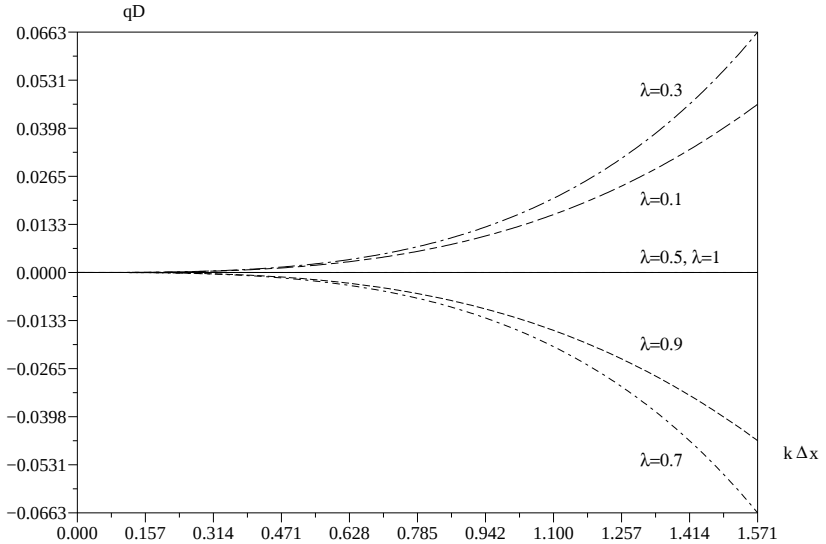


FIGURE 4.11 – Facteur de dispersion q_D du schéma décentré (en fonction de $k\Delta x$).

Pour $\lambda = 0.5$ (et pour $\lambda = 1$, bien sûr), il n'y a pas de dispersion du schéma décentré.

⋆ **Interprétation du schéma décentré.**

On fixe $f \equiv 0$ et on considère la caractéristique passant par le point (x_j, t^{n+1}). On suppose que $c > 0$ et que la condition CFL est satisfaite avec $\lambda < 1$. La caractéristique coupe la droite $t = t^n$ en $x = x_j^*$ (cf. Figure 4.12).

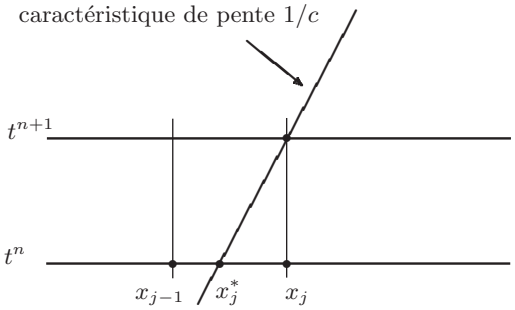


FIGURE 4.12 – Le schéma décentré vu comme interpolation linéaire le long des caractéristiques.

La pente de la caractéristique valant $1/c$ dans le plan (x,t), on a $\Delta t/(x_j - x_j^*) = 1/c$ ce qui donne $x_j^* = x_j - c\Delta t$. La solution (continue) est constante le long de la caractéristique, on a donc $u_j^{n+1} \simeq u(x_j, t^{n+1}) = u(x_j^*, t^n)$. On calcule cette dernière valeur par interpolation linéaire entre x_{j-1} et x_j :

$$u(x_j^*, t^n) \simeq u_{j-1}^n \left(\frac{x_j^* - x_j}{x_{j-1} - x_j} \right) + u_j^n \left(\frac{x_j^* - x_{j-1}}{x_j - x_{j-1}} \right).$$

On retrouve ainsi le schéma décentré $u_j^{n+1} = (1-\lambda)u_j^n + \lambda u_{j-1}^n$. Par conséquent, le schéma décentré est obtenu par interpolation linéaire le long des caractéristiques.

4.2.5 Schéma de Lax-Wendroff

Il s'agit d'un schéma à 2 pas de temps, d'ordre 2 en espace et en temps.

On introduit les quantités $u_{j+1/2}^{n+1/2}$ aux points $x_{j+1/2} = x_j + \Delta x/2$ et $t^{n+1/2} = t^n + \Delta t/2$.

- On approche l'équation de transport en $(x_j, t^{n+1/2})$ par un <u>schéma centré</u> sur les dérivées en temps et en espace :
$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\,\frac{u_{j+1/2}^{n+1/2} - u_{j-1/2}^{n+1/2}}{\Delta x} = f(x_j, t^{n+1/2}). \tag{4.62}$$
- On approche ensuite l'équation de transport en $(x_{j+1/2}, t^n)$ par un <u>schéma de Lax</u> :
$$\frac{u_{j+1/2}^{n+1/2} - \frac{1}{2}(u_j^n + u_{j+1}^n)}{\Delta t/2} + c\,\frac{u_{j+1}^n - u_j^n}{\Delta x} = f(x_{j+1/2}, t^n). \tag{4.63}$$

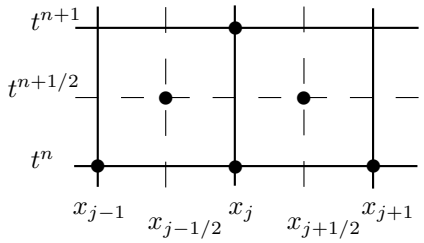


FIGURE 4.13 – Cône numérique du schéma de Lax-Wendroff.

En injectant (4.63) dans (4.62), on obtient l'expression suivante du schéma de Lax-Wendroff :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}+c\left(\frac{u_{j+1}^n-u_{j-1}^n}{2\Delta x}\right)-\frac{c^2\Delta t}{2}\left(\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}\right)=f_j^{n+1/2}-\frac{c\Delta t}{2}\left(\frac{f_{j+1/2}^n-f_{j-1/2}^n}{\Delta x}\right). \quad (4.64)$$

En posant

$$a_{-1}=\frac{\lambda}{2}(1+\lambda),\quad a_0=1-\lambda^2,\quad a_1=\frac{\lambda}{2}(\lambda-1), \quad (4.65)$$

le schéma de Lax-Wendroff s'écrit encore

$$u_j^{n+1}=a_{-1}\,u_{j-1}^n+a_0\,u_j^n+a_1\,u_{j+1}^n+\Delta t\left(f_j^{n+1/2}-\frac{\lambda}{2}(f_{j+1/2}^n-f_{j-1/2}^n)\right). \quad (4.66)$$

Le schéma de Lax-Wendroff est obtenu à partir du schéma centré (qui est toujours instable...) auquel on a rajouté un terme stabilisant (voir plus loin l'étude de stabilité) égal à $-c^2\frac{\Delta t}{2}\left(\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{\Delta x^2}\right)$. Ce terme est un équivalent discret de $-c^2\frac{\Delta t}{2}\frac{\partial^2 u}{\partial x^2}$. Il s'agit d'un terme de dissipation d'ordre 1 en Δt quand la solution est régulière (u_{xx} borné). Il y a dissipation lorsque le système perd de l'énergie. Par exemple, considérons l'équation $u_t+cu_x-c^2\frac{\Delta t}{2}u_{xx}=0$. En multipliant par u puis en intégrant, on obtient

$$\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}}u^2\,dx+c^2\frac{\Delta t}{2}\int_{\mathbb{R}}(u_x)^2\,dx=0. \quad (4.67)$$

Par conséquent, on a $\frac{1}{2}\frac{d}{dt}\int_{\mathbb{R}}u^2\,dx<0$ (pour $u_x\not\equiv 0$), ce qui signifie que l'énergie décroit au cours du temps.

⋆ **Consistance.**

On note $L=\partial/\partial t+c\partial/\partial x-f$ l'opérateur de l'équation de transport et $L_{\Delta x,\Delta t}$ l'opérateur approché du schéma de Lax-Wendroff défini par

$$\begin{aligned}L_{\Delta x,\Delta t}v(x,t) &= \frac{u(x,t+\Delta t)-u(x,t)}{\Delta t}+c\,\frac{u(x+\Delta x,t)-u(x-\Delta x,t)}{2\Delta x}\\ &\quad -c^2\frac{\Delta t}{2}\left(\frac{u(x+\Delta x,t)-2u(x,t)+u(x-\Delta x,t)}{\Delta x^2}\right)\\ &\quad -f(x,t+\Delta t/2)+c\frac{\Delta t}{2}\left(\frac{f(x+\Delta x/2,t)-f(x-\Delta x/2,t)}{\Delta x}\right).\end{aligned} \quad (4.68)$$

Pour v et f suffisamment régulières, on montre facilement que

$$\left(L_{\Delta x,\Delta t}v-Lv\right)(x,t)=\frac{\Delta t}{2}\left(v_{tt}(x,t)-c^2v_{xx}-f_t+cf_x\right)+\mathcal{O}(\Delta t^2+\Delta x^2). \quad (4.69)$$

Le schéma de Lax-Wendroff est (au moins) d'ordre 1 en temps et d'ordre 2 en espace. En fait, si on se restreint aux solutions u de l'équation de transport i.e. telles que $Lu\equiv 0$, alors le schéma est d'ordre 2 en temps. En effet, pour u régulière telle que $u_t+cu_x=f$, il vient en dérivant l'équation de transport par rapport à t et par rapport à x : $u_{tt}+cu_{xt}=f_t$ et $u_{xt}+cu_{xx}=f_x$. En éliminant u_{xt}, on obtient

$$u_{tt}-c^2u_{xx}=f_t-cf_x.$$

L'estimation (4.69) se réduit alors

$$\left(L_{\Delta x,\Delta t}u-Lu\right)(x,t)=\mathcal{O}(\Delta t^2+\Delta x^2) \quad (4.70)$$

pour u régulière telle que $Lu\equiv 0$.

⋆ **Cône de dépendance numérique.**

Le cône de dépendance numérique devant contenir la caractéristique associée (cf. Figure 4.13), on obtient la condition CFL nécessaire

$$|\lambda| \leq 1.$$

De plus, si $f \equiv 0$ et si $|\lambda| = 1$ alors $u_j^n = u(x_j, t^n)$ pour tout j et n.

⋆ **Analyse de von Neumann/Fourier.**

On prend $f \equiv 0$. Avec la donnée initiale $u_j^0 = \xi_0 \mathbf{e}^{ikj\Delta x}$, on cherche la solution du schéma de Lax-Wendroff sous la forme

$$u_j^n = \xi_n \mathbf{e}^{ikj\Delta x}. \tag{4.71}$$

D'après la relation (4.64), on obtient

$$\xi_{n+1} = \left[1 - i\lambda \sin k\Delta x + \lambda^2(\cos k\Delta x - 1)\right] \xi_n. \tag{4.72}$$

Le coefficient d'amplification du schéma de Lax-Wendroff est donné par

$$g_{LW}(\lambda, k) = 1 - i\lambda \sin k\Delta x + \lambda^2(\cos k\Delta x - 1) \tag{4.73}$$

et on a

$$|g_{LW}(\lambda, k)|^2 = 1 - \lambda^2(1 - \lambda^2)(1 - \cos k\Delta x)^2. \tag{4.74}$$

- Si $\lambda > 1$ alors $|g_{LW}(\lambda, k)| > 1$ (pour $k\Delta x \notin \pi\mathbb{Z}$) et le schéma de Lax-Wendroff est instable.
- Si $\lambda \leq 1$ alors $|g_{LW}(\lambda, k)| \leq 1$ et le schéma de Lax-Wendroff est stable au sens de Von Neumann/Fourier.

⋆ **Non positivité du schéma de Lax-Wendroff.**

Il n'y a pas de principe du maximum discret pour le schéma de Lax-Wendroff. En fait, il n'y a pas de positivité pour $0 < \lambda < 1$ c'est-à-dire que $u_j^0 \geq 0 \not\Rightarrow u_j^n \geq 0, \forall j, \forall n$. Par exemple, pour $f \equiv 0$ avec $u_j^0 = 1$ si $j > j^*$ et $u_j^0 = 0$ sinon, on a $u_{j^*}^1 = a_1 u_{j^*+1}^0 = a_1 < 0$ pour $0 < \lambda < 1$.

⋆ **Stabilité L^2.**

Proposition 4.3 *Soit $f \equiv 0$ et $\mathbf{u}^0 \in l^2_{\Delta x}$. Si $|\lambda| \leq 1$ alors $\mathbf{u}^n \in l^2_{\Delta x}$ et on a*

$$\|\mathbf{u}^n\|_{2,\Delta x} \leq \|\mathbf{u}^0\|_{2,\Delta x}, \text{ pour tout } n \in \mathbb{N}. \tag{4.75}$$

Démonstration. Remarquons tout d'abord qu'on ne peut pas appliquer le Lemme 4.1 avec (4.66) car u_j^{n+1} n'est pas une combinaison convexe des u_j^n. En effet, on a bien que $a_{-1} + a_0 + a_1 = 1$ et par exemple pour $0 \leq \lambda \leq 1$, on a $a_{-1}, a_0 \geq 0$ mais $a_1 \leq 0$.

Il est clair que si $\mathbf{u}^n \in l^2_{\Delta x}$ alors $\mathbf{u}^{n+1} \in l^2_{\Delta x}$ d'après (4.66). Montrons sous la condition $|\lambda| \leq 1$ que $\|\mathbf{u}^{n+1}\|_{2,\Delta x} \leq \|\mathbf{u}^n\|_{2,\Delta x}$. On élève au carré la relation (4.66) et on somme sur $j \in \mathbb{Z}$. En utilisant le fait que $\sum_{j\in\mathbb{Z}} u_{j+1}^n u_j^n = \sum_{j\in\mathbb{Z}} u_j^n u_{j-1}^n$, on obtient

$$\sum_{j\in\mathbb{Z}} (u_j^{n+1})^2 = (1 + \alpha_1) \sum_{j\in\mathbb{Z}} (u_j^n)^2 + \alpha_2 \sum_{j\in\mathbb{Z}} u_j^n u_{j+1}^n + \alpha_3 \sum_{j\in\mathbb{Z}} u_{j-1}^n u_{j+1}^n, \tag{4.76}$$

avec $1 + \alpha_1 = a_{-1}^2 + a_0^2 + a_1^2$, $\alpha_2 = 2a_0(a_{-1} + a_1)$ et $\alpha_3 = 2a_{-1}a_1$, soit

$$\alpha_1 = \frac{3}{2}\lambda^2(\lambda^2 - 1), \quad \alpha_2 = 2(1 - \lambda^2)\lambda^2, \quad \alpha_3 = \frac{\lambda^2}{2}(\lambda^2 - 1).$$

Par ailleurs, on a $u_j^n u_{j+1}^n = \frac{1}{2}\left((u_j^n)^2 + (u_{j+1}^n)^2 - (u_{j+1}^n - u_j^n)^2\right)$. En sommant sur $j \in \mathbb{Z}$, il vient

$$\sum_{j\in\mathbb{Z}} u_j^n u_{j+1}^n = \sum_{j\in\mathbb{Z}} (u_j^n)^2 - \frac{1}{2} \sum_{j\in\mathbb{Z}} (u_{j+1}^n - u_j^n)^2.$$

On a de même

$$\sum_{j\in\mathbb{Z}} u_{j-1}^n u_{j+1}^n = \sum_{j\in\mathbb{Z}} (u_j^n)^2 - \frac{1}{2}\sum_{j\in\mathbb{Z}} (u_{j+1}^n - u_{j-1}^n)^2.$$

En injectant les deux relations précédentes dans (4.76), on obtient

$$\sum_{j\in\mathbb{Z}} (u_j^{n+1})^2 - \sum_{j\in\mathbb{Z}} (u_j^n)^2 = (\alpha_1+\alpha_2+\alpha_3)\sum_{j\in\mathbb{Z}} (u_j^n)^2 - \frac{\alpha_2}{2}\sum_{j\in\mathbb{Z}} (u_{j+1}^n - u_j^n)^2 - \frac{\alpha_3}{2}\sum_{j\in\mathbb{Z}} (u_{j+1}^n - u_{j-1}^n)^2. \quad (4.77)$$

Or $\alpha_1+\alpha_2+\alpha_3 = 0$ et on a

$$\begin{aligned}\sum_j (u_{j+1}^n - u_{j-1}^n)^2 &\leq \sum_j (|u_{j+1}^n - u_j^n| + |u_j^n - u_{j-1}^n|)^2 \\ &\leq 2\sum_j (|u_{j+1}^n - u_j^n|^2 + |u_j^n - u_{j-1}^n|^2) \\ &= 4\sum_j (u_{j+1}^n - u_j^n)^2.\end{aligned}$$

Pour $|\lambda| \leq 1$, on a $\alpha_3 \leq 0$ et $\alpha_2 \geq 0$. L'inégalité précédente combinée à (4.77), conduit à

$$\sum_j (u_j^{n+1})^2 - \sum_j (u_j^n)^2 \leq (-\frac{\alpha_2}{2} - 2\alpha_3)\sum_j (u_{j+1}^n - u_j^n)^2 = 0, \quad (4.78)$$

ce qui termine la démonstration. □

⋆ **Diffusion du schéma de Lax-Wendroff.**

Le module du facteur d'amplification est donné par (cf. (4.74))

$$|g_{LW}(\lambda,k)| = \Big(1 - \lambda^2(1-\lambda^2)(1-\cos k\Delta x)^2\Big)^{1/2}.$$

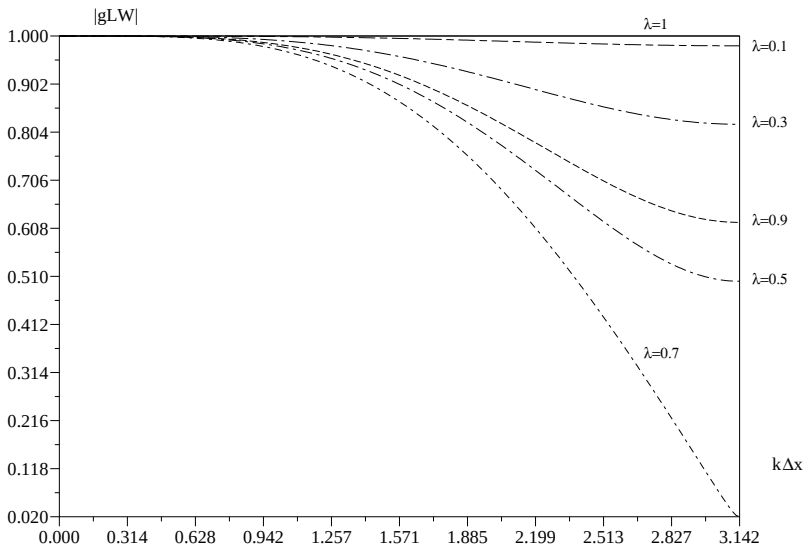


FIGURE 4.14 – Facteur d'amplification $|g_{LW}|$ du schéma de Lax-Wendroff (en fonction de $k\Delta x$).

La diffusion est maximale pour $\lambda = 1/\sqrt{2} \simeq 0.7$ (à $k\Delta x$ fixé). Il n'y a pas de diffusion avec $\lambda = 1$.

⋆ **Dispersion du schéma de Lax-Wendroff.**

On note $q_{LW}(\lambda, k) = k\Delta t - \varphi_k$ avec $\varphi_k = -\arg(g_{LW}(\lambda, k))$, le facteur de dispersion du schéma de Lax-Wendroff. Ce facteur vaut

$$q_{LW}(\lambda, k) = \lambda k \Delta x - \arctan\left(\frac{\lambda \sin k\Delta x}{1 + \lambda^2(\cos k\Delta x - 1)}\right). \tag{4.79}$$

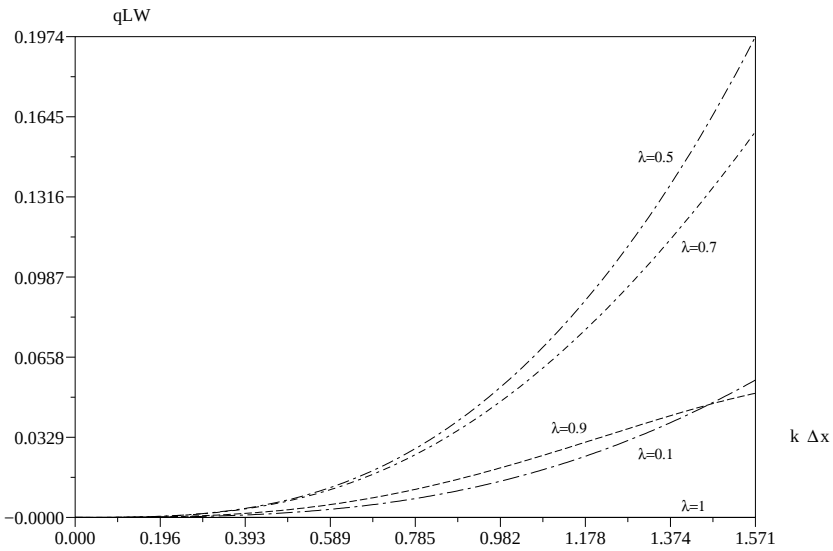


FIGURE 4.15 – Facteur de dispersion q_{LW} du schéma de Lax-Wendroff (en fonction de $k\Delta x$).

4.2.6 Comparaison des différents schémas

Parmi les trois schémas présentés, le schéma de Lax-Wendroff diffuse le moins. C'est le schéma de Lax qui diffuse le plus (cf. Figure 4.17). Le schéma de Lax donne généralement d'assez médiocres résultats numériques (cf. Figure 4.16) à moins que le maillage soit très fin. La figure 4.16 montre les solutions approchées de l'équation $u_t + u_x = 0$ obtenues avec les trois schémas ($\lambda = 0.7$). La donnée initiale est une "marche descendante" et la solution exacte (indiquée sur les figures) est donc la marche translatée. On observe bien que le principe du maximum discret n'est pas respecté avec le schéma de Lax-Wendroff.

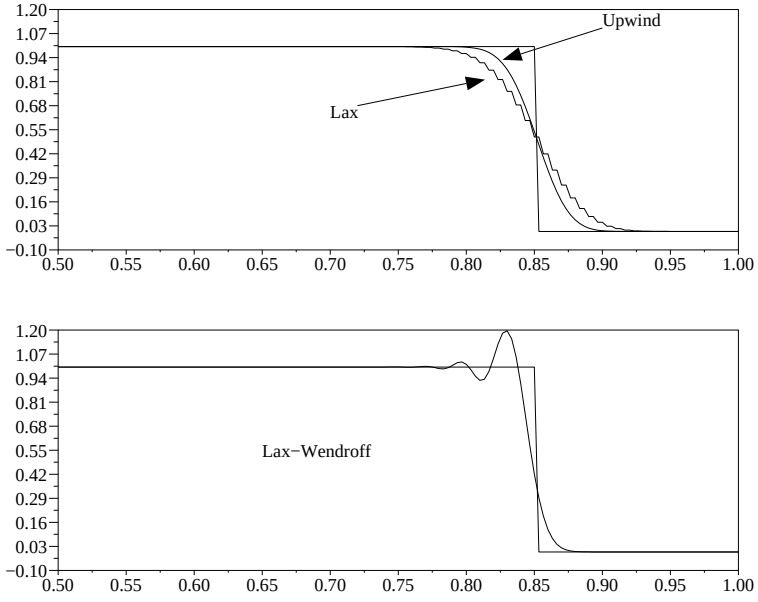


FIGURE 4.16 – Solutions de $u_t + u_x = 0$ pour différents schémas. La donnée initiale est une "marche descendante"

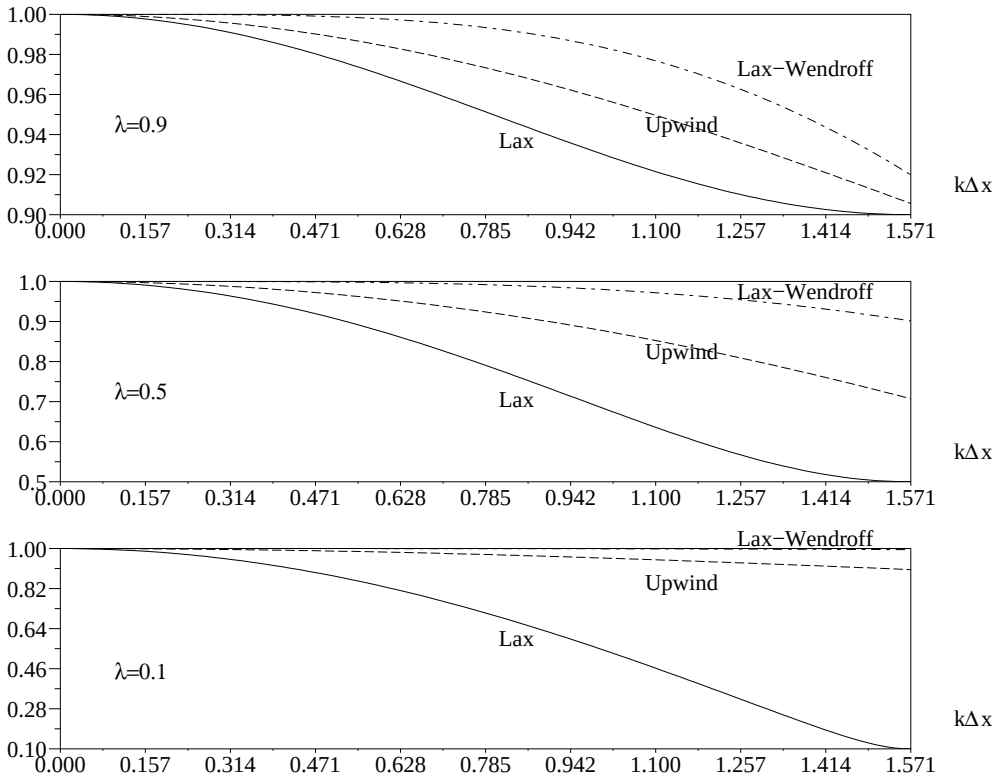


FIGURE 4.17 – Comparaison de la diffusion des schémas pour différentes valeurs de λ.

Les schémas décentré et de Lax-Wendroff ont des dispersions comparables. Le schéma de Lax disperse le plus (cf. Figure 4.18)

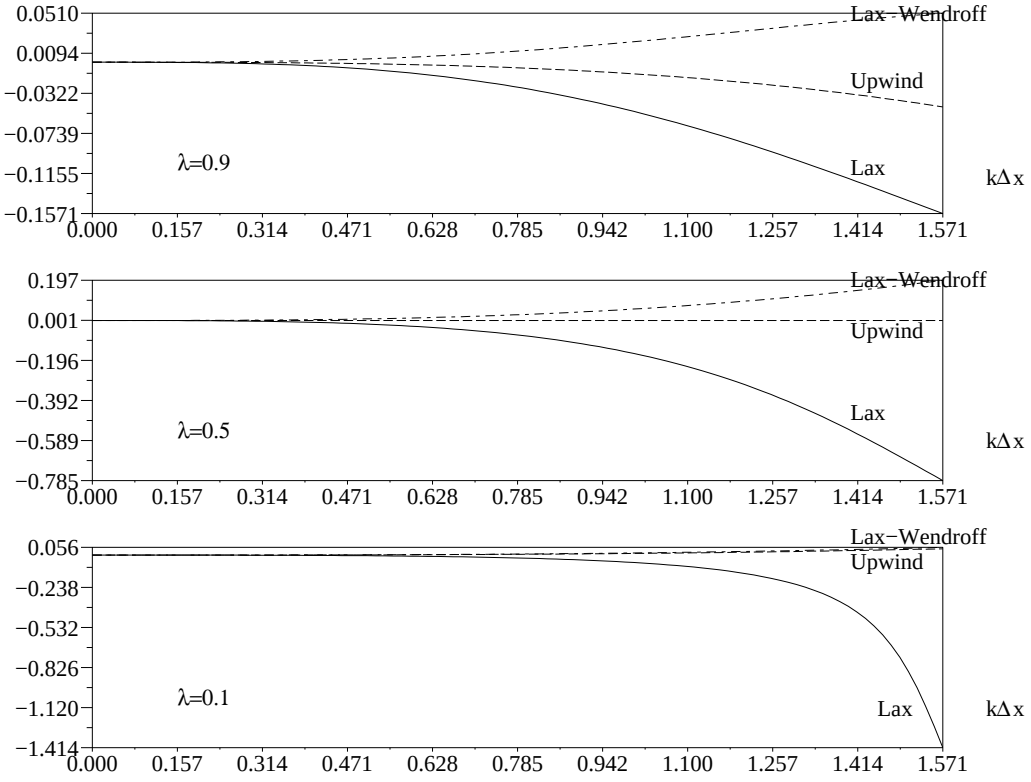


FIGURE 4.18 – Comparaison de la dispersion des schémas pour différentes valeurs de λ.

4.2.7 Quelques schémas pour le 2D

L'équation de transport en dimension 2 d'espace consiste à chercher $u = u(x, y, t)$ pour $(x, y) \in \mathbb{R}^2$ vérifiant

$$(P) \left\{ \begin{array}{lcl} u_t + \mathbf{v} \cdot \nabla u & = & 0 \quad \text{dans } \mathbb{R}^2,\ t > 0 \\ u(x, y, 0) & = & u_0(x, y), \quad (x, y) \in \mathbb{R}^2. \end{array} \right.$$

où $\mathbf{v} = (v_1, v_2)$ est une vitesse donnée. On note $P_{ij} = (i\Delta x, j\Delta y)$, $i, j \in \mathbb{Z}$ les noeuds du maillage du plan $\mathbb{R}^2$ tout entier et $u_{ij}^n \simeq u(P_{ij}, t^n)$ une approximation de la solution exacte aux noeuds P_{ij} et à l'instant $t^n = n\Delta t$.

⋆ Schéma décentré 2D (Donor Cell Upwind).

Il s'agit d'une extension directe au cas 2D du schéma décentré 1D. Il s'écrit

$$\begin{array}{rcl} \dfrac{u_{ij}^{n+1} - u_{ij}^n}{\Delta t} & + & v_1^+ \dfrac{(u_{ij}^n - u_{i-1,j}^n)}{\Delta x} + v_1^- \dfrac{(u_{i+1,j}^n - u_{i,j}^n)}{\Delta x} \\ & + & v_2^+ \dfrac{(u_{ij}^n - u_{i,j-1}^n)}{\Delta y} + v_2^- \dfrac{(u_{i,j+1}^n - u_{i,j}^n)}{\Delta y} = 0. \end{array}$$

Ce schéma décentré est d'ordre 1 en temps et en espace et il est stable sous la condition CFL :

$$\lambda_{DCU} \equiv \left| \frac{v_1 \Delta t}{\Delta x} \right| + \left| \frac{v_2 \Delta t}{\Delta y} \right| \leq 1. \tag{4.80}$$

Ce schéma diffuse toujours y compris dans le cas où on prend la condition CFL, $\lambda_{DCU} = 1$ (sauf si la vitesse $\mathbf{v}$ a une composante nulle, auquel cas on résout un problème 1D...). La Figure 4.19 montre la solution calculée à $t = 0.22$ par ce schéma décentré avec une condition CFL égale à 1 ($\lambda_{DCU} = 1$). La vitesse est choisie égale à $\mathbf{v} = (1, 2)$. On constate la diffusion numérique de la solution approchée.

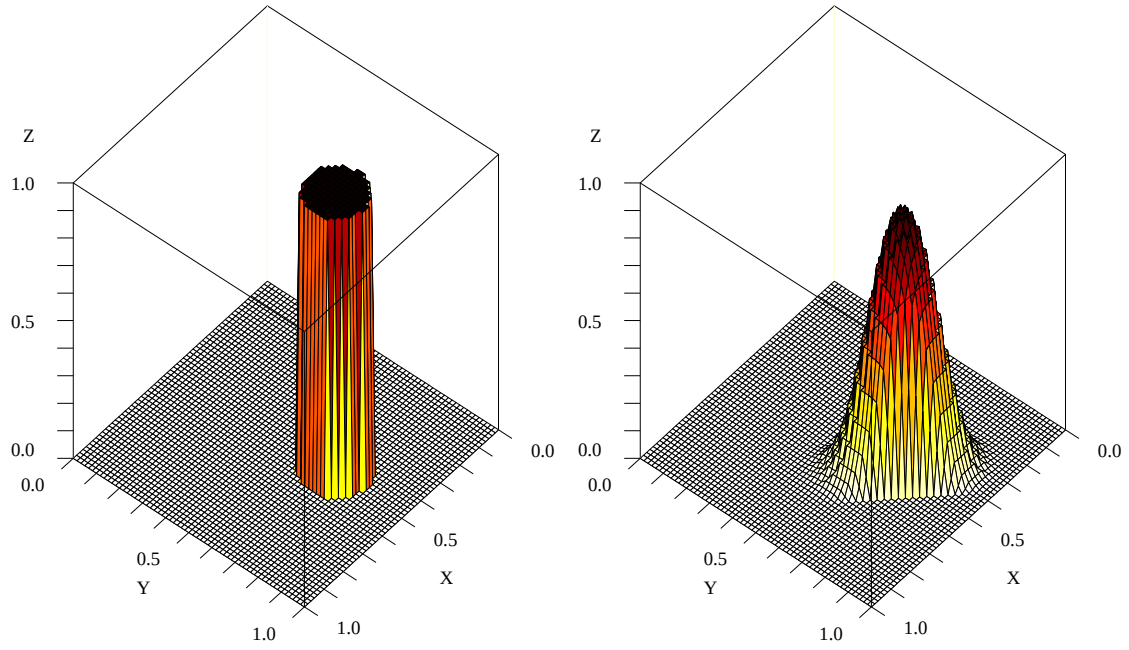


FIGURE 4.19 – Equation de transport $u_t + \mathbf{v} \cdot \nabla u = 0$ avec $\mathbf{v} = (1, 2)$: solution exacte (à gauche) et schéma *DCU* (à droite) à l'instant $t = 0.22$ avec $\lambda_{DCU} = 1$ et une grille 61×61.

⋆ Schéma décentré modifié (Corner Transport Upwind).

Soit u la solution supposée régulière de l'équation de transport 2D. Par développement de Taylor et en tenant compte de l'équation de transport, on a

$$\begin{aligned}\frac{u(x,y,t^{n+1})-u(x,y,t^n)}{\Delta t} &= u_t(x,y,t^n)+\frac{\Delta t}{2}\,u_{tt}(x,y,t^n)+\cdots\\ &= -v_1\,u_x-v_2\,u_y\\ &\quad +\frac{\Delta t}{2}\left(v_1^2\,u_{xx}+v_1v_2\,u_{yx}+v_2v_1\,u_{xy}+v_2^2\,u_{yy}\right)+\cdots \qquad (4.81)\end{aligned}$$

Le schéma décentré modifié s'obtient à partir de la discrétisation de la relation précédente. Il s'écrit

$$\begin{aligned}\frac{u_{ij}^{n+1}-u_{ij}^n}{\Delta t} &= -\frac{v_1}{\Delta x}\left(u_{ij}^n-u_{i-1,j}^n\right)-\frac{v_2}{\Delta y}\left(u_{ij}^n-u_{i,j-1}^n\right)\\ &\quad +\frac{\Delta t}{2}\left[\frac{v_1}{\Delta x}\left(\frac{v_2}{\Delta y}(u_{ij}^n-u_{i,j-1}^n)-\frac{v_2}{\Delta y}(u_{i-1,j}^n-u_{i-1,j-1}^n)\right)\right.\\ &\quad \left.+\frac{v_2}{\Delta y}\left(\frac{v_1}{\Delta x}(u_{ij}^n-u_{i-1,j}^n)-\frac{v_1}{\Delta x}(u_{i,j-1}^n-u_{i-1,j-1}^n)\right)\right].\end{aligned}$$

Le premier terme de la relation précédente correspond au schéma DCU précédent alors que le dernier terme est associé aux dérivées croisées $v_1v_2\,u_{yx}+v_2v_1\,u_{xy}$. Ce schéma est d'ordre 1 en temps et en espace et il est stable sous la condition CFL suivante :

$$\lambda_{CTU}\equiv\max\left(\left|\frac{v_1\Delta t}{\Delta x}\right|,\left|\frac{v_2\Delta t}{\Delta y}\right|\right)\leq 1. \qquad (4.82)$$

Cette condition CFL est meilleure que celle du schéma DCU précédent (dans le cas où $\mathbf{v}$ n'a pas de composante nulle). Ce schéma diffuse moins que le précédent (cf. Figure 4.20). De plus, si $v_1\neq 0$ et $v_2\neq 0$, on peut toujours choisir Δx et Δy pour que $|v_1\Delta t/\Delta x|=1$ et $|v_2\Delta t/\Delta y|=1$ et alors dans ce cas, il n'y a aucune diffusion numérique : la solution approchée du schéma TCU coïncide avec la solution exacte.

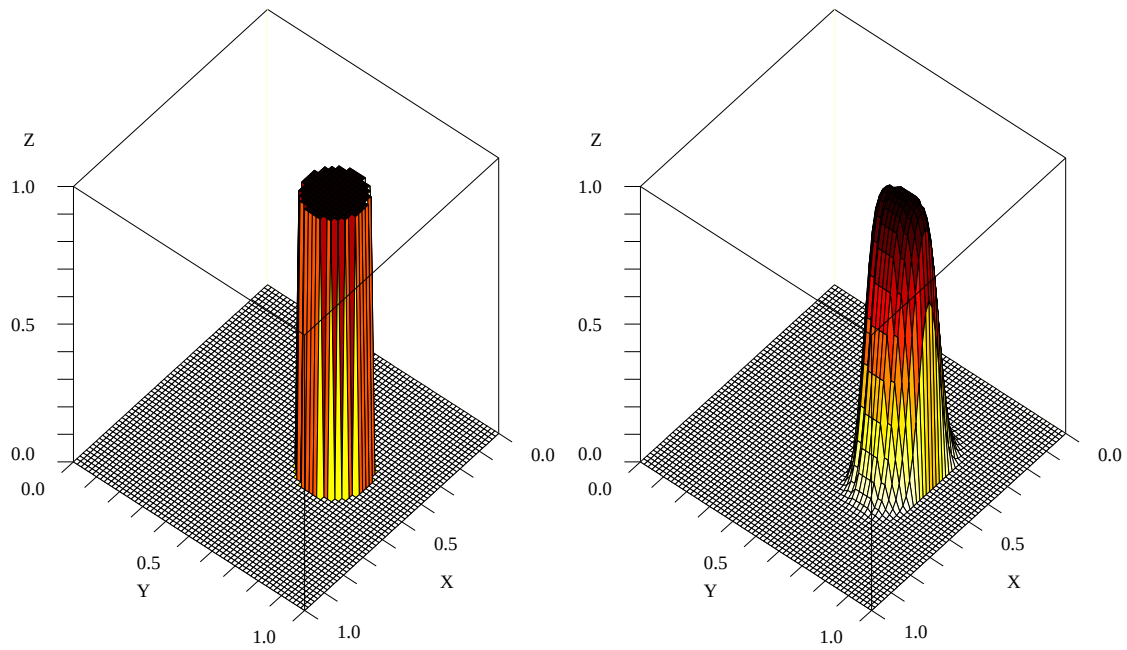


FIGURE 4.20 – Equation de transport $u_t+\mathbf{v}\cdot\nabla u=0$ avec $\mathbf{v}=(1,2)$: solution exacte (à gauche) et schéma *CTU* (à droite) à l'instant $t=0.22$ avec $\lambda_{CTU}=1$ et une grille 61×61.

⋆ Schéma décentré à pas fractionnaire (Godounov splitting).

Le principe de la méthode des pas fractionnaires (voir le paragraphe "Directions alternées" du Chap. 3) est de résoudre uniquement des problèmes monodimensionnels (en espace). Supposons que l'on connaisse une approximation u^n de la solution u à l'instant t^n. On détermine u^{n+1} de la façon suivante. On commence par résoudre

$$(P_1) \left\{ \begin{array}{l} q_t + v_1\, q_x = 0, \quad t \in (t^n, t^{n+1}) \\ q(t^n) = u^n. \end{array} \right.$$

On obtient ainsi $q^* = q(t^{n+1})$, puis on résout

$$(P_2) \left\{ \begin{array}{l} q_t + v_2\, q_y = 0, \quad t \in (t^n, t^{n+1}) \\ q(t^n) = q^*, \end{array} \right.$$

et on choisit $u^{n+1} = q(t^{n+1})$. Par exemple, en choisissant le schéma décentré 1D pour résoudre (P_1) puis (P_2), on obtient le schéma à pas fractionnaire de Godounov :

$$\begin{aligned} q^*_{ij} &= u^n_{ij} - \frac{\Delta t}{\Delta x}\left(v_1^+(u^n_{ij} - u^n_{i-1,j}) + v_1^-(u^n_{i+1,j} - u^n_{i,j})\right) \\ u^{n+1}_{ij} &= q^*_{ij} - \frac{\Delta t}{\Delta y}\left(v_2^+(q^*_{ij} - q^*_{i,j-1}) + v_2^-(q^*_{i,j+1} - q^*_{i,j})\right). \end{aligned}$$

Ce schéma d'ordre 2 en temps et est stable sous la même condition CFL que le schéma CTU, c'est-à-dire si :

$$\lambda_{GS} \equiv \max\left(\left|\frac{v_1\Delta t}{\Delta x}\right|, \left|\frac{v_2\Delta t}{\Delta y}\right|\right) \leq 1. \tag{4.83}$$

Comme précédemment, on peut toujours choisir une grille pour que les conditions CFL dans chaque direction valent 1 et dans ce cas, il n'y a pas de diffusion.

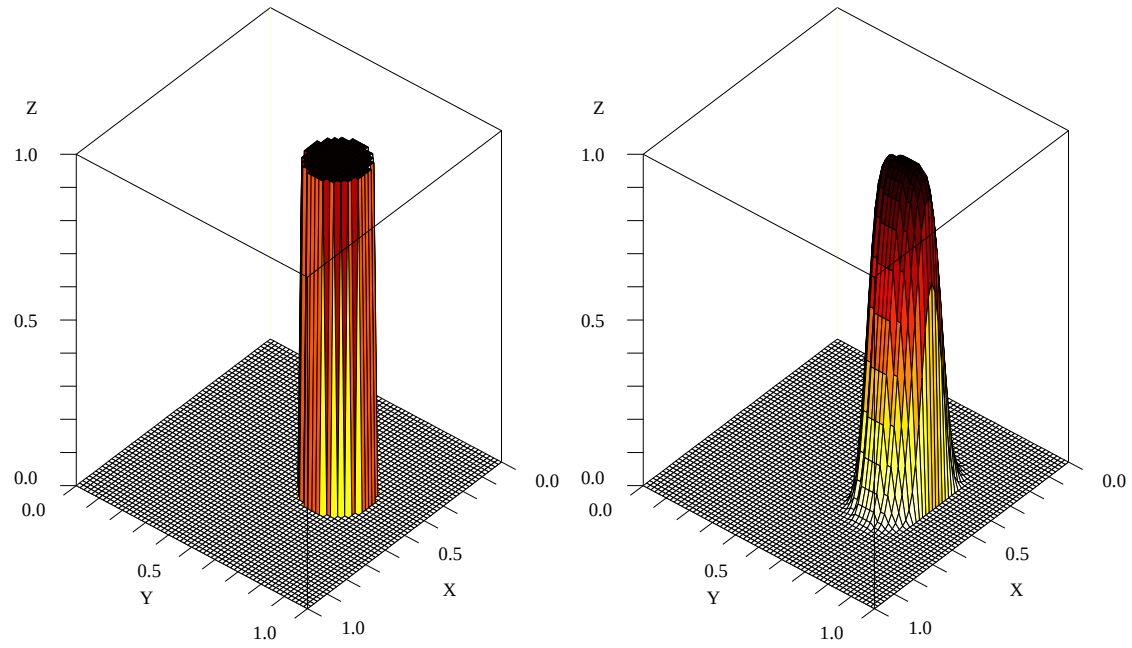


FIGURE 4.21 – Equation de transport $u_t + \mathbf{v}\cdot\nabla u = 0$ avec $\mathbf{v} = (1, 2)$: solution exacte (à gauche) et schéma *Godounov splitting* (à droite) à l'instant $t = 0.22$ avec $\lambda_{GS} = 1$ et une grille 61×61.

Chapitre 5

EDP hyperboliques non-linéaires - Lois de conservation

5.1 Introduction

Dans tout ce chapitre, on se place en dimension 1 d'espace. Soient f et u_0 des fonctions données avec $f \in C^1(\mathbb{R}, \mathbb{R})$. On considère la loi de conservation scalaire suivante pour $u = u(x,t)$:

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) &= 0 \quad \text{pour } x \in \mathbb{R}, t > 0 \qquad (5.1) \\ u(x,0) &= u_0(x) \quad \text{pour } x \in \mathbb{R}. \qquad (5.2) \end{aligned}$$

Cette équation correspond à une modélisation très simplifiée des phénomènes suivants :

1. Aéronautique (écoulement supersonique) : équation de Burgers avec $f(u) = u^2/2$.
2. Ecoulement en milieux poreux : équation de Buckley-Leverett avec $f(u) = \dfrac{u^2}{u^2 + (1-u)^2 \mu_w/\mu_o}$ où μ_w et μ_o sont les viscosités de l'eau et de l'huile (pétrole) respectivement.
3. Ecoulement en eau peu profonde : équation de Saint-Venant.
4. Dynamique des gaz : systèmes hyperboliques.

La particularité des problèmes non-linéaires comme (5.1),(5.2) réside dans l'apparition et la propagation de discontinuités et ceci même si la donnée initiale u_0 est régulière. Pour mieux comprendre ce phénomène, on introduit les courbes caractéristiques associées à (5.1),(5.2).

Courbes caractéristiques : on appelle courbe caractéristique associée à (5.1),(5.2) et partant d'un point $x_0 \in \mathbb{R}$, la courbe $\mathcal{C} : t \mapsto X(t)$, $t \geq 0$ où X vérifie

$$\begin{cases} \dfrac{dX}{dt}(t) = a(u(X(t),t)), & t > 0 \\ X(0) = x_0. \end{cases}$$

avec $a = f'$. On a la propriété suivante.

Proposition 5.1 *Soit $u_0 \in C^0(\mathbb{R})$. Si $u \in C^1(\mathbb{R} \times (0,+\infty)) \cap C^0(\mathbb{R} \times [0,+\infty))$ vérifie (5.1),(5.2) alors u est constante le long des caractéristiques. De plus, la caractéristique issue de x_0 est une droite de pente $a^{-1}(u_0(x_0))$ dans le plan (x,t).*

Preuve. Pour tout $t > 0$, on a

$$\begin{aligned} \frac{d}{dt}\left(u(X(t),t)\right) &= \frac{\partial u}{\partial x}(X(t),t)\frac{dX}{dt}(t) + \frac{\partial u}{\partial t}(X(t),t) \\ &= \frac{\partial u}{\partial x}(X(t),t) f'(u(X(t),t)) + \frac{\partial u}{\partial t}(X(t),t) \\ &= \frac{\partial f(u)}{\partial x}(X(t),t)) + \frac{\partial u}{\partial t}(X(t),t) = 0 \end{aligned}$$

donc, pour tout $t > 0$, on obtient

$$u(X(t), t) = u(X(0), 0) = u_0(x_0), \tag{5.3}$$

i.e. u est constante le long des caractéristiques. De plus, on a $\dfrac{dX}{dt}(t) = a(u(X(t), t)) = a(u_0(x_0)) = \text{Cte}$ et par conséquent

$$X(t) = x_0 + t\, a(u_0(x_0)). \tag{5.4}$$

Ainsi, la caractéristique issue de x_0 est bien une droite de pente $a^{-1}(u_0(x_0))$ dans le plan (x, t). □

Nous allons voir plusieurs concepts de solution du problème (5.1),(5.2).

5.2 Solutions classiques

Définition 5.1 *Soit $u_0 \in C^0(\mathbb{R})$. On dit que u est solution classique de* (5.1),(5.2) *si*

i. $u \in C^1(\mathbb{R} \times (0, +\infty)) \cap C^0(\mathbb{R} \times [0, +\infty))$.

ii. u *vérifie* (5.1),(5.2).

D'après la Proposition 5.1, toute solution classique est constante le long des caractéristiques qui sont des droites. Le problème est qu'en général, il n'y a pas de solution classique pour tout temps même si la donnée initiale u_0 est très régulière. En effet, supposons qu'il existe deux points $x_1 < x_2$ tels que

$$a^{-1}(u_0(x_1)) < a^{-1}(u_0(x_2)).$$

Alors toujours d'après la Proposition 5.1, les deux courbes caractéristiques $\mathcal{C}_1$ et $\mathcal{C}_2$ issues respectivement de x_1 et x_2 sont des droites qui se coupent en un point P (cf. Figure 5.1)

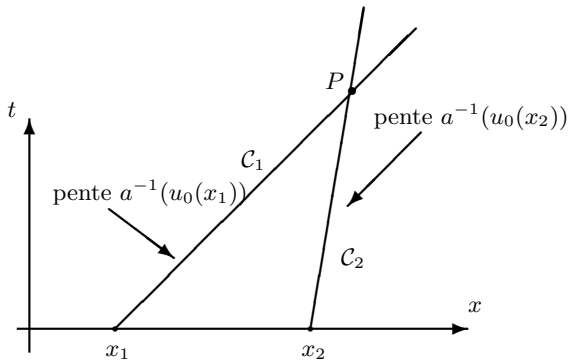


FIGURE 5.1 – Apparition de singularité : rencontre des caractéristiques.

Si u est solution classique de (5.1),(5.2) alors u est constante le long de $\mathcal{C}_1$ et de $\mathcal{C}_2$. On doit donc avoir $u_0(x_1) = u_0(x_2)$ ce qui est impossible en général. Par conséquent, la solution ne peut être continue en P.

Cependant, dans le cas linéaire où $f(u) = c(x)u$ avec c une fonction lipschitzienne sur $\mathbb{R}$, les caractéristiques ne peuvent se croiser. En effet, les caractéristiques vérifient

$$\frac{dX}{dt}(t) = c(X(t)), \quad t > 0 \tag{5.5}$$

$$X(0) = x_0 \tag{5.6}$$

et d'après le théorème de Cauchy-Lipschitz, le problème (5.5), (5.6) admet une unique solution et par conséquent les caractéristiques ne peuvent se rencontrer. L'apparition de discontinuités est réellement liée à la nonlinéarité de l'équation.

On a néanmoins le résultat d'existence suivant pour le cas nonlinéaire général.

Proposition 5.2 *Soient $f \in C^2(\mathbb{R})$ et $u_0 \in C^1(\mathbb{R})$ bornée ainsi que sa dérivée. Il existe $T^* > 0$ tel que le problème (5.1),(5.2) admet une unique solution classique $u \in C^1(\mathbb{R} \times [0, T^*))$ donnée par*

$$u(x,t) = u_0(x_0(x,t)), \text{ pour tout } (x,t) \in \mathbb{R} \times [0,T^*), \tag{5.7}$$

où $x_0(x,t)$ est l'unique solution de $x_0 + f'(u_0(x_0))\, t = x$. De plus, si $\inf_{\mathbb{R}} \dfrac{df'(u_0)}{dx} \geq 0$ alors $T^* = +\infty$, sinon on peut choisir $T^* = \left(- \inf_{\mathbb{R}} \dfrac{df'(u_0)}{dx} \right)^{-1} > 0$.

Preuve. Existence. Puisque $f \in C^2(\mathbb{R})$ et que u_0 et sa dérivée sont bornées, la dérivée $\dfrac{df'(u_0)}{dx}$ est bornée sur $\mathbb{R}$. On peut alors définir T^* comme dans la proposition. Pour $t \in [0, T^*)$ fixé, on pose $F(y) = y + f'(u_0(y))\, t$ pour tout $y \in \mathbb{R}$. La fonction F est dérivable sur $\mathbb{R}$ et on a $F'(y) = 1 + \frac{df'(u_0)}{dx}(y)\, t \geq 1 + \left(\inf_{\mathbb{R}} \frac{df'(u_0)}{dx} \right) t > 0$ pour tout $t \in [0, T^*)$. La fonction F est donc strictement croissante sur $\mathbb{R}$. La donnée initiale u_0 étant bornée, on a $\lim_{y \to \pm\infty} F(y) = \pm\infty$ et par conséquent pour tout $t \in [0, T^*)$ et tout $x \in \mathbb{R}$, il existe un unique $x_0(x,t)$ tel que $x_0 + f'(u_0(x_0))\, t = x$. On vérifie alors facilement que la fonction u définie par (5.7) satisfait les équations (5.1),(5.2).

Unicité. Soit u_1 et u_2 deux solutions de (5.1),(5.2). D'après la Proposition 5.1, les deux solutions sont constantes le long des caractéristiques. Autrement dit, on a

$$u_i(X_i(t), t) = u_0(X_i(0)) \text{ pour } i = 1, 2, \quad t \in [0, T^*), \tag{5.8}$$

où X_i sont de classe C^1 sur $[0, T^*)$ et vérifient

$$\frac{dX_i}{d\tau}(\tau) = f'(u_i(X_i(\tau), \tau)), \quad \tau \in [0, T^*) \tag{5.9}$$

$$X_i(0) = x_0 \tag{5.10}$$

avec x_0 l'unique solution de $x_0 + f'(u_0(x_0))\, t = x$ pour $x \in \mathbb{R}$ et $t \in [0, T^*)$ donnés. En combinant (5.8), (5.9) et (5.10) on obtient que $X_i(t) = x$ pour $i = 1, 2$ et par conséquent $u_1(x,t) = u_0(x_0) = u_1(x,t)$ et ceci quels que soient $x \in \mathbb{R}$ et $t \in [0, T^*)$. □

5.3 Solutions faibles

L'apparition possible de singularités (discontinuités) rend la définition de solutions classiques inadaptées dans de nombreux cas. La notion de solutions faibles permet d'obtenir des solutions non continues. Rappelons tout d'abord quelques définitions d'espaces. Pour un ouvert $\Omega \subset \mathbb{R}^d$, on note

$$L^\infty_{loc}(\Omega) = \{v \text{ mesurable}, \ v_{|K} \in L^\infty(\Omega), \ \forall K \text{ compact } \subset \Omega\} \tag{5.11}$$

$$= \{v \text{ mesurable}, \ \varphi v \in L^\infty(\Omega), \ \forall \varphi \in C^\infty_0(\Omega)\}. \tag{5.12}$$

$$C^1_0(\Omega) = \{v \in C^1(\Omega) \text{ telle que } v(x) = 0 \ \ \forall x \in \Omega \setminus K, \ \ K \text{ compact } \subset \Omega\}. \tag{5.13}$$

On rappelle que le support d'une fonction φ est l'adhérence de l'ensemble des points x tels que $\varphi(x) \neq 0$. Pour $\varphi \in C^1_0(\mathbb{R} \times (0, +\infty))$, on a $\varphi(x, 0) = 0$. En revanche, pour $\varphi \in C^1_0(\mathbb{R} \times [0, +\infty))$, on a $\lim_{t \to +\infty} \varphi(x,t) = 0$, $\forall x \in \mathbb{R}$ mais $\varphi(x,0) \neq 0$ en général.

Définition 5.2 *Soit $u_0 \in L^\infty_{loc}(\mathbb{R})$. On dit que u est solution faible de (5.1),(5.2) si $u \in L^\infty_{loc}(\mathbb{R} \times (0, +\infty))$ et vérifie*

$$\int_0^\infty \int_{\mathbb{R}} u \frac{\partial \varphi}{\partial t} + f(u) \frac{\partial \varphi}{\partial x} \, dxdt + \int_{\mathbb{R}} u_0(x) \varphi(x, 0) \, dx = 0 \quad \forall \varphi \in C^1_0(\mathbb{R} \times [0, +\infty)). \tag{5.14}$$

On vérifie sans peine les propriétés suivantes qui font le lien entre solutions classiques et solutions faibles.

Proposition 5.3

1. *Toute solution classique est solution faible.*
2. *Toute solution faible vérifie l'équation $u_t + f(u)_x = 0$ au sens des distributions dans $\mathbb{R} \times (0, +\infty)$.*
3. *Toute solution faible appartenant à $C^1(\mathbb{R} \times (0, +\infty)) \cap C^0(\mathbb{R} \times [0, +\infty))$ est solution classique .*

5.4 Relations de Rankine-Hugoniot

On a vu que des discontinuités pouvaient apparaître. On va voir qu'elles ne peuvent être n'importe quoi pour avoir une solution faible.

Supposons que pour un certain intervalle de temps, une solution faible u soit discontinue le long d'une courbe Σ paramétrée par $(\xi(t), t)$ avec $\xi \in C^1$. La dérivée $\xi'(t)$ est appelée *vitesse de propagation de la discontinuité*. On suppose également que u admet à gauche et à droite de la discontinuité des limites u^- et u^+ et enfin que u est C^1 ailleurs. Soit M un point de Σ et $D \subset \mathbb{R} \times (0, \infty)$ une petite boule ouverte centrée en M. On décompose $D = D^- \cup D^+ \cup (\Sigma \cap D)$ où D^- (resp. D^+) est la partie de D située à gauche (resp. à droite) de Σ lorsqu'on parcourt celle-ci avec t (cf. Figure 5.2).

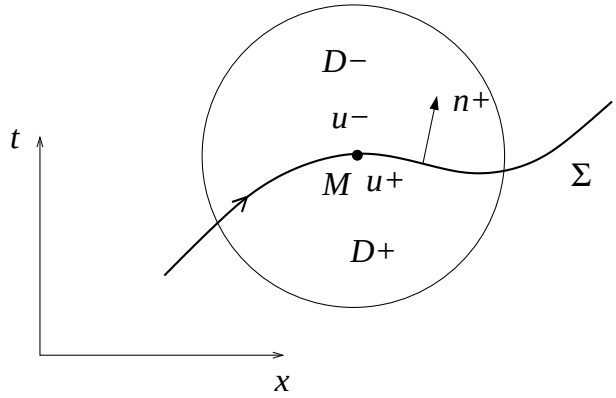


FIGURE 5.2 – Courbe de discontinuités et sous-domaines.

Soit $\varphi \in C_0^\infty(D)$. La fonction u étant solution faible de (5.1),(5.2), on a alors

$$\int_D \left(u \frac{\partial \varphi}{\partial t} + f(u) \frac{\partial \varphi}{\partial x} \right) dxdt = 0.$$

On décompose l'intégrale sur D en la somme de deux intégrales sur D^- et D^+ et on applique la formule de Green sur chaque intégrale. On obtient ainsi avec D^+

$$\int_{D^+} u \frac{\partial \varphi}{\partial t}\, dxdt = -\int_{D^+} \frac{\partial u}{\partial t} \varphi\, dxdt + \int_{\partial D^+} u \varphi n^+_{(t)}\, d\sigma,$$

où $n^+ = (n^+_{(x)}, n^+_{(t)})^\top$ est la normale unitaire extérieure à ∂D^+. Or

$$\int_{\partial D^+} u\varphi\, n^+_{(t)}\, d\sigma = \int_{\Sigma \cap D} u^+ \varphi\, n^+_{(t)}\, d\sigma.$$

Par ailleurs, Σ est la courbe de niveau de F définie par $F(x,t) = x - \xi(t)$ donc $\nabla_{(x,t)} F$ est un vecteur normal à Σ. On a $\nabla_{(x,t)} F = (\partial F/\partial x, \partial F/\partial t)^\top = (1, -\xi'(t))^\top$ et donc $n^+_{(x)} = 1/|\nabla F|$ et $n^+_{(t)} = -\xi'(t)/|\nabla F|$ avec $|\nabla F| = \sqrt{1 + (\xi'(t))^2}$. De plus,

$$\begin{aligned} \int_{D^+} f(u) \frac{\partial \varphi}{\partial x}\, dxdt &= -\int_{D^+} \frac{\partial f(u)}{\partial x} \varphi\, dxdt + \int_{\partial D^+} f(u) \varphi\, n^+_{(x)}\, d\sigma \\ &= -\int_{D^+} \frac{\partial f(u)}{\partial x} \varphi\, dxdt + \int_{\Sigma \cap D} f(u^+) \varphi\, \frac{d\sigma}{|\nabla F|}. \end{aligned}$$

Ainsi, on a

$$\int_{D^+} \left(u \frac{\partial \varphi}{\partial t} + f(u) \frac{\partial \varphi}{\partial x} \right) dxdt = -\int_{D^+} \left(\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} \right) \varphi\, dxdt - \int_{\Sigma \cap D} u^+ \xi'(t) \varphi\, \frac{d\sigma}{|\nabla F|} + \int_{\Sigma \cap D} f(u^+) \varphi\, \frac{d\sigma}{|\nabla F|}.$$

Or $\dfrac{\partial u}{\partial t} + \dfrac{\partial f(u)}{\partial x} = 0$ car $u \in C^1(D^+)$. Par conséquent, on obtient

$$\int_{D^+} \left(u\frac{\partial \varphi}{\partial t} + f(u)\frac{\partial \varphi}{\partial x} \right) dxdt = \int_{\Sigma \cap D} \left(f(u^+) - u^+ \xi'(t) \right) \varphi \, \frac{d\sigma}{|\nabla F|}.$$

On montre de même,

$$\int_{D^-} \left(u\frac{\partial \varphi}{\partial t} + f(u)\frac{\partial \varphi}{\partial x} \right) dxdt = -\int_{\Sigma \cap D} \left(f(u^-) - u^- \xi'(t) \right) \varphi \, \frac{d\sigma}{|\nabla F|}$$

et donc en sommant, on obtient

$$\int_{\Sigma \cap D} \left((f(u^+) - f(u^-)) - \xi'(t)(u^+ - u^-) \right) \varphi \, \frac{d\sigma}{|\nabla F|} = 0.$$

La relation précédente étant vraie quel que soit $\varphi \in C_0^\infty(D)$, on en déduit que

$$\xi'(t)(u^+ - u^-) = f(u^+) - f(u^-)$$

que l'on écrit sous la forme

$$\xi'(t)[u] = [f(u)].$$

Cette relation (condition nécessaire) est appelée relation de Rankine-Hugoniot. En fait, on peut montrer le résultat suivant.

Proposition 5.4 *Soit u une fonction de classe C^1 par morceaux. Plus précisément, on suppose qu'il existe un nombre fini de courbes régulières $\Sigma = \{(\xi(t), t)\} \subset \mathbb{R} \times (0, +\infty)$ en dehors desquelles u est de classe C^1 et à travers lesquelles u admet un saut $[u]$ (sur les courbes Σ, u admet une limite à gauche u^+ et une limite à droite u^- et $[u] = u^+ - u^-$). Alors u est solution faible de (5.1),(5.2) si et seulement si*

1. *u vérifie les équations (5.1),(5.2) là où elle est de classe C^1.*
2. *les relations de Rankine-Hugoniot*

$$\xi'(t)[u] = [f(u)], \tag{5.15}$$

sont satisfaites le long des courbes de discontinuité Σ.

On remarquera que si u est continue sur $\mathbb{R} \times [0, +\infty)$ alors les relations de Rankine-Hugoniot sont satisfaites. Les relations de Rankine-Hugoniot permettent de construire des solutions faibles (discontinues).

Example. Solutions de l'équation de Burgers.

On considère l'équation de Burgers

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\frac{u^2}{2}) &= 0, \quad x \in \mathbb{R},\ t > 0, \\ u(x,0) &= u_0(x) \end{aligned}$$

avec u_0 continue donnée par

$$u_0(x) = \begin{cases} 1, & x \leq 0 \\ 1 - x, & 0 \leq x \leq 1 \\ 0, & x \geq 1. \end{cases}$$

1) Solution continue.

On cherche une solution continue u de l'équation de Burgers sous la forme $u(x,t) = u_0(x_0)$ où $x_0 = x_0(x,t)$ vérifie la relation implicite $x = x_0 + f'(u_0(x_0))\, t$ (cf. Proposition 5.2). Pour l'équation de Burgers, x_0 vérifie

$$x = x_0 + u_0(x_0)\, t. \tag{5.16}$$

1. Si $x_0 \leq 0$ alors $u_0(x_0) = 1$ et $x = x_0 + t$ ce qui implique $x \leq t$. Réciproquement, si $x \leq t$ alors $x_0 = x - t$ vérifie (5.16) et $u(x,t) = u_0(x_0) = 1$.
2. Si $0 \leq x_0 \leq 1$ alors $u_0(x_0) = 1 - x_0$ et $x = x_0 + t(1 - x_0)$. Si $t < 1$ alors $t \leq x \leq 1$. Réciproquement, si $t < 1$ et t$\leq x \leq 1$ alors $x_0 = \dfrac{x-t}{1-t}$ vérifie (5.16) et $u(x,t) = u_0(x_0) = 1 - \dfrac{x-t}{1-t} = \dfrac{1-x}{1-t}$.
3. Si $x_0 \geq 1$ alors $u_0(x_0) = 0$ et $x = x_0$ ce qui implique $x \geq 1$. Réciproquement, si $x \geq 1$ alors $u(x,t) = 0$.

On a ainsi trouvé une solution u continue, définie pour tout $x \in \mathbb{R}$ et tout $t \in [0, T_{\max})$ avec $T_{\max} = 1$, par

$$u(x,t) = \begin{cases} 1, & x \leq t \\ \dfrac{1-x}{1-t}, & t \leq x \leq 1 \\ 0, & x \geq 1. \end{cases} \tag{5.17}$$

Remarque. Dans le cas $0 \leq x_0 \leq 1$, si $t > 1$ on ne peut pas obtenir de solution. En effet, pour $t > 1$, si $1 \leq x \leq t$ alors $u(x,t) = \dfrac{1-x}{1-t}$ ce qui contredit le cas 3 ($x_0 \geq 1$).

2) Solution faible : extension au delà de $T_{\max}$.

Les relations de Rankine-Hugoniot permettent de construire une solution faible (discontinue) au delà de $T_{\max}$. Pour cela, on procède de la façon suivante. Pour l'équation de Burgers, les caractéristiques sont des droites de pentes $1/u_0$ dans le plan (x,t) (cf. Figure 5.3).

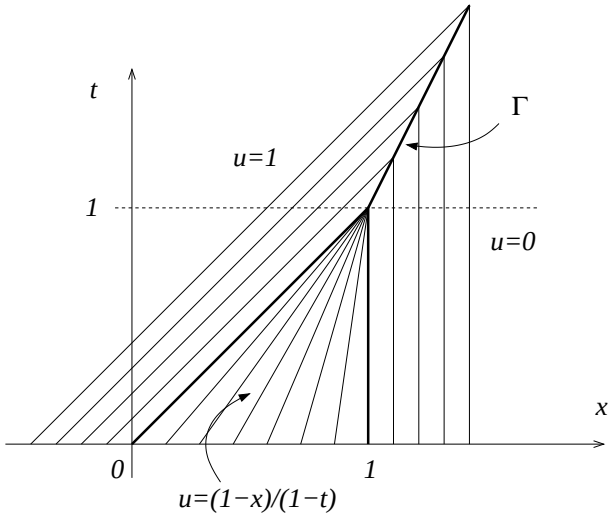


FIGURE 5.3 – Caractéristiques et solution faible globale pour l'équation de Burgers.

On prolonge u au delà de $t \geq 1$ par $u \equiv 1$ et $u \equiv 0$. La droite Γ des discontinuités sépare alors les régions du plan (x,t) avec $t \geq 1$ où $u \equiv 1$ et $u \equiv 0$ (cf. Figure 5.3). Dans le plan (x,t), cette droite a pour équation $t = \alpha x + \beta$. Plus précisément, on a

$$\Gamma = \{(\xi(t), t),\ t \geq 1 \text{ et } \xi(t) = (t - \beta)/\alpha\}.$$

La relation de Rankine-Hugoniot donne $\xi'(t)[u] = [f(u)]$ soit $1/\alpha[u] = 1/2[u^2]$ ou bien encore $1/\alpha(u^+ - u^-) = 1/2((u^+)^2 - (u^-)^2)$. Puisque $u^+ \equiv 1$ et $u^- \equiv 0$, on obtient nécessairement $\alpha = 2$ pour que la relation de Rankine-Hugoniot soit vérifiée le long de Γ. De plus, on déduit du fait que $\xi(1) = 1$, la valeur $\beta = -1$ et par conséquent on a

$$\Gamma = \{(x,t),\ t \geq 1 \text{ et } t = 2x - 1\}.$$

Pour $t \geq 1$, on obtient ainsi la solution

$$u(x,t) = \begin{cases} 1, & 1 \leq t \leq 2x-1 \\ 0, & t \geq 2x-1. \end{cases} \tag{5.18}$$

La fonction u définie sur $\mathbb{R} \times [0, +\infty)$ par (5.17) et (5.18) est, par construction et par application de la Proposition 5.4, une solution faible.

Remarque. Même en prenant une donnée initiale plus régulière, il y a apparition de discontinuités. Ce phénomène est vraiment du à la nonlinéarité de l'équation de Burgers.

5.5 Solutions d'entropie

L'introduction des solutions entropiques est motivée par la non unicité des solutions faibles. Par exemple, pour l'équation de Burgers avec la donnée initiale u_0 suivante (problème de Riemann)

$$u_0(x) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x > 0, \end{cases} \tag{5.19}$$

la solution u est telle que

$$u(x,t) = \begin{cases} 0 & \text{si } x < 0 \\ 1 & \text{si } x > t. \end{cases} \tag{5.20}$$

Pour la région $0 \leq x \leq t$, on peut construire deux solutions. La première solution est construite à partir de la courbe de discontinuité $\Gamma = \{(x,t), t = 2x\}$ et elle est donnée par

$$u_1(x,t) = \begin{cases} 0 & \text{si } x < t/2 \\ 1 & \text{si } x > t/2. \end{cases} \tag{5.21}$$

On vérifie facilement que les relations de Rankine-Hugoniot sont satisfaites et que u_1 est solution faible de l'équation de Burgers avec la donnée initiale (5.19). Une deuxième solution est donnée par

$$u_2(x,t) = \frac{x}{t} \quad \text{pour } 0 \leq x \leq t \tag{5.22}$$

et u_2 vérifiant (5.20). Les relations de Rankine-Hugoniot sont également satisfaites puisque u_2 est continue. La fonction u_2 est donc aussi une solution faible de l'équation de Burgers avec la donnée initiale (5.19).

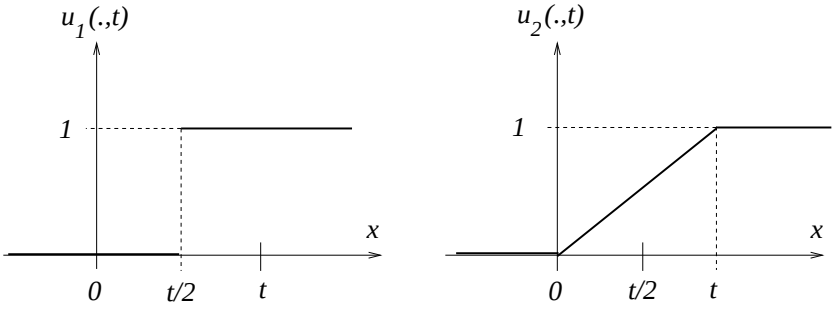


FIGURE 5.4 – Deux solutions faibles du problème de Riemann pour l'équation de Burgers.

La question est à présent de savoir quelle est la "bonne solution" ? La réponse est donnée par la notion d'entropie et de solution entropique.

Définition 5.3 *Soient U, $F \in C^1(\mathbb{R},\mathbb{R})$. Le couple (U,F) est un couple entropie - flux d'entropie si*

i) U est convexe.

ii) $F' = U'f'$ (F est une primitive de $U'f'$)

L'intérêt de cette définition est que si u vérifie (5.1) i.e. $u_t + f(u)_x = 0$, alors en utilisant le point *ii)* de la définition, on obtient

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} = 0.$$

Le point *ii)* de la définition peut être pris au sens des distributions si U et F ne sont pas C^1. Dans ces conditions, les fonctions

$$\begin{array}{rcl} U(u) & = & |u-k|, \quad k \in \mathbb{R} \\ F(u) & = & \text{signe}(u-k)(f(u)-f(k)) \end{array} \tag{5.23}$$

forment un couple entropie - flux d'entropie.

On peut introduire les solutions entropiques à partir des solutions de viscosité obtenue en perturbant la loi de conservation par un terme de viscosité/dissipation. Les lois de conservation sont d'ailleurs généralement des simplifications d'une réalité physique plus complexe. Considérons l'équation perturbée suivante avec $\varepsilon > 0$ donné,

$$(P_\varepsilon) \left\{ \begin{array}{rcl} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x} f(u) & = & \varepsilon \dfrac{\partial^2 u}{\partial x^2} \quad \text{pour } x \in \mathbb{R}, t > 0 \\ u(x,0) & = & u_0(x) \quad \text{pour } x \in \mathbb{R}. \end{array} \right.$$

On veut retrouver une solution de la loi de conservation (5.1),(5.2) lorsque $\varepsilon \to 0$.

Proposition 5.5 *On suppose que (P_ε) admet une solution régulière u_ε telle que*

i) il existe une conctante $C > 0$ indépendante de ε telle que $\|u_\varepsilon\|_{L^\infty(\mathbb{R}\times]0,+\infty[)} \leq C$,

ii) $u_\varepsilon \to u$ quand $\varepsilon \to 0$, p.p. dans $\mathbb{R}\times]0,+\infty[$.

Alors u est solution faible de (5.1),(5.2) et satisfait la condition d'entropie

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \leq 0 \quad \text{au sens des distributions dans } \mathbb{R}\times]0,+\infty[, \tag{5.24}$$

pour tout couple (U,F) d'entropie - flux d'entropie.

Remarque. L'inégalité d'entropie (5.24) signifie que

$$\int_0^{+\infty}\int_{\mathbb{R}} \left(U(u)\frac{\partial \varphi}{\partial t} + F(u)\frac{\partial \varphi}{\partial x} \right) dx\,dt \geq 0 \qquad \forall \varphi \in C_0^\infty(\mathbb{R}\times]0,+\infty[),\ \varphi \geq 0.$$

La condition d'entropie est un moyen de reconnaitre parmi les solutions faibles, les solutions d'origine physique. Si u est à support compact et vérifie la condition d'entropie (5.24) alors en intégrant par rapport à x, on obtient formellement

$$\int_{\mathbb{R}} \frac{\partial U(u)}{\partial t}\, dx \leq 0$$

c'est-à-dire que $\displaystyle\int_{\mathbb{R}} U(u)\, dx$ est décroissante en temps. Ceci correspond au deuxième principe de la thermodynamique.

Démonstration de la Proposition 5.5. Soit $\varphi \in C_0^\infty(\mathbb{R}\times[0,+\infty[)$. La solution u_ε de (P_ε) étant régulière, on peut alors multiplier par φ l'équation vérifiée par u_ε et intégrer par partie. On obtient

$$\int_0^{+\infty}\int_{\mathbb{R}} u_\varepsilon \varphi_t\, dx\,dt + \int_{\mathbb{R}} u_0(x)\varphi(x,0)\, dx + \int_0^{+\infty}\int_{\mathbb{R}} f(u_\varepsilon)\varphi_x\, dx\,dt + \varepsilon \int_0^{+\infty}\int_{\mathbb{R}} u_\varepsilon \varphi_{xx}\, dx\,dt = 0.$$

Par les hypothèses *i)*, *ii)* et le Théorème de convergence dominée de Lebesgue, on a, quand $\varepsilon \to 0$,

$$\begin{aligned} \int_0^{+\infty}\int_{\mathbb{R}} u_\varepsilon \varphi_t \, dx\, dt &\to \int_0^{+\infty}\int_{\mathbb{R}} u\varphi_t \, dx\, dt \\ \int_0^{+\infty}\int_{\mathbb{R}} f(u_\varepsilon)\varphi_x \, dx\, dt &\to \int_0^{+\infty}\int_{\mathbb{R}} f(u)\varphi_x \, dx\, dt, \end{aligned}$$

et d'autre part

$$\varepsilon \int_0^{+\infty}\int_{\mathbb{R}} u_\varepsilon \varphi_{xx} \, dx\, dt \to 0 \quad \text{quand } \varepsilon \to 0.$$

On conclut par densité de $C_0^\infty(\mathbb{R}\times[0,+\infty[)$ dans $C_0^1(\mathbb{R}\times[0,+\infty[)$ que u est solution faible de (5.1),(5.2). Pour établir que u vérifie la condition d'entropie (5.24), on montre d'abord que

$$\frac{\partial U(u_\varepsilon)}{\partial t} + \frac{\partial F(u_\varepsilon)}{\partial x} \le \varepsilon \frac{\partial^2 U(u_\varepsilon)}{\partial x^2}$$

au sens des distributions dans $\mathbb{R}\times]0,+\infty[$ et on passe à la limite avec $\varepsilon \to 0$. □

Définition 5.4 *Une fonction u est solution entropique de la loi de conservation* (5.1),(5.2) *si*

i) u est solution faible de (5.1),(5.2)

ii) u vérifie la condition d'entropie

$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} \le 0 \quad \textit{au sens des distributions dans } \mathbb{R}\times]0,+\infty[, \tag{5.25}$$

pour tout couple (U,F) d'entropie - flux d'entropie.

Remarques.

1. Si u est solution classique alors u est solution entropique et pour tout couple (U,F) d'entropie - flux d'entropie, on a
$$\frac{\partial U(u)}{\partial t} + \frac{\partial F(u)}{\partial x} = 0 \quad \text{au sens des distributions dans } \mathbb{R}\times]0,+\infty[.$$
2. Soit $u \in C^1$ par morceaux, vérifiant (5.1),(5.2) là où elle est C^1 et satisfaisant les relations de Rankine-Hugoniot le long des courbes de discontinuités $\Sigma = \{(\xi(t),t)\}$. Alors u est une solution entropique *si et seulement si* $\xi'[U(u)] \ge [F(u)]$ pour tout couple (U,F) d'entropie - flux d'entropie (cf. démonstration de la Proposition 5.4).

On peut alors montrer qu'il y a unicité de la solution entropique. Le résultat suivant est dû à Kruzkhov (1956).

Théorème 5.1 *(*Kruzkhov*) Soient $u_0 \in L^\infty(\mathbb{R})$ et $f \in C^1(\mathbb{R})$. Le problème* (5.1),(5.2) *admet une et une seule solution entropique $u \in L^\infty(\mathbb{R}\times]0,+\infty[)$.*

Critères d'entropie.

On donne à présent quelques critères pour que la condition d'entropie soit satisfaite.

Proposition 5.6 Conditions d'entropie d'Oleinik.

Soit u une solution faible de la loi de conservation (5.1), (5.2) *telle que les relations de Rankine-Hugoniot soient satisfaites le long de courbes de discontinuité $\Sigma = \{(\xi(t),t)\}$. On note $s = \xi'(t)$ la vitesse de propagation des discontinuités. La relation d'entropie*

$$s[U(u)] \ge [F(u)] \tag{5.26}$$

est vérifiée pour tout couple (U, F) d'entropie - flux d'entropie si et seulement si une des 3 conditions suivantes est satisfaite :

$$(\mathcal{O}_1) \quad \begin{cases} i) & f(\alpha u_- + (1-\alpha)u_+) \geq \alpha f(u_-) + (1-\alpha) f(u_+) \quad \textit{si } u_+ > u_- \\ ii) & f(\alpha u_- + (1-\alpha)u_+) \leq \alpha f(u_-) + (1-\alpha) f(u_+) \quad \textit{si } u_+ < u_- \end{cases}$$

pour tout $0 \leq \alpha \leq 1$.

$$(\mathcal{O}_2) \quad s \geq \frac{f(u_+) - f(k)}{u_+ - k} \quad \textit{pour tout } k \in \mathbb{R} \textit{ compris entre } u_- \textit{ et } u_+.$$

$$(\mathcal{O}_3) \quad s \leq \frac{f(u_-) - f(k)}{u_- - k} \quad \textit{pour tout } k \in \mathbb{R} \textit{ compris entre } u_- \textit{ et } u_+.$$

La condition $(\mathcal{O}_1)$-*i)* exprime le fait que f est au-dessus de la corde reliant u_- à u_+ dans le cas où $u_+ > u_-$. La condition $(\mathcal{O}_1)$-*ii)* dit que f est au-dessous de la corde reliant u_+ à u_- dans le cas où $u_+ < u_-$. Autrement dit, les discontinuités admissibles pour que la condition d'entropie soit vérifiée, sont telles que f doit être au-dessus (resp. au-dessous) de la corde reliant u_- à u_+ dans le cas où $u_- < u_+$ (resp. dans le cas où $u_+ < u_-$) (voir Figures 5.5 et 5.6). En fait, on va montrer un résultat un peu plus précis à savoir que si la relation d'entropie (5.26) est vérifiée alors les trois conditions $(\mathcal{O}_1)$, $(\mathcal{O}_2)$ et $(\mathcal{O}_3)$ sont satisfaites.

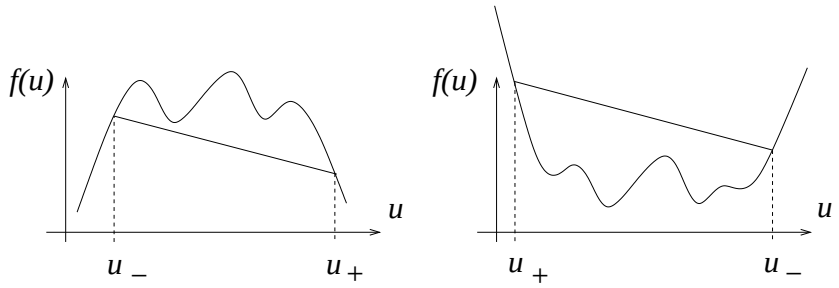


FIGURE 5.5 – Discontinuités admissibles pour la condition d'entropie.

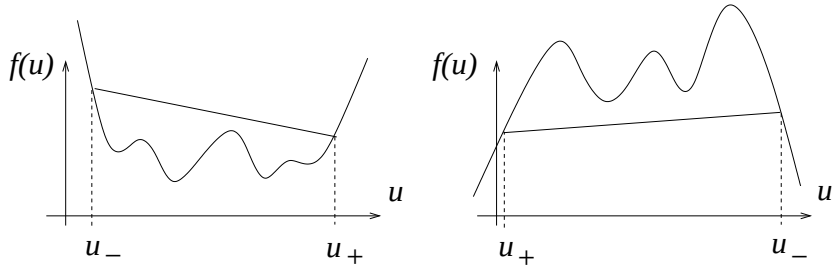


FIGURE 5.6 – Discontinuités non admissibles pour la condition d'entropie.

Démonstration des conditions nécessaires de la Proposition 5.6. On montre seulement une partie de la Proposition en établissant que les trois conditions $(\mathcal{O}_1)$, $(\mathcal{O}_2)$ et $(\mathcal{O}_3)$ sont nécessaires pour que la relation d'entropie (5.26) soit vérifiée. Supposons que la relation d'entropie (5.26) soit satisfaite pour tout couple (U, F) d'entropie - flux d'entropie. On prend alors

$$\begin{aligned} U(u) &= |u - k|, \quad k \in \mathbb{R} \\ F(u) &= \text{signe}(u-k)(f(u) - f(k)). \end{aligned}$$

Par hypothèse, on a

$$[F(u)] - s[U(u)] \leq 0 \quad \text{avec } s = \frac{[f(u)]}{[u]}. \tag{5.27}$$

Montrons que $(\mathcal{O}_1)$ est vérifiée. Soit $\alpha \in [0,1]$ fixé. On prend $k = \alpha u_- + (1-\alpha)u_+$ et on a

$$\begin{aligned} u_+ - k &= u_+ - \alpha u_- - (1-\alpha)u_+ = \alpha(u_+ - u_-) \\ u_- - k &= u_- - \alpha u_- - (1-\alpha)u_+ = (\alpha - 1)(u_+ - u_-). \end{aligned}$$

On obtient alors

$$\begin{aligned} [U(u)] &= U(u_+) - U(u_-) = |u_+ - k| - |u_- - k| \\ &= (2\alpha - 1)|u_+ - u_-| = (2\alpha - 1)(u_+ - u_-)\text{signe}(u_+ - u_-) \end{aligned}$$

et

$$\begin{aligned} [F(u)] &= F(u_+) - F(u_-) = \text{signe}(u_+ - u_-)\,(f(u_+) - f(k)) - \text{signe}(u_- - u_+)\,(f(u_-) - f(k)) \\ &= \text{signe}(u_+ - u_-)\,(f(u_+) + f(u_-) - 2f(k))\,. \end{aligned}$$

Ainsi, (5.27) devient

$$\left(f(u_+) + f(u_-) - 2f(k) - \frac{f(u_+) - f(u_-)}{u_+ - u_-}(2\alpha - 1)(u_+ - u_-) \right) \text{signe}(u_+ - u_-) \leq 0$$

soit

$$2\left((1-\alpha)f(u_+) + \alpha f(u_-) - f(k)\right)\text{signe}(u_+ - u_-) \leq 0,$$

ce qui implique la condition $(\mathcal{O}_1)$.

Montrons à présent que $(\mathcal{O}_2)$ est vérifiée. On suppose que $u_- < u_+$ (le cas contraire se traite de la même façon) et soit $k \in \mathbb{R}$ tel que $u_- < k < u_+$. On a

$$\begin{aligned} [U(u)] &= |u_+ - k| - |u_- - k| = u_+ + u_- - 2k \\ [F(u)] &= f(u_+) + f(u_-) - 2f(k). \end{aligned}$$

On obtient donc $s(u_+ + u_- - 2k) \geq f(u_+) + f(u_-) - 2f(k)$ et $f(u_-) = f(u_+) - s(u_+ - u_-)$. Ainsi,

$$2s(u_+ - k) \geq 2(f(u_+) - f(k))$$

ce qui implique $(\mathcal{O}_2)$. On montre de façon analogue, que la condition $(\mathcal{O}_3)$ est également satisfaite. □

Corollaire 5.1 *Soit $f \in C^1(\mathbb{R})$ strictement convexe (resp. strictement concave) et u une solution faible de* (5.1),(5.2)*, C^1 par morceaux. Alors u est solution entropique de* (5.1),(5.2) *si et seulement si en tout point de discontinuité, on a*

$$u_+ < u_- \quad (\textit{resp. } u_+ > u_-).$$

Remarque. Dans le cas où f est strictement convexe, la condition $u_+ < u_-$ est équivalente à

$$f'(u_+) < s < f'(u_-) \tag{5.28}$$

où s est la vitesse de propagation des discontinuités (il suffit d'utiliser le théorème des valeurs intermédiaires et le fait que f' est strictement croissante). C'est la condition de choc de Lax. Cette condition signifie que dans la cas où f est strictement convexe, la vitesse de propagation des discontinuités est toujours comprise entre les vitesses des caractéristiques évaluées de chaque côté de la discontinuité. Par conséquent, les lignes caractéristiques "rentrent" dans les discontinuités (voir Figure 5.7).

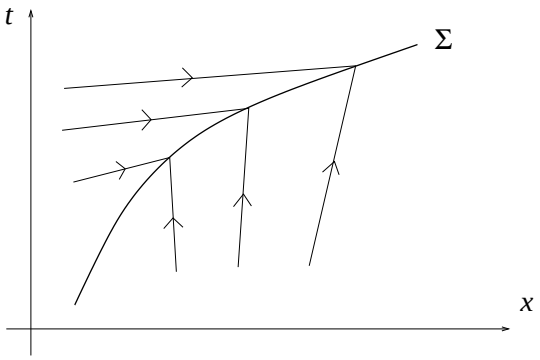


FIGURE 5.7 – Condition de choc de Lax.

5.6 Problème de Riemann

Le problème de Riemann pour la loi de conservation scalaire (5.1),(5.2) est caractérisé par le choix de la donnée initiale constante par morceaux (une marche). Plus précisément, le problème de Riemann s'écrit

$$\frac{\partial u}{\partial t} + \frac{\partial}{\partial x} f(u) = 0 \quad \text{pour } x \in \mathbb{R}, t > 0 \tag{5.29}$$

$$u(x,0) = u_0(x) = \begin{cases} u_g, & x < 0 \\ u_d, & x > 0, \end{cases} \tag{5.30}$$

où u_g et u_d sont deux constantes. On notera désormais

$$a = f'.$$

On cherche à résoudre exactement le problème de Riemann (5.29), (5.30). Ce problème intervient de façon essentiel dans la compréhension du comportement d'une loi de conservation. Il est à la base de nombreuses méthodes numériques de résolution des lois de conservation, en particulier celles qui seront décrites dans la prochaine section. On va tout d'abord considérer le cas où f est strictement convexe, puis on traitera le cas général.

5.6.1 Cas où f est strictement convexe

On suppose que f est strictement convexe et donc que a est strictement croissante. On cherche des solutions faibles u qui sont C^1 par morceaux et entropiques c'est-à-dire :

- les relations de Rankine-Hugoniot sont vérifiées : $s[u] = [f(u)]$ (avec $s = \xi'(t)$).
- en tout point de discontinuité $u_+ < u_-$.

a) Si $u_g = u_d$, l'unique solution entropique est $u \equiv u_g = u_d$.

b) Si $u_d < u_g$, l'unique solution entropique est

$$u(x,t) = \begin{cases} u_g, & x < st \\ u_d, & x > st \end{cases} \qquad \text{avec } s = \frac{f(u_d) - f(u_g)}{u_d - u_g}. \tag{5.31}$$

Cette solution présente une discontinuité (un choc) le long de la courbe $x = st$ et correspond à une onde de chocs ('shockwave') (voir Figure 5.8).

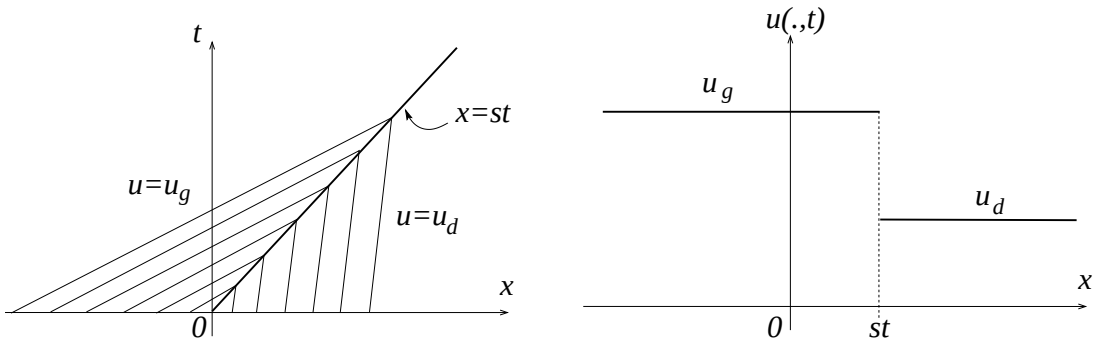


FIGURE 5.8 – Solution "onde de chocs" pour le problème de Riemann (cas $u_d < u_g$).

c) Si $u_g < u_d$ alors la discontinuité n'est pas admissible pour les solutions entropiques (cf. Corollaire 5.1). Il faut donc trouver une solution continue pour $t > 0$. On cherche la solution sous forme d'une "onde de raréfaction" ("rarefaction wave") définie par

$$u(x,t) = \begin{cases} u_g, & \text{si } x/t \le a(u_g) \\ v(x/t), & \text{si } a(u_g) \le x/t \le a(u_d) \\ u_d, & \text{si } x/t \ge a(u_d) \end{cases} \quad , \quad t > 0. \tag{5.32}$$

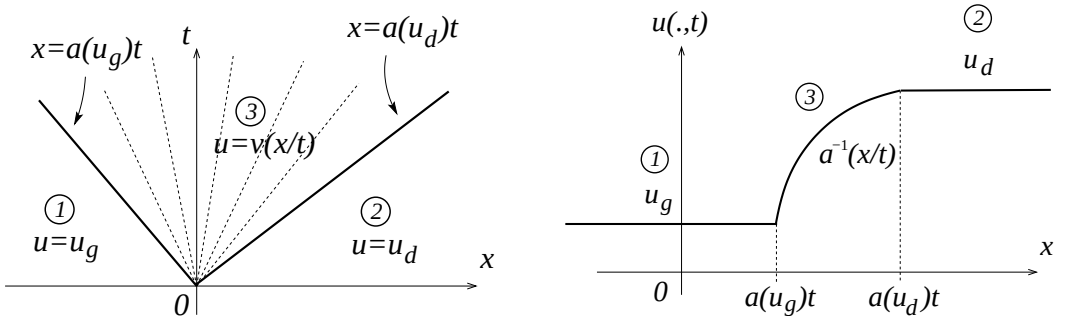


FIGURE 5.9 – Solution "onde de raréfaction" pour le problème de Riemann (cas $u_g < u_d$).

Pour déterminer la fonction v, on écrit que u est solution dans le cône ③. On a $u_t = -\frac{x}{t^2}v'(\frac{x}{t})$, $u_x = \frac{1}{t}v'(\frac{x}{t})$, et par conséquent $-\frac{x}{t}v'(\frac{x}{t}) + f'(v)v'(\frac{x}{t}) = 0$. On obtient ainsi

$$a(v(\frac{x}{t})) = \frac{x}{t}. \tag{5.33}$$

Comme a est strictement croissante, on obtient

$$\boxed{v(\frac{x}{t}) = a^{-1}(\frac{x}{t}).} \tag{5.34}$$

5.6.2 Cas général

On considère à présent le cas général où f n'est pas nécessairement convexe.

a) Si $u_g = u_d$ alors l'unique solution entropique est donnée par $u \equiv u_d = u_g$.

b) Si $u_g < u_d$, on considère alors l'enveloppe convexe de f dans $[u_g, u_d]$. L'intervalle $[u_g, u_d]$ se décompose en sous-intervalles I où f est strictement convexe, en alternance avec des sous-intervalles J où f est remplacée par une fonction affine. Pour fixer les idées, on suppose que l'intervalle $[u_g, u_d]$ se décompose

en intervalles $[u_{2i}, u_{2i+1}]$, $i = 1, \cdots, M-1$, où f est strictement convexe et les intervalles $[u_{2i-1}, u_{2i}]$, $i = 1, \cdots, M$ où f est remplacée par une fonction affine (cf. Figure 5.10).

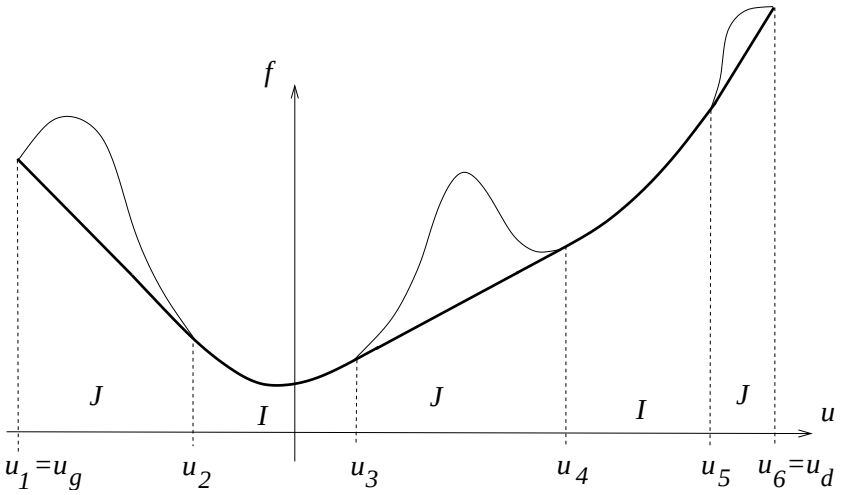


FIGURE 5.10 – Enveloppe convexe de f pour le problème de Riemann.

Dans I où f est strictement convexe, f' est strictement croissante donc $f'(u_0(\cdot))$ est croissante car u_0 est croissante. Ainsi les pentes des caractéristiques $\dfrac{1}{f'(u_0)}$ sont décroissantes. Dans ce cas, il n'y a pas de chocs puisque les caractéristiques ne se croisent pas pour $t > 0$. La solution est continue dans I et correspond à une "onde de raréfaction".

Dans J, la fonction f est remplacée par une fonction affine f_c. D'après les relations de Rankine-Hugoniot, il y a M courbes de discontinuité qui sont les droites définies par (voir la Figure 5.11)

$$x = s_{2i-1}t \quad \text{avec } s_{2i-1} = \frac{f(u_{2i}) - f(u_{2i-1})}{u_{2i} - u_{2i-1}} \quad \text{pour } i = 1, \cdots, M.$$

Par construction de l'enveloppe convexe de f, on a

$$\begin{aligned} &s_1 = a(u_2), \\ &s_3 = a(u_4) = a(u_3), \\ &\quad\vdots \\ &s_{2M-2} = a(u_{2M-1}) = a(u_{2M-2}), \\ &s_{2M-1} = a(u_{2M-1}). \end{aligned} \tag{5.35}$$

On remarquera que la condition d'entropie est vérifiée puisque f est au-dessus de f_c.

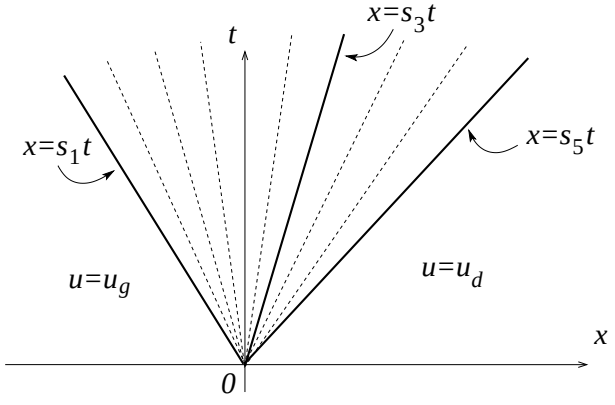


FIGURE 5.11 – Caractéristiques du problème de Riemann dans le cas général (avec l'exemple de la Figure 5.10).

La solution entropique est donnée par (voir Figure 5.12)

$$u(x,t) = \begin{cases} u_g, & x < s_1 t \\ v_i(x/t), & s_{2i-1}t < x < s_{2i+1}t, \quad i = 1, \cdots, M-1 \\ u_d, & x > s_{2M-1}t. \end{cases} \tag{5.36}$$

Les fonctions v sont déterminées de la même façon que dans la section précédente où f était strictement convexe. Les fonctions v_i vérifient (cf. (5.34))

$$a(v_i(\zeta)) = \zeta \quad \text{pour } \zeta \in [s_{2i-1}, s_{2i+1}] = [a(u_{2i}), a(u_{2i+1})] \tag{5.37}$$

Puisque a est strictement croissante sur l'intervalle $[u_{2i}, u_{2i+1}]$, l'équation (5.37) admet une unique solution

$$v_i(\zeta) = a^{-1}(\zeta), \quad \text{pour } \zeta \in [s_{2i-1}, s_{2i+1}] = [a(u_{2i}), a(u_{2i+1})]. \tag{5.38}$$

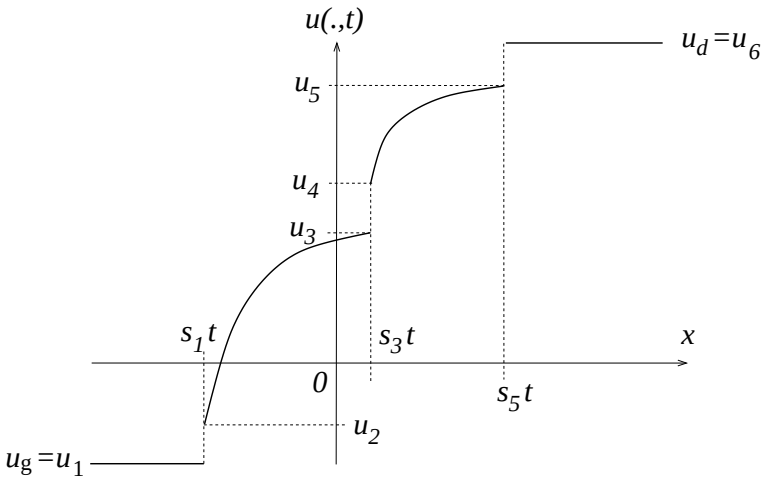


FIGURE 5.12 – Solution entropique du problème de Riemann (cas général).

c) Si $u_g > u_d$, on procède de la même façon avec l'enveloppe concave de f.

On note désormais $w_R(\frac{x}{t}, u_g, u_d)$ la solution entropique du problème de Riemann (5.29),(5.30) pour $t > 0$. La solution entropique du problème de Riemann possède la propriété suivante.

Proposition 5.7 *La solution entropique du problème de Riemann* (5.29), (5.30) *possède les propriétés suivantes.*

i) La fonction $z \mapsto f(w_R(\cdot, ug, u_d))$ *est continue en* $z = 0$ *c'est-à-dire*

$$f(w_R(0^-, u_g, u_d)) = f(w_R(0^+, u_g, u_d)). \tag{5.39}$$

ii) On note $I(u_g, u_d)$ *l'intervalle ouvert d'extrémités* u_g, u_d. *Alors*

$$signe(u_d - u_g) f(w_R(0, u_g, u_d)) = \min_{k \in I(u_g, u_d)} \left(signe(u_d - u_g) f(k) \right). \tag{5.40}$$

Démonstration.

i) Si $w_R(\cdot, u_g, u_d)$ est continue en $z = 0$ alors le résultat est évident. Si $w_R(\cdot, u_g, u_d)$ est discontinue en $z = 0$ alors la vitesse de propagation des discontinuités en 0 est nulle i.e. $s = 0$ car la donnée initiale

u_0 est discontinue en 0. Les relations de Rankine-Hugoniot impliquent alors que $f(w_R(0^-, u_g, u_d)) = f(w_R(0^+, u_g, u_d))$.

ii) On prouve le résultat dans le cas où f est strictement convexe. Le cas général est un peu plus compliqué mais se traite fondamentalement de la même façon en décomposant l'intervalle $[u_g, u_d]$ en sous-intervalles où f est strictement convexe, en alternance avec des intervalles où f est remplacée par une fonction affine.

– Suppose d'abord que $u_d < u_g$. La solution entropique u du problème de Riemann est donnée par (5.31). Si $0 < s = \dfrac{f(u_d) - f(u_g)}{u_d - u_g}$ alors $w_R(0, u_g, u_d) = u(0,t) = u_g$ et $f(w_R(0, u_g, u_d)) = f(u_g) \geq f(k)$ pour tout $k \in [u_d, u_g]$ car f étant convexe, elle est située en dessous de la corde reliant $f(u_d)$ à $f(u_g)$. On a donc dans ce cas $f(w_R(0, u_g, u_d)) = \max\limits_{k\in[u_d,u_g]} f(k)$. De même, si $0 > s = \dfrac{f(u_d) - f(u_g)}{u_d - u_g}$ alors $f(w_R(0, ug, u_d)) = f(u_d) \geq f(k)$ pour tout $k \in [u_d, u_g]$.

– Supposons maintenant que $u_d > u_g$. La solution entropique u du problème de Riemann est dans ce cas donnée par (5.32). Si $a(u_g) \geq 0$ alors $u(0,t) = u_g$ et puisque que f' est strictement croissante, on a $0 \leq f'(u_g) \leq f'(k)$ pour tout $k \in [u_g, u_d]$. Autrement dit, f est croissante sur $[u_g, u_d]$ et donc $f(u(0,t)) = f(u_g) \leq f(k)$ pour tout $k \in [u_g, u_d]$ c'est-à-dire $f(u(0,t)) = \min\limits_{k\in[u_g,u_d]} f(k)$. On procède de la même façon si $a(u_d) \leq 0$. Enfin, si $a(u_g) < 0 < a(u_d)$ alors $u(0,t) = v(0) = a^{-1}(0) = (f')^{-1}(0) = \gamma \in [u_g, u_d]$. On obtient ainsi $f'(\gamma) = 0$ et comme f est strictement convexe, on a $f(u(0,t)) = f(\gamma) = \min\limits_{k\in[u_g,u_d]} f(k)$, ce qui termine la démonstration dans le cas où f est strictement convexe.

□

5.7 Schémas d'approximations aux Différences Finies

5.7.1 Introduction et généralités

Soient $\Delta x > 0$ et $\Delta t > 0$ les pas de discrétisation respectivement en espace et temps. On note désormais

$$\lambda = \frac{\Delta t}{\Delta x}. \tag{5.41}$$

On introduit les points $x_j = j\Delta x$, $j \in \mathbb{Z}$ et les instants $t^n = n\Delta t$, $n \geq 0$. On s'intéresse alors aux approximations $u_j^n \simeq u(x_j, t^n)$ de la solution u de la loi de conservation (5.1),(5.2) et on considère des schémas de discrétisation *explicites* qui s'écrivent sous la forme générale suivante, pour $j \in \mathbb{Z}$ et $n \in \mathbb{N}$:

$$u_j^{n+1} = H(u_{j-k}^n, \cdots, u_j^n, \cdots, u_{j+k}^n), \tag{5.42}$$

avec $k \in \mathbb{N}$ et $H : \mathbb{R}^{2k+1} \to \mathbb{R}$ une fonction continue.

Définition 5.5 *Soit u_j^0, $j \in \mathbb{Z}$, donnés. Un schéma explicite de type* (5.42) *est dit* <u>conservatif</u> *s'il s'écrit sous la forme*

$$u_j^{n+1} = u_j^n - \lambda \left(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n \right) \tag{5.43}$$

où

$$g_{j+\frac{1}{2}}^n = g\left(u_{j-k+1}^n, \cdots, u_{j+k}^n\right) \tag{5.44}$$

et la fonction $g : \mathbb{R}^{2k} \to \mathbb{R}$ avec $k \in \mathbb{N}$, est continue et définit le <u>flux numérique</u> *du schéma conservatif* (5.43).

Explication heuristique du schéma conservatif (5.43).
Si on intègre l'équation $u_t + f(u)_x = 0$ entre $x_{j-\frac{1}{2}}$ et $x_{j+\frac{1}{2}}$ et entre t^n et t^{n+1}, on obtient

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t^{n+1})\, dx - \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x, t^n)\, dx + \int_{t^n}^{t^{n+1}} f(u(x_{j+\frac{1}{2}}, t))\, dt - \int_{t^n}^{t^{n+1}} f(u(x_{j-\frac{1}{2}}, t))\, dt = 0.$$

On note $\overline{u}_j^n$ la valeur moyenne de u dans l'intervalle $(x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$ évaluée en t^n i.e.

$$\overline{u}_j^n = \frac{1}{\Delta x}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t^n)\,dt$$

et on a

$$\overline{u}_j^{n+1} = \overline{u}_j^n - \frac{1}{\Delta x}\left(\int_{t^n}^{t^{n+1}} f(u(x_{j+\frac{1}{2}},t))\,dt - \int_{t^n}^{t^{n+1}} f(u(x_{j-\frac{1}{2}},t))\,dt\right).$$

En comparant avec le schéma conservatif (5.43), on voit que

$$u_j^n \simeq \overline{u}_j^n \quad \text{avec} \quad g_{j+\frac{1}{2}}^n \simeq \frac{1}{\Delta t}\int_{t^n}^{t^{n+1}} f(u(x_{j+\frac{1}{2}},t))\,dt.$$

Définition 5.6 *Le schéma conservatif* (5.43) *est dit consistant si*

$$g(u,\cdots,u) = f(u).$$

Proposition 5.8 *Un schéma* (5.43) *conservatif et consistant avec $g \in C^2$, est au moins d'ordre 1 en espace et en temps.*

Démonstration. Pour simplifier, on considère le cas $k=1$ c'est-à-dire avec $g = g(s_1,s_2)$ (la généralisation n'est pas difficile). On considère l'opérateur $L_{\Delta x,\Delta t}$ défini par

$$L_{\Delta x,\Delta t}u(x,t) = \frac{u(x,t+\Delta t)-u(x,t)}{\Delta t} + \frac{g(u(x,t),u(x+\Delta x,t)) - g(u(x-\Delta x,t),u(x,t))}{\Delta x}.$$

Par développement de Taylor, on a

$$\begin{aligned}
g(u(x,t),u(x+\Delta x,t)) &= g(u(x,t),u(x,t)) + \Delta x\frac{\partial g}{\partial s_2}(u(x,t),u(x,t))\cdot\frac{\partial u}{\partial x}(x,t) + \mathcal{O}(\Delta x^2),\\
g(u(x-\Delta x,t),u(x,t)) &= g(u(x,t),u(x,t)) - \Delta x\frac{\partial g}{\partial s_1}(u(x,t),u(x,t))\cdot\frac{\partial u}{\partial x}(x,t) + \mathcal{O}(\Delta x^2).
\end{aligned}$$

On obtient ainsi,

$$\begin{aligned}
L_{\Delta x,\Delta t}u(x,t) &= u_t(x,t) + \mathcal{O}(\Delta t) + \left(\frac{\partial g}{\partial s_2}(u(x,t),u(x,t)) + \frac{\partial g}{\partial s_1}(u(x,t),u(x,t))\right)u_x(x,t) + \mathcal{O}(\Delta x)\\
&= u_t(x,t) + \frac{\partial}{\partial x}\Big(g(u(x,t),u(x,t))\Big) + \mathcal{O}(\Delta t + \Delta x)\\
&= u_t(x,t) + \frac{\partial}{\partial x}\Big(f(u)\Big)(x,t) + \mathcal{O}(\Delta t + \Delta x).
\end{aligned}$$

□

5.7.2 Schéma de Godounov

On approche $u(x_{j+\frac{1}{2}},t)$ par $u_{j+\frac{1}{2}}^n$ pour tout $t \in (t^n,t^{n+1})$ où $u_{j+\frac{1}{2}}^n$ est la solution (exacte) du problème de Riemann suivant, évaluée en $x = x_{j+\frac{1}{2}}$:

$$(P_R)\begin{cases} \dfrac{\partial w}{\partial t} + \dfrac{\partial f(w)}{\partial x} = 0, \quad t \in]t^n,t^{n+1}[\\ w(x,t^n) = \begin{cases} u_j^n, & x < x_{j+\frac{1}{2}}\\ u_{j+1}^n, & x > x_{j+\frac{1}{2}}\end{cases}\end{cases}$$

La solution entropique de (P_R) s'écrit

$$w(x,t) = w_R\Big(\frac{x-x_{j+\frac{1}{2}}}{t-t^n}, u_j^n, u_{j+1}^n\Big), \tag{5.45}$$

où $w_R(\cdot, u_j^n, u_{j+1}^n)$ est la solution entropique du problème de Riemann (5.29), (5.30) avec $u_g = u_j^n$, $u_d = u_{j+1}^n$. On choisit

$$u_{j+\frac{1}{2}}^n = w(x_{j+\frac{1}{2}}, t) = w_R(0^-, u_j^n, u_{j+1}^n) \tag{5.46}$$

et de même on prend (cf. Figure 5.13)

$$u_{j-\frac{1}{2}}^n = w_R(0^+, u_{j-1}^n, u_j^n). \tag{5.47}$$

Le schéma de Godounov s'écrit alors

$$u_j^{n+1} = u_j^n - \lambda \left(g_{j+\frac{1}{2}}^n - g_{j-\frac{1}{2}}^n \right) \tag{5.48}$$

avec

$$g_{j+\frac{1}{2}}^n = g_{\text{godounov}} \left(u_j^n, u_{j+1}^n \right) = f(w_R(0, u_j^n, u_{j+1}^n)) = f(u_{j+\frac{1}{2}}^n). \tag{5.49}$$

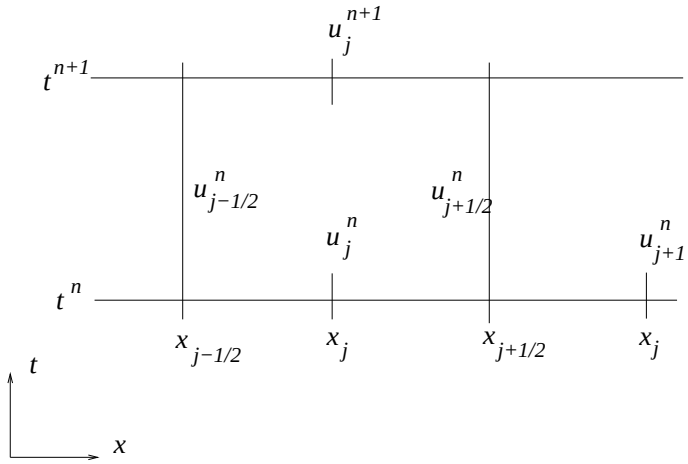


FIGURE 5.13 – Schéma de Godounov.

Remarques.

- La relation (5.49) a bien un sens puisque la fonction $z \mapsto f(w_R(z, \alpha, \beta))$ est continue en $z = 0$ (cf. Proposition 5.7).
- A chaque pas de temps, il faut calculer les solutions du problème de Riemann (P_R).

D'après la Proposition 5.7, le flux numérique de Godounov peut s'écrire :

$$g_{\text{godounov}}(u_j, u_{j+1}) = \begin{cases} \min\limits_{k \in [u_j, u_{j+1}]} f(k), & \text{si } u_j < u_{j+1} \\ \max\limits_{k \in [u_{j+1}, u_j]} f(k), & \text{si } u_j > u_{j+1}. \end{cases} \tag{5.50}$$

Finalement, u_j^{n+1} est aussi défini comme la valeur moyenne sur l'intervalle $[x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}]$ de la solution exacte w du problème de Riemann (P_R), à l'instant t^{n+1} :

$$u_j^{n+1} = \frac{1}{\Delta x} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} w(x, t^{n+1})\, dx. \tag{5.51}$$

Cas particuliers.

1. *Schéma upwind.* Si f est monotone, alors

$$g_{\text{godounov}}(s_1, s_2) = \begin{cases} f(s_1) & \text{si } f' > 0 \\ f(s_2) & \text{si } f' < 0 \end{cases} \tag{5.52}$$

2. Si f est strictement convexe sur $\mathbb{R}$, alors

$$g_{\text{godounov}}(s_1, s_2) = \begin{cases} \min(f(s_1), f(s_2)) & \text{si } s_1 < s_2 \text{ et } \overline{s} \notin [s_1, s_2] \\ f(\overline{s}) & \text{si } s_1 < \overline{s} < s_2 \\ \max(f(s_1), f(s_2)) & \text{si } s_1 \geq s_2 \end{cases} \tag{5.53}$$

où $\overline{s}$ est tel que $f'(\overline{s}) = 0$.

Proposition 5.9 *Sous la condition CFL suivante*

$$\lambda \max\left\{|f'(v)|,\ v \in I(u_j^n, u_{j+1}^n)\right\} \leq 1, \quad \textit{pour tout } j \in \mathbb{Z}$$

le schéma de Godounov converge vers la solution entropique de la loi de conservation.

Le schéma de Godounov est un schéma d'ordre 1 (il est conservatif et consistant), qui produit une régularisation près des discontinuités.

5.7.3 Schéma d'Engquist-Osher

Le flux numérique du schéma d'Engquist-Osher est donné par

$$g^{EO}(s_1, s_2) = \frac{1}{2}\left(f(s_1) + f(s_2) - \int_{s_1}^{s_2} |f'(s)| ds\right). \tag{5.54}$$

Cas particuliers.

i) Si f est monotone, on retrouve le schéma de Godounov *upwind* dont le flux numérique est donné par (5.52).

ii) Si f est strictement convexe sur $\mathbb{R}$, le flux numérique d'Engquist-Osher est alors donné par

$$g^{EO}(s_1, s_2) = f(s_1) + f(\min(s_2, \bar{s})) - f(\min(s_1, \bar{s})), \tag{5.55}$$

où le point $\bar{s}$ est l'unique point stationnaire de f i.e. tel que $f'(\bar{s}) = 0$.

Remarque. Les expressions (5.52) et (5.55) restent valables pour des fonctions f telles qu'il existe un unique $\bar{s}$ tel que $f'(\bar{s}) = 0,\ f'(s) > 0,\ \forall s > \bar{s}$ et $f'(s) < 0,\ \forall s < \bar{s}$.

Chapitre 6

Equations de Stokes

6.1 Introduction

Les équations de Stokes et de Navier-Stokes[1] modélisent la dynamique de fluides *visqueux incompressibles*. Dans ce chapitre on s'intéressera essentiellement à la discrétisation des équations linéaires de Stokes mais on va commencer par introduire les équations de Navier-Stokes (non linéaires) et on indiquera comment on peut en déduire en particulier les équations de Stokes. Le paragraphe suivant décrit un schéma de discrétisation par *Différences Finies* des équations de Stokes. Ce schéma sera étendu aux équations de Navier-Stokes au chapitre suivant.

Dans tout ce chapitre, on désigne par $\Omega \subset \mathbb{R}^N (N \leq 3)$ un domaine représentant une région de l'espace occupé par un fluide. Le domaine Ω sera toujours supposé borné et régulier. La dynamique d'un fluide *visqueux incompressible* peut être décrite par les équations de Navier-Stokes où les inconnues sont la vitesse $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ et la pression $p = p(\mathbf{x}, t)$ du fluide au point $\mathbf{x} = (x_1, \cdots, x_N) \in \Omega$ et à l'instant t. La vitesse est une fonction vectorielle $\mathbf{u} = (u_1, \cdots, u_N) \in \mathbb{R}^N$ avec $u_i = u_i(x_1, \cdots, x_N, t)$ et la pression p est une fonction scalaire. Les équations de Navier-Stokes sont :

$$\begin{aligned} \rho\,(\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u}) - \nu\Delta\mathbf{u} + \nabla p &= \mathbf{f} \quad \text{dans } \Omega \times \mathbb{R}^+, & (6.1)\\ \operatorname{div}\mathbf{u} &= 0 \quad \text{dans } \Omega \times \mathbb{R}^+. & (6.2) \end{aligned}$$

L'équation (6.2) traduit l'incompressibilité du fluide. La densité $\rho > 0$ du fluide est constante et $\nu > 0$ désigne la viscosité dynamique du fluide. Enfin, $\mathbf{f} = (f_1, \cdots, f_N)$ représente une densité massique de forces extérieures (la gravité par exemple). Les différents opérateurs différentiels intervenant dans les équations de Navier-Stokes sont définis par :

$$\nabla = \left(\frac{\partial}{\partial x_1}, \cdots, \frac{\partial}{\partial x_N}\right), \quad (\mathbf{u}\cdot\nabla)\mathbf{u} = \sum_{i=1}^{N} u_i \frac{\partial \mathbf{u}}{\partial x_i} \in \mathbb{R}^N,$$
$$\Delta\mathbf{u} = \sum_{i=1}^{N} \frac{\partial^2 \mathbf{u}}{\partial x_i^2} \in \mathbb{R}^N, \quad \operatorname{div}\mathbf{u} = \sum_{i=1}^{N} \frac{\partial u_i}{\partial x_i} \in \mathbb{R}.$$

Aux équations de Navier-Stokes (6.1), (6.2), on ajoute une condition limite de Dirichlet sur le bord $\partial\Omega$:

$$\mathbf{u}(\mathbf{x}, t) = \mathbf{g}(\mathbf{x}, t) \quad \text{pour } \mathbf{x} \in \partial\Omega,\ t > 0. \tag{6.3}$$

Dans la plupart des cas, on choisira la condition de non-glissement $\mathbf{u} = 0$ sur $\partial\Omega$. Enfin, on ajoute une condition initiale sur la vitesse :

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{u}_0(\mathbf{x}) \quad \text{pour } x \in \partial\Omega, \tag{6.4}$$

où $\mathbf{u}_0$ est une fonction donnée.

1. Henri Navier (1785-1836), George Stokes (1819-1903)

6.2 Adimensionalisation

Il est utile de travailler avec des équations de Navier-Stokes adimensionnalisées. En particulier, l'adimensionnalisation faite ci-dessous permettra d'obtenir les équations de Stokes et d'Euler comme cas *limites* (formelles) des équations de Navier-Stokes. Pour l'adimensionnalisation, on introduit une vitesse caractéristique $U \in \mathbb{R}$ de l'écoulement étudié (par exemple liée à une condition limite non-homogène) ainsi qu'une longueur caractéristique L (par exemple le diamètre de Ω). On considère alors le temps caractéristique $T = L/U$ et on pose

$$\tilde{x} = \frac{x}{L}, \quad \tilde{t} = \frac{t}{T}, \quad \tilde{\mathbf{u}}(\tilde{\mathbf{x}}, \tilde{t}) = \frac{\mathbf{u}(\mathbf{x}, t)}{U}, \quad \tilde{p}(\tilde{\mathbf{x}}, \tilde{t}) = \frac{p(\mathbf{x}, t)}{\rho U^2}. \tag{6.5}$$

Les nouvelles vitesse et pression $\tilde{\mathbf{u}}$ et $\tilde{p}$ vérifient alors

$$\rho\left(\frac{U}{L}\tilde{\mathbf{u}}_{\tilde{t}} + \frac{U^2}{L}(\tilde{\mathbf{u}}\cdot\nabla)\tilde{\mathbf{u}}\right) - \nu\frac{U}{L^2}\Delta\tilde{\mathbf{u}} + \frac{\rho U^2}{L}\nabla\tilde{p} = \mathbf{f} \quad \text{dans } \tilde{\Omega}\times\mathbb{R}^+.$$

Les (nouveaux) opérateurs différentiels ∇ et Δ ci-dessus sont relatifs à la (nouvelle) variable $\tilde{\mathbf{x}}$. On obtient ainsi

$$\tilde{\mathbf{u}}_t + (\tilde{\mathbf{u}}\cdot\nabla)\tilde{\mathbf{u}} - \frac{1}{Re}\Delta\tilde{\mathbf{u}} + \nabla\tilde{p} = \tilde{\mathbf{f}} \quad \text{dans } \tilde{\Omega}\times\mathbb{R}^+, \tag{6.6}$$

$$\operatorname{div}\tilde{\mathbf{u}} = 0 \quad \text{dans } \tilde{\Omega}\times\mathbb{R}^+, \tag{6.7}$$

avec $\tilde{\mathbf{f}} = \dfrac{L}{\rho U^2}\mathbf{f}$ et Re est le nombre de Reynolds défini par

$$Re = \frac{LU}{\nu}\rho. \tag{6.8}$$

Le nombre $\tilde{\nu} = \nu/\rho$ représente la viscosité cinématique. Par exemple, on a

$$\tilde{\nu} = 0.15\cdot 10^{-4} \text{ m/s pour l'air}$$
$$\tilde{\nu} = 10^{-6} \text{ m/s pour l'eau.}$$

Le tableau suivant indique quelques valeurs du nombre de Reynolds.

	U	L	$Re = LU/\tilde{\nu}$
bactérie (dans l'eau)	100 μm/s	0.1 μm	10^{-5}
protozoaire	10^{-1} cm/s	10^{-2} cm	10^{-1}
guêpe	2 cm/s	2 cm	26
papillon	1 m/s	5 cm	3333
pigeon	5 m/s	30 cm	10^5
poisson (hareng)	1.67 m/s	30 cm	$5\cdot 10^5$
poisson (saumon)	12.5 m/s	1 m	$1.25\cdot 10^7$
automobile	100 km/h	3 m	$5\cdot 10^6$
avion (airbus A330)	860 km/h	60 m	$\simeq 10^9$

Le nombre de Reynolds caractérise le type d'écoulement étudié. Mathématiquement, il prend en compte le terme de viscosité à travers le laplacien de la vitesse. Il induit le caractère *elliptique* de l'équation de Navier-Stokes.

6.3 Réductions des équations

A partir des équations de Navier-Stokes, on peut obtenir les équations de Stokes et d'Euler selon que le nombre de Reynolds Re est petit ou grand.

⋆ Pour $Re \ll 1$, les effets dus à la viscosité sont dominants. Si on pose $p' = LU\rho\,\tilde{p} = \nu Re\,\tilde{p}$ et $\mathbf{f}' = \nu Re\,\tilde{\mathbf{f}}$, l'équation (6.6) devient

$$\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} - \frac{1}{Re}\Delta\mathbf{u} + \frac{1}{\nu Re}\nabla p' = \frac{1}{\nu Re}\mathbf{f}'.$$

En faisant tendre Re vers 0, on obtient alors les équations de Stokes stationnaires (on oublie les *primes*) :

$$-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \quad \text{dans } \Omega, \tag{6.1}$$
$$\operatorname{div}\mathbf{u} = 0 \quad \text{dans } \Omega. \tag{6.2}$$

⋆ Pour $Re \gg 1$, le terme de convection nonlinéaire $(\mathbf{u}\cdot\nabla)\mathbf{u}$ est dominant ; dans ce cas, en faisant tendre Re vers $+\infty$ dans l'équation (6.6) (ou bien en prenant directement $\nu = 0$ dans (6.1)), on obtient les équations d'Euler (sans les *tildes*) :

$$\mathbf{u}_t + (\mathbf{u}\cdot\nabla)\mathbf{u} + \nabla p = \mathbf{f} \quad \text{dans } \Omega\times\mathbb{R}^+, \tag{6.3}$$
$$\operatorname{div}\mathbf{u} = 0 \quad \text{dans } \Omega\times\mathbb{R}^+. \tag{6.4}$$

Dans le paragraphe suivant, on s'intéresse à la résolution numérique du problème de Stokes (6.1),(6.2) par *Différences Finies*.

6.4 Discrétisation des équations de Stokes par *Différences Finies*

On considère le problème de Stokes *stationnaire* qui consiste à chercher $\mathbf{u} = \mathbf{u}(\mathbf{x})$ et $p = p(\mathbf{x})$ vérifiant

$$-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \quad \text{dans } \Omega, \tag{6.5}$$
$$\operatorname{div}\mathbf{u} = 0 \quad \text{dans } \Omega, \tag{6.6}$$
$$\mathbf{u} = \mathbf{g} \quad \text{sur } \partial\Omega. \tag{6.7}$$

On impose la condition limite de Dirichlet $\mathbf{u} = \mathbf{g}$ sur le bord du domaine. On suppose désormais que Ω est un ouvert borné connexe et que la donnée $\mathbf{g}$ est à flux nul sur $\partial\Omega$, c'est-à-dire qu'on suppose que

$$\int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n}\,d\sigma = 0 \tag{6.8}$$

où $\mathbf{n}$ est la normale extérieure à $\partial\Omega$. Cette condition est *nécessaire* pour que le problème de Stokes admette une unique solution $\mathbf{u}$ et une fonction p définie à une *constante* (additive) près. La condition sur $\mathbf{g}$ est une condition de compatibilité avec la condition d'incompressibilité $\operatorname{div}\mathbf{u} = 0$. En effet, on a

$$0 = \int_\Omega \operatorname{div}\mathbf{u}\,d\mathbf{x} = \int_{\partial\Omega} \mathbf{u}\cdot\mathbf{n}\,d\sigma = \int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n}\,d\sigma.$$

6.4.1 Introduction

On se place en dimension 2 d'espace avec $\Omega =]0,1[\times]0,1[$ et on choisit $\mathbf{g} = 0$ et $\nu = 1$. Le domaine Ω est discrétisé par un maillage uniforme défini par les points

$$P_{ij} = (ih,\ jh)\,, \quad i,j = 0,\cdots,N+1 \quad \text{avec } h = \frac{1}{N+1}. \tag{6.9}$$

On cherche alors les approximations

$$u_{ij}^{(1)} \simeq u_1(P_{ij}), \quad u_{ij}^{(2)} \simeq u_2(P_{ij}), \quad p_{ij} \simeq p(P_{ij}) \tag{6.10}$$

où les $u^{(i)}$ désignent les deux composantes de la vitesse exacte $\mathbf{u}$.

Une façon "naturelle" de discrétiser les équations de Stokes (6.5),(6.6) est d'utiliser des différences centrées :

$$\begin{aligned} -\Delta_h u_{ij}^{(1)} + \frac{1}{2h}\left(p_{i+1,j} - p_{i-1,j}\right) &= f_{ij}^{(1)} \\ -\Delta_h u_{ij}^{(2)} + \frac{1}{2h}\left(p_{i,j+1} - p_{i,j-1}\right) &= f_{ij}^{(2)} \\ \frac{1}{2h}\left(u_{i+1,j}^{(1)} - u_{i-1,j}^{(1)}\right) + \frac{1}{2h}\left(u_{i,j+1}^{(2)} - u_{i,j-1}^{(2)}\right) &= 0 \end{aligned} \tag{6.11}$$

où

$$\Delta_h u_{ij} = \frac{1}{h^2}\left(u_{i+1,j} - 2u_{ij} + u_{i-1,j}\right) + \frac{1}{h^2}\left(u_{i,j+1} - 2u_{ij} + u_{i,j-1}\right) \tag{6.12}$$

pour $i, j = 1, \cdots, N$.

Les conditions limites (avec $\mathbf{g} = 0$) se traduisent par

$$u_{0,j} = u_{N+1,j} = u_{i,0} = u_{i,N+1} = 0 \text{ pour } i, j = 0, \cdots, N+1.$$

Il y a $3N^2$ équations et $2N^2$ (pour $\mathbf{u}$)+$(N+2)^2$ (pour p) $= 3N^2 + 2(N+2)$ inconnues. Par conséquent, il y a plus d'inconnues que d'équations. Le système (6.11) ne possède pas de solution en général . Il n'est donc pas possible de prendre le même maillage pour la vitesse $\mathbf{u}$ et la pression p.

6.4.2 Schéma MAC pour le problème de Stokes

Pour obtenir le même nombre d'équations que d'inconnues, il faut s'arranger pour que les points de discrétisations de la pression p soient ceux auquels on discrétise l'équation d'incompressibilité. On discrétise l'équation $\operatorname{div} \mathbf{u} = 0$ par :

$$\frac{1}{2h}\left(u_{i,j}^{(1)} + u_{i,j+1}^{(1)} - u_{i+1,j}^{(1)} - u_{i+1,j+1}^{(1)}\right) + \frac{1}{2h}\left(u_{i,j}^{(2)} + u_{i+1,j}^{(2)} - u_{i,j+1}^{(2)} - u_{i+1,j+1}^{(2)}\right) = 0 \tag{6.13}$$

pour $i, j = 0, \cdots, N$.

Cette discrétisation représente l'équation $\operatorname{div} \mathbf{u} = 0$ aux points (cf. Figure 6.1)

$$P_{i+1/2,j+1/2} = \{(i+1/2)h,\ (j+1/2)h\}.$$

On cherche les approximations de la pression aux points où l'équation de la divergence est discrétisée c'est-à-dire les approximations

$$p_{i+1/2,j+1/2} \simeq p(P_{i+1/2,j+1/2}), \tag{6.14}$$

pour $i, j = 0, \cdots, N$. On introduit les valeurs moyennes de la pression :

$$p_{i+1/2,j} = \frac{1}{2}\left(p_{i+1/2,j+1/2} + p_{i+1/2,j-1/2}\right) \tag{6.15}$$

$$p_{i,j+1/2} = \frac{1}{2}\left(p_{i+1/2,j+1/2} + p_{i-1/2,j+1/2}\right). \tag{6.16}$$

L'équation de Stokes est alors discrétisée par

$$-\Delta_h u_{ij}^{(1)} + \frac{1}{h}\left(p_{i+1/2,j} - p_{i-1/2,j}\right) = f_{ij}^{(1)} \tag{6.17}$$

$$-\Delta_h u_{ij}^{(2)} + \frac{1}{h}\left(p_{i,j+1/2} - p_{i,j-1/2}\right) = f_{ij}^{(2)} \tag{6.18}$$

pour $i,j=1,\cdots,N$. En tenant compte de (6.15),(6.16), on obtient

$$-\Delta_h u_{ij}^{(1)} + \frac{1}{2h}\left(p_{i+1/2,j+1/2} + p_{i+1/2,j-1/2} - p_{i-1/2,j+1/2} - p_{i-1/2,j-1/2}\right) = f_{ij}^{(1)} \tag{6.19}$$

$$-\Delta_h u_{ij}^{(2)} + \frac{1}{2h}\left(p_{i+1/2,j+1/2} + p_{i-1/2,j+1/2} - p_{i+1/2,j-1/2} - p_{i-1/2,j-1/2}\right) = f_{ij}^{(2)} \tag{6.20}$$

pour $i,j=1,\cdots,N$.

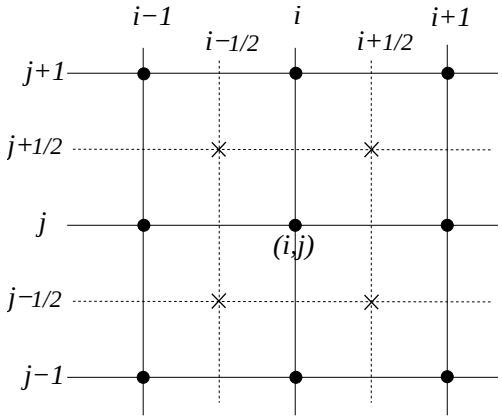


FIGURE 6.1 – Schéma MAC : • degré de liberté pour la vitesse ; x degré de liberté pour la pression.

Si on fait à présent le bilan du nombre d'équations et d'inconnues, on obtient $2N^2$ équations de Stokes (6.19),(6.20) et $(N+1)^2$ équations pour l'équation de la divergence (6.13), soit au total

$$2N^2+(N+1)^2 \quad \text{équations.}$$

Par ailleurs, on a $2N^2$ inconnues pour la vitesse et $(N+1)^2$ inconnues pour la pression, soit au total

$$2N^2+(N+1)^2 \quad \text{inconnues.}$$

Le schéma MAC fournit donc bien le même nombre d'équations que d'inconnues.

6.4.3 Forme matricielle du schéma de MAC

On ordonne les noeuds du maillage en partant du bas vers le haut et de la gauche vers la droite. Pour le vecteur vitesse, on stocke la première puis la deuxième composante :

$$\mathcal{U} = \Big(\underbrace{u_{11}^{(1)},u_{21}^{(1)},\cdots,u_{N2}^{(1)}}_{\text{1ère ligne pour } u_1},\cdots,\underbrace{u_{1N}^{(1)},\cdots,u_{NN}^{(1)}}_{\text{ligne } N \text{ pour } u_1},\ \underbrace{u_{11}^{(2)},\cdots,u_{N2}^{(2)}}_{\text{1ère ligne pour } u_2},\cdots,\underbrace{u_{1N}^{(2)},\cdots,u_{NN}^{(2)}}_{\text{ligne } N \text{ pour } u_2}\Big)^\top \in \mathbb{R}^{2N^2}. \tag{6.21}$$

De même pour la pression :

$$\mathcal{P} = \left(\underbrace{p_{\frac12,\frac12},p_{\frac32,\frac12},\cdots,p_{N+\frac12,\frac12}}_{\text{1ère ligne pour } p},\ \underbrace{p_{\frac12,\frac32},\cdots,p_{N+\frac12,\frac32}}_{\text{ligne 2 pour } p},\cdots,\ \underbrace{p_{\frac12,N+\frac12},\cdots,p_{N+\frac12,N+\frac12}}_{\text{ligne } N \text{ pour } p}\right)^\top \in \mathbb{R}^{(N+1)^2}. \tag{6.22}$$

Les relations (6.19), (6.20) s'écrivent matriciellement

$$A\mathcal{U}+B\mathcal{P}=\mathcal{F}. \tag{6.23}$$

⋆ Le vecteur $\mathcal{F}$ a la même structure que $\mathcal{U}$:

$$\mathcal{F} = \left(\underbrace{f_{11}^{(1)}, f_{21}^{(1)}, \cdots, f_{N2}^{(1)}, \cdots, f_{1N}^{(1)}, \cdots, f_{NN}^{(1)}}_{f^{(1)}}, \underbrace{f_{11}^{(2)}, \cdots, f_{N2}^{(2)}, \cdots, f_{1N}^{(2)}, \cdots, f_{NN}^{(2)}}_{f^{(2)}} \right)^{\top} \in \mathbb{R}^{2N^2}. \tag{6.24}$$

⋆ La matrice A est de taille $2N^2 \times 2N^2$ et s'écrit

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_1 \end{pmatrix}, \tag{6.25}$$

où la matrice A_1 est la matrice du Laplacien de taille $N^2 \times N^2$:

$$A_1 = \frac{1}{h^2} \begin{pmatrix} R & -I_d & & & 0 \\ -I_d & R & -I_d & & \\ & \ddots & \ddots & \ddots & \\ & & -I_d & R & -I_d \\ 0 & & & -I_d & R \end{pmatrix} \text{ avec } R = \begin{pmatrix} 4 & -1 & & & 0 \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ 0 & & & -1 & 4 \end{pmatrix} \text{ de taille } N \times N \tag{6.26}$$

et I_d désigne la matrice identité de taille $N \times N$.

⋆ La matrice B provient de la discrétisation du gradient de la pression. Elle est de taille $2N^2 \times (N+1)^2$ et s'écrit

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}. \tag{6.27}$$

La matrice B_1 (resp. B_2) correspond à la dérivée par rapport à x (resp. y) de la pression. Les matrices B_1 et B_2 sont toutes deux de taille $N^2 \times (N+1)^2$:

$$B_1 = \frac{1}{2h} \begin{pmatrix} B_{11} & B_{11} & & & 0 \\ & B_{11} & B_{11} & & \\ & & \ddots & \ddots & \\ 0 & & & B_{11} & B_{11} \end{pmatrix}, \quad B_2 = \frac{1}{2h} \begin{pmatrix} B_{22} & -B_{22} & & & 0 \\ & B_{22} & -B_{22} & & \\ & & \ddots & \ddots & \\ 0 & & & B_{22} & -B_{22} \end{pmatrix}. \tag{6.28}$$

Les matrices B_{11} et B_{22} sont toutes deux de taille $N \times (N+1)$ et valent

$$B_{11} = \begin{pmatrix} -1 & +1 & & & 0 \\ & -1 & +1 & & \\ & & \ddots & \ddots & \\ 0 & & & -1 & +1 \end{pmatrix}, \quad B_{22} = \begin{pmatrix} -1 & -1 & & & 0 \\ & -1 & -1 & & \\ & & \ddots & \ddots & \\ 0 & & & -1 & -1 \end{pmatrix}. \tag{6.29}$$

La relation de divergence (6.13) peut s'écrire matriciellement à l'aide de la matrice B précédente :

$$B^{\top} \mathcal{U} = 0. \tag{6.30}$$

En regroupant les relations (6.23) et (6.30), on obtient le système

$$\begin{pmatrix} A & B \\ B^{\top} & 0 \end{pmatrix} \begin{pmatrix} \mathcal{U} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} \mathcal{F} \\ 0 \end{pmatrix}. \tag{6.31}$$

Proposition 6.1 *Pour tout vecteur $\mathcal{F}$, le système de Stokes* (6.31) *admet au moins une solution $(\mathcal{U}, \mathcal{P})$. De plus, le vecteur vitesse $\mathcal{U}$ est unique.*

Démonstration. On note $\mathcal{M}$ la matrice carrée d'ordre $2N^2 + (N+1)^2$ du système de Stokes (6.31) :

$$\mathcal{M} = \begin{pmatrix} A & B \\ B^\top & 0 \end{pmatrix}.$$

Existence. On rappelle que $\operatorname{Im}\mathcal{M} = \left(\operatorname{Ker}\mathcal{M}^\top\right)^\perp$ et puisque $\mathcal{M}$ est symétrique, $\operatorname{Im}\mathcal{M} = (\operatorname{Ker}\mathcal{M})^\perp$ où par définition

$$(\operatorname{Ker}\mathcal{M})^\perp = \left\{ \mathbf{v} \in \mathbb{R}^{2N^2+(N+1)^2} \mid (\mathbf{v}, \mathbf{w}) = 0, \quad \forall \mathbf{w} \in \operatorname{Ker}\mathcal{M} \right\}.$$

Quelque soit le vecteur $\mathcal{F}$, on veut montrer que $(\mathcal{F}, 0)^\top \in \operatorname{Im}\mathcal{M} = (\operatorname{Ker}\mathcal{M})^\perp$ c'est-à-dire que $\left((\mathcal{F}, 0)^\top, \mathbf{w}\right) = 0$ pour tout $\mathbf{w} \in \operatorname{Ker}\mathcal{M}$. On décompose $\mathbf{w} = (\mathbf{w}_1, \mathbf{w}_2)^\top$ avec $\mathbf{w}_1 \in \mathbb{R}^{2N^2}$ et $\mathbf{w}_2 \in \mathbb{R}^{(N+1)^2}$. Puisque $\mathbf{w} \in \operatorname{Ker}\mathcal{M}$, on a $\mathcal{M}w = 0$ c'est-à-dire

$$A\mathbf{w}_1 + B\mathbf{w}_2 = 0 \tag{6.32}$$
$$B^\top \mathbf{w}_1 = 0. \tag{6.33}$$

On multiplie (6.32) par $\mathbf{w}_1$:

$$(A\mathbf{w}_1, \mathbf{w}_1) + (B\mathbf{w}_2, \mathbf{w}_1) = 0.$$

Or $(B\mathbf{w}_2, \mathbf{w}_1) = \left(\mathbf{w}_2, B^\top \mathbf{w}_1\right) = 0$ grâce à (6.33). Par conséquent, on obtient $(A\mathbf{w}_1, \mathbf{w}_1) = 0$ et donc $\mathbf{w}_1 = 0$ car la matrice A est définie positive (il s'agit de la matrice du Laplacien). Ainsi, on obtient $\left((\mathcal{F}, 0)^\top, \mathbf{w}\right) = (\mathcal{F}, \mathbf{w}_1) = 0$, pour tout $\mathbf{w} \in \operatorname{Ker}\mathcal{M}$, ce qui prouve que $(\mathcal{F}, 0)^\top \in \operatorname{Im}\mathcal{M}$.

Unicité de $\mathcal{U}$. On note par $\mathcal{U}$ et $\mathcal{P}$ les différences en vitesse et pression de deux solutions du système de Stokes (6.31). On veut montrer que $\mathcal{U} = 0$ et $\mathcal{P} = 0$. On a

$$A\mathcal{U} + B\mathcal{P} = 0 \tag{6.34}$$
$$B^\top \mathcal{U} = 0. \tag{6.35}$$

En procédant exactement comme avec le système (6.32),(6.33), on obtient $\mathcal{U} = 0$ ce qui prouve l'unicité de la vitesse. □

6.5 Résolution du système discrétisé de Stokes

Pour résoudre le système de Stokes (6.31), il faut modifier le système puisque la matrice $\mathcal{M}$ n'est pas inversible (mais le système (6.31) admet au moins une solution...). Pour cela, on utilise une méthode de pénalisation en considérant le système modifié suivant

$$\begin{pmatrix} A & B \\ B^\top & -\varepsilon I_d \end{pmatrix} \begin{pmatrix} \mathcal{U} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} \mathcal{F} \\ 0 \end{pmatrix}, \tag{6.36}$$

avec un paramètre $\varepsilon > 0$ petit. On résout alors ce système en découplant les équations vitesse/pression. C'est devenu possible avec l'introduction de la pénalisation :

$$\left(A + \frac{1}{\varepsilon} BB^\top\right) \mathcal{U} = \mathcal{F} \tag{6.37}$$
$$\mathcal{P} = \frac{1}{\varepsilon} B^\top \mathcal{U}. \tag{6.38}$$

La pénalisation a rendu inversible la matrice du système modifié (6.36). En fait, pour tout $\varepsilon > 0$ la matrice $A_\varepsilon = A + \dfrac{1}{\varepsilon} BB^\top$ est définie positive et le système (6.37),(6.38) admet une unique solution $(\mathcal{U}_\varepsilon, \mathcal{P}_\varepsilon)$.

Conditionnement de la matrice A_ε.

La méthode de pénalisation souffre de l'inconvénient d'être mal conditionnée avec ε petit. Le résultat suivant montre que le conditionnement de la matrice A_ε tend vers $+\infty$ quand ε tend vers 0.

Proposition 6.2 *Soit* $A_\varepsilon = A + \dfrac{1}{\varepsilon} BB^\top$ *avec* $\varepsilon > 0$. *On a*

$$K_2(A_\varepsilon) = \|A_\varepsilon\|_2 \|{A_\varepsilon}^{-1}\|_2 \geq \frac{1}{\varepsilon} \frac{\|BB^\top\|_2}{\sigma} \tag{6.39}$$

où $\sigma = \min_{\mathbf{v} \in Ker\, B^\top \setminus \{0\}} \dfrac{(A\mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2}$.

Démonstration. Le conditionnement spectral de A_ε est donné par $K_2(A_\varepsilon) = \dfrac{\max(\lambda_\varepsilon)}{\min(\lambda_\varepsilon)}$ où λ_ε désigne les valeurs propres de A_ε. On rappelle les relations de Courant-Fischer donnant la plus petite et la plus grande valeurs propres de A_ε à partir du quotient de Rayleigh R_{A_ε} :

$$\begin{aligned} R_{A_\varepsilon}(\mathbf{v}) &= \frac{(A_\varepsilon \mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2}, \\ \min(\lambda_\varepsilon) &= \min_{\mathbf{v} \in \mathbb{R}^{2N^2} \setminus \{0\}} R_{A_\varepsilon}(\mathbf{v}), \\ \max(\lambda_\varepsilon) &= \max_{\mathbf{v} \in \mathbb{R}^{2N^2} \setminus \{0\}} R_{A_\varepsilon}(\mathbf{v}). \end{aligned}$$

On a

$$R_{A_\varepsilon}(\mathbf{v}) = \frac{(A_\varepsilon \mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2} = \frac{(A\mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2} + \frac{1}{\varepsilon} \frac{(BB^\top \mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2}. \tag{6.40}$$

On en déduit

$$R_{A_\varepsilon}(\mathbf{v}) \geq \frac{1}{\varepsilon} \frac{(BB^\top \mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2} = \frac{1}{\varepsilon} R_{BB^\top}(\mathbf{v})$$

et par conséquent

$$\|A_\varepsilon\|_2 = \max(\lambda_\varepsilon) = \max_{\mathbf{v} \neq 0} R_{A_\varepsilon}(\mathbf{v}) \geq \frac{1}{\varepsilon} \max_{\mathbf{v} \neq 0} R_{BB^\top}(\mathbf{v}) = \frac{1}{\varepsilon} \|BB^\top\|_2. \tag{6.41}$$

Par ailleurs, pour $\mathbf{v} \in \operatorname{Ker} B^\top \setminus \{0\}$ la relation (6.40) se réduit à

$$R_{A_\varepsilon}(\mathbf{v}) = \frac{(A\mathbf{v},\, \mathbf{v})}{\|\mathbf{v}\|_2^2}. \tag{6.42}$$

On obtient ainsi

$$\min(\lambda_\varepsilon) = \min_{\mathbf{v} \neq 0} R_{A_\varepsilon}(\mathbf{v}) \leq \min_{\mathbf{v} \in \operatorname{Ker} B^\top \setminus \{0\}} R_{A_\varepsilon}(\mathbf{v}) = \sigma \tag{6.43}$$

En combinant (6.41), (6.42) et (6.43), on obtient bien l'estimation (6.39) du conditionnement de A_ε. □

6.6 Conditions de Dirichlet non-homogènes

On considère à présent la condition de Dirichlet non-homogène

$$\mathbf{u} = \mathbf{g} = (g_1,\ g_2)^\top \quad \text{sur le bord } \partial\Omega. \tag{6.44}$$

Dans la méthode des Différences Finies, la matrice A du système de Stokes (6.31) (ou (6.36)) est inchangée. Seul le second membre $\begin{pmatrix} \mathcal{F} \\ 0 \end{pmatrix}$ est modifié et s'écrit à présent $\begin{pmatrix} \mathcal{F}_1 \\ \mathcal{F}_2 \end{pmatrix}$ avec

$$\mathcal{F}_1 = \mathcal{F} + (G_1(P_{11}), \cdots, G_1(P_{NN}),\ G_2(P_{11}), \cdots, G_2(P_{NN}))^\top$$

et (pour $i = 1, 2$)

$$G_i(P) = \frac{1}{h^2} \sum_{\substack{Q \in \mathcal{V}(P) \\ Q \in \partial\Omega}} g_i(Q) \tag{6.45}$$

où $\mathcal{V}(P)$ est l'ensemble des noeuds du maillage $\overline{\Omega}_h$ qui sont les plus proches voisins de P. Si $\mathcal{V}(P) \cap \partial\Omega = \emptyset$ alors par convention $G_i(P) = 0$.

De plus,

$$\begin{aligned} \mathcal{F}_2 &= \big(H_1(P_{1/2,1/2}), H_1(P_{3/2,1/2}), \cdots, H_1(P_{N+1/2,N+1/2})\big)^\top \\ &\quad + \big(H_2(P_{1/2,1/2}), H_2(P_{3/2,1/2}), \cdots, H_2(P_{N+1/2,N+1/2})\big)^\top \in \mathbb{R}^{(N+1)^2} \end{aligned}$$

et pour $k = 1, 2$

$$H_k(P_{i+1/2,j+1/2}) = -\frac{1}{2h} \sum_{\substack{Q \in \mathcal{V}(P_{i+1/2,j+1/2}) \\ Q \in \partial\Omega}} \varepsilon_{ij}^k g_k(Q) \tag{6.46}$$

où $\varepsilon_{ij}^k = \varepsilon_{ij}^k(Q) \in \{+1, -1\}$ dépend du voisin Q de $P_{i+1/2,j+1/2}$ et de k. Plus précisément (cf. (6.13) et Fig. (6.2)), pour $k = 1$, on choisit

$$\begin{aligned} \varepsilon_{ij}^1 &= +1 \quad \text{si } Q = P_{i,j} \text{ ou } Q = P_{i,j+1} \\ \varepsilon_{ij}^1 &= -1 \quad \text{si } Q = P_{i+1,j} \text{ ou } Q = P_{i+1,j+1} \end{aligned}$$

De même, pour $k = 2$, on a

$$\begin{aligned} \varepsilon_{ij}^2 &= +1 \quad \text{si } Q = P_{i,j} \text{ ou } Q = P_{i+1,j} \\ \varepsilon_{ij}^2 &= -1 \quad \text{si } Q = P_{i,j+1} \text{ ou } Q = P_{i+1,j+1} \end{aligned}$$

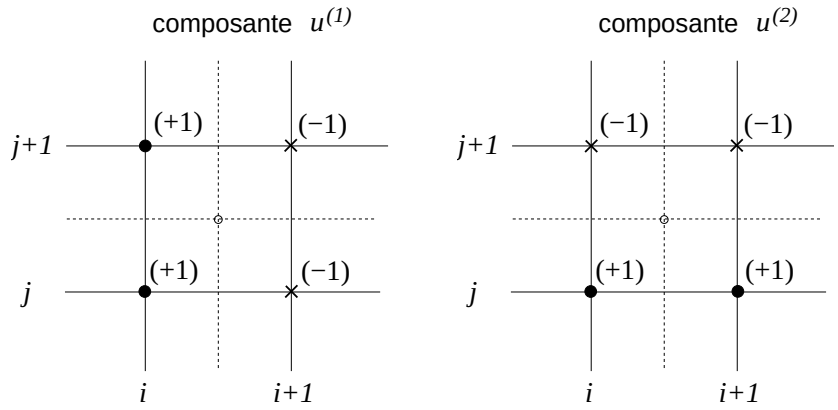


FIGURE 6.2 – Coefficients de $u^{(1)}$ (gauche) et $u^{(2)}$ (droite) dans l'équation de la divergence discrétisée près du bord.

Résolution du système pénalisé. Avec les conditions de Dirichlet non-homogènes, le système pénalisé s'écrit

$$\begin{pmatrix} A & B^\top \\ B & -\varepsilon I_d \end{pmatrix} \begin{pmatrix} \mathcal{U} \\ \mathcal{P} \end{pmatrix} = \begin{pmatrix} \mathcal{F}_1 \\ \mathcal{F}_2 \end{pmatrix}, \tag{6.47}$$

avec $\varepsilon > 0$ petit. On découple alors les équations vitesse/pression :

$$\left(A + \frac{1}{\varepsilon} B B^\top\right) \mathcal{U} = \mathcal{F}_1 + \frac{1}{\varepsilon} B \mathcal{F}_2 \tag{6.48}$$

$$\mathcal{P} = \frac{1}{\varepsilon} \left(B^T \mathcal{U} - \mathcal{F}_2\right) \tag{6.49}$$

6.7 Traitement pratique des conditions limites

Pour construire la matrice du système (6.31) et le second membre avec les conditions de Dirichlet non homogènes (6.44), on peut procéder d'une façon bien adaptée aux langages à script de type Matlab, Scilab, Octave ..., où on dispose de commandes de haut niveaux pour la manipulation des matrices et vecteurs. L'idée est de construire la matrice du système pour tous les noeuds y compris ceux du bord, puis de mettre à jour le second membre en calculant les termes (6.45), (6.46), et enfin de supprimer les lignes et les colonnes de la matrice correspondants à des noeuds du bord.

1. Commençons par construire la matrice du système avec la taille $2(N+2)^2+(N+1)^2\times 2(N+2)^2+(N+1)^2$ c'est-à-dire en tenant compte des points du bord :

$$M=\begin{pmatrix}\tilde{A} & \tilde{B}\\ \tilde{B}^\top & 0\end{pmatrix}. \tag{6.50}$$

• La matrice $\tilde{A}$ est de taille $2(N+2)^2\times 2(N+2)^2$ et est définie par

$$\tilde{A}=\begin{pmatrix}\tilde{A}_1 & 0\\ 0 & \tilde{A}_1\end{pmatrix} \text{ avec } \tilde{A}_1=\frac{\nu}{h^2}\begin{pmatrix} 4 & -1 & & -1 & & 0\\ -1 & 4 & -1 & & \ddots & \\ & \ddots & \ddots & \ddots & & -1\\ -1 & & \ddots & \ddots & \ddots & \\ & \ddots & & -1 & 4 & -1\\ 0 & & -1 & & -1 & 4\end{pmatrix}$$

La matrice $\tilde{A}_1$ est *pentadiagonale* de taille $(N+2)^2\times(N+2)^2$. Les deux sur et sous-diagonales de $\tilde{A}_1$ les plus éloignées de la diagonale principale sont à une distance $N+2$ de celle-ci.

• La matrice $\tilde{B}$ est de taille $\left[2(N+2)^2\times(N+1)^2\right]\times\left[2(N+2)^2\times(N+1)^2\right]$ et est définie par

$$B=\begin{pmatrix}\tilde{B}_1\\ \tilde{B}_2\end{pmatrix}, \tag{6.51}$$

avec les matrices $\tilde{B}_1$ et $\tilde{B}_2$ toutes deux de taille $(N+2)^2\times(N+1)^2$:

$$\tilde{B}_1=\frac{1}{2h}\begin{pmatrix}\tilde{B}_{11} & & 0\\ \tilde{B}_{11} & \ddots & \\ & \ddots & \tilde{B}_{11}\\ 0 & & \tilde{B}_{11}\end{pmatrix},\quad \tilde{B}_2=\frac{1}{2h}\begin{pmatrix}-\tilde{B}_{22} & & 0\\ \tilde{B}_{22} & \ddots & \\ & \ddots & -\tilde{B}_{22}\\ 0 & & \tilde{B}_{22}\end{pmatrix}. \tag{6.52}$$

Les matrices $\tilde{B}_{11}$ et $\tilde{B}_{22}$ sont toutes deux de taille $(N+2)\times(N+1)$ et valent

$$\tilde{B}_{11}=\begin{pmatrix}+1 & & 0\\ -1 & \ddots & \\ & \ddots & +1\\ & & -1\end{pmatrix},\quad \tilde{B}_{22}=\begin{pmatrix}-1 & & 0\\ -1 & \ddots & \\ & \ddots & -1\\ & & -1\end{pmatrix}. \tag{6.53}$$

2. Mise à jour du second membre et de la matrice. On construit le tableau d'indices des noeuds du bord pour la vitesse dans la matrice S. Si on stocke successivement les noeuds des bords $\{y = 0\}$, $\{x = 0\}$, $\{x = 1\}$, $\{y = 1\}$ et en tenant compte de l'ordonnancement choisi des noeuds, on a par exemple en Scilab

```
ibord=[[1:Nt], [1:Nt-2]*Nt+1, [2:(Nt-1)]*Nt, [(Nt-1)*Nt+1:Nt*Nt]];
```

avec `Nt=N+2`.

On désigne par `M` la matrice de Stokes (6.50), par `b` le vecteur valant $\begin{pmatrix}\mathcal{F}\\0\end{pmatrix}$ et par `ubd` le vecteur contenant les valeurs de la vitesse $\mathbf{g}$ sur le bord. On modifie alors le second membre de la façon suivante :

```
// indices du bord répétés (pour les deux composantes de la vitesse)
ibordS=[ibord ibord+Nt*Nt];

// traitement du second membre: on bascule à droite les termes connus à gauche.
b=b-M(:,ibordS)*ubd;
```

Il ne reste plus qu'à supprimer les lignes et colonnes de `M` et `b` correspondants à des noeuds du bord :

```
// On efface les lignes et colonnes de M correspondants à des noeuds du bord:
// on recupère un système à  2*N*N inconnues (au lieu de 2*(N+2)*(N+2))
M(:,ibordS)=[]; M(ibordS,:)=[];
b(ibordS)=[];
```

6.8 Exemples

• Si on choisit $\mathbf{f}(\mathbf{x}) = \begin{pmatrix}-6x_1x_2^2 - 2x_1^3\\6x_1^2x_2 + 2x_2^3\end{pmatrix}$ et $\mathbf{g}(\mathbf{x}) = \begin{pmatrix}x_1^3x_2^2 + 1/3\\-x_1^2x_2^3 + 2/3\end{pmatrix}$ sur le bord $\partial\Omega$, alors la solution exacte du problème de Stokes est donnée (avec $\nu = 1$) par :

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix}x_1^3x_2^2 + 1/3\\-x_1^2x_2^3 + 2/3\end{pmatrix} \text{ et } p = Constante.$$

La figure (6.3) représente le champ de vitesse numériquement calculé avec la méthode MAC sur un maillage 91×91. La résolution du système pénalisé a été faite avec $\varepsilon = 10^{-7}$.

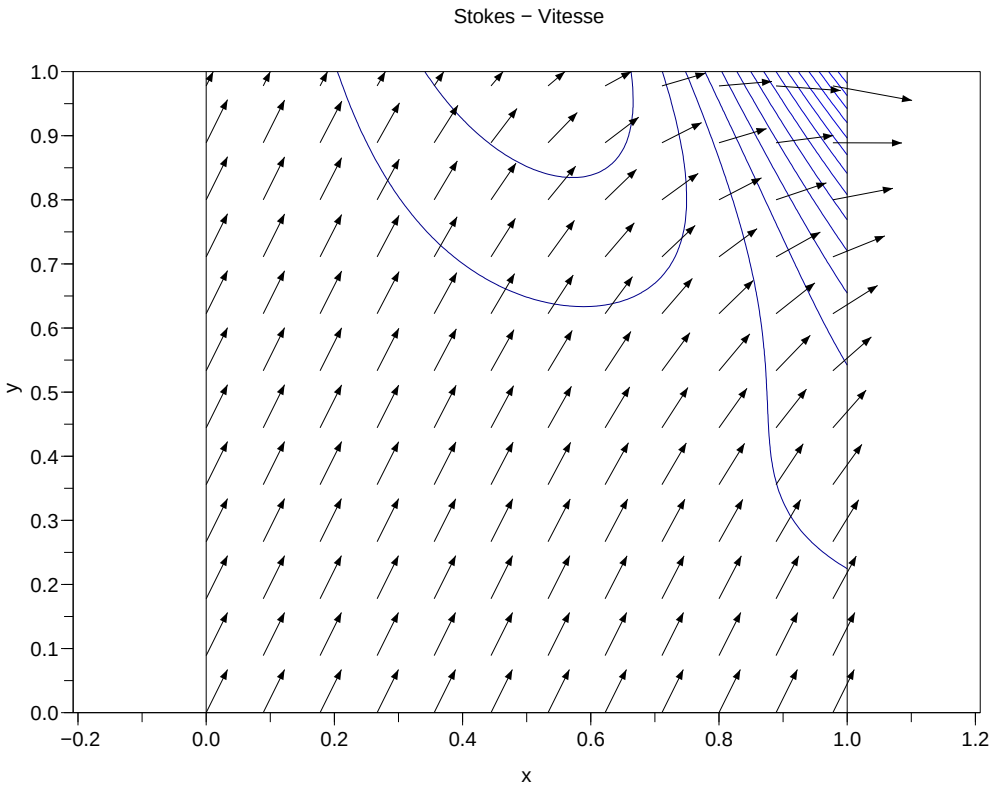


FIGURE 6.3 – Equations de Stokes - Vitesse.

Cette solution exacte permet d'évaluer l'ordre de convergence de la méthode. On trouve numériquement un ordre $\mathcal{O}(h^2)$.

• *Cavité entrainée.* Dans cet exemple, on impose une vitesse nulle partout sur le bord du carré sauf sur le bord supérieur $\{y = 1\}$ où la vitesse est horizontale. Plus précisément, on choisit

$$\mathbf{u} = \mathbf{g} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ sur les bords } \{x = 0\}, \{x = 1\} \text{ et } \{y = 0\}.$$

$$\mathbf{u} = \mathbf{g} = \begin{pmatrix} 15 \\ 0 \end{pmatrix} \text{ sur le bord } \{y = 1\}.$$

Avec une force de gravité verticale $\mathbf{f} = \begin{pmatrix} 0 \\ -30 \end{pmatrix}$ et la viscosité $\nu = 10^{-2}$, on obtient le résultat reproduit sur la figure suivante (maillage 91×91).

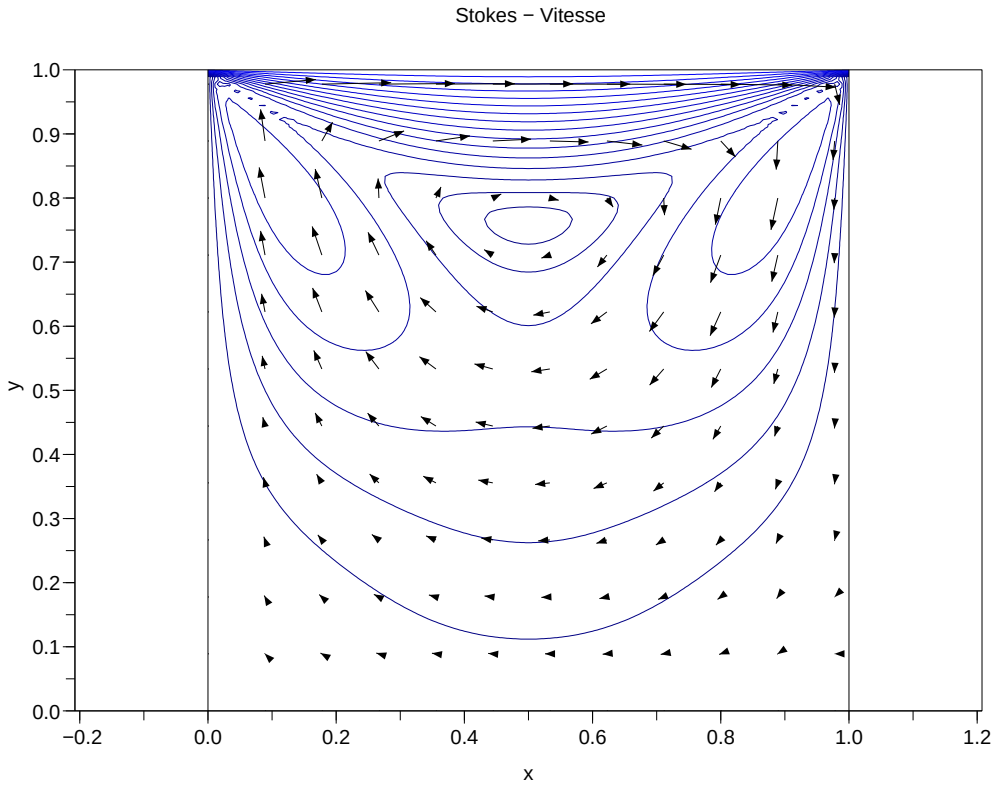


FIGURE 6.4 – Equations de Stokes - Cavité entrainée.

Chapitre 7

Equations de Navier-Stokes

7.1 Introduction

On considère les équations de Navier-Stokes incompressibles posées dans un domaine $\Omega \subset \mathbb{R}^2$. Afin de ne pas compliquer la description de la méthode numérique, on choisit désormais le carré unité comme domaine c'est-à-dire $\Omega = (0,1) \times (0,1)$. On va décrire un schéma de type MAC (Marker And Cell) pour les équations de Navier-Stokes incompressibles. Il s'agit d'une extension du schéma décrit au chapitre précédent pour les équations de Stokes. Ce schéma est semi-implicite en temps et il est obtenu en linéarisant (en temps) le terme nonlinéaire des équations de Navier-Stokes.

Dans les équations de Navier-Stokes incompressibles, les inconnues sont la vitesse $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ et la pression $p = p(\mathbf{x}, t)$ du fluide au point $\mathbf{x} = (x, y) \in \Omega$ et à l'instant t. La vitesse est une fonction vectorielle $\mathbf{u} = (u, v) \in \mathbb{R}^2$ avec $u = u(x, y, t)$, $v = v(x, y, t)$ et la pression p est une fonction scalaire. Les inconnues $\mathbf{u}$ et p vérifient les équations de Navier-Stokes sous la forme adimensionnée suivante :

$$\begin{aligned} \mathbf{u}_t + (\mathbf{u} \cdot \nabla)\mathbf{u} - \nu \Delta \mathbf{u} + \nabla p &= \mathbf{f} \quad \text{dans } \Omega \times \mathbb{R}^+, && (7.1) \\ \operatorname{div} \mathbf{u} &= 0 \quad \text{dans } \Omega \times \mathbb{R}^+. && (7.2) \\ \mathbf{u} &= \mathbf{g} \quad \text{sur } \partial\Omega \times \mathbb{R}^+ && (7.3) \\ \mathbf{u}(\cdot, 0) &= \mathbf{u}_0 \quad \text{dans } \Omega. && (7.4) \end{aligned}$$

L'équation (6.2) traduit l'incompressibilité du fluide. Le paramètre $\nu > 0$ prend en compte la viscosité du fluide et $\mathbf{f} = (f_1, f_2)$ représente les forces extérieures (la gravité par exemple). La fonction $\mathbf{g} = \mathbf{g}(\mathbf{x}, t)$ dans la condition limite de Dirichlet (7.3) sur le bord du domaine doit être à flux nul sur $\partial\Omega$, c'est-à-dire qu'on suppose que

$$\int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} \, d\sigma = 0 \tag{7.5}$$

où $\mathbf{n}$ est la normale extérieure à $\partial\Omega$. Cette condition est *nécessaire* pour que le problème de Navier-Stokes admette une solution. La condition sur $\mathbf{g}$ est une condition de compatibilité avec la condition d'incompressibilité $\operatorname{div} \mathbf{u} = 0$ car

$$0 = \int_\Omega \operatorname{div} \mathbf{u} \, d\mathbf{x} = \int_{\partial\Omega} \mathbf{u} \cdot \mathbf{n} \, d\sigma = \int_{\partial\Omega} \mathbf{g} \cdot \mathbf{n} \, d\sigma.$$

7.2 Semi-discrétisation en temps

Soit $\Delta t > 0$, le pas de discrétisation en temps et on note les instants $t^n = n\Delta t$, $n \in \mathbb{N}$. On définit les approximations en temps de la vitesse $\mathbf{u}^n(\mathbf{x}) \simeq \mathbf{u}(\mathbf{x}, t^n)$ et de la pression $p^n(\mathbf{x}) \simeq p(\mathbf{x}, t^n)$.

Connaissant $\mathbf{u}^{n-1}$, on détermine $\mathbf{u}^n$ et p^n par le schéma

$$\begin{aligned} \frac{\mathbf{u}^n - \mathbf{u}^{n-1}}{\Delta t} + (\mathbf{u}^{n-1}\cdot\nabla)\mathbf{u}^n - \nu\Delta\mathbf{u}^n + \nabla p^n &= \mathbf{f}^n \quad \text{dans } \Omega \qquad (7.6)\\ \operatorname{div}\mathbf{u}^n &= 0 \quad \text{dans } \Omega \qquad (7.7)\\ \mathbf{u}^n &= \mathbf{g}^n \quad \text{sur } \partial\Omega \qquad (7.8) \end{aligned}$$

Il s'agit d'un problème *linéaire* en $\mathbf{u}^n$ et p^n de type Stokes avec un terme de convection (linéaire) provenant de la linéarisation en temps du terme $(\mathbf{u}\cdot\nabla)\mathbf{u}$ dans les équations de Navier-Stokes.

7.3 Discrétisation totale

On choisit (pour simplifier) la donnée de Dirichlet $\mathbf{g} = 0$. On discrétise à présent en espace le problème semi-discrétisé (7.6)–(7.8). Le domaine $\Omega = (0,1)\times(0,1)$ est discrétisé par une grille uniforme définie par les points

$$P_{ij} = (ih,\ jh)\,, \quad i,j = 0,\cdots,N+1 \quad \text{avec } h = \frac{1}{N+1}. \qquad (7.9)$$

Les deux composantes de la vitesse exacte sont $u = u(\mathbf{x},t)$ et $v = v(\mathbf{x},t)$ i.e. $\mathbf{u} = (u,v)$. On cherche alors les approximations de la vitesse

$$u_{ij}^n \simeq u(P_{ij},t^n), \quad v_{ij}^n \simeq v(P_{ij},t^n). \qquad (7.10)$$

• Le Laplacien est discrétisé par un schéma à 5 points. On introduit

$$\Delta_h u_{ij} = \frac{1}{h^2}\left(u_{i+1,j} - 2u_{ij} + u_{i-1,j}\right) + \frac{1}{h^2}\left(u_{i,j+1} - 2u_{ij} + u_{i,j-1}\right) \qquad (7.11)$$

pour $i,j = 1,\cdots,N$.

• Le terme de convection $(\mathbf{u}^{n-1}\cdot\nabla)\mathbf{u}^n$ est discrétisé par un schéma décentré *upwind*.

On a $(\mathbf{u}^{n-1}\cdot\nabla)\mathbf{u}^n = \left(\mathbf{u}^{n-1}\cdot\nabla u^n,\ \mathbf{u}^{n-1}\cdot\nabla v^n\right)$ et

$$\mathbf{u}^{n-1}\cdot\nabla w = u^{n-1}\frac{\partial w}{\partial x} + v^{n-1}\frac{\partial w}{\partial y} \quad (\text{avec } w = u^n \text{ou } v^n) \qquad (7.12)$$

Le schéma *upwind* tient compte du signe des composantes de $\mathbf{u}^{n-1}$ pour approcher les dérivées premières. On définit les opérateurs aux différences décentrées

$$D_{x,h}^+ w_{ij} = \frac{(w_{ij} - w_{i-1,j})}{h}, \quad D_{x,h}^- w_{ij} = \frac{(w_{i+1,j} - w_{i,j})}{h},$$

avec des formules similaires pour $D_{y,h}^+$ et $D_{y,h}^-$. Les approximations *upwind* s'écrivent alors

$$u^{n-1}\frac{\partial w}{\partial x}(P_{ij}) \simeq \begin{cases} u_{ij}^{n-1} D_{x,h}^+ w_{ij} & \text{si } u_{ij}^{n-1} > 0,\\ u_{ij}^{n-1} D_{x,h}^- w_{ij} & \text{sinon.} \end{cases} \qquad (7.13)$$

$$v^{n-1}\frac{\partial w}{\partial y}(P_{ij}) \simeq \begin{cases} v_{ij}^{n-1} D_{y,h}^+ w_{ij} & \text{si } v_{ij}^{n-1} > 0,\\ v_{ij}^{n-1} D_{y,h}^- w_{ij} & \text{sinon.} \end{cases} \qquad (7.14)$$

En notant $u^+ = \max(u,0)$, $u^- = \min(u,0)$, on peut alors réécrire (7.13) et (7.14) :

$$u^{n-1}\frac{\partial w}{\partial x}(P_{ij}) \simeq (u_{ij}^{n-1})^+ D_{x,h}^+ w_{ij} + (u_{ij}^{n-1})^- D_{x,h}^- w_{ij} \qquad (7.15)$$

$$v^{n-1}\frac{\partial w}{\partial y}(P_{ij}) \simeq (v_{ij}^{n-1})^+ D_{y,h}^+ w_{ij} + (v_{ij}^{n-1})^- D_{y,h}^- w_{ij} \qquad (7.16)$$

et on obtient l'approximation

$$\mathbf{u}^{n-1} \cdot \nabla w(P_{ij}) \simeq C_h(\mathbf{u}^{n-1}, w_{ij}),$$

avec

$$C_h(\mathbf{u}^{n-1}, w_{ij}) = (u_{ij}^{n-1})^+ D_{x,h}^+ w_{ij} + (u_{ij}^{n-1})^- D_{x,h}^- w_{ij} + (v_{ij}^{n-1})^+ D_{y,h}^+ w_{ij} + (v_{ij}^{n-1})^- D_{y,h}^- w_{ij}. \quad (7.17)$$

En tenant compte du fait que $|u| = u^+ - u^-$, on obtient

$$\begin{aligned} C_h(\mathbf{u}^{n-1}, w_{ij}) &= \frac{1}{h}\Big(-(u_{ij}^{n-1})^+ w_{i-1,j} + (|u_{ij}^{n-1}| + |v_{ij}^{n-1}|)w_{ij} + (u_{ij}^{n-1})^- w_{i+1,j} \\ &\quad -(v_{ij}^{n-1})^+ w_{i,j-1} + (v_{ij}^{n-1})^- w_{i,j+1}\Big). \end{aligned} \quad (7.18)$$

• On discrétise l'équation $\operatorname{div} \mathbf{u}^n = 0$ par :

$$\frac{1}{2h}\left(u_{i,j}^n + u_{i,j+1}^n - u_{i+1,j}^n - u_{i+1,j+1}^n\right) + \frac{1}{2h}\left(v_{i,j}^n + v_{i+1,j}^n - v_{i,j+1}^n - v_{i+1,j+1}^n\right) = 0 \quad (7.19)$$

pour $i, j = 0, \cdots, N$.

• Comme pour le problème de Stokes, on ne peut pas prendre les mêmes degrés de libertés pour la vitesse et pour la pression sinon on n'obtient pas le même nombre d'équations que d'inconnues. Il faut s'arranger pour que les points de discrétisations de la pression p soient ceux auquels on discrétise l'équation d'incompressibilité. La discrétisation (7.19) représente l'équation $\operatorname{div} \mathbf{u}^n = 0$ aux points (cf. Figure 6.1 du Chapitre 6)

$$P_{i+1/2,j+1/2} = \{(i+1/2)h,\ (j+1/2)h\}.$$

On cherche alors les approximations de la pression aux points où l'équation de la divergence est discrétisée c'est-à-dire les approximations

$$p_{i+1/2,j+1/2}^n \simeq p(P_{i+1/2,j+1/2}, t^n), \quad (7.20)$$

pour $i, j = 0, \cdots, N$.

Les valeurs moyennes de la pression sont notées

$$p_{i+1/2,j}^n = \frac{1}{2}\left(p_{i+1/2,j+1/2}^n + p_{i+1/2,j-1/2}^n\right) \quad (7.21)$$

$$p_{i,j+1/2}^n = \frac{1}{2}\left(p_{i+1/2,j+1/2}^n + p_{i-1/2,j+1/2}^n\right). \quad (7.22)$$

On approche alors le gradient de la pression en $P_{i+1/2,j+1/2}$ par

$$\nabla p(P_{i+1/2,j+1/2}) \simeq (\delta_{x,h} p_{i+1/2,j+1/2},\ \delta_{y,h} p_{i+1/2,j+1/2})$$

avec

$$\begin{aligned} \delta_{x,h} p_{i+1/2,j+1/2} &= \frac{1}{h}\left(p_{i+1/2,j}^n - p_{i-1/2,j}^n\right) \\ &= \frac{1}{2h}\left(p_{i+1/2,j+1/2}^n + p_{i+1/2,j-1/2}^n - p_{i-1/2,j+1/2}^n - p_{i-1/2,j-1/2}^n\right) \end{aligned} \quad (7.23)$$

et

$$\begin{aligned} \delta_{y,h} p_{i+1/2,j+1/2} &= \frac{1}{h}\left(p_{i,j+1/2}^n - p_{i,j-1/2}^n\right) \\ &= \frac{1}{2h}\left(p_{i+1/2,j+1/2}^n + p_{i-1/2,j+1/2}^n - p_{i+1/2,j-1/2}^n - p_{i-1/2,j-1/2}^n\right) \end{aligned} \quad (7.24)$$

L'équation de Navier-Stokes est alors discrétisée par

$$\frac{u_{ij}^n}{\Delta t} + C_h(\mathbf{u}^{n-1}, u_{ij}^n) - \nu\Delta_h u_{ij}^n + \delta_{x,h} p_{i+1/2,j+1/2}^n = f_{(1),ij}^n + \frac{u_{ij}^{n-1}}{\Delta t} \tag{7.25}$$

$$\frac{v_{ij}^n}{\Delta t} + C_h(\mathbf{u}^{n-1}, v_{ij}^n) - \nu\Delta_h v_{ij}^n + \delta_{y,h} p_{i+1/2,j+1/2}^n = f_{(2),ij}^n + \frac{v_{ij}^{n-1}}{\Delta t}. \tag{7.26}$$

pour $i, j = 1, \cdots, N$.

Le schéma MAC fournit le même nombre d'équations que d'inconnues. En effet, il y a $2N^2$ équations de Navier-Stokes (7.25),(7.26) et $(N+1)^2$ équations pour l'équation de la divergence (7.19), soit au total $2N^2 + (N+1)^2$ équations. Par ailleurs, on a $2N^2$ inconnues pour la vitesse et $(N+1)^2$ inconnues pour la pression, soit au total $2N^2 + (N+1)^2$ inconnues.

7.4 Forme matricielle du schéma semi-implicite

On ordonne les noeuds du maillage en partant du bas vers le haut et de la gauche vers la droite. Pour le vecteur vitesse, on stocke la première puis la deuxième composante :

$$\mathcal{U}^n = \Big(\underbrace{u_{11}^n, u_{21}^n, \cdots, u_{N2}^n}_{\text{1ère ligne pour } u^n}, \cdots, \underbrace{u_{1N}^n, \cdots, u_{NN}^n}_{\text{ligne } N \text{ pour } u^n}, \underbrace{v_{11}^n, \cdots, v_{N2}^n}_{\text{1ère ligne pour } v^n}, \cdots, \underbrace{v_{1N}^n, \cdots, u_{NN}^n}_{\text{ligne } N \text{ pour } v^n} \Big)^\top \in \mathbb{R}^{2N^2}. \tag{7.27}$$

De même pour la pression :

$$\mathcal{P}^n = \left(\underbrace{p_{\frac12,\frac12}^n, p_{\frac32,\frac12}^n, \cdots, p_{N+\frac12,\frac12}^n}_{\text{1ère ligne pour } p^n}, \underbrace{p_{\frac12,\frac32}^n, \cdots, p_{N+\frac12,\frac32}^n}_{\text{ligne 2 pour } p^n}, \cdots, \underbrace{p_{\frac12,N+\frac12}^n, \cdots, p_{N+\frac12,N+\frac12}^n}_{\text{ligne } N \text{ pour } p^n} \right)^\top \in \mathbb{R}^{(N+1)^2}. \tag{7.28}$$

Les relations (7.25), (7.26) s'écrivent matriciellement

$$\big(I_d + A + C^{n-1}\big)\,\mathcal{U}^n + B\mathcal{P}^n = \Delta t \mathcal{F}^n + \mathcal{U}^{n-1}. \tag{7.29}$$

$\star$ Le vecteur $\mathcal{F}^n \in \mathbb{R}^{2N^2}$ a la même structure que $\mathcal{U}$:

$$\mathcal{F}^n = \left(\underbrace{f_{(1),11}^n, \cdots, f_{(1),N2}^n, \cdots, f_{(1),1N}^n, \cdots, f_{(1),NN}^n}_{f_{(1)}^n}, \underbrace{f_{(2),11}^n, \cdots, f_{(2),N2}^n, \cdots, f_{(2),1N}^n, \cdots, f_{(2),NN}^n}_{f_{(2)}^n} \right)^\top. \tag{7.30}$$

$\star$ La matrice A est de taille $2N^2 \times 2N^2$ et s'écrit

$$A = \begin{pmatrix} A_1 & 0 \\ 0 & A_1 \end{pmatrix}, \tag{7.31}$$

où la matrice A_1 est la matrice du Laplacien de taille $N^2 \times N^2$:

$$A_1 = \nu\frac{\Delta t}{h^2} \begin{pmatrix} R & -I_d & & & 0 \\ -I_d & R & -I_d & & \\ & \ddots & \ddots & \ddots & \\ & & -I_d & R & -I_d \\ 0 & & & -I_d & R \end{pmatrix} \text{ avec } R = \begin{pmatrix} 4 & -1 & & & 0 \\ -1 & 4 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 4 & -1 \\ 0 & & & -1 & 4 \end{pmatrix} \text{ de taille } N\times N \tag{7.32}$$

et I_d désigne la matrice identité de taille $N \times N$.

$\star$ La matrice C^{n-1} provient de la discrétisation de l'opérateur de convection $\mathbf{u}^{n-1} \cdot \nabla$ par le schéma *upwind* (7.18). Elle est de taille $2N^2 \times 2N^2$ et s'écrit

$$C^{n-1} = \begin{pmatrix} C_1^{n-1} & 0 \\ 0 & C_1^{n-1} \end{pmatrix} \tag{7.33}$$

où la matrice C_1^{n-1} est de taille $N^2 \times N^2$ et s'écrit

$$C_1^{n-1} = \frac{\Delta t}{h} \begin{pmatrix} C_{1,1}^{n-1} & I_1 & & & 0 \\ J_2 & C_{1,2}^{n-1} & I_2 & & \\ & \ddots & \ddots & \ddots & \\ & & J_{N-1} & C_{1,N-1}^{n-1} & I_{N-1} \\ 0 & & & J_N & C_{1,N}^{n-1} \end{pmatrix}. \tag{7.34}$$

Les matrices I_j et J_j sont diagonales et de taille $N \times N$ avec

$$I_j = \begin{pmatrix} (v_{1j}^{n-1})^- & & 0 \\ & \ddots & \\ 0 & & (v_{Nj}^{n-1})^- \end{pmatrix}, \quad J_j = \begin{pmatrix} -(v_{1j}^{n-1})^+ & & 0 \\ & \ddots & \\ 0 & & -(v_{Nj}^{n-1})^+ \end{pmatrix}. \tag{7.35}$$

Les matrices $C_{1,j}^{n-1}$ sont de taille $N \times N$ avec

$$C_{1,j}^{n-1} = \begin{pmatrix} |u_{1j}^{n-1}| + |v_{1j}^{n-1}| & (u_{1j}^{n-1})^- & & & 0 \\ -(u_{2j}^{n-1})^+ & |u_{2j}^{n-1}| + |v_{2j}^{n-1}| & (u_{2j}^{n-1})^- & & \\ & \cdots & \cdots & \cdots & \\ & & -(u_{N-1,j}^{n-1})^+ & |u_{N-1,j}^{n-1}| + |v_{N-1,j}^{n-1}| & (u_{N-1,j}^{n-1})^- \\ 0 & & & -(u_{Nj}^{n-1})^+ & |u_{Nj}^{n-1}| + |v_{Nj}^{n-1}| \end{pmatrix} \tag{7.36}$$

$\star$ La matrice B provient de la discrétisation du gradient de la pression. Elle est de taille $2N^2 \times (N+1)^2$ et s'écrit

$$B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}. \tag{7.37}$$

La matrice B_1 (resp. B_2) correspond à la dérivée par rapport à x (resp. y) de la pression. Les matrices B_1 et B_2 sont toutes deux de taille $N^2 \times (N+1)^2$:

$$B_1 = \frac{\Delta t}{2h} \begin{pmatrix} B_{11} & B_{11} & & & 0 \\ & B_{11} & B_{11} & & \\ & & \ddots & \ddots & \\ 0 & & & B_{11} & B_{11} \end{pmatrix}, \quad B_2 = \frac{\Delta t}{2h} \begin{pmatrix} B_{22} & -B_{22} & & & 0 \\ & B_{22} & -B_{22} & & \\ & & \ddots & \ddots & \\ 0 & & & B_{22} & -B_{22} \end{pmatrix}. \tag{7.38}$$

Les matrices B_{11} et B_{22} sont toutes deux de taille $N \times (N+1)$ et valent

$$B_{11} = \begin{pmatrix} -1 & +1 & & & 0 \\ & -1 & +1 & & \\ & & \ddots & \ddots & \\ 0 & & & -1 & +1 \end{pmatrix}, \quad B_{22} = \begin{pmatrix} -1 & -1 & & & 0 \\ & -1 & -1 & & \\ & & \ddots & \ddots & \\ 0 & & & -1 & -1 \end{pmatrix}. \tag{7.39}$$

La relation de divergence (7.19) peut s'écrire matriciellement à l'aide de la matrice B précédente :

$$B^\top \mathcal{U}^n = 0. \tag{7.40}$$

En regroupant les relations (7.29) et (7.40), on obtient le système

$$\begin{pmatrix} I_d + A + C^{n-1} & B \\ B^\top & 0 \end{pmatrix} \begin{pmatrix} \mathcal{U}^n \\ \mathcal{P}^n \end{pmatrix} = \begin{pmatrix} \Delta t \mathcal{F}^n + \mathcal{U}^{n-1} \\ 0 \end{pmatrix}. \tag{7.41}$$

7.5 Résolution du système discrétisé de Navier-Stokes

Comme pour le système de Stokes (6.31), on peut montrer que le système de Navier-Stokes discrétisé (7.41) admet au moins un solution mais que la matrice du système n'est pas inversible (il n'y a pas unicité de la pression). Comme pour Stokes, on utilise une méthode de pénalisation pour résoudre (7.41) en considérant le système modifié suivant

$$\begin{pmatrix} I_d + A + C^{n-1} & B \\ B^\top & -\varepsilon I_d \end{pmatrix} \begin{pmatrix} \mathcal{U}^n \\ \mathcal{P}^n \end{pmatrix} = \begin{pmatrix} \Delta t \mathcal{F}^n + \mathcal{U}^{n-1} \\ 0 \end{pmatrix}, \tag{7.42}$$

avec un paramètre $\varepsilon > 0$ petit. On découple alors les équations vitesse/pression :

$$\left(I_d + A + C^{n-1} + \frac{1}{\varepsilon} B B^\top \right) \mathcal{U}^n = \Delta t \mathcal{F}^n + \mathcal{U}^{n-1} \tag{7.43}$$

$$\mathcal{P}^n = \frac{1}{\varepsilon} B^\top \mathcal{U}^n. \tag{7.44}$$

Grâce au schéma *upwind* utilisé, la matrice C^{n-1} est positive et par conséquent la matrice du système (7.43) est semi-définie positive.

7.6 Conditions de Dirichlet non-homogènes - Traitement des conditions limites

Si on considère des conditions de Dirichlet non-homogènes, on peut procéder comme pour les équations de Stokes (cf. Chapitre 6). On construit les matrices en tenant compte de tous les noeuds, y compris ceux du bord. On met ensuite à jour les seconds membres et les matrices en éliminant les lignes et colonnes correspondants à des noeuds du bord. Ces mises à jour sont identiques à ce qui a été fait pour les équations de Stokes. La seule construction supplémentaire porte sur la matrice C^{n-1}. On construit la matrice "étendue" $\tilde{D}^{n-1}$ de taille $2(N+2)^2 \times 2(N+2)^2$ définit par

$$\tilde{C}^{n-1} = \begin{pmatrix} \tilde{C}_1^{n-1} & 0 \\ 0 & \tilde{C}_1^{n-1} \end{pmatrix}$$

où la matrice $\tilde{C}_1^{n-1}$ est de taille $(N+2)^2 \times (N+2)^2$ et s'écrit

$$\tilde{C}_1^{n-1} = \frac{\Delta t}{h} \begin{pmatrix} |u_{00}^{n-1}| + |v_{00}^{n-1}| & (u_{00}^{n-1})^- & & & & 0 \\ -(u_{10}^{n-1})^+ & |u_{10}^{n-1}| + |v_{10}^{n-1}| & (u_{10}^{n-1})^- & & & \\ & \cdots & \cdots & & \cdots & \\ & & -(u_{N,N+1}^{n-1})^+ & |u_{N,N+1}^{n-1}| + |v_{N,N+1}^{n-1}| & (u_{N,N+1}^{n-1})^- & \\ 0 & & & -(u_{N+1,N+1}^{n-1})^+ & |u_{N+1,N+1}^{n-1}| + |v_{N+1,N+1}^{n-1}| & \end{pmatrix}$$

La matrice $\tilde{C}_1^{n-1}$ est *tridiagonale*.

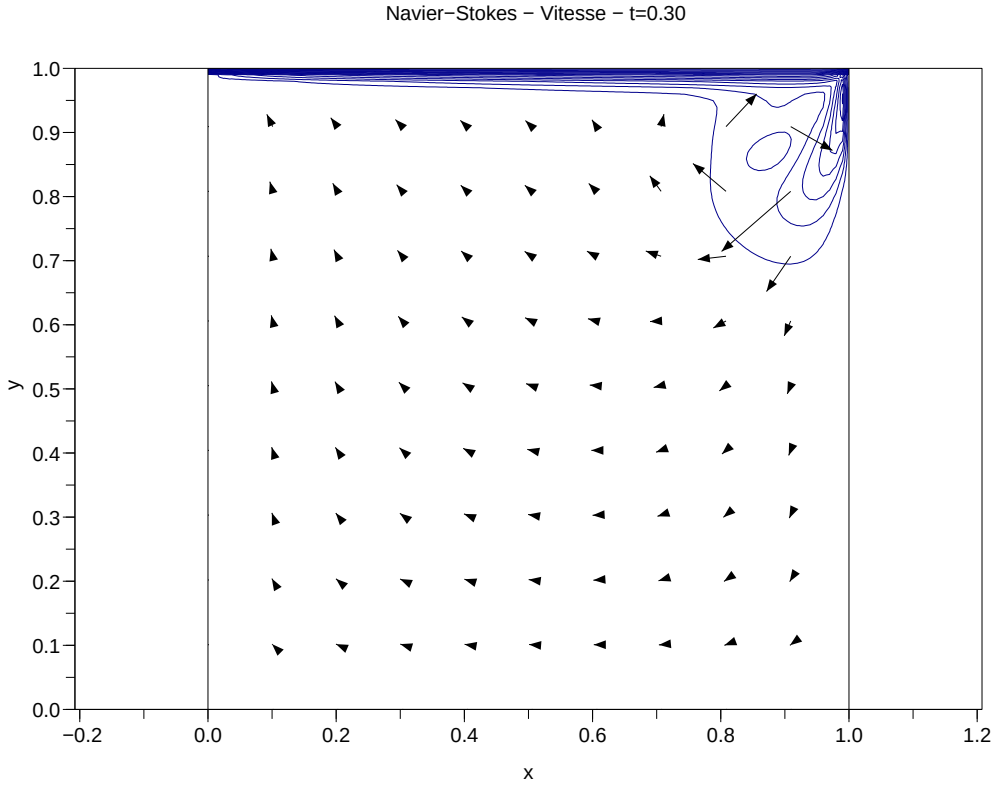
7.7 Exemples

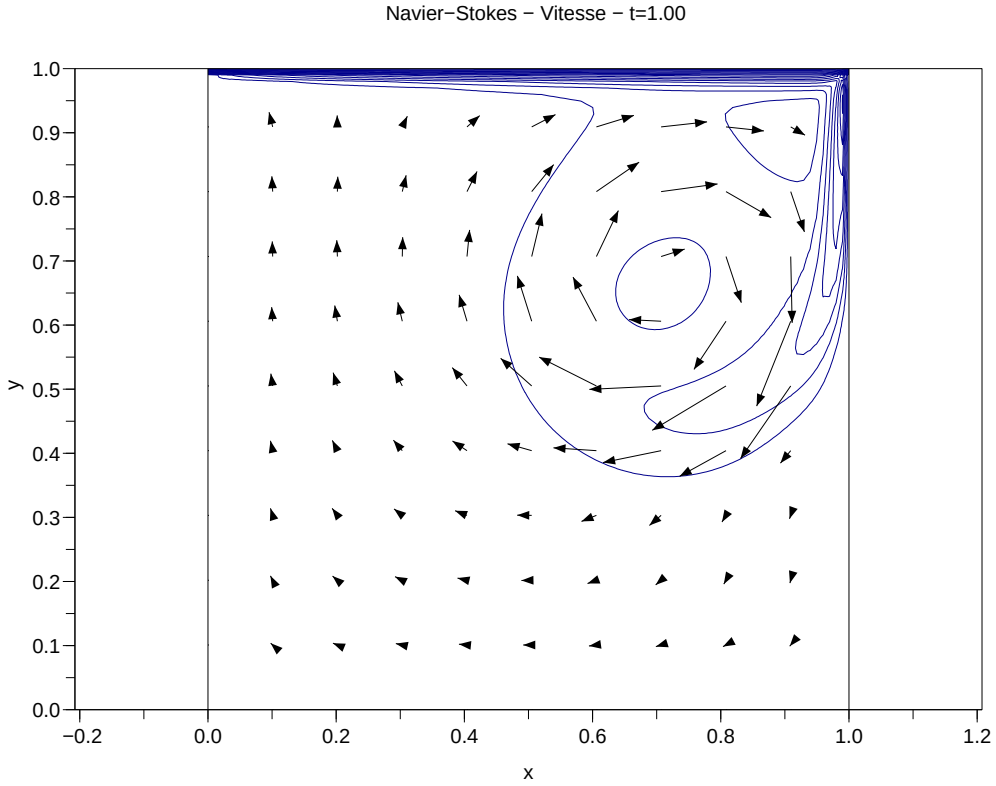
On choisit l'exemple de la cavité entrainée avec

$$\mathbf{u} = \mathbf{g} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \text{ sur les bords } \{x = 0\}, \{x = 1\} \text{ et } \{y = 0\}.$$

$$\mathbf{u} = \mathbf{g} = \begin{pmatrix} 15 \\ 0 \end{pmatrix} \text{ sur le bord } \{y = 1\}.$$

La force vaut $\mathbf{f} = \begin{pmatrix} 0 \\ -30 \end{pmatrix}$ et la viscosité $\nu = 10^{-2}$. La figure (7.1) représente la vitesse à différents instants calculée par le schéma semi-implicite sur un maillage 91×91 et avec un pas de temps $\Delta t = 0.05$.


Navier−Stokes − Vitesse − t=0.30
y
x
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2


Navier−Stokes − Vitesse − t=1.00
y
x
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

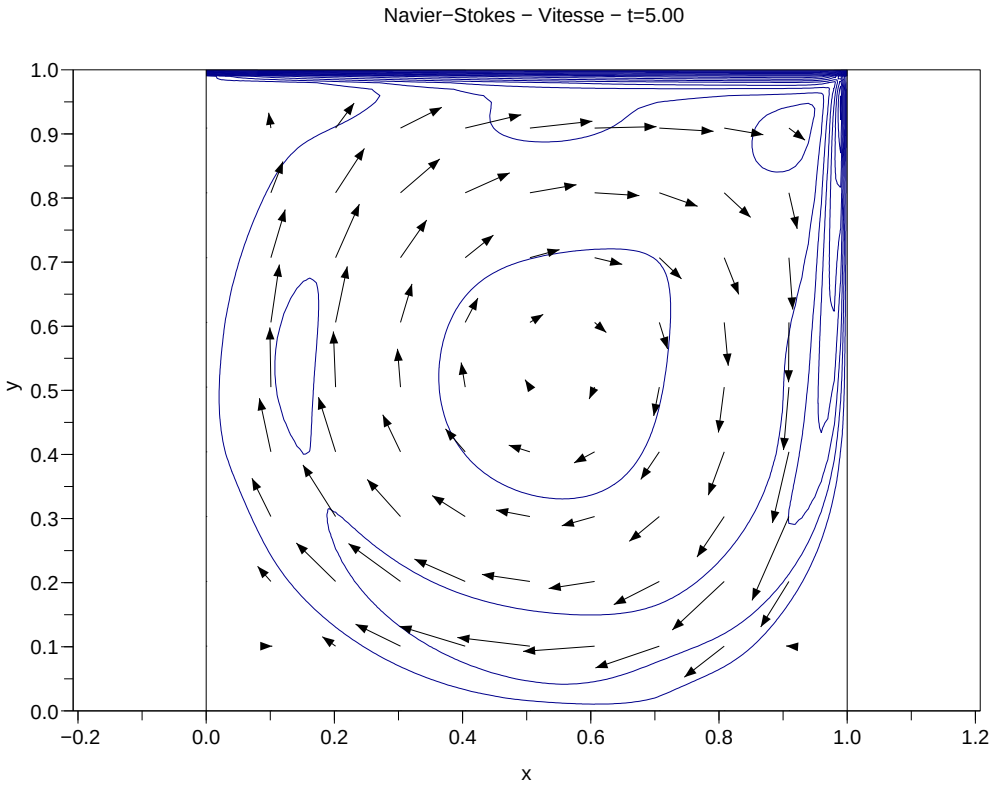


FIGURE 7.1 – Equations de Navier-Stokes - Cavité entrainée ; vitesse à différents instants.

Bibliographie

[A] Ouvrages sur les EDP elliptiques, paraboliques, hyperboliques linéaires

[1] H. Brézis, *Analyse fonctionnelle ; théorie et applications*, Masson, 1987.

[2] P.G. Ciarlet, *Introduction à l'analyse numérique matricielle et à l'optimisation*, Masson, 1990.

[3] L.C. Evans, *Partial differential equations*, 1998.

[4] Protter, Weinberger, *Maximum principle in differential equations*, 1967.

[5] M. Renardy, R.C. Rogers, *An introduction to partial differential equations*, 1993.

[B] Ouvrages sur les Différences Finies pour les EDP elliptiques, paraboliques et hyperboliques linéaires

[6] G.D. Smith, *Numerical solution of partial differential equations : finite difference methods*, 1985.

[7] R.D. Richtmyer, K.W. Morton, *Difference methods for initial-value problems*, 1967.

[8] A.R. Mitchell, D.F. Griffiths, *The finite difference method in partial differential equations*, 1980.

[9] S. Godounov, V. Riabenki *Schémas aux différences*, Editions MIR, 1977 (Version anglaise *Difference Schemes*, North-Holland, 1987).

[10] H.P. Langtangen, *Computational partial differential equations*, 1999.

[11] G. Evans, J. Blackledge, P. Yardley, *Numerical methods for partial differential equations*, 2000.

[12] A. Quarteroni, R. Sacco, F. Saleri, *Numerical mathematics*, 2000.

[13] J.W. Thomas, *Numerical partial differential equations - Finite difference methods*, 1995.

[14] R. Dautray, J.L. Lions, *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Chapitre XX, 1985.

[15] G.E. Forsythe, W.R. Wasow, *Finite -difference methods for partial differential equations*, 1960.

[16] L. Fox Editor, *Numerical solution of ordinary and partial differential equations*, 1962.

[17] G.I. Marchuk, *Methods of numerical mathematics*, 1975.

[18] R.J. Leveque, *Finite volume methods for hyperbolic problems*, 2002.

[C] Ouvrages sur les Différences Finies pour les lois de conservation

[19] R.J. Leveque, *Finite volume methods for hyperbolic problems*, 2002.

[20] E. Godlewski, P.A. Raviart, *Hyperbolic systems of conservation laws*, 1991.

[21] E. Godlewski, P.A. Raviart, *Numerical approximation of hyperbolic systems of conservation laws*, 1996.

[22] D. Serre, *Systèmes de lois de conservation*, vol. I, 1996.

[23] D. Kroner, *Numerical schemes for conservation laws*, 1997.