Différences Finies et Volumes Finis

Master Mathématiques et Applications

Bruno Després

2018

Introduction

On peut considérer que les méthodes numériques pour les équations aux dérivées partielles (EDP) d'évolution s'appuient sur deux piliers. Le premier pilier en est l'analyse fonctionnelle et la théorie des espaces fonctionnels, le second pilier s'appuie sur les modèles d'EDP et leurs liens avec la modélisation des phénomènes réels. Cette discipline est liée de très près également au développement des moyens de calculs informatiques. Pour autant la construction et l'analyse numérique de méthodes numériques efficaces pour les EDP d'évolution s'appuient sur des règles propres qui forment l'objet de ces notes pour le **cours de base du M2-Mathématiques de la modélisation**[1].

Un problème modèle central dans ces notes est issu de la modélisation des phénomènes réels et de la pratique de l'art de l'ingénieur. Il est de type transport-diffusion et s'écrit

$$\partial_t u + \mathbf{a} \cdot \nabla u - \Delta u = 0.$$

Cependant on considèrera le plus souvent séparément l'équation de transport ou d'advection $\partial_t u + \mathbf{a} \cdot \nabla u = 0$, qui est de type hyperbolique, et l'équation de la chaleur $\partial_t u - \Delta u = 0$, qui est de type parabolique. Les équations de convection-diffusion, non linéaires cette fois, sont aussi très utilisées en traitement de l'image, par exemple en suivant les modèles de Perona-Malik : $\partial_t u = \nabla \cdot (g \nabla u) = g \Delta u + \nabla g \cdot \nabla u$ avec g une fonction non linéaire compliquée de u ; pour $g = u$ on retrouve une équation pour les écoulements en milieux poreux. Nous ne considérerons dans la suite que des équations à coefficients constants et donnés.

On s'appuiera sur les deux notions fondamentales que sont la **stabilité** et la **consistance** pour construire et justifier les méthodes de Différences Finies et Volumes Finis qui seront étudiées dans ces notes. Les méthodes d'Eléments Finis sont évoquées rapidement au chapitre 2. Les méthodes de Différences Finies sont simples à construire et leur théorie sert de socle à la plupart des méthodes numériques non stationnaires. Les méthodes de Volumes Finis peuvent être vues comme des méthodes de Différences Finies sur maillage tordu. Elles sont également simples de construction et sont à la base de la plupart des codes industriels et de recherche de CFD (Computational Fluid Dynamics).

Ce texte est rédigé avec deux niveaux de lecture. Tout ce qui concerne la construction des méthodes numériques est en taille normale. Les parties en taille réduite apportent des détails complémentaires pour justifier certains éléments ou pour mener à bien les diverses preuves. Elles doivent être laissées de côté en première lecture. De même il est conseillé de passer directement au deuxième chapitre qui présente des principes de construction de schémas numériques.

1. Des coquilles/erreurs peuvent subsister. Merci de les signaler par mail à **despres@ann.jussieu.fr**

Table des matières

Chapitre 1

Cadre fonctionnel et modèles

Pour toute méthode de discrétisation numérique d'une équation aux dérivées partielles, une question fondamentale est de montrer la convergence de la solution numérique vers la solution exacte, et mieux d'obtenir des estimations quantitatives optimales pour l'erreur. Pour cela, nous aurons besoin d'un cadre fonctionnel.
Ce chapitre peut être laissé de côté en première lecture.

1.1 Cadre fonctionnel

On renvoie à [6].

Définition 1 (**Espace de Banach**). *Un espace de Banach réel V est un espace vectoriel réel, muni d'une norme $u \mapsto \|u\|$ définie pour tout $u \in V$, et complet pour cette norme. Les propriétés de la norme sont*
- $\|u\| \geq 0$ *pour tout* $u \in V$,
- $\|u\| = 0$ *si et seulement si* $u = 0$,
- $\|\lambda u\| = |\lambda| \, \|u\|$ *pour tout* $\lambda \in \mathbb{R}$,
- $\|u + v\| \leq \|u\| + \|v\|$ *pour tous* $u, v \in V$.

L'espace V est appelé un **espace de Hilbert** dans le cas où la norme est associée à un produit scalaire

$$\|u\| = \sqrt{(u, u)}$$

avec $(u, v) \in \mathbb{R}$ étant le produit scalaire de u et v. Pour mémoire, les propriétés d'un produit scalaire réel sont
- le produit scalaire est une forme bilinéaire,
- $(u, u) \geq 0$ pour tout $u \in V$,
- $(u, u) = 0$ si et seulement si $u = 0$,
- $(u, v) = (v, u)$ pour tous $u, v \in V$.

1.1.1 Espaces de Lebesgue L^p

Soit Ω un ouvert régulier de $\mathbb{R}^d$, borné ou non.

Définition 2 (Espaces de Lebesgue). *Soit $p \in [1, \infty]$.*
- *Pour $1 \leq p < \infty$, l'espace $L^p(\Omega)$ est constitué des fonctions mesurables telles que $\int_\Omega |u(\mathbf{x})|^p \, d\mathbf{x} < \infty$. La norme dans $L^p(\Omega)$ est*

$$\|u\|_{L^p(\Omega)} = \left(\int_\Omega |u(\mathbf{x})|^p \, d\mathbf{x} \right)^{\frac{1}{p}}.$$

- *Pour $p = \infty$, l'espace $L^\infty(\Omega)$ est constitué des fonctions mesurables et bornées. La norme dans $L^\infty(\Omega)$ est*

$$\|u\|_{L^\infty(\Omega)} = \sup \{\lambda; \quad mes\,(|u(\mathbf{x})| > \lambda) \neq 0\} < \infty.$$

— *Les espaces de Lebesgue sont des espaces de Banach.*

Les dérivées partielles d'une fonction sont notées

$$u^{(k_1,\cdots,k_d)} = \frac{\partial^{k_1+\cdots+k_d}}{\partial_{x_1}^{k_1}\dots\partial_{x_d}^{k_d}}u, \quad \text{avec } 0 \leq k_i \text{ pour tout } i = 1,\dots,d.$$

On renvoie à [6] pour une définition rigoureuse de la dérivation au sens des distributions d'une fonction mesurable.

Définition 3. *L'ensemble des fonctions mesurables de $L^p(\Omega)$ dont toutes les dérivées sont également dans $L^p(\Omega)$ jusqu'à un ordre de dérivation totale de $q \in \mathbb{N}$ est noté $W^{q,p}(\Omega)$.*
Pour $1 \leq p \leq \infty$ une norme dans $W^{q,p}(\Omega)$ est

$$\|u\|_{W^{q,p}(\Omega)} = \sum_{k_1+\cdots+k_d\leq q} ||u^{(k_1,\cdots,k_d)}||_{L^p(\Omega)}.$$

Le sous espace vectoriel de $W^{q,p}(\Omega)$ constitué des fonctions à suport compact dans Ω est noté $W_0^{q,p}(\Omega) \subset W^{q,p}(\Omega)$.

1.1.2 Inégalités

Soient deux nombres positifs $p \in [1,\infty]$ et $q \in [1,\infty]$ (l'infini est autorisé) tels que

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Nous dirons que p et q sont **conjugués**.

Lemme 1 (Inégalité de Hölder)**.** *Soient $u \in L^p(\Omega)$ et $v \in L^q(\Omega)$ où p et q sont des nombres conjugués. Alors*

$$\left|\int_\Omega u(\mathbf{x})v(\mathbf{x})d\mathbf{x}\right| \leq \|u\|_{L^p(\Omega)} \times \|v\|_{L^q(\Omega)}.$$

Dans le cas $p = q = 2$, l'inégalité de Hölder est identique à l'inégalité de Cauchy-Schwarz. Le cas $p = \infty$ et $q = 1$ est immédiat.

1.1.3 Fonctions à variation bornée

Le cadre des fonctions à variation bornée permet de manipuler des fonctions discontinues, ce qui est très utile pour l'analyse numérique des équations de transport. On renvoie à [18, 19].
Pour un vecteur $\varphi = (\varphi_1,\cdots,\varphi_d)$ on notera

$$|\varphi| = \sqrt{\varphi_1^2 + \cdots \varphi_d^2}.$$

L'espace des fonctions à dérivée bornée, à valeur vectorielle, à support compact, et bornées par 1, sera noté

$$W_{b,0}^{1,\infty}(\mathbb{R}^d) = \left\{\varphi \in W_0^{1,\infty}(\mathbb{R}^d),\ |\varphi(\mathbf{x})| \leq 1\ \forall \mathbf{x}\right\}$$

Définition 4 (Variation totale)**.** *Soit $u \in L^1_{loc}(\mathbb{R}^d)$. Le nombre éventuellement infini*

$$|u|_{\mathrm{BV}(\mathbb{R}^d)} = \sup_{\varphi\in W_{b,0}^{1,\infty}(\mathbb{R}^d)} \left(-\int_{\mathbb{R}^d} u(\mathbf{x})\nabla\cdot\varphi(\mathbf{x})dx\right)$$

sera appelé la variation totale de u.

Exemple 1 (En dimension un d'espace)**.** *Soit $u \in W^{1,1}(\mathbb{R})$. Alors*

$$|u|_{\mathrm{BV}} = \|u'\|_{L^1(\mathbb{R})} = \int_{\mathbb{R}} |u'(x)|dx.$$

Cela vient de la formule d'intégration par parties

$$-\int_{\mathbb{R}} u(x)\varphi'(x)dx = \int_{\mathbb{R}} u'(x)\varphi(x)dx.$$

Le supremum sur tous les φ tels que $|\varphi| \leq 1$ montre que

$$\sup_{|\varphi|\leq 1}\left(\int_{\mathbb{R}} u'(x)\varphi(x)dx\right) = \int_{\mathbb{R}} |u'(x)|\,dx = \|u'\|_{L^1(\mathbb{R})}.$$

Exemple 2 (En dimension deux d'espace). *Soit le carré unité* $C = \{\mathbf{x} = (x_1, x_2),\ 0 < x_1, x_2 < 1\} \subset \mathbb{R}^2$. *La fonction indicatrice de* C *est notée* $\mathbf{1}_C$ *avec* $\mathbf{1}_C(\mathbf{x}) = 1$ *si* $\mathbf{x} \in C$; $\mathbf{1}_C(\mathbf{x}) = 0$ *dans le cas contraire.*
Alors $|\mathbf{1}_C|_{\mathrm{BV}} = 4$.

En effet on a a pour tout $\varphi \in W^{1,\infty}_{b,0}(\mathbb{R}^2)$

$$-\int_{\mathbb{R}^2} \mathbf{1}_C(\mathbf{x}) \nabla \cdot \varphi(\mathbf{x}) dx = -\int_{\mathbf{x}\in C} \nabla \cdot \varphi(\mathbf{x}) dx = \int_{\mathbf{x}\in \partial C} \varphi(\mathbf{x}) \cdot \mathbf{n}_S dx \leq 4.$$

La borne est atteinte pour une suite bien choisie de fonctions φ_n.
On remarque par ailleurs que la valeur 4 est la valeur du périmètre du carré C, ce que nous noterons

$$|C| = |\mathbf{1}_C|_{\mathrm{BV}}.$$

Définition 5 (Espace BV). *L'espace des fonctions de* $L^1(\mathbb{R}^d)$ *à variation totale bornée est noté* $\mathrm{BV}(\mathbb{R}^{\mathrm{d}})$. *Une norme associée est*

$$\|u\|_{\mathrm{BV}(\mathbb{R}^{\mathrm{d}})} = |u|_{\mathrm{BV}(\mathbb{R}^{\mathrm{d}})} + \|u\|_1.$$

On a l'inclusion[1] *dense* $W^{1,1}(\mathbb{R}^d) \subset \mathrm{BV}(\mathbb{R}^{\mathrm{d}})$.

L'exemple 1 montre l'inclusion en dimension un d'espace. La densité de l'inclusion sera montrée dans un cas particulier à la section 5.3. L'inclusion est stricte $W^{1,1}(\mathbb{R}^d) \neq \mathrm{BV}(\mathbb{R}^{\mathrm{d}})$ comme conséquence de la définition et des exemples.
Soit $u \geq 0$ une fonction mesurable positive ou nulle. On définit l'ensemble de niveau

$$E_\lambda = \left\{\mathbf{x} \in \mathbb{R}^d,\ u(\mathbf{x}) > \lambda\right\} \subset \mathbb{R}^d.$$

Le périmètre de E_λ est

$$|E_\lambda| = \left|\mathbf{1}_{E_\lambda}\right|_{\mathrm{BV}(\mathbb{R}^{\mathrm{d}})} = \sup_{\varphi \in W^{1,\infty}_{b,0}(\mathbb{R}^d)^d} \left(-\int_{E_\lambda} \nabla \cdot \varphi(\mathbf{x}) dx\right),$$

où $\mathbf{1}_{E_\lambda}$ est la fonction indicatrice de E_λ. Pour toute fonction positive ou nulle, on

$$u(\mathbf{x}) = \int_0^\infty \mathbf{1}_{E_\lambda}(\mathbf{x}) d\lambda \qquad p.p.$$

Lemme 2 (Formule de la coaire : voir [18]). *Soit* $u \in \mathrm{BV}(\mathbb{R}^{\mathrm{d}})$ *une fonction positive ou nulle,* $u \geq 0$. *Alors*

$$|u|_{\mathrm{BV}(\mathbb{R}^{\mathrm{d}})} = \int_0^\infty |E_\lambda| d\lambda. \tag{1.1}$$

1.2 Quelques modèles

Les modèles considérés sont linéaires. Ils servent souvent de briques de base pour des modèles plus élaborés.

1.2.1 Equation de transport

L'équation du transport libre à vitesse constante s'écrit en tout dimension

$$\partial_t u + \mathbf{c}.\nabla u = 0, \qquad t > 0, \quad \mathbf{x} \in \mathbb{R}^d.$$

1. Notons aussi que $BV(\mathbb{R}) \subset L^\infty(\mathbb{R})$ en dimension un d'espace. Une preuve rapide est la suivante. Soit $u \in BV(\mathbb{R})$: on se donne trois nombres $x_0 \in \mathbb{R}$, $\varepsilon > 0$ et $\mu > 0$ et on considère la fonction continue négative ou nulle

$$\varphi(x) = -\begin{cases} 0 & \text{pour } x \leq x_0, \\ \frac{x-x_0}{\varepsilon} & \text{pour } x_0 \leq x \leq x_0 + \varepsilon, \\ 1 - \mu(x - x_0 - \varepsilon), & \text{pour } x_0 + \varepsilon \leq x \leq x_0 + \varepsilon + \frac{1}{\mu}, \\ 0 & \text{pour } x_0 + \varepsilon + \frac{1}{\mu} \leq x. \end{cases}$$

On a bien $|\varphi| \leq 1$. On a aussi $-\int_{\mathbb{R}} u(\mathbf{x})\varphi'(x)dx \leq \mathrm{BV}(u)$. Un calcul montre que $-\int_{\mathbb{R}} u(\mathbf{x})\varphi'(x)dx = \frac{1}{\varepsilon}\int_{x_0}^{x_0+\varepsilon} u(x)dx - \mu\int_{x_0+\varepsilon}^{x_0+\varepsilon+\frac{1}{\mu}} u(x)dx$. Donc $\int_{x_0}^{x_0+\varepsilon} (u(x) - \mathrm{BV}(u))\, dx \leq \varepsilon\mu \int_{x_0+\varepsilon}^{x_0+\varepsilon+\frac{1}{\mu}} u(x)dx$. Comme $u \in L^1(\mathbb{R})$, on peut passer à la limite $\mu \to 0$ pour le deuxième terme qui tend vers zéro : $\lim_{\mu=0^+} \mu \int_{x_0+\varepsilon}^{x_0+\varepsilon+\frac{1}{\mu}} u(x)dx = 0$. Donc $\int_{x_0}^{x_0+\varepsilon} (u(x) - \mathrm{BV}(u))\, dx \leq 0$. Cela étant arbitraire par rapport à x_0 et ε qui peut être aussi petit que souhaité, alors $u(x) \leq BV(u)$ presque partout. De même on montre en prenant $\psi = -\varphi$ que $-BV(u) \leq u(x)$ presque partout. Donc $u \in L^\infty(\mathbb{R})$.

La fonction $(t, \mathbf{x}) \mapsto u(t, \mathbf{x})$ est l'inconnue : t est la variable de temps, et $\mathbf{x} = (x_1, \cdots, x_d) \in \mathbb{R}^d$ est la variable d'espace. L'opérateur gradient est défini par

$$\nabla u = \left(\frac{\partial}{\partial x_1} u, \cdots, \frac{\partial}{\partial x_d} u \right).$$

Le champ $\mathbf{x} \mapsto \mathbf{c}(\mathbf{x}) \in \mathbb{R}^d$ est donné. Il est appelé champ de vitesse pour des raisons qui paraîtront évidentes dans la suite.

Dimension $d = 1$

On considère tout d'abord le cas en dimension $d = 1$ pour une vitesse constante que l'on note $a \in \mathbb{R}$. Il s'agit de l'équation d'advection

$$\partial_t u + a \partial_x u = 0, \qquad t > 0, \quad x \in \mathbb{R}. \tag{1.2}$$

On supposera que $a > 0$. L'autre cas $a < 0$ est symétrique et se déduit du cas $a > 0$. On munit l'équation d'une condition initiale à $t = 0$

$$u(0, x) = u_0(x). \tag{1.3}$$

Lemme 3. *L'unique solution de (1.2) avec la condition initiale (1.3) est*

$$u(t, x) = u_0(x - at). \tag{1.4}$$

Démonstration. Cette propriété peut se démontrer dans tout type d'espace fonctionnel. Par souci de simplicité on considère une donnée initiale régulière $u_0 \in C^1(\mathbb{R})$. Prenons la fonction définie par (1.4). On a $\partial_t u = -a u_0'(x - at)$ et $\partial_x u = u_0'(x - at)$. Donc $\partial_t u + a \partial_x u = -a u_0' + a u_0' = 0$ ce qui montre que (1.4) est bien une solution.

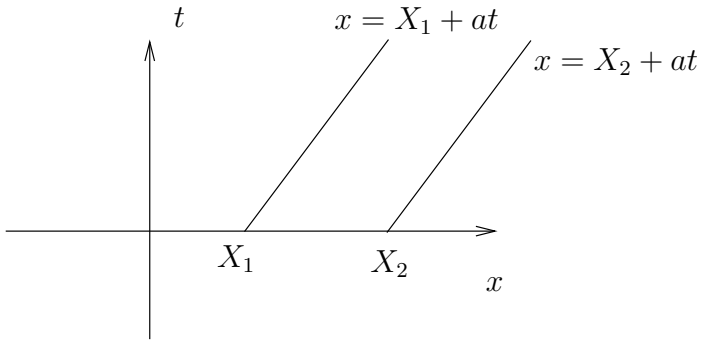


FIGURE 1.1 – La solution de l'équation d'advection est constante le long des droites caractéristiques $x = X + at$.

Montrons à présent l'unicité. Soient u_1 et u_2 deux solutions de classe $C^1(\mathbb{R})$ éventuellement différentes, avec la même donnée initiale

$$u_1(0, x) = u_2(0, x) = u_0(x).$$

Soit $x \mapsto \varphi_0(x)$ une fonction dérivable, positive ou nulle, à support compact : $\varphi_0(x) = 0$ for $|x| \geq A$. On note $\varphi(t, x) = \varphi_0(x - at)$ qui est solution de l'équation d'advection. Posons $v = (u_1 - u_2)^2 \varphi \geq 0$. On commence par vérifier que v est aussi solution de l'équation d'advection

$$\partial_t v + a \partial_x v = 2 (u_1 - u_2) \varphi \left(\partial_t (u_1 - u_2) + a \partial_x (u_1 - u_2) \right) + (u_1 - u_2)^2 \left(\partial_t \varphi + a \partial_x \varphi \right) = 0.$$

Par construction v est à support compact ce qui n'était pas nécessairement le cas de u_1 ni de u_2. Donc

$$0 = \int_{\mathbb{R}} \left(\partial_t v + a \partial_x v \right) dx = \int_{\mathbb{R}} \partial_t v dx + a \int_{-A+at}^{A+at} \partial_x v dx = \frac{d}{dt} \int_{\mathbb{R}} v dx.$$

Or $v(0,x)=0$. Donc $\int_{\mathbb{R}} v(T,x)dx = 0$ pour tout $T>0$. Comme $v \geq 0$, il s'ensuit que $v \equiv 0$. Le support de v pouvant être aussi grand que souhaitée, cela montre que $u_1 = u_2$.

□

Soit un champ de vitesse de transport $x \mapsto c(x) \in \mathbb{R}$ et u une solution de l'équation du transport

$$\partial_t u + c(x)\partial_x u = 0, \quad u(x,0) = u_0(x).$$

Nous construisons les courbes caractéristiques

$$\begin{cases} y'(t;X) = c(y(t;X)), \\ y(0;X) = X. \end{cases}$$

On utilise souvent des notations simplifiées. Par exemple en notant les courbes caractéristiques $x(t)$ à la place de $x = y(t;X)$.

Proposition 1. *Supposons c Lipschitzienne et bornée. Alors il existe une et une seule solution de l'équation des courbes caractéristiques ($x \in \mathbb{R}$, $t \geq 0$).*

Démonstration. C'est une conséquence du théorème de Cauchy-Lipshitz. □

Proposition 2. *Sous les mêmes hypothèses, une solution de l'équation du transport est*

$$u(x,t) = u_0(X), \quad x = y(t;X).$$

Démonstration. On a $u(x,t) = u(y(t;X),t)$. Dérivant par rapport à t, X étant fixe, on obtient

$$0 = \frac{d}{dt}u_0(X) = \frac{d}{dt}u(y(t;X),t) = y'(t;X)\partial_x u(y(t;X),t) + \partial_t u(y(t;X),t)$$

$$= \partial_t u(y(t;X),t) + c(y(t;X))\partial_x u(y(t;X),t).$$

Cela est vrai pour tout (t,X), c'est vrai pour tout $x = y(t;X)$ et tout t. La preuve est terminée. □

Dimension $d \geq 2$ en domaine borné

Nous nous concentrons à présent sur les conditions au bord qu'il faut considérer en domaine borné, car cela constituera un bon point de départ pour la construction de schémas numériques pour cette équation.
Soit $\Omega \subset \mathbb{R}^d$ un ouvert borné régulier. On note le champ de vitesse $\mathbf{x} \mapsto \mathbf{a}(\mathbf{x})$. On supposera que $\mathbf{a} \in C^1(\overline{\Omega})$ est à divergence nulle

$$\nabla.\mathbf{a} = 0.$$

De ce fait l'équation admet une formulation conservative

$$\partial_t u + \nabla \cdot (\mathbf{a}u) = \partial_t u + \mathbf{a} \cdot \nabla u + (\nabla \cdot \mathbf{a})\, u = 0.$$

Le bord de Ω est séparé en deux parties $\Gamma = \Gamma^- \cup \Gamma^+$ avec

$$\Gamma^- = \{\mathbf{x} \in \Gamma, \ \ \mathbf{a}\cdot\mathbf{n} < 0\}, \qquad \Gamma^+ = \{\mathbf{x} \in \Gamma, \ \ \mathbf{a}\cdot\mathbf{n} \geq 0\}.$$

Nous considérons le problème avec condition initiale et condition au bord

$$\begin{cases} \partial_t u + \mathbf{a}\cdot\nabla u = 0, & \mathbf{x} \in \Omega, \quad t > 0, \\ u(0,\mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \\ u(t,\mathbf{x}) = u^-(t,\mathbf{x}), & \mathbf{x} \in \Gamma^-. \end{cases} \tag{1.5}$$

On note immédiatement qu'il n'y a pas de condition sur le bord Γ^+.

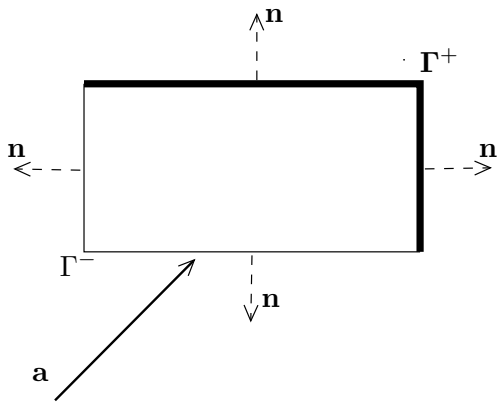


FIGURE 1.2 – Sur cet exemple le champ de vitesse $\mathbf{a}$ est orienté en diagonale : la partie Γ^+ du bord surlignée en gras est constitué des parties du bord en haut et à droite ; la partie Γ^- du bord correspond aux parties du bord en bas et à gauche.

Lemme 4. *Soient deux fonctions u_1 et u_2 solutions régulières de (1.5). Supposons que u_1 ont u_2 ont la même condition initiale, et ont la même condition sur le bord Γ^-. Alors $u_1 = u_2$.*

Démonstration. La différence $e = u_1 - u_2$ est solution de

$$\begin{cases} \partial_t e + \mathbf{a}\cdot\nabla e = 0, & \mathbf{x}\in\Omega, \quad t>0, \\ e(0,\mathbf{x}) = 0, & \mathbf{x}\in\Omega, \\ e(t,\mathbf{x}) = 0, & \mathbf{x}\in\Gamma^-. \end{cases}$$

Posons $E(t) = \frac{1}{2}\|e(t)\|_2^2$. Alors

$$E'(t) = \int_\Omega e\partial_t e dx = -\int_\Omega e\mathbf{a}\cdot\nabla e dx = -\int_\Omega \nabla\cdot\left(\mathbf{a}\frac{e^2}{2}\right)dx$$

$$= -\int_{\Gamma^-}\mathbf{a}\cdot\mathbf{n}\frac{e^2}{2}d\sigma - \int_{\Gamma^+}\mathbf{a}\cdot\mathbf{n}\frac{e^2}{2}d\sigma = -\int_{\Gamma^+}\mathbf{a}\cdot\mathbf{n}\frac{e^2}{2}d\sigma \le 0.$$

Notons que l'on a utilisé que $e = 0$ on Γ^-. Or $E(0) = 0$ donc $E(t) = 0$ pour tout temps $t > 0$. Cela montre que $u_1 = u_2$. □

Il est important de bien comprendre pourquoi le bord Γ^+ ne joue finalement aucun rôle dans la preuve d'unicité.

Courbes caractéristiques "en avant" dans un domaine borné

A présent nous construisons la solution à partir des courbes caractéristiques $t \mapsto \mathbf{y}(t,\mathbf{X})$ définies par

$$\frac{d}{dt}\mathbf{y}(t,\mathbf{X}) = \mathbf{a}(\mathbf{X}) \text{ avec la donnée initiale } \mathbf{y}(0,\mathbf{x}) = \mathbf{X}.$$

Ces courbes sont correctement construites dans le cadre du théorème de Cauchy-Lipschitz pour $\mathbf{a}\in C^1(\overline{\Omega})$.
Soit une fonction u constante le long des caractéristiques

$$u(\mathbf{y}(\mathbf{X}),t) = u_0(\mathbf{X}).$$

Elle vérifie

$$\frac{d}{dt}u = \partial_t u + \frac{d}{dt}\mathbf{y}(t,\mathbf{X}).\nabla u = \partial_t u + \mathbf{a}.\nabla u = 0.$$

Pour un $\mathbf{x}$ donné et un t donné, on peut ainsi déterminer la valeur de $u(t,\mathbf{x})$ une fois que le pied de la caractéristique X a été défini en résolvant l'équation

$$\mathbf{y}(t,\mathbf{X}) = \mathbf{x}. \tag{1.6}$$

Il s'ensuit qu'il est nécessaire d'inverser l'équation (1.6) pour obtenir le point de départ $\mathbf{X}\in\Omega\cup\Gamma^-$ de la caractéristique qui arrive en $(t,\mathbf{x})\in\mathbb{R}^+\times\Omega$. Pour rendre la discussion légèrement plus simple, on peut construire les caractéristiques "en arrière".

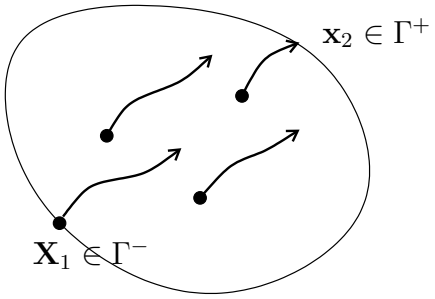


FIGURE 1.3 – La fonction u est constante le long des caractéristiques dont le point de départ est noté sous la forme de cercle noir : le point $\mathbf{X}_1 \in \Gamma^-$ est un point de départ ; le point $\mathbf{x}_2 \in \Gamma^+$ n'est pas un point de départ.

Courbes caractéristiques "en arrière" dans un domaine borné

Les courbes caractéristiques en arrière sont construites à partir de la position au temps final

$$\frac{d}{dt}\mathbf{X}(t,\mathbf{x}) = -\mathbf{a}(\mathbf{X}) \text{ pour } t>0, \quad \text{avec } \mathbf{X}(0,\mathbf{x}) = \mathbf{x} \in \Omega.$$

Bien sûr $\mathbf{X}(t,\mathbf{x})$ est aussi le point de départ de la caractéristique en avant discutée précédemment. Nous définissons le temps (de sortie)

$$T(\mathbf{x}) = \inf(t) \text{ tel que } \mathbf{X}(t,\mathbf{x}) \in \partial\Omega.$$

Si $\mathbf{X}(t,\mathbf{x}) \in \Omega$ pour tout $t>0$, on posera $T(\mathbf{x}) = +\infty$. Par définition $T(\mathbf{x}) > 0$ pour tout $\mathbf{x} \in \Omega$.
La construction de la solution u au point (t,x) s'appuie sur deux cas.

Premier cas : $t < T(\mathbf{x})$. On pose

$$u(t,\mathbf{x}) = u_0(\mathbf{X}(t,\mathbf{x})). \tag{1.7}$$

Deuxième cas : $T(\mathbf{x}) \le t$. Pour le temps $t = T(\mathbf{x})$ la courbe caractéristique rencontre le bord, nécessairement en Γ^-. On pose

$$u(t,\mathbf{x}) = u^-\left(t - T(\mathbf{x}), \mathbf{X}\left(T(\mathbf{x}),\mathbf{x}\right)\right). \tag{1.8}$$

Par construction la fonction u (1.7-1.8) satisfait la condition initiale

$$u(0,\mathbf{x}) = u_0(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega, \qquad \text{(c'est à dire (1.7) à } t=0\text{)},$$

et la condition au bord

$$u(t,\mathbf{x}) = u^-(t,\mathbf{x}), \quad \forall \mathbf{x} \in \Gamma^-, \qquad \left(\text{c'est à dire (1.8) pour } \mathbf{x} \in \Gamma^-\right).$$

Il reste à vérifier que u est bien solution, et en quel sens, de l'équation de transport. On a un premier résultat, qui est partiel cependant car il y une restriction sur le temps.

Lemme 5. *Supposons que $u_0 \in C^1(\Omega)$. Soit un point de l'espace temps (t,x) tel que $t < T(\mathbf{x})$. Alors la fonction u (1.9) est localement C^1 et est solution de*

$$\partial_t u + \mathbf{a}.\nabla u = 0 \qquad \forall \mathbf{x} \in \Omega \quad \forall t < T(\mathbf{x}).$$

Démonstration. On a par construction

$$\mathbf{X}(t-h, \mathbf{X}(h,\mathbf{x})) = \mathbf{X}(t,\mathbf{x}) \text{ pour de petits } h>0,$$

donc

$$u(t-h, \mathbf{X}(t-h\mathbf{X}(h,\mathbf{x})) = u(t,\mathbf{x}) \text{ pour de petits } h>0. \tag{1.9}$$

La transformation $(t,\mathbf{x}) \mapsto \mathbf{X}(t,\mathbf{x})$ est C^1 localement autour de $(t,\mathbf{x})$ dans le cas $t < T(\mathbf{x})$. Par dérivation de (1.9) on obtient $\frac{d}{dh}u(t-h, \mathbf{X}(t-h\mathbf{X}(h,\mathbf{x})) = 0$, ou encore

$$-\partial_t u - \frac{d}{dh}\mathbf{X}(t-h\mathbf{X}(h,\mathbf{x}) \cdot \nabla u = 0.$$

Par ailleurs $\frac{d}{dh}\mathbf{X}(t-h\mathbf{X}(h,\mathbf{x}) = \mathbf{a}(\mathbf{X}(t-h\mathbf{X}(h,\mathbf{x}))$, donc pour $h=0$ on obtient $-\partial_t u - \mathbf{a}.\nabla u = 0$. □

La restriction est pour $t \ge T(\mathbf{x})$, qui peut faire apparaitre des pertes dans le caractère régulier de la solution. Par exemple le temps de sortie $\mathbf{x} \mapsto T(\mathbf{x})$ peut même ne pas être continu, comme dans l'exemple de la figure 1.4.
De manière générale il est possible de considérer que la fonction définie par formulation **Lagrangienne** (1.9) est une solution généralisée de la formulation **Eulérienne** de l'équation du transport. On pourra consulter [2].

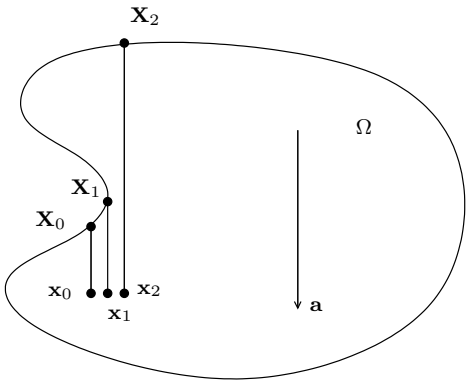


FIGURE 1.4 – La vitesse $\mathbf{a}$ est verticale et constante. La fonction $\mathbf{x} \mapsto \mathbf{X}(T(\mathbf{x}), \mathbf{x}))$ n'est pas continue au point $\mathbf{x}_1$. Le temps de sortie $T(\mathbf{x})$ est également discontinu en $\mathbf{x}_1$.

1.2.2 Equation de la chaleur

L'opérateur Laplacien est défini en dimension d par

$$\Delta u = \nabla \cdot \nabla u = \frac{\partial^2}{\partial x_1^2} u + \cdots + \frac{\partial^2}{\partial x_d^2} u.$$

Soit le problème de la chaleur en dimension $d = 2$ avec une condition de Neumann

$$\left\{ \begin{array}{lll} \partial_t u - \Delta u = 0, & t > 0, & \mathbf{x} \in \Omega, \\ \nabla u \cdot \mathbf{n} = 0, & t > 0, & \mathbf{x} \in \Gamma, \\ u(0, \mathbf{x}) = u_0(\mathbf{x}) & \mathbf{x} \in \Omega. & \end{array} \right. \tag{1.10}$$

Ce problème est bien posé. Il existe une et une seule solution de la formulation variationnelle associée : voir [17, 19].
On considère l'énergie quadratique $E(t) = \frac{1}{2} \|u(t)\|^2_{L^2(\Omega)}$. On a $E'(t) = \int_\Omega u \partial_t u dx = \int_\Omega u \Delta u dx$. Une intégration par parties montre que $E'(t) = -\int_\Omega \nabla u \cdot \nabla u dx + \int_\Gamma u \nabla u \cdot \mathbf{n} d\sigma = -\int_\Omega |\nabla u|^2 \, dx$. Une intégration en temps montre que

$$E(T) + \int_0^T \int_\Omega |\nabla u(t, \mathbf{x})|^2 \, dx dt = E(0).$$

L'unicité est alors immédiate pour les solutions régulières.

Lemme 6. *Soient deux solutions u_1 et u_2 pour la même condition initiale u_0. Alors $u_1 = u_2$.*

Démonstration. Soit $u = u_1 - u_2$, qui est alors solution du même problème avec une condition initiale nulle. L'identité précédente montre que $E(T) \leq E(0) = 0$, donc $u \equiv 0$, ce qui montre l'unicité de la solution. □

Les liens entre (1.10) et les problèmes variationnels stationnaires sont immédiats après utilisation d'une procédure d'Euler implicite pour la discrétisation de la dérivée en temps. Soit $\Delta t > 0$ un pas de temps destiné in fine à tendre vers 0. On approche (1.10) par une succession de problèmes stationnaires u^n

$$\left\{ \begin{array}{lll} u^{n+1} - \Delta t \Delta u^{n+1} = u^n, & \mathbf{x} \in \Omega, & \\ \nabla u^{n+1} \cdot \mathbf{n} = 0, & t > 0, & \mathbf{x} \in \Gamma, \\ u^0 = u_0 & \mathbf{x} \in \Omega. & \end{array} \right. \tag{1.11}$$

Exercice 1. *On considère que $u_0 \in L^2(\Omega)$. Montrer que la formulation variationnelle de (1.11) admet une unique solution dans $H^1(\Omega)$ pour tout $n \in \mathbb{N}$.*

On renvoie à [9] pour les aspects complémentaires.

1.2.3 Principe du maximum

Les équations d'advection et de diffusion satisfont le principe du maximum que nous étudions ici pour le problème dans le plan

$$\begin{cases} \partial_t u + \mathbf{a}\cdot\nabla u - k\Delta u = 0, & \mathbf{x}\in\mathbb{R}^2, \quad t>0 \\ u(0,\mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x}\in\mathbb{R}^2, \end{cases} \tag{1.12}$$

pour $\mathbf{a}\in\mathbb{R}^2$ et $k\geq 0$.
Soit $\varphi:\mathbb{R}\to\mathbb{R}$ une fonction de classe C^2 et convexe : $\varphi''\geq 0$. On supposera que $\varphi(0)=0$ et que u est négligeable à l'infini.

Lemme 7 (Estimation a priori)**.** *On a*

$$\int_{\mathbb{R}^2}\varphi\left(u(t,\mathbf{x})\right)d\mathbf{x} \leq \int_{\mathbb{R}^2}\varphi\left(u_0(\mathbf{x})\right)d\mathbf{x}. \tag{1.13}$$

Démonstration. On a

$$\frac{d}{dt}\int_{\mathbb{R}^2}\varphi\left(u(t,\mathbf{x})\right)d\mathbf{x} = \int_{\mathbb{R}^2}\partial_t u(t,\mathbf{x})\varphi'\left(u(t,\mathbf{x})\right)d\mathbf{x} = \int_{\mathbb{R}^2}\left(k\Delta u - \mathbf{a}\cdot\nabla u\right)\varphi'\left(u(t,\mathbf{x})\right)d\mathbf{x}$$

$$= \nabla\cdot\left(k\nabla\int_{\mathbb{R}^2}\varphi\left(u(t,\mathbf{x})\right)d\mathbf{x} - \mathbf{a}\int_{\mathbb{R}^2}\varphi\left(u(t,\mathbf{x})\right)d\mathbf{x}\right) - k\int_{\mathbb{R}^2}\left|\nabla u(t,\mathbf{x})\right|^2\varphi''\left(u(t,\mathbf{x})\right)d\mathbf{x}.$$

Comme u tend vers 0 pour $|\mathbf{x}|\to\infty$ et que $\varphi(0)=0$, on peut intégrer dans tout le domaine car les termes à l'infini disparaissent. On obtient

$$\frac{d}{dt}\int_{\mathbb{R}^2}\varphi\left(u(t,\mathbf{x})\right)d\mathbf{x}\leq 0.$$

Cela termine la preuve après intégration en temps. □

Soit $u_0\in L^\infty(\mathbb{R}^2)$, et pour simplifier positive et à support compact : $0\leq u_0\leq\|u_0\|_{L^\infty(\mathbb{R}^2)}$.

Lemme 8. *On a pour tout $t>0$*

$$0\leq u(t,\mathbf{x})\leq\|u_0\|_{L^\infty(\mathbb{R}^2)}, \qquad \mathbf{x}\in\mathbb{R}^2. \tag{1.14}$$

Démonstration. Soit la fonction $\varphi:\ \mathbb{R}\to\mathbb{R}$

$$\varphi_-(v) = \max\left(-v,0\right)^3 = \max\left(-v^3,0\right).$$

Cette fonction est convexe. Sa dérivée seconde est continue et nulle en $v=0$. Donc φ_- est C^2. De plus $\varphi\geq 0$ et $\varphi(v)=0$ si et seulement si $v\geq 0$. Du fait de la positivité de la donnée initiale, l'estimation a priori fournit : $\int_{\mathbb{R}^2}\varphi_-\left(u(t,\mathbf{x})\right)\leq 0$. Donc $\int_{\mathbb{R}^2}\varphi_-\left(u(t,\mathbf{x})\right)\leq 0$ et au final $u(t,\mathbf{x})\geq 0$.
Soit à présent

$$\varphi_+(v) = \max\left(0, v-\|u_0\|_{L^\infty(\mathbb{R}^2)}\right)^3 = \max\left(0, \left(v-\|u_0\|_{L^\infty(\mathbb{R}^2)}\right)^3\right),$$

qui est une fonction convexe et de dérivée seconde continue (et nulle en $v=\|u_0\|_{L^\infty(\mathbb{R}^2)}$). On a alors

$$\int_{\mathbb{R}^2}\varphi_+\left(u(t,\mathbf{x})\right)d\mathbf{x}\leq\int_{\mathbb{R}^2}\varphi_+\left(u_0(\mathbf{x})\right)d\mathbf{x}=0$$

ce qui montre *in fine* que $u\leq\|u_0\|_{L^\infty(\mathbb{R}^2)}$. □

1.2.4 Systèmes de Friedrichs

Soient deux matrices $A_1, A_2\in\mathbb{R}^{n\times n}$. On fait l'hypothèse majeure que les matrices sont symétriques

$$A_1 = A_1^t \text{ et } A_2 = A_2^t.$$

On considère le système de Friedrichs à coefficients constants

$$\partial_t\mathbf{U} + A_1\partial_{x_1}\mathbf{U} + A_2\partial_{x_2}\mathbf{U} = 0, \quad t>0,\ \mathbf{x}=(x_1,x_2)\in\mathbb{R}^2. \tag{1.15}$$

La fonction inconnue est $\mathbf{U}(t,\mathbf{x})\in\mathbb{R}^n$. La condition initiale s'écrit $\mathbf{U}(0,\mathbf{x})=\mathbf{U}_0(\mathbf{x})$ pour tout $\mathbf{x}\in\mathbb{R}^2$, où la fonction $\mathbf{U}_0$ est la donnée initiale. Les systèmes de Friedrichs sont accompagnés d'une identité d'énergie quadratique.

Proposition 3. *Les systèmes de Friedrichs conservent l'énergie quadratique : $\frac{d}{dt}\|\mathbf{U}(t,\cdot)\|_{L^2(\mathbb{R})^n}=0$.*

Démonstration. On considère une solution de (1.15), suffisamment régulière d'une part. Le produit scalaire avec $\mathbf{U}$ donne

$$\partial_t \mathbf{U}\cdot\mathbf{U} + A_1\partial_{x_1}\mathbf{U}\cdot\mathbf{U} + A_2\partial_{x_2}\mathbf{U}\cdot\mathbf{U} = 0.$$

On a $\partial_t \mathbf{U}\cdot\mathbf{U} = \frac{1}{2}\partial_t|\mathbf{U}|^2$. Par ailleurs

$$\frac{1}{2}\partial_{x_1}\left(A_1\mathbf{U}\cdot\mathbf{U}\right) = \frac{1}{2}A_1\partial_{x_1}\mathbf{U}\cdot\mathbf{U} + \frac{1}{2}A_1U\cdot\partial_{x_1}\mathbf{U} = \frac{1}{2}A_1\partial_{x_1}\mathbf{U}\cdot\mathbf{U} + \frac{1}{2}\mathbf{U}\cdot A_1^t\partial_{x_1}\mathbf{U}. = \left(\frac{A_1+A_1^t}{2}\partial_{x_1}\mathbf{U}\right)\cdot\mathbf{U}.$$

Or A_1 est symétrique. Donc $\frac{1}{2}\partial_{x_1}\left(A_1\mathbf{U}\cdot\mathbf{U}\right) = (A_1\partial_{x_1}\mathbf{U})\cdot\mathbf{U}$. De même $\frac{1}{2}\partial_{x_2}\left(A_2\mathbf{U}\cdot\mathbf{U}\right) = (A_2\partial_{x_1}\mathbf{U})\cdot\mathbf{U}$ car A_2 est aussi symétrique. On a donc

$$\frac{1}{2}\partial_t|\mathbf{U}|^2 + \frac{1}{2}\partial_{x_1}\left(A_1\mathbf{U}\cdot\mathbf{U}\right) + \frac{1}{2}\partial_{x_2}\left(A_2\mathbf{U}\cdot\mathbf{U}\right) = (\partial_t\mathbf{U} + A_1\partial_{x_1}\mathbf{U} + A_2\partial_{x_2}\mathbf{U})\cdot\mathbf{U} = 0.$$

D'où après intégration en espace pour une fonction assez petite à l'infini (en espace) $\frac{d}{dt}\int_{\mathbb{R}^2}|\mathbf{U}|^2dx = 0$, d'où l'on déduit le résultat. □

1.2.5 Termes sources ou de couplage

Le couplage de certains modèles d'EDP avec des termes sources ou de couplage peut générer de nouvelles questions, tant en terme d'analyse des modèles que de construction pour les méthodes numériques. C'est particulièrement vrai lorsqu'il a y interaction forte entre les termes sources et les opérateurs aux dérivés partielles. Nous illustrons ce comportement sur le modèle suivant.
Soit le système des ondes linéaires avec deux paramètres $\epsilon > 0$ et $\sigma > 0$

$$\begin{cases} \partial_t p + \frac{1}{\varepsilon}\nabla\cdot\mathbf{u} = 0, & t>0, \quad \mathbf{x}\in\mathbb{R}^2, \\ \partial_t\mathbf{u} + \frac{1}{\varepsilon}\nabla p + \frac{\sigma}{\varepsilon^2}\mathbf{u} = 0, & t>0, \quad \mathbf{x}\in\mathbb{R}^2, \end{cases} \tag{1.16}$$

avec les conditions initiales $p(0) = p_0$ et $\mathbf{u}(0) = \mathbf{u}_0$. Les inconnues sont d'une part $p\in\mathbb{R}$ qui est un scalaire et d'autre part $\mathbf{u}\in\mathbb{R}^2$ qui est un vecteur.

Exercice 2. *Montrer formellement l'identité d'énergie*

$$\frac{d}{dt}\int_{\mathbb{R}^2}\left(p(\mathbf{x},t)^2 + |\mathbf{u}(\mathbf{x},t)|^2\right)dx = -\frac{\sigma}{\varepsilon^2}\int_{\mathbb{R}^2}|\mathbf{u}(\mathbf{x},t)|^2dx.$$

Un phénomène particulièrement intéressant apparait dans le régime où $\varepsilon > 0$ est petit. Pour le mettre en évidence nous considérons un développement de Hilbert, c'est à dire que nous développons a priori chacune des quantités présentes en fonction de ε sous la forme

$$p = p^0 + \varepsilon p^1 + \varepsilon^2 p^2 + O(\varepsilon^3)$$

et

$$\mathbf{u} = \mathbf{u}^0 + \varepsilon\mathbf{u}^1 + \varepsilon^2\mathbf{u}^2 + O(\varepsilon^3).$$

Dans ces expressions p et $\mathbf{u}$ dépendent de ε car sont solutions d'un système d'EDP qui dépend de ε. Cependant nous considérons que p^0, p^1, p^2, $\mathbf{u}^0$, $\mathbf{u}^1$ et $\mathbf{u}^2$ sont eux indépendants du paramètre ε. Ceci est un développement a priori ou Ansatz.

Lemme 9. *La limite formelle p^0 vérifie l'équation de la chaleur*

$$\partial_t p^0 - \frac{1}{\sigma}\Delta p^0 = 0, \quad t>0 \text{ et } \mathbf{x}\in\mathbb{R}^2. \tag{1.17}$$

Démonstration. En plongeant ce développement dans le système (1.16) et en organisant en puissance de ε on obtient pour la première équation

$$\frac{1}{\varepsilon}\left(\nabla\cdot\mathbf{u}^0\right) + \left(\partial_t p^0 + \nabla\cdot\mathbf{u}^1\right) + O(\varepsilon) = 0$$

et pour la deuxième équation (la puissance en ε du terme résiduel n'est pas la même)

$$\frac{1}{\varepsilon^2}\left(\sigma\mathbf{u}^0\right) + \frac{1}{\varepsilon}\left(\sigma\mathbf{u}^1 + \nabla p^0\right) + O(1) = 0.$$

En identifiant les coefficients en puissance de ε, on obtient

$$\nabla\cdot\mathbf{u}^0 = 0 \text{ et } \sigma\mathbf{u}^0 = 0 \Longrightarrow \mathbf{u}^0 = 0,$$

puis $\partial_t p^0 + \nabla\cdot\mathbf{u}^1 = 0$ et $\sigma\mathbf{u}^1 + \nabla p^0 = 0$ d'où l'on déduit le résultat après élimination de $\mathbf{u}^1$. □

Il s'ensuit que le **système hyperbolique avec terme source** (1.16) admet une limite asymptotique (1.17) qui est **parabolique sans terme source**. Un tel phénomène de **changement de type** est tout à fait caractéristique de l'interaction de termes sources avec des opérateurs aux dérivées partielles.

Chapitre 2

Quelques principes de construction

Nous considérons plusieurs types de discrétisation numérique en distinguant suivant que la grille est cartésienne ou quelconque, suivant le type d'équation (transport ou chaleur) et suivant la méthode d'approximation (Différences Finies, Eléments Finis et Volumes Finis).
L'indice abstrait signalant une approximation numérique sera noté h. En pratique h est souvent égal au pas d'espace Δx. Plus généralement h pourra désigner l'ensemble des paramètres numériques, par exemple $h = (\Delta x, \Delta t)$.

2.1 Approximation numérique en dimension $d = 1$

Nous considérons une grille de pas d'espace uniforme $\Delta x > 0$ et de pas de temps $\Delta t > 0$. Comme sur la figure 2.1, les points de grille en espace seront notés $x_j = j\Delta x$ pour $j \in \mathbb{Z}$ et les points de grille en temps seront notés $t_n = n\Delta t$ pour $n \in \mathbb{N}$.

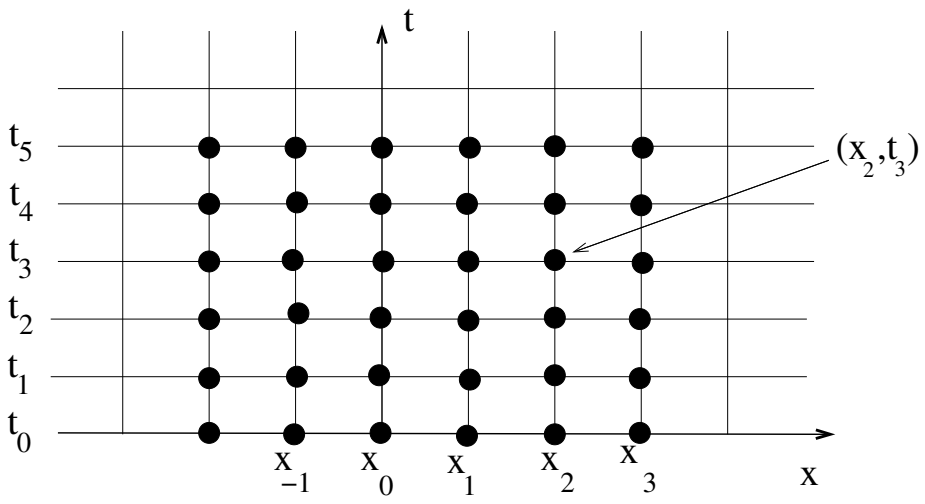


FIGURE 2.1 – Grille Différences Finies

La valeur de la solution exacte u en (x_j, t_n) est

$$u(x_j, t_n).$$

La solution numérique au point (x_j, t_n) sera notée u_j^n. A priori $u_j^n \neq u(x_j, t_n)$ En rassemblant toutes les valeurs

pour un temps donné, on définit la solution au temps t_n

$$v^n = (u(x_j, t_n))_{j\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}.$$

On notera la solution numérique au temps t_n par

$$u^n = \left(u_j^n\right)_{j\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}.$$

Nous considérerons que la donnée initiale u_0 est connue, et pour simplifier que c'est une fonction continue. Aussi la discrétisation sur la grille de la condition initiale est immédiate

$$u_j^0 = u_0(x_j), \qquad j \in \mathbb{Z}. \tag{2.1}$$

Le principe de discrétisation consiste à utiliser l'opérateur aux dérivées partielles pour établir une relation de récurrence qui permette de calculer successivement la solution numérique à chaque pas de temps t_n.

2.1.1 Equation du transport

Approximation par Différences Finies

Principe 1 (Différences Finies)**.** *Le principe de construction des méthodes de Différences Finies consiste à discrétiser les opérateurs différentiels ∂_t et ∂_x, en faisant toutes hypothèses de régularité nécessaires pour justifier les divers ordres d'approximations.*

On a par exemple pour la dérivation en temps

$$\partial_t u(x_j, t_n) = \frac{u(x_j, t_n) - u(x_j, t_n)}{\Delta t} + O(\Delta t).$$

Concernant la dérivation en espace on a

$$\begin{cases} \partial_x u(x_j, t_n) &= \dfrac{u(x_j, t_n) - u(x_{j-1}, t_n)}{\Delta x} + O(\Delta x) & \text{(décentrement à gauche)}, \\ \partial_x u(x_j, t_n) &= \dfrac{u(x_{j+1}, t_n) - u(x_{j-1}, t_n)}{2\Delta x} + O(\Delta x^2) & \text{(approximation centrée)}, \\ \partial_x u(x_j, t_n) &= \dfrac{u(x_{j+1}, t_n) - u(x_j, t_n)}{\Delta x} + O(\Delta x) & \text{(décentrement à droite)}. \end{cases}$$

Supposons que la vitesse est positive $a > 0$ dans l'équation du transport. Pour des raisons de **stabilité**, il faut privilégier la discrétisation en espace décentrée à gauche

$$\frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t} + a\frac{u(x_j, t_n) - u(x_{j-1}, t_n)}{\Delta x} = O(\Delta x, \Delta t). \tag{2.2}$$

En abandonnant le terme de résidu et en remplaçant la valeur exacte par l'approximation numérique, on obtient le schéma de différences finies décentré

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0, \qquad j \in \mathbb{Z}, \quad n \geq 0. \tag{2.3}$$

Ce schéma prend aussi le nom de schéma **upwind**, car le décentrement va chercher l'information en remontant le courant, ou encore en remontant le vent. L'erreur de troncature visible dans (2.2) fait que ce schéma est dit d'ordre un (en temps et en espace).

Principe 2 (Ordre d'un schéma : ce principe sera précisé et généralisé aux définitions 10, 9 et 11 et à la remarque 12)**.** *Soit une équation aux dérivées partielles du premier ordre en temps et d'ordre quelconque en espace. Soit un schéma numérique donné. Supposons que l'insertion des valeurs ponctuelles de la solution exacte dans le schéma permet d'obtenir un résidu de la forme $O(\Delta x^p + \Delta t^q)$. Alors on dira que le schéma est d'***ordre p en espace et q en temps***.*

Une illustration numérique avec le schéma upwind est la suivante. Nous considérons une donnée initiale $u_0(x) = 1$ si $0.2 < x < 0.6$ et $u_0(x) = 0$ ailleurs, et des conditions périodiques aux bords. La solution exacte est $u(x,t) = u_0(x - at)$ aussi

$$u(x, 0.3) = 1 \text{ pour } 0.5 < x < 0.8, \text{ et } u(x, 0.3) = 0 \text{ ailleurs.}$$

Nous notons que cette solution est discontinue.
Nous résolvons numériquement avec 100 mailles sur un intervalle de longueur 1, soit $\Delta x = 0.01$. Les résultats calculés avec le schéma upwind sont présentés à la figure 2.2 pour $\nu = a\frac{\Delta t}{\Delta x}$ avec trois valeurs du paramètre $\nu = 1.1$, $\nu = 0.1$ et $\nu = 0.7$. Pour $\nu = 1.1$, on observe une solution numérique **violemment oscillante**, on dira **instable**. En revanche la solution numérique semble proche de la solution exacte pour $\nu = 0.1$ et $\nu = 0.7$.
A partir de la forme explicite du schéma upwind,

$$u_j^{n+1} = (1 - \nu)u_j^n + \nu u_{j-1}^n$$

on retrouve aisément que $\nu \leq 1$ est une condition suffisante pour éliminer les violentes oscillations numériques du cas $\nu = 1.1$. En effet

$$\nu \leq 1 \Longrightarrow \sup_j \left|u_j^{n+1}\right| \leq \sup_j \left|u_j^n\right|. \tag{2.4}$$

Le phénomène de **stabilité/instabilité** sera étudié systématiquement au chapitre suivant. On démontrera aussi la convergence numérique sous des conditions générales.

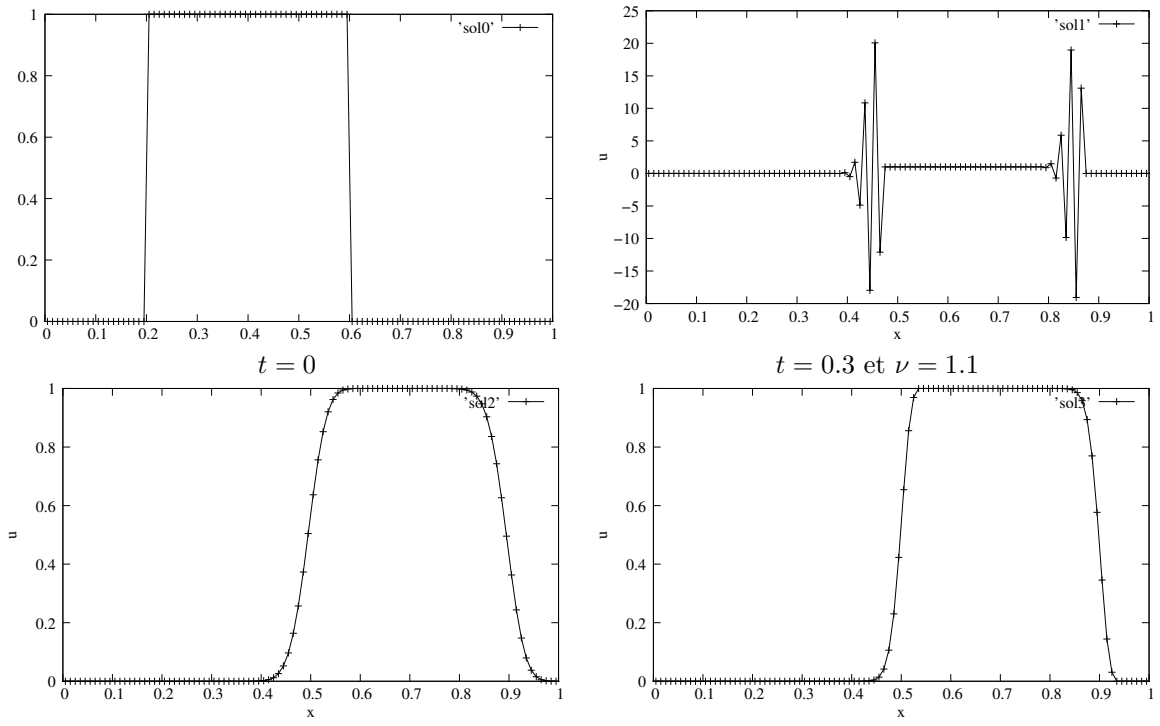


FIGURE 2.2 – Donnée initiale en haut à gauche, solution numérique au temps $t = 0.3$ pour trois valeurs différentes du paramètre $\nu = a\frac{\Delta t}{\Delta x}$. On observe une instabilité en haut à droite, et une solution numérique "correcte" en bas.

Approximation par Éléments Finis

Principe 3 (Méthode des éléments finis)**.** *La discrétisation numérique par méthode des éléments finis s'appuie d'une part sur l'établissement d'une formulation variationnelle des équations, et d'autre part sur le choix d'un espace d'approximation de Galerkin.*

Nous présentons l'application de ce principe sur l'équation simplifiée stationnaire

$$\frac{d}{dx}u = f, \qquad x \in \mathbb{R}. \tag{2.5}$$

pour un second membre donné f. La formulation faible que nous considérons est

$$\int_{\mathbb{R}} \frac{d}{dx}u(x)v(x)dx = \int_{\mathbb{R}} f(x)v(x)dx, \qquad u \in V,\ \forall v \in V. \tag{2.6}$$

A priori l'espace vérifie $V \subset H^1(\mathbb{R})$, ce qui fait que les intégrales sont bien définies (i.e. sont convergentes). Pour une raison de symétrie qui fait partie intégrante des approximations de Galerkin, les fonctions tests sont à prendre dans le même espace. Il faut faire attention cependant car la forme bilinéaire définie dans (2.6) n'est pas coercive. Cependant cela n'empêche pas d'appliquer l'approximation de Galerkin en dimension finie pour obtenir une discrétisation numérique.

Lemme 10. *L'approximation Eléments Finis de type P^1 de l'opérateur différentiel $\frac{d}{dx}$ est centrée.*

Démonstration. L'approximation de Galerkin discrète la plus simple de type P^1 s'appuie sur $V_h = \text{Vect}\,(\varphi_j)_{j\in\mathbb{Z}} \subset V$ avec

$$\begin{cases} \varphi_j(x) = 0 & \text{pour } x \le (j-1)\Delta x \text{ ou } x \ge (j+1)\Delta x, \\ \varphi_j(x) = \dfrac{x-(j-1)\Delta x}{\Delta x} & \text{pour } (j-1)\Delta x \le x \le j\Delta x, \\ \varphi_j(x) = \dfrac{(j+1)\Delta x - x}{\Delta x} & \text{pour } j\Delta x \le x \le (j+1)\Delta x. \end{cases} \tag{2.7}$$

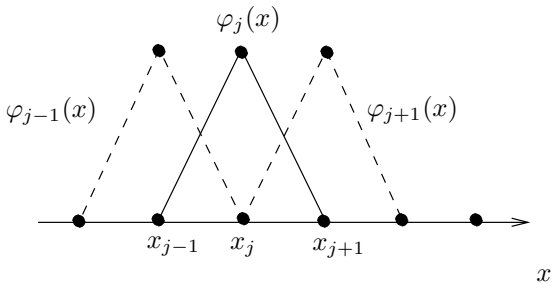


FIGURE 2.3 – Fonction chapeau φ_j et les deux fonctions voisines φ_{j-1} et φ_{j+1}

La formulation discrète est

$$\int_{\mathbb{R}} \frac{d}{dx}u_h(x)v_h(x)dx = \int_{\mathbb{R}} f(x)v_h(x)dx, \qquad u_h \in V_h,\ \forall v_h \in V_h, \tag{2.8}$$

ou encore

$$\int_{\mathbb{R}} \frac{d}{dx}u_h(x)\varphi_j(x)dx = \int_{\mathbb{R}} f(x)\varphi_j(x)dx, \qquad \forall j. \tag{2.9}$$

L'approximation numérique est u_h

$$u_h = \sum_{i\in\mathbb{Z}} u_i\varphi_i.$$

On obtient

$$\sum_{i\in\mathbb{Z}} \left(\int_{\mathbb{R}} \varphi_i'(x)\varphi_j(x)dx \right) u_i = \int_{\mathbb{R}} f(x)\varphi_j(x)dx, \qquad \forall j.$$

Posons $a_{i,j} = \int_{\mathbb{R}} \varphi_i'(x)\varphi_j(x)dx$. Des calculs élémentaires montrent que

$$\begin{cases} a_{i,j} = 0 & i \le j-2, \\ a_{i,j} = 0 & i \ge j+2, \\ a_{j+1,j} = \displaystyle\int_{j\Delta x}^{(j+1)\Delta x} \frac{1}{\Delta x} \times \frac{(j+1)\Delta x - x}{\Delta x} dx = \frac{1}{2}, \\ a_{j-1,j} = \displaystyle\int_{(j-1)\Delta x}^{j\Delta x} \frac{-1}{\Delta x} \times \frac{x - j\Delta x}{\Delta x} dx = -\frac{1}{2}, \\ a_{j,j} = \displaystyle\int_{\mathbb{R}} \frac{d}{dx}\left(\frac{\varphi_j^2}{2}\right) dx = 0. \end{cases}$$

On obtient une approximation numérique sous la forme

$$\frac{u_{j+1} - u_{j-1}}{2} = \int_{\mathbb{R}} f\varphi_j, \qquad j \in \mathbb{Z}.$$

Posons par commodité $f_j = \frac{1}{\Delta x}\int_{\mathbb{R}} f\varphi_j$. On écrit alors

$$\frac{u_{j+1} - u_{j-1}}{2\Delta x} = f_j, \qquad j \in \mathbb{Z}. \tag{2.10}$$

En comparant avec l'équation de départ (2.5), cela montre bien que l'approximation numérique obtenue par éléments finis est centrée. □

Ce principe s'étend naturellement à l'approximation par méthode variationnelle en espace-temps de $\partial_t u + a\partial_x u = 0$ qui s'écrit

$$\int_{\mathbb{R}}\int_{\mathbb{R}} (\partial_t u + a\partial_x u)\, v(x,t)dxdt = 0, \qquad u \in V,\ \forall v \in V.$$

Les fonctions discrètes en temps sont

$$\begin{cases} \psi_n(x) = 0 & \text{pour } t \le (n-1)\Delta t \text{ ou } t \ge (n+1)\Delta t, \\ \psi_n(x) = \dfrac{t - (n-1)\Delta t}{\Delta t} & \text{pour } (n-1)\Delta t \le t \le n\Delta t, \\ \psi_n(x) = \dfrac{(n+1)\Delta t - t}{\Delta t} & \text{pour } n\Delta t \le t \le (t+1)\Delta t. \end{cases}$$

L'approximation numérique est

$$u_h(x,t) = \sum_{i,m} u_i^m \varphi_i(x)\psi_m(t).$$

La variante discrète s'écrit

$$\int_{\mathbb{R}}\int_{\mathbb{R}} (\partial_t u_h + a\partial_x u_{\Delta x,\Delta t})\, \varphi_j(x)\psi_n(t)dxdt = 0, \qquad \forall j, n.$$

On obtient

$$\sum_{j,n} \left(\int_{\mathbb{R}}\int_{\mathbb{R}} (\varphi_i'(x)\psi_m(t) + a\varphi_i(x)\psi_m'(t))\, \varphi_j(x)\psi_n(t)dxdt \right) u_i^m = 0, \qquad \forall j, n,$$

ou encore

$$\sum_{j,n} \left(\int_{\mathbb{R}} a_{i,j}b_{m,n} + ab_{i,j}a_{m,n} \right) u_i^m = 0, \qquad \forall j, n.$$

Les coefficients sont

$$b_{i,j} = \int_{\mathbb{R}} \varphi_i(x)\varphi_j(x)dx = \begin{cases} \frac{2}{3} & \text{pour } i = j, \\ \frac{1}{6} & \text{pour } i = j \pm 1, \\ 0 & \text{pour } i \neq j-1, j, j+1. \end{cases}$$

On obtient finalement le schéma

$$\frac{\frac{1}{6}u_{j-1}^{n+1} + \frac{2}{3}u_j^{n+1} + \frac{1}{6}u_{j+1}^{n+1} - \frac{1}{6}u_{j-1}^{n-1} - \frac{2}{3}u_j^{n-1} - \frac{1}{6}u_{j+1}^{n-1}}{\Delta t} \tag{2.11}$$
$$+a\frac{\frac{1}{6}u_{j+1}^{n-1} + \frac{2}{3}u_{j+1}^{n} + \frac{1}{6}u_{j+1}^{n+1} - \frac{1}{6}u_{j-1}^{n-1} - \frac{2}{3}u_{j-1}^{n} - \frac{1}{6}u_{j-1}^{n+1}}{\Delta t} = 0.$$

On remarque que ce schéma est centré en temps et en espace. Il est aussi **implicite** car on ne peut pas calculer directement u^{n+1}.
Une autre possibilité consiste à utiliser une approximation d'éléments finis pour la partie en espace, et à se contenter d'une discrétisation explicite pour la dérivée en temps. On obtient

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + a\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0. \tag{2.12}$$

Dans les trois cas (2.10), (2.11) et (2.12), l'approximation de $\frac{d}{dx}$ par éléments finis est centrée.

Approximation par Volumes Finis

Principe 4. *La discrétisation numérique par méthodes de volumes finies s'appuie : a) sur une écriture sous forme divergente des équations ; b) sur une intégration des équations dans un volume de contrôle s'appuyant sur un maillage : c) sur la construction de flux numériques pour clore la construction.*

Une forme divergente des équations signale que les différents termes sont rangés "à l'intérieur" des opérateurs différentiels. Pour l'advection c'est le cas car on peut écrire $\partial_t(u) + \partial_x(au) = 0$.
L'étape b) peut se réaliser en intégrant dans un volume espace-temps ou uniquement espace, avec le même résultat. Par souci de simplicité, nous intégrons dans un volume de type espace.
Le **volume** (ou maille, ou cellule) d'indice j est situé entre les bords de volume $x_{j-\frac{1}{2}} = \left(j - \frac{1}{2}\right)\Delta x$ et $x_{j+\frac{1}{2}} = \left(j + \frac{1}{2}\right)\Delta x$. La longueur (volume en 3D) de la maille est $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$: on remarque que les longueurs de mailles peuvent être variables ce qui autorise plus de souplesse pour la mise en oeuvre.
L'intégration dans la maille fournit

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \left(\partial_t u + a\partial_x u\right) dx = \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_t u dx + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a\partial_x u dx = 0. \tag{2.13}$$

La première intégrale est aussi $\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_t u dx = \frac{d}{dt}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t,x)dx$. La quantité $\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t,x)dx$ représente la masse de l'inconnue u dans la maille. Puis nous définissons la valeur moyenne de cette même quantité au temps t_n

$$v_j^n = \frac{\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(x,t_n)dx}{\Delta x_j}.$$

On peut remarquer qu'aucune approximation n'a pour l'instant été réalisée. Une approximation de type Différences Finies de l'opérateur $\frac{d}{dt}$ permet d'obtenir

$$\frac{d}{dt}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t,x)dx = \Delta x_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(\Delta t) \tag{2.14}$$

qui est correct dès que u est suffisamment régulier. Il n'y a donc pas de difficulté véritable avec la discrétisation de la dérivée temporelle.
A présent nous considérons

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a\partial_x u(n\Delta t, x)dx.$$

On intègre dans la maille la forme divergente

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_x au(n\Delta t, x)dx = au(n\Delta t, x_{j+\frac{1}{2}}) - au(n\Delta t, x_{j-\frac{1}{2}}).$$

Le terme de bord $a\ u(n\Delta t, x_{j+\frac{1}{2}})$ est le **flux** que nous devons discrétiser lors de l'étape c). L'idée est d'obtenir une représentation précise de $u(n\Delta t, x_{j+\frac{1}{2}})$ à partir de combinaisons bien choisies des v_j^n.

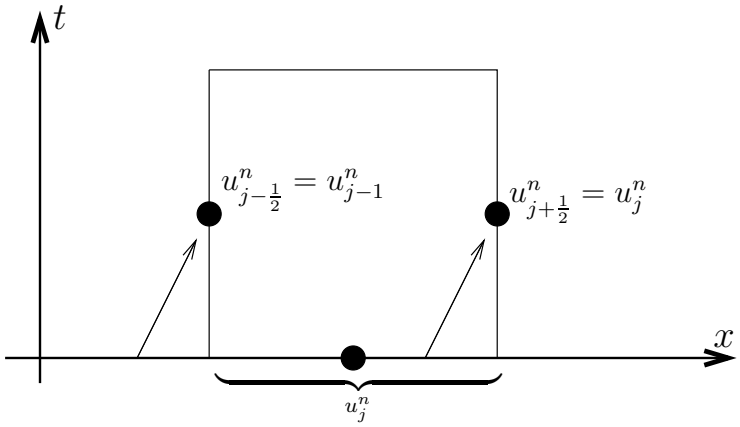


FIGURE 2.4 – La valeur en $x_{j+\frac{1}{2}}$ est décentré en suivant le signe de la vitesse $a > 0$, ce qui revient à remonter le long des caractéristiques.

Le choix usuel (de base) consiste à décentrer cette quantité suivant le sens des caractéristiques, donc suivant le signe de la vitesse a. Pour $a > 0$, on prendra

$$u(n\Delta t, x_{j+\frac{1}{2}}) = v_j^n + O(\Delta x), \qquad \forall j.$$

D'où

$$\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} a\partial_x u(n\Delta t, x)dx = a(v_j^n - v_{j-1}^n) + O(\Delta x). \tag{2.15}$$

On trouve en insérant (2.14) et (2.15) dans (2.13)

$$\Delta x_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + a(v_j^n - v_{j-1}^n) = O(\Delta x) + O(\Delta t).$$

Abandonnant le résidu à droite, nous obtenons le schéma de Volumes Finis

$$\Delta x_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + a(u_j^n - u_{j-1}^n) = 0. \tag{2.16}$$

Ce schéma est d'ordre un en temps et en espace.
Pour le cas de l'équation d'advection, il est aisé de comparer le résultat de ces trois constructions.

Lemme 11. *Soit une grille uniforme : $\Delta x_j = \Delta x$ pour tout j.*
Les schémas de Volumes Finis (2.16) et de Différences Finies (2.3) sont identiques et décentrés, et sont donc différents des schémas d'Elements Finis centrés tels que (2.11) et (2.12).

Exercice 3. *Montrer que le schéma de Lax-Wendroff défini par (2.17) est d'ordre deux en temps et en espace.*

$$u_j^{n+1} = (1-\nu^2)u_j^n + \frac{\nu+\nu^2}{2}u_{j-1}^n + \frac{\nu^2-\nu}{2}u_{j+1}^n. \tag{2.17}$$

Exercice 4. *Montrer que le schéma de Beam-Warming défini par (2.18) est d'ordre deux en temps et en espace.*

$$u_j^{n+1} = \left(1 - \frac{3}{2}\nu + \frac{1}{2}\nu^2\right) u_j^n + (2\nu - \nu^2)u_{j-1}^n + \frac{\nu^2-\nu}{2}u_{j-2}^n. \tag{2.18}$$

2.1.2 Équation de la chaleur

Nous appliquons à présent les principes de construction de Différences Finies, d'Elements Finis et de Volumes Finis à l'équation de la chaleur sur la droite réelle

$$\partial_t u - \partial_{xx} u = 0, \qquad x \in \mathbb{R}, t > 0.$$

Les notations discrètes de points et de mailles sont conservées. Le pas de temps est noté $\Delta t > 0$, et $\Delta x > 0$ est le pas d'espace.

Approximation par Différences Finies

Le schéma explicite de Différences Finis prend la forme

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = 0, \qquad \forall j \in \mathbb{Z}. \tag{2.19}$$

Exercice 5. *Montrer que ce schéma est d'ordre un en temps et deux en espace.*

Une illustration numérique calculée avec le schéma (2.19) est la suivante. Soit une donnée initiale $u_0(x) = \cos(2\pi x)$ et des conditions périodiques aux bords. La solution exacte est

$$u(x,t) = \cos(2\pi x)e^{-4\pi^2 t}.$$

Pour un temps de $T = \frac{\log 2}{4\pi^2} \approx 0.0175581$, on obtient $u(x,T) = \frac{1}{2}u_0(x)$.
Nous résolvons ce problème avec 100 mailles sur un intervalle de longueur 1, soit $\Delta x = 0.01$. Les résultats calculés avec le schéma (2.19) pour un paramètre $\nu = \frac{\Delta t}{\Delta x^2}$ sont présentés à la figure 2.5, pour trois valeurs du paramètre $\nu = 1.1$, $\nu = 0.1$ et $\nu = 0.45$.
A partir de la forme explicite du schéma upwind,

$$u_j^{n+1} = (1-2\nu)u_j^n + \nu u_{j-1}^n + \nu u_{j+1}^n$$

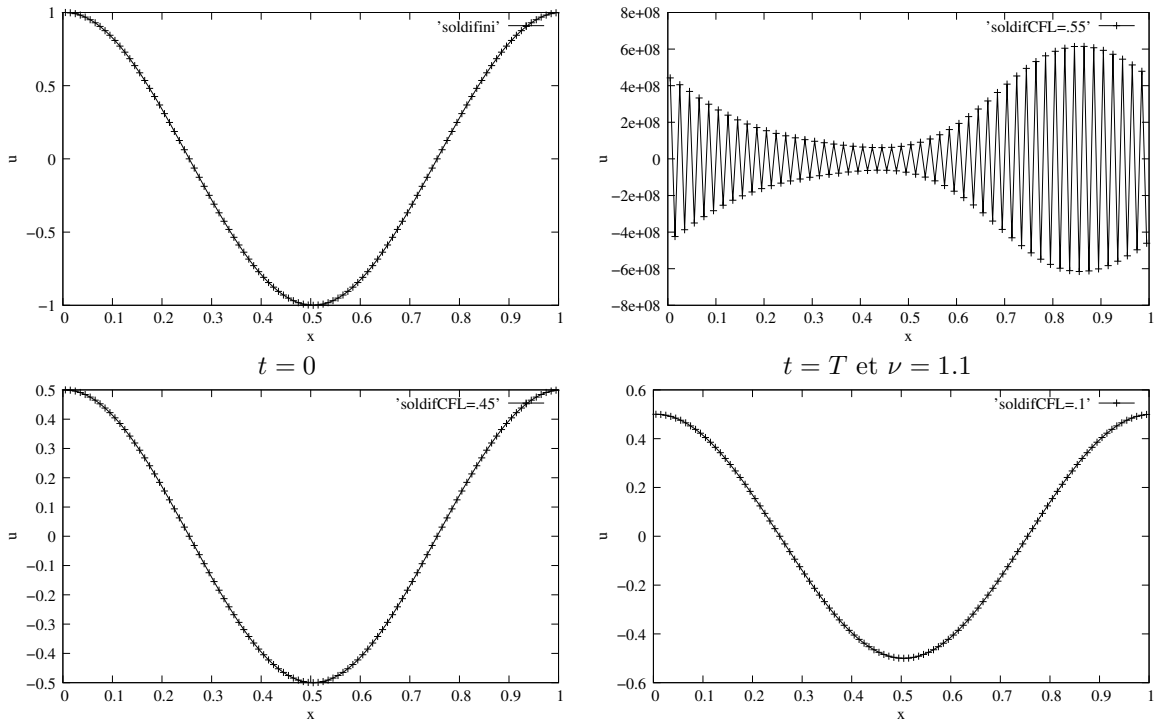
on retrouve aisément que $\nu \leq \frac{1}{2}$ est une condition suffisante pour éliminer les violentes oscillations numériques du cas $\nu = 0.55$. En effet

$$\nu \leq 1 \Longrightarrow \sup_j \left|u_j^{n+1}\right| \leq \sup_j \left|u_j^n\right|.$$

Approximation par Éléments Finis

La méthode des Eléments Finis s'appuie sur une formulation variationnelle que nous développons tout d'abord pour l'équation stationnaire $-u''(x) = f$ avec $f \in L^2(\mathbb{R})$ pour fixer les idées. On a

$$\int_{\mathbb{R}} u'(x)v'(x)dx = \int f(x)v(x)dx, \text{ pour tout } v \text{ dans un espace bien choisi de type } H^1(\mathbb{R}).$$



$t = 0$ $\qquad$ $t = T$ et $\nu = 1.1$

$t = T$ et $\nu = 0.45$ $\qquad$ $t = T$ et $\nu = 0.1$

FIGURE 2.5 – Donnée initiale en haut à gauche, solution numérique au temps $T = \frac{\log 2}{4\pi^2}$. On observe une instabilité en haut à droite, et une solution numérique "correcte" en bas. Les amplitudes de l'instabilité sont sans commune mesure avec l'amplitude de la solution exacte.

Soit une fonction test P^1 définie par (2.7). La formulation discrète est

$$\int_{\mathbb{R}} u_h' \varphi_j' dx = \int f \varphi_j dx.$$

Considérons comme auparavant que $u_h = \sum_i u_i \varphi_i$. On obtient

$$\sum_i \left(\int_{\mathbb{R}} \varphi_i'(x) \varphi_j'(x) dx \right) u_i = \int f \varphi_j dx, \qquad \forall j.$$

Posons $c_{i,j} = \int_{\mathbb{R}} \varphi_i'(x) \varphi_j'(x) dx$. On obtient

$$\begin{cases} c_{i,j} = 0 & i \leq j-2, \\ c_{i,j} = 0 & i \geq j+2, \\ c_{j+1,j} = \displaystyle\int_{j\Delta x}^{(j+1)\Delta x} \frac{1}{\Delta x} \times \frac{-1}{\Delta x} dx = -\frac{1}{\Delta x}, \\ c_{j-1,j} = \displaystyle\int_{(j-1)\Delta x}^{j\Delta x} \frac{-1}{\Delta x} \times \frac{1}{\Delta x} dx = -\frac{1}{\Delta x}, \\ c_{j,j} = \displaystyle\int_{(j-1)\Delta x}^{(j+1)\Delta x} \frac{1}{\Delta x^2} dx = \frac{2}{\Delta x}. \end{cases}$$

Nous posons par commodité $f_j = \frac{1}{\Delta x} \int_{\mathbb{R}} f \varphi_j$. On obtient alors le schéma

$$-\frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x} = \Delta x f_j, \qquad \forall j.$$

Une discrétisation de type Différences Finies explicite du terme $\partial_t u$ permet d'obtenir le schéma

$$\Delta x \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1} - 2u_j + u_{j-1}}{\Delta x} = 0, \qquad n \in \mathbb{N}, \quad j \in \mathbb{Z}. \tag{2.20}$$

On retrouve le schéma aux Différences Finies (2.19).

Approximation par Volumes Finis

Considérons à présent une discrétisation par la méthode des Volumes Finis pour la forme divergente

$$\partial_t u + \partial_x g = 0, \qquad g = -\partial_x u.$$

Nous intégrons en espace entre $x_{j-\frac{1}{2}}$ et $x_{j+\frac{1}{2}}$

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t,x)dx + \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \partial_x g(t,x)dx = 0.$$

D'où

$$\frac{d}{dt} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(t,x)dx + \partial_x g(t, x_{j+\frac{1}{2}}) - \partial_x g(t, x_{j-\frac{1}{2}}) = 0. \tag{2.21}$$

Le centre des mailles est noté x_j avec

$$x_j = \frac{x_{j-\frac{1}{2}} + x_{j+\frac{1}{2}}}{2}.$$

Comme auparavant la valeur moyenne de u dans la maille est notée

$$v_j^n = \frac{\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} u(n\Delta t, x)dx}{\Delta x_j} = u(n\Delta t, x_j) + O(\Delta x_j^2), \tag{2.22}$$

et la dérivation en temps est approchée par la différence finie explicite (2.14). Une discrétisation naturelle du flux $\partial_x u(t, x_{j+\frac{1}{2}})$ est

$$\partial_x u(n\Delta t, x_{j+\frac{1}{2}}) = \frac{u(n\Delta t, x_{j+1}) - u(n\Delta t, x_j)}{x_{j+1} - x_j} + O(x_{j+1} - x_j). \tag{2.23}$$

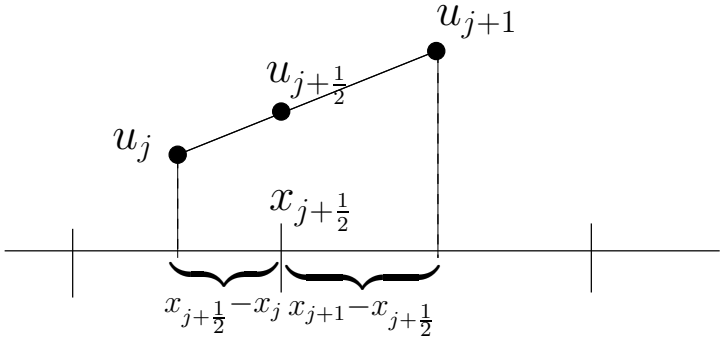


FIGURE 2.6 – Interpolation en $x_{j+\frac{1}{2}}$ de la dérivée en espace.

On peut remarquer que si le maillage est uniforme

$$x_{j+1} - x_{j+\frac{1}{2}} = x_{j+\frac{1}{2}} - x_j \iff x_{j+\frac{3}{2}} - x_{j+\frac{1}{2}} = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}},$$

alors l'erreur d'interpolation est du second ordre

$$\partial_x u(n\Delta t, x_{j+\frac{1}{2}}) = \frac{u(n\Delta t, x_{j+1}) - u(n\Delta t, x_j)}{x_{j+1} - x_j} + O\left((x_{j+1} - x_j)^2\right)$$

et donc a priori plus précis que (2.23). En remplaçant la valeur exacte par la valeur moyenne (2.22), nous obtenons

$$\partial_x u(n\Delta t, x_{j+\frac{1}{2}}) = \frac{v_{j+1}^n - v_j^n}{x_{j+1} - x_j} + O\left(\max(\Delta x_{j+1}, \Delta x_j)\right)$$

D'où à partir de (2.21)

$$\Delta x_j \frac{v_j^{n+1} - v_j^n}{\Delta t} - \frac{v_{j+1}^n - v_j^n}{x_{j+1} - x_j} + \frac{v_j^n - v_{j-1}^n}{x_j - x_{j-1}} = O\left(\max\left(\Delta x_{j+1}, \Delta x_j, \Delta t\right)\right).$$

Il reste à abandonner le résidu et à remplacer la solution exacte par la solution numérique pour obtenir

$$\Delta x_j \frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{u_{j+1}^n - u_j^n}{x_{j+1} - x_j} + \frac{u_j^n - u_{j-1}^n}{x_j - x_{j-1}} = 0. \tag{2.24}$$

Proposition 4. *Soit une grille uniforme : $\Delta x_j = \Delta x$ pour tout j. Alors les schémas de Différences Finies (2.19), d'Eléments Finis (2.20) et de Volumes Finis (2.24) sont identiques.*

2.2 Approximation numérique en dimension $d \geq 2$

Nous passons en revue quelques principes qui permettent d'étendre les schémas précédents en dimension supérieure.
La présentation sera faite en dimension $d = 2$, cependant les principes restent les mêmes en dimension $d = 3$ et plus.

2.2.1 Méthodes de Différences Finies

Soit une grille cartésienne uniforme dont les points en espace sont notés

$$\mathbf{x}_{i,j} = (i\Delta x, j\Delta x), \qquad i, j \in \mathbb{Z}.$$

Les pas de temps sont toujours $t_n = n\Delta t$ pour $n \in \mathbb{N}$. La solution numérique au point d'espace-temps $(\mathbf{x}_{i,j}, t_n)$ sera notée $u_{i,j}^n$.

Principe 5. *Une extension bidimensionnelle immédiate d'un schéma de Différences Finies monodimensionnel consiste à additionner les discrétisations dans les diverses directions spatiales.*

Soit par exemple l'équation d'advection bidimensionnelle

$$\partial_t u + a\partial_x u + b\partial_y u = 0, \qquad (x, y) \in \mathbb{R}^2. \tag{2.25}$$

En supposant $a > 0$ et $b < 0$, un schéma bidimensionnel explicite construit à partir de (2.3) s'écrit

$$\frac{u_{i,j}^{n+1} - u_{i,j}^n}{\Delta t} + a\frac{u_{i,j}^n - u_{i-1,j}^n}{\Delta x} + b\frac{u_{i,j+1}^n - u_{i,j}^n}{\Delta x} = 0. \tag{2.26}$$

Ce schéma est d'ordre un en temps et en espace.

Principe 6. *Une extension bidimensionnelle par splitting directionnel d'un schéma de Différences Finies monodimensionnel consiste à décomposer le schéma en deux étapes monodimensionnelles.*

Toujours pour la même équation (2.25) et avec les mêmes hypothèses $a > 0$ et $b < 0$, on aura le schéma explicite en deux étapes

$$\text{Première étape : } \frac{u_{i,j}^{n+\frac{1}{2}} - u_{i,j}^{n}}{\Delta t} + a\frac{u_{i,j}^{n} - u_{i-1,j}^{n}}{\Delta x} = 0$$

suivi de

$$\text{Deuxième étape : } \frac{u_{i,j}^{n+1} - u_{i,j}^{n+\frac{1}{2}}}{\Delta t} + b\frac{u_{i,j+1}^{n+\frac{1}{2}} - u_{i,j}^{n+\frac{1}{2}}}{\Delta x} = 0.$$

Une telle décomposition peut sembler surprenante à première vue. Cependant en additionnant ces deux étapes on obtient

$$\frac{u_{i,j}^{n+1} - u_{i,j}^{n}}{\Delta t} + a\frac{u_{i,j}^{n} - u_{i-1,j}^{n}}{\Delta x} + b\frac{u_{i,j+1}^{n+\frac{1}{2}} - u_{i,j}^{n+\frac{1}{2}}}{\Delta x} = 0, \tag{2.27}$$

dans lequel on retrouve la discrétisation des dérivées en x et en y mais avec un centrage en temps intermédiaire pour la dérivée discrète en y.
L'extention de ces principes est immédiate pour tout type d'équation qui admet une discrétisation de Différences Finies en dimension $d = 1$ d'espace.

2.2.2 Méthode de Volumes Finis pour l'équation d'advection

Ces idées ont été développées à partir des travaux initiaux de Hill-Reed et Lesaint-Raviart [32, 28] pour la discrétisation de problèmes en neutronique. La motivation initiale était d'utiliser des maillages quelconques avec des données numériques constantes par maille, i.e. P^0, car cela est adapté à la prise en compte d'une physique complexe.
Soit $\mathbf{a} \in \mathbb{R}^2$ un champ de vitesse constant en espace et en temps. Soit $\Omega \subset \mathbb{R}^2$ un ouvert borné polygonal, une illustration se trouve à la figure 2.7.

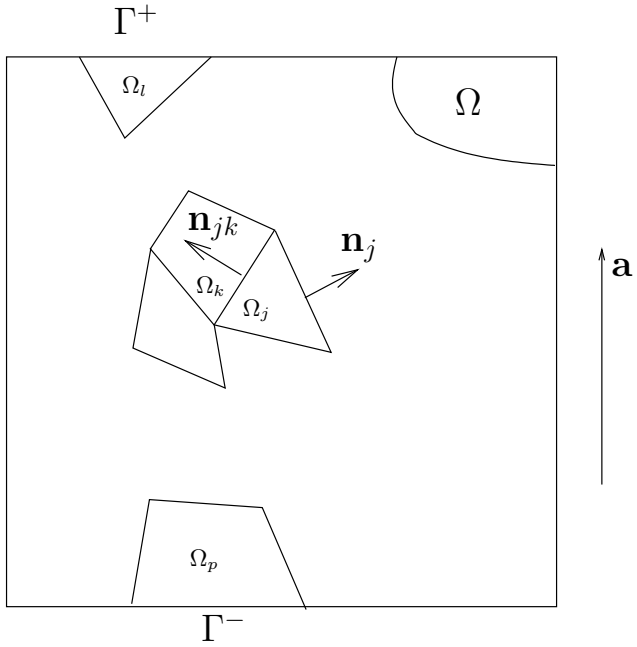


FIGURE 2.7 – Domaine et maillage

Soit un maillage de Ω. Ce maillage est une collection de mailles **polygonales** Ω_j considérées comme des ouverts disjoints, c'est à dire $\Omega_i \cap \Omega_j = \emptyset$ pour $i \neq j$. La condition de recouvrement s'écrit

$$\overline{\Omega} = \cup_j \overline{\Omega_j}.$$

L'aire de Ω_j est notée $s_j > 0$ et

$$\text{Aire}\,(\Omega) = \sum_j s_j.$$

La normale sortante de Ω_j est $\mathbf{n}_j$. L'interface entre Ω_j et Ω_k est $\Sigma_{jk} = \Sigma_{kj}$. Sur cette interface $\mathbf{n}_j$ sera aussi noté $\mathbf{n}_{jk}$. La longueur de l'interface est $l_{jk} = l_{kj}$. Elle peut être nulle pour $\Sigma_{jk} = \emptyset$ ce qui signale que les mailles ne sont pas voisines. Par construction

$$\mathbf{n}_{jk} + \mathbf{n}_{kj} = 0 \text{ pour } l_{jk} > 0.$$

La frontière de Ω_j est alors égale à la collection de segments

$$\overline{\partial\Omega_j} = \overline{\cup_k \Sigma_{jk} \cup \Gamma_j^- \cup \Gamma_j^+}$$

où

$$\Gamma_j^- = \partial\Omega_j \cap \Gamma^-, \quad \text{c'est à dire } \mathbf{a}\cdot\mathbf{n}_j < 0 \text{ sur } \Gamma_j^-,$$

et

$$\Gamma_j^+ = \partial\Omega_j \cap \Gamma^+, \quad \text{c'est à dire } \mathbf{a}\cdot\mathbf{n}_j \geq 0 \text{ sur } \Gamma_j^+.$$

La longueur de $\Gamma_j^\pm$ sera notée $l_j^\pm$.

Définition 6 (Longueur caractéristique du maillage)**.** *Il est utile de définir une longueur caractéristique qui mesure la finesse du maillage : on la notera h. On la définit comme*

$$h = \max\left(\max_{jk} l_{jk}, \max_j l_j^-, \max_j l_j^+\right). \tag{2.28}$$

À partir de ces notations, nous sommes en mesure de construire un schéma de Volumes Finies en suivant le principe 4. Une intégration de l'équation dans la maille Ω_j donne

$$\int_{\Omega_j} (\partial_t u + \nabla\cdot(\mathbf{f}(u)))\, dx = 0.$$

Ici

$$\mathbf{f}(u) = \mathbf{a}u$$

est le **flux exact**. Séparant la dérivée en temps des dérivées en espace

$$\frac{d}{dt}\int_{\Omega_j} u dx + \int_{\Omega_j} \nabla\cdot\mathbf{f}(u) dx = 0. \tag{2.29}$$

La fonction u sera supposée aussi régulière que nécessaire, ce qui permet de justifier tous les développements de Taylor qui seront réalisés. Le terme en temps est

$$\left(\frac{d}{dt}\int_{\Omega_j} u dx\right)(n\Delta t) = s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(h^2\Delta t). \tag{2.30}$$

Ici v_j^n est la valeur moyenne de u au temps t_n

$$v_j^n = \frac{\int_{\Omega_j} u(t_n, \mathbf{x}) dx}{s_j}. \tag{2.31}$$

Exercice 6. *En notant* $\mathbf{x}_j$ *le centre de masse de la maille, montrer que*

$$v_j^n = u(n\Delta t, \mathbf{x}_j) + O(h^2). \tag{2.32}$$

Considérons à présent la deuxième partie de (2.29), que nous intégrons directement dans la maille. En supposant que Ω_j est situé strictement à l'intérieur du domaine, on obtient

$$\int_{\Omega_j} \nabla \cdot \mathbf{f}\left(u(n\Delta t, \mathbf{x})\right) dx = \int_{\partial\Omega_j} \mathbf{f}\ \left(u(n\Delta t, \mathbf{x})\right) \cdot \mathbf{n}_j d\sigma = \sum_k \left(l_{jk}\ \mathbf{a}\cdot\mathbf{n}_{jk}\right) v_{jk}^n. \tag{2.33}$$

Ici v_{jk}^n est la valeur moyenne de u sur l'interface Σ_{jk} au temps t_n

$$v_{jk}^n = \frac{\int_{\Sigma_{jk}} u(n\Delta t, \mathbf{x}) d\sigma}{l_{jk}}.$$

Il est temps d'appliquer l'étape c) du principe de construction 4 des schémas de Volumes Finis. Nous nous appuyons sur les droites caractéristiques. Pour un champ de vitesse $\mathbf{a}$ est orienté de Ω_j vers Ω_k, on considére que $v_{jk}^n \approx v_j^n$. Cela est illustré à la figure 2.8.

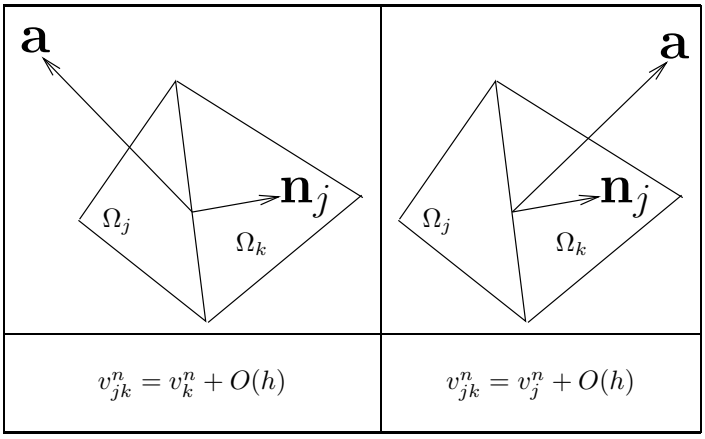


FIGURE 2.8 – On recherche une approximation décentrée suivant le signe de $\mathbf{a}\cdot\mathbf{n}$ de la valeur moyenne à l'interface entre deux mailles.

La même idée est utilisée sur chaque interface. Sur le bord entrant Γ_j^-, on utilise la donnée au bord u^-

$$u_j^{-,n} = \frac{\int_{n\Delta t}^{(n+1)\Delta t} \int_{\Gamma_j^-} u^-(s, \mathbf{x}) d\sigma ds}{\Delta t\ l_j^-}. \tag{2.34}$$

Si $\mathbf{a}\cdot\mathbf{n}_{jk} = 0$, alors la valeur choisie de v_{jk}^n n'a pas d'importance car elle est multipliée $l_{jk}\mathbf{a}\cdot\mathbf{n}_{jk} = 0$ dans (2.33). On obtient

$$s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + O(h^2\Delta t)$$

$$+ \sum_{k,\ \mathbf{a}\cdot\mathbf{n}_{jk}>0} l_{jk}\ \mathbf{a}\cdot\mathbf{n}_{jk}\left(v_j^n + O(h)\right) + \sum_{k,\ \mathbf{a}\cdot\mathbf{n}_{jk}<0} l_{jk}\ \mathbf{a}\cdot\mathbf{n}_{jk}\left(v_k^n + O(h)\right)$$

$$+ l_j^-\ \mathbf{a}\cdot\mathbf{n}_j v_j^{-,n} + l_j^+\ \mathbf{a}\cdot\mathbf{n}_j\left(v_j^n + O(h)\right) = 0.$$

En abandonnant les résidus $O(\cdot)$ et en remplaçant systématiquement les moyennes de la solutions exactes par la solution numérique on obtient

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_{k,\, \mathbf{a}\cdot\mathbf{n}_{jk}>0} l_{jk}\, \mathbf{a}\cdot\mathbf{n}_{jk} u_j^n + \sum_{k,\, \mathbf{a}\cdot\mathbf{n}_{jk}<0} l_{jk}\, \mathbf{a}\cdot\mathbf{n}_{jk} u_k^n + l_j^-\, \mathbf{a}\cdot\mathbf{n}_j u_j^{-,n} + l_j^+\, \mathbf{a}\cdot\mathbf{n}_j u_j^n = 0. \tag{2.35}$$

Notons que le flux numérique sur chaque bord peut se représenter par la formule symétrique

$$F_{j,k}(u,v) = \frac{\mathbf{a}\cdot\mathbf{n}_{jk} + |\mathbf{a}\cdot\mathbf{n}_{jk}|}{2} u + \frac{\mathbf{a}\cdot\mathbf{n}_{kj} - |\mathbf{a}\cdot\mathbf{n}_{kj}|}{2} v. \tag{2.36}$$

Ce flux numérique est une approximation numérique du flux exact, au sens où

$$F_{j,k}(u,u) = \mathbf{f}(u)\cdot\mathbf{n}_{jk}. \tag{2.37}$$

Cette propriété est appelée la **consistance du flux numérique**. Une autre propriété du flux numérique est

$$F_{j,k}(u,v) + F_{j,k}(v,u) = 0 \qquad \forall u, v \in \mathbb{R}. \tag{2.38}$$

Avec ces notations le schéma peut se récrire

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_k l_{jk} F_{jk}(u_j^n, u_k^n) + l_j^- F_{j,j^-}(u_j^n, u_j^{-,n}) + l_j^+ F_{j,j^+}(u_j^n, u_j^{+,n}) = 0 \tag{2.39}$$

où nous avons utilisé la même convention d'écriture pour les flux sur les bords extérieurs indicés j^- et j^+ (le terme $u_j^{+,n}$ est artificiel et ne joue pas sur la valeur du flux numérique).
Soit $M^n = \sum_j s_j u_j^n$ la masse totale dans le domaine de calcul.

Lemme 12. *Le schéma (2.35) est conservatif, au sens où la variation de masse totale se détermine en fonction des flux sur les bords sortant et entrant*

$$\frac{M^{n+1} - M^n}{\Delta t} + \sum_j l_j^-\, \mathbf{a}\cdot\mathbf{n}_j u_j^{-,n} + \sum_j l_j^+\, \mathbf{a}\cdot\mathbf{n}_j u_j^n = 0.$$

Démonstration. Considérant (2.39), il suffit de sommer sur toutes les mailles et de montrer que la contribution des flux internes s'annule. On a

$$\sum_j \sum_k l_{jk} F_{jk}(u_j^n, u_k^n) = \sum_{j,k} \big(l_{jk} F_{jk}(u_j^n, u_k^n) + l_{kj} F_{kj}(u_k^n, u_j^n) \big) = 0$$

en vertu de (2.38). La preuve est terminée. □

La stabilité et la convergence de ce schéma seront établies au chapitre 5.

2.2.3 Méthode de Volumes Finis pour l'équation de la chaleur

Nous considérons l'équation de la chaleur en dimension $d = 2$ avec une condition de Neumann homogène

$$\begin{cases} \partial_t u + \nabla\cdot\mathbf{g}(\nabla u) = 0, & \mathbf{x}\in\Omega, \\ \mathbf{g}(\nabla u)\cdot\mathbf{n} = 0, & \mathbf{x}\in\Gamma \end{cases}$$

pour le flux $\mathbf{g}(\nabla u) = -\nabla u$.
Nous utilisons les notations précédentes sur le maillage. La méthode d'intégration de Volumes Finis est similaire au cas de l'advection, cependant il apparaitra une condition de compatibilité sur le maillage, ce qui traduit une différence fondamentale entre ces deux problèmes.

Après intégration dans Ω_j, discrétisation explicite de la dérivée temporelle et expression des flux aux bords, on obtient

$$s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} + \sum_k l_{jk}\, w_{jk}^n = O(h^2 \Delta t) \tag{2.40}$$

où v_j^n représente la valeur moyenne (2.31) de la solution exacte dans la maille et w_{jk} est la valeur moyenne du flux exact sur l'interface

$$l_{jk} w_{jk}^n = -\int_{\Sigma_{jk}} \nabla u(n\Delta t, x) d\sigma \cdot \mathbf{n}_{jk} = -\nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot \mathbf{n}_{jk} + O(h^2)$$

où $\mathbf{x}_{jk}$ est défini comme le milieu du bord. On a

$$u(n\Delta t, \mathbf{x}_k) = u(n\Delta t, \mathbf{x}_{jk}) + \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{x}_k - \mathbf{x}_{jk}) + O(h^2),$$

et

$$u(n\Delta t, \mathbf{x}_j) = u(n\Delta t, \mathbf{x}_{jk}) + \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{x}_j - \mathbf{x}_{jk}) + O(h^2).$$

En soustrayant nous obtenons

$$u(n\Delta t, \mathbf{x}_k) - u(n\Delta t, \mathbf{x}_j) = \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\mathbf{x}_k - \mathbf{x}_j) + O(h^2).$$

Posons

$$d_{jk} = |\mathbf{x}_k - \mathbf{x}_j| \text{ et } \mathbf{m}_{jk} = \frac{\mathbf{x}_k - \mathbf{x}_j}{d_{jk}} \text{ avec } |\mathbf{m}_{jk}| = 1.$$

Pour continuer la construction, nous ajoutons des conditions sur le maillage.

Hypothèse 1 (Sur le maillage)**.** *Nous supposons qu'il existe une constante $C > 0$ indépendante de h telle que*

$$\inf_{(j,k)} d_{jk} \geq Ch \tag{2.41}$$

où h est la longueur caractéristique (2.28). De plus nous supposons que le segment qui relie les centres de maille est orthogonal au bras

$$\mathbf{m}_{jk} = \mathbf{n}_{jk}, \qquad \forall j, k. \tag{2.42}$$

Un contre-exemple est proposé à la figure 2.9.
Grâce à (2.41) et (2.42) on peut écrire après division par d_{jk}

$$\nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot \mathbf{n}_{jk} = \frac{u(n\Delta t, \mathbf{x}_k) - u(n\Delta t, \mathbf{x}_j)}{d_{jk}} + O(h). \tag{2.43}$$

C'est donc que

$$w_{jk}^n = \frac{u(n\Delta t, \mathbf{x}_k) - u(n\Delta t, \mathbf{x}_j)}{d_{jk}} + O(h).$$

Or on peut approcher à l'ordre deux les valeurs ponctuelles par les valeurs moyennes grâce à (2.32), d'où

$$w_{jk}^n = \frac{v_k^n - v_j^n}{d_{jk}} + O(h).$$

On reporte cette expression dans (2.40). Abandonnant les résidus et remplaçant les moyennes de la solution exacte par la solution numérique, on obtient le schéma numérique de Volumes Finis

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} - \sum_k l_{jk} \frac{u_k^n - u_j^n}{d_{jk}} = 0, \quad \forall j. \tag{2.44}$$

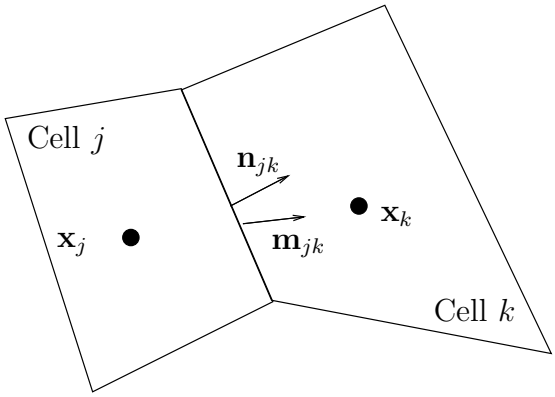


FIGURE 2.9 – Exemple d'un maillage satisfaisant localement (2.41), mais ne satisfaisant pas la condition d'alignement (2.42) car $\mathbf{m}_{jk} \neq \mathbf{n}_{jk}$.

On remarque que la condition au bord de Neumann est automatiquement prise en compte car la somme sur les mailles k exclut le bord. Cette construction permet d'identifier un flux numérique

$$G_{jk}(u,v) = \frac{u-v}{d_{jk}}, \qquad \text{avec } G_{jk}(u,v) + G_{jk}(v,u) = 0. \tag{2.45}$$

Il s'ensuit que le schéma (2.44) se récrit sous la forme générale

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_k l_{jk} G_{jk}\left(u_j^n, u_k^n\right) = 0, \quad \forall j. \tag{2.46}$$

Lemme 13. *Le schéma (2.46) est conservatif, au sens où la variation de masse totale est nulle.*

Exercice 7. *Le montrer.*

La construction de ce schéma est soumise à la contrainte que les centres de mailles $(\mathbf{x}_j)$ doivent satisfaire l'hypothèse 1. Cette hypothèse est en pratique une contrainte extrêmement forte sur le maillage. Les maillages cartésiens voire cartésiens à pas variable tels que celui de la figure 2.10 vérifie cette contrainte. Cependant il n'y a pas de raison qu'un maillage quelconque la satisfasse. Cela montre qu'il y a des liens forts entre la méthode considérée de discrétisation de l'équation de la chaleur et la géométrie du maillage sur lequel s'appuie la discrétisation.
On décrit dans ce qui suit une solution élégante qui relaxe en partie cette contrainte pour les maillages en triangles. Nous allons définir un point $\widehat{\mathbf{x}}_j$ attaché à la maille Ω_j et nous étudions quelques propriétés de la valeur de la solution exacte interpolée en ce point

$$\widehat{v}_j^n = u(n\Delta t, \widehat{\mathbf{x}}_j).$$

Nous définissons par ailleurs des quantités géométriques

$$\widehat{d}_{jk} = |\widehat{\mathbf{x}}_k - \widehat{\mathbf{x}}_j| \text{ et } \widehat{\mathbf{m}}_{jk} = \frac{\widehat{\mathbf{x}}_k - \widehat{\mathbf{x}}_j}{\widehat{d}_{jk}} \text{ tel que } |\widehat{\mathbf{m}}_{jk}| = 1. \tag{2.47}$$

Hypothèse 2. *Supposons alors qu'il existe une constante universelle $C > 0$ indépendante de h avec les propriétés suivantes :* $\sup_j |\widehat{\mathbf{x}}_j - \mathbf{x}_j| \leq Ch$ *;* $\inf_{(j,k)} \widehat{d}_{jk} \geq Ch$ *;* $\widehat{\mathbf{m}}_{jk} = \mathbf{n}_{jk}$ *pour tout* (j,k).

On a pour une solution u suffisamment régulière

$$s_j \frac{v_j^{n+1} - v_j^n}{\Delta t} = s_j \frac{w_j^{n+1} - w_j^n}{\Delta t} + O(h^3).$$

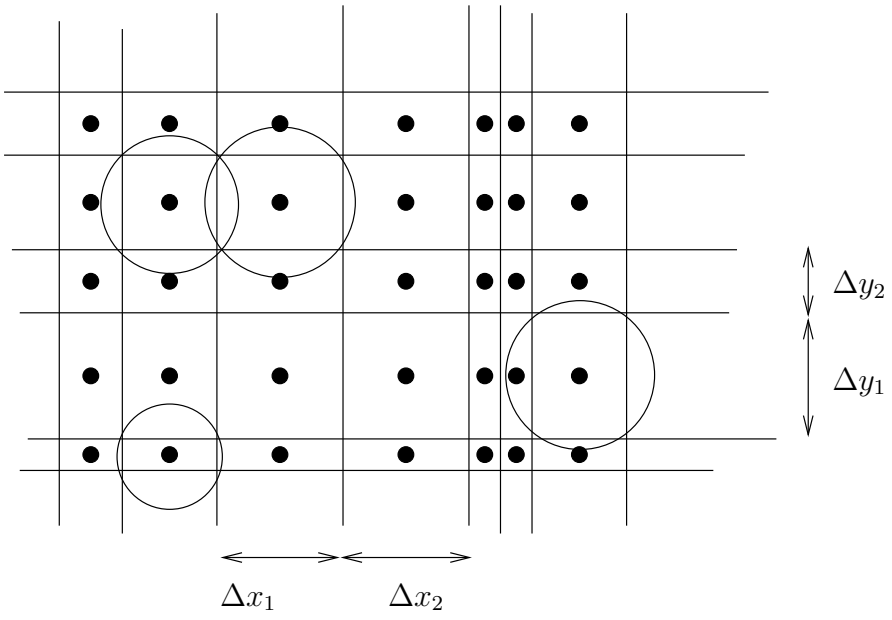


FIGURE 2.10 – Exemple d'un maillage en quadrangles satisfaisant les conditions de l'hypothèse 2. Ce maillage est fortement contraint.

Donc on peut récrire (2.40) sous la forme

$$s_j \frac{w_j^{n+1} - w_j^n}{\Delta t} - \sum_k l_{jk}\, w_{jk}^n = O(h^2 \Delta t) + O(h^3). \tag{2.48}$$

Or nous pouvons approcher w_{jk}^n par une combinaison linéaire de w_j^n et w_k^n. En effet on a

$$u(n\Delta t, \widehat{\mathbf{x}}_k) - u(n\Delta t, \widehat{\mathbf{x}}_j) = \nabla u(n\Delta t, \mathbf{x}_{jk}) \cdot (\widehat{\mathbf{x}}_k - \widehat{\mathbf{x}}_j) + O(h^2).$$

On trouve en utilisant les points a) et b) plus haut

$$\nabla u(n\Delta t, \widehat{\mathbf{x}}_{jk}) \cdot \widehat{\mathbf{m}}_{jk} = \frac{u(n\Delta t, \widehat{\mathbf{x}}_k) - u(n\Delta t, \widehat{\mathbf{x}}_j)}{\widehat{d}_{jk}} + O(h).$$

Le reste de la construction étant similaire, on obtient en suivant les mêmes principes

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_k l_{jk} \widehat{G}_{jk}\left(u_j^n, u_k^n\right) = 0, \quad \forall j \tag{2.49}$$

pour le flux numérique

$$\widehat{G}_{jk}(u, v) = \frac{u - v}{\widehat{d}_{jk}}. \tag{2.50}$$

La convergence de la variante implicite de ce schéma sera établie au chapitre 5.
Pour un maillage donné, les conditions énoncées à l'hypothèse 2 sont des conditions suffisantes pour que le schéma de Volumes Finis (2.49) soit construit en accord avec le principe 4.
Il est absolument remarquable qu'une solution simple à mettre en oeuvre existe pour un maillage en triangles dont tous les angles sont strictement inférieurs à $\frac{\pi}{2}$.

Lemme 14. *Soit un maillage constitué de triangles. Supposons que les angles des triangles soient tous strictement inférieurs à $\frac{\pi}{2} - \epsilon$ pour un ϵ indépendant de h. Soit $\widehat{\mathbf{x}}_j$ le centre du cercle circonscrit à la maille d'indice j.*
Alors $\widehat{\mathbf{x}}_j \in \Omega_j$ pour tout j, et les autres conditions de l'hypothèse 2 sont vérifiées.

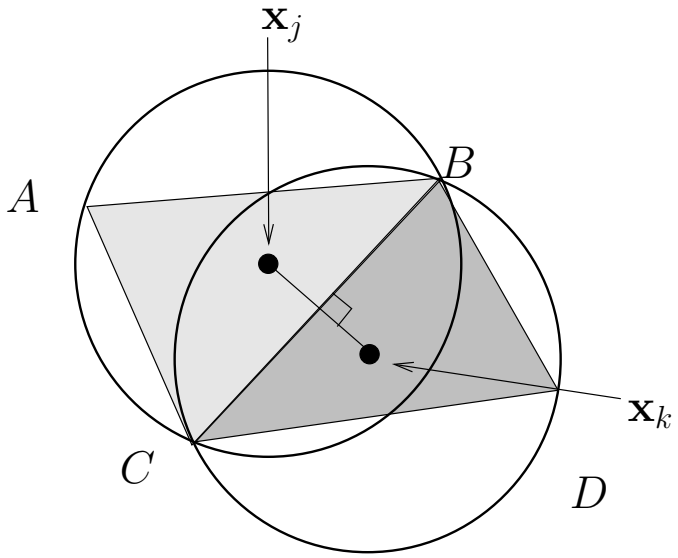


FIGURE 2.11 – Triangles et cercles circonscrits.

Exercice 8. *Démontrer ce résultat en partant de la figure 2.11.*

Les triangles de la figure 2.11 constituent un cas particulier de maillage de Delaunay [16]. La discrétisation de l'équation de la chaleur en Volumes Finis se conduit aussi pour les maillages de Delaunay-Voronoi pour lesquels on renvoie à une référence initiale [22]. Voir aussi [15] pour une utilisation de la condition d'orthogonalité entre les centres de mailles et les bras visible à la figure 2.11 dans le cadre des méthodes de Volumes Finis pour l'équation de la chaleur.

2.2.4 Méthodes de Volumes Finis pour les systèmes de Friedrichs

On reprend les notations sur le maillage introduites précédemment et on considère une solution régulière du système de Friedrichs (1.15). La valeur moyenne de la solution exacte dans la maille est notée

$$\mathbf{V}_j^n = \frac{\int_{\Omega_j} \mathbf{U}(\mathbf{x}, t) dx}{s_j}.$$

La valeur moyenne sur un bras de la solution exacte est notée

$$\mathbf{V}_{jk}^n = \frac{\int_{\Sigma_{jk}} \mathbf{U}(\mathbf{x}, t) d\sigma}{l_{jk}}.$$

Après intégration en temps de l'équation (1.15), discrétisation explicite de la dérivée temporelle et expression des termes de flux au bord, on obtient

$$s_j \frac{\mathbf{V}_j^{n+1} - \mathbf{V}_j^n}{\Delta t} + \sum_k l_{jk} A_{jk} \mathbf{V}_{jk}^n = O(h^2 \Delta t) \tag{2.51}$$

où les matrices de bord sont définies exactement par

$$A_{jk} = A_1 \mathbf{n}_{jk}^1 + A_2 \mathbf{n}_{jk}^2 = -A_{kj}, \qquad \mathbf{n}_{jk} = \left(\mathbf{n}_{jk}^1, \mathbf{n}_{jk}^2\right) \in \mathbb{R}^2. \tag{2.52}$$

Le terme de flux vient d'une utilisation de la formule de Stockes sous la forme

$$\int_{\Omega_j} \left(\partial_{x_1}\left(A_1\mathbf{U}\right) + \partial_{x_2}\left(A_2\mathbf{U}\right)\right) dx = \int_{\partial\Omega_j} \left(\mathbf{n}_j^1 A_1\mathbf{U} + \mathbf{n}_j^2 A_2\mathbf{U}\right) d\sigma = \sum_k l_{jk} A_{jk}\mathbf{V}_{jk}.$$

L'expression (2.51) est exacte car aucune approximation n'a été réalisée pour l'instant.
Si on peut déterminer une expression des termes d'interfaces $\mathbf{V}_{jk}^n$ en fonction des valeurs moyennes $\mathbf{V}_j^n$ et en tenant compte de la structure matricielle du problème, cela permet de proposer une façon de terminer la construction de la méthode. C'est ce que l'on appelle communément un **solveur de Riemann**. Il se trouve qu'il est beaucoup plus judicieux en pratique de chercher à déterminer une valeur pour le produit $A_{jk}\mathbf{V}_{jk}^n$ en fonction de $\mathbf{V}_j^n$ et de $\mathbf{V}_k^n$. En effet les exemples usuels montrent que les matrices A_{jk} peuvent être non inversible ($\det A_{jk} = 0$).
On considère dans ce qui suit un mode de construction simple qui s'appuie sur une décomposition en partie positive et partie négative de la matrice de bord sous la forme

$$A_{jk} = A_{jk}^+ + A_{jk}^- \tag{2.53}$$

où $A_{jk}^+ = \left(A_{jk}^+\right)^t \geq 0$ est une matrice symétrique **positive ou nulle** et $A_{jk}^- = \left(A_{jk}^-\right)^t \leq 0$ est une matrice symétrique **négative ou nulle**, tout en conservant $A_{jk}^+ = A_{kj}^-$ et $A_{jk}^- = A_{kj}^+$. Une telle décomposition est aisée à réaliser pour des matrice symétriques, cependant elle n'est pas unique ce qui explique en partie la profusion de **solveurs de Riemann**. Pour fixer les idées on part d'une diagonalisation

$$A_{jk} = \sum_{p=1}^{n} \lambda_{jk}^p \mathbf{w}_{jk}^p \otimes \mathbf{w}_{jk}^p$$

où les vecteurs propres $\mathbf{w}_{jk}^p$ sont orthonormés. On choisit alors

$$A_{jk}^+ = \sum_{\lambda_{jk}^p > 0} \lambda_{jk}^p \mathbf{w}_{jk}^p \otimes \mathbf{w}_{jk}^p \text{ et } A_{jk}^- = \sum_{\lambda_{jk}^p < 0} \lambda_{jk}^p \mathbf{w}_{jk}^p \otimes \mathbf{w}_{jk}^p. \tag{2.54}$$

On a la formule

$$A_{jk}\mathbf{V}_{jk}^n = A_{jk}^+\mathbf{V}_j^n + A_{jk}^-\mathbf{V}_k^n + O(h) \tag{2.55}$$

qui sert pour définir le flux numérique. En effet on pose

$$\mathbf{f}_{jk}\left(\mathbf{U},\mathbf{V}\right) = A_{jk}^+\mathbf{U} + A_{jk}^-\mathbf{V}. \tag{2.56}$$

En abandonnant les différents termes d'erreur et en remplaçant la solution exacte par la solution numérique, on obtient le schéma de Volumes Finis explicite

$$s_j \frac{\mathbf{U}_j^{n+1} - \mathbf{U}_j^n}{\Delta t} + \sum_k l_{jk}\mathbf{f}_{jk}\left(\mathbf{U}_j^n, \mathbf{U}_k^n\right) = 0. \tag{2.57}$$

On peut faire quelques remarques.

Remarque 1. *On peut se demander pourquoi ne pas prendre un flux numérique plus simple, par exemple* $\mathbf{f}_{jk}\left(\mathbf{U},\mathbf{V}\right) = A_{jk}\frac{\mathbf{U}+\mathbf{V}}{2}$. *Il se trouve qu'un tel choix mène à un schéma instable, et c'est pour cela qu'il n'est jamais retenu.*

Remarque 2. *Si le problème est scalaire c'est à dire* $n = 1$, *alors le système de Friedrichs est identique à l'équation d'advection. On peut alors vérifier que le flux (2.56) est identique au schéma décentré défini par le flux (2.36).*

Remarque 3. *On peut remarquer que la formule (2.55) ne permet pas de définir de valeur intermédiaire car la matrice* A_{jk} *peut ne pas être inversible.*

La stabilité et la convergence de ce schéma peuvent être établies avec les estimations développées au chapitre 5, ce qui justifie cette construction.

Chapitre 3

Transformée de Fourier et Schémas de différences finis

Les schémas ont été abondamment étudiés dans la littérature depuis le début des méthodes numériques [5, 11, 20, 34, 35]. Pour simplifier on considère dans ce chapitre des schémas **explicites à un pas**. Cependant l'ordre peut être arbitrairement élevé.
Ces schémas prennent la forme

$$u_j^{n+1} = \sum_{r=k-p}^{k} \alpha_r u_{j+r}^n \tag{3.1}$$

où les $p+1$ coefficients $(\alpha_r)_{k-p\leq r\leq k}$ caractérisent la méthode et dépendent des paramètres numériques tels que les pas de temps Δt et d'espace Δx. On conviendra que $\alpha_r = 0$ pour $r > k$ ou $r < k-p$. On parle aussi de **schémas compacts** car le stencil est le plus petit possible compte tenu des propriétés d'approximation obtenues.

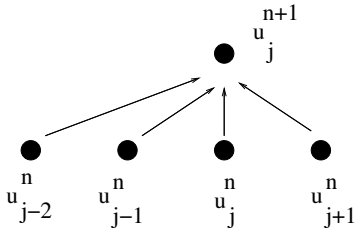


FIGURE 3.1 – Représentation graphique d'un schéma à 4 points pour lequel $k = 1$ et $p = 3$.

3.1 Transformations de Fourier continue et discrète

Un outil d'analyse important pour cette famille de schémas est la **transformation de Fourier**.
On commence par caractériser l'EDP à l'aide de la transformée de Fourier. Le schéma (3.1) est une discrétisation d'une équation aux dérivées partielles que l'on prend sous la forme

$$\partial_t u = Au,$$

et dont l'inconnue est $u(t,x)$ avec une donnée initiale $u(0,x) = u_0(x)$. La transformée de Fourier de $w \in L^2(\mathbb{R})$ est

$$\widehat{w}(\theta) = \int_{\mathbb{R}} w(x) e^{-\mathbf{i}\theta x} dx, \qquad \mathbf{i}^2 = -1,$$

avec la transformée inverse

$$w(x) = \frac{1}{2\pi}\int_{\mathbb{R}} \widehat{w}(\theta) e^{\mathbf{i}\theta x} d\theta$$

et la formule de Plancherel $\|w\|^2_{L^2(\mathbb{R})} = \frac{1}{2\pi}\|\widehat{w}\|^2_{L^2(\mathbb{R})}$. Un opérateur A à coefficients constants peut se caractériser par son symbole en Fourier au moyen de son symbole.

Définition 7 (Symbole d'un opérateur)**.** *La fonction $\theta \mapsto \mu(\theta)$ telle que $Ae^{\mathbf{i}\theta x} = \mu(\theta)e^{\mathbf{i}\theta x}$ est le symbole de A.*

On a la représentation intégrale de la solution

$$u(t,x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{\mu(\theta)t + \mathbf{i}\theta x}\widehat{u}_0(\theta) d\theta \tag{3.2}$$

où $\widehat{u}_0$ est la transformée de Fourier de la donnée initiale u_0. En effet

$$\partial_t u - Au = \frac{1}{2\pi}\int_{\mathbb{R}} (\partial_t - A)\, e^{\mu(\theta)t + \mathbf{i}\theta x}\widehat{u}_0(\theta) d\theta = \frac{1}{2\pi}\int_{\mathbb{R}} (\mu(\theta) - \mu(\theta))\, e^{\mu(\theta)t + \mathbf{i}\theta x}\widehat{u}_0(\theta) d\theta = 0.$$

Par ailleurs on a bien

$$u(0,x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{\mathbf{i}\theta x}\widehat{u}_0(\theta) d\theta = u_0(x).$$

Remarque 4. *Pour l'équation d'advection $A = -a\partial_x$ avec $a \in \mathbb{R}$, le symbole est $\mu(\theta) = -\mathbf{i}a\theta$.*
Pour l'équation de diffusion, $A = D\partial_{xx}$ avec un coefficient de diffusion $D \geq 0$, le symbole est $\mu(\theta) = -D\theta^2$.
On note que dans les deux cas

$$\left|e^{\mu(\theta)t}\right| \leq 1 \text{ pour } t \geq 0 \text{ et } \theta \in \mathbb{R}. \tag{3.3}$$

Notons à présent la solution numérique au temps $t_n = n\Delta t$ comme un vecteur infini disposé en colonne

$$U_h^n = \begin{pmatrix} \cdots \\ u_{-1}^n \\ u_0^n \\ u_1^n \\ \cdots \end{pmatrix} \in \mathbb{R}^{\mathbb{Z}}.$$

Le schéma numérique peut se mettre sous la forme

$$U_h^{n+1} = MU_h^n$$

où la matrice doublement infinie $M = (m_{ij})_{i,j\in\mathbb{Z}}$ a pour coefficients

$$m_{ij} = \alpha_r \text{ avec } r = j - i. \tag{3.4}$$

On dit que M est une matrice **bande**. Seules $p+1$ bandes de M ne sont pas nulles. La matrice M caractérise l'opérateur d'itération $J_{h,\Delta t}$. Par exemple la matrice doublement infinie à deux bandes du schéma upwind est

$$M = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1-\nu & \nu & 0 & 0 & \cdot \\ \cdot & 0 & 1-\nu & \nu & 0 & \cdot \\ \cdot & 0 & 0 & 1-\nu & \nu & \cdot \\ \cdot & 0 & 0 & 0 & 1-\nu & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

Notons que les coefficients de M dépendent de h et Δt, ce qui fait que l'on aurait pu noter avec plus de précision $M_{h,\Delta t}$.

Proposition 5. *La matrice M commute avec sa matrice transposée.*

Démonstration. Cette propriété est une conséquence directe de la structure bande. Les coefficients de $P = MM^t$ sont $p_{ij} = \sum_k m_{ik}m_{jk} = \sum_k \alpha_{k-i}\alpha_{k-j}$. Les coefficients de $Q = M^tM$ sont

$$q_{ij} = \sum_k m_{ki}m_{kj} = \sum_k \alpha_{i-k}\alpha_{j-k} = \sum_{l;\ k=i+j-l} \alpha_{l-j}\alpha_{l-i} = p_{ij}$$

ce qui montre le résultat. □

Exercice 9. *Vérifier que M commute aussi avec l'opérateur de décalage (translation) d'un indice. Vérifier que deux matrices bandes commutent.*

Comme le signale le lemme 23, un bon cadre alors est le cadre quadratique (Hilbertien). Cependant une différence importante avec la situation évoquée au lemme 23 est que nous sommes à présent en **dimension infinie**. On pose

$$V_h = l^2 = \left\{ U = (u_i)_{i\in\mathbb{Z}},\ \sum_i |u_i|^2 < \infty \right\}$$

muni d'une norme pondérée par le pas d'espace

$$\|U\|_h^2 = \Delta x \sum_{i\in\mathbb{Z}} |u_i|^2. \tag{3.5}$$

La projection/interpolation sur les points de grille est naturellement [1]

$$\Pi_h :\ C^0(\mathbb{R}) \cap L^2(\mathbb{R}) \to V_h$$

avec $\Pi_h u = (u(i\Delta x))_{i\in\mathbb{Z}}$. Soit la transformation de Fourier discrète $\widehat{u}(\theta) = \Delta x \sum_{j\in\mathbb{Z}} u_j e^{-i\theta j\Delta x}$ qui est une fonction $\frac{2\pi}{\Delta x}$-périodique appartenant à $L^2\left(-\frac{\pi}{\Delta x}, \frac{\pi}{\Delta x}\right)$ La représentation en Fourier de la solution est

$$u_j = \frac{1}{2\pi} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} \widehat{u}(\theta) e^{\mathbf{i}\theta j\Delta x} d\theta, \qquad j \in \mathbb{Z},$$

avec la formule de Plancherel adaptée

$$\|U\|_h^2 = \frac{1}{2\pi} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} |\widehat{u}(\theta)|^2 \, d\theta. \tag{3.6}$$

Définition 8 (Symbole du schéma)**.** *Le symbole du schéma est la fonction*

$$\lambda(\theta) = \sum_{r=k-p}^{k} \alpha_r e^{\mathbf{i}\theta r}, \qquad \theta \in \mathbb{R},$$

en notant bien que λ dépend de h et Δt par l'intermédiaire des α_r.

Le symbole est en fait la valeur propre de M. Les vecteurs propres (on parle plutôt de vecteurs propres généralisés voir [25]) de M sont $U(\theta) = \left(e^{\mathbf{i}\theta j}\right)_{j\in\mathbb{Z}}$ avec

$$MU(\theta) = \lambda(\theta)U(\theta).$$

En effet

$$\left(MU(\theta)\right)_i = \left(\sum_{r=k-p}^{p} \alpha_r e^{\mathbf{i}(j+r)\theta}\right) = \left(\sum_{r=k-p}^{p} \alpha_r e^{\mathbf{i}r\theta}\right) e^{\mathbf{i}j\theta} = \lambda(\theta)e^{\mathbf{i}j\theta} = \lambda(\theta)\left(U(\theta)\right)_i \qquad i \in \mathbb{Z}.$$

On note que les vecteurs propres sont des modes de Fourier, et qu'ils ne dépendent pas des paramètres de discrétisation. En revanche la valeur propre en dépend par l'intermédiaire des coefficients α_r.

1. Soit une fonction $w \in L^2(\mathbb{R})$, constante sur tout morceau $\left]\left(i-\frac{1}{2}\right)\Delta x, \left(i+\frac{1}{2}\right)\Delta x\right[$. Comme w est continue autour de chaque point $i\Delta x$ on peut définir $\Pi_h w$ sans problème. On note que $\|w\|_{L^2(\mathbb{R})} = \|\Pi_h w\|_h$ ce qui est la raison du poids Δx dans la norme (3.5).

3.2 Stabilité

Lemme 15 (Stabilité au sens de von Neumann)**.** *Le schéma numérique (3.1) est stable au sens de Von Neumann ssi*

$$\sup_{\theta\in\mathbb{R}} |\lambda(\theta)| \leq 1. \tag{3.7}$$

La stabilité au sens de Von Neumann est ici équivalente à la stabilité uniforme dans l^2.

Démonstration. On définit $\widehat{u}_h^n(\theta) = \Delta x \sum_{j\in\mathbb{Z}} u_j^n e^{\mathbf{i}\theta j\Delta x}$ où $(u_j^n) = u_h^n$. Donc

$$\widehat{u}_h^{n+1}(\theta) = \Delta x \sum_{j\in\mathbb{Z}} \left(\sum_{r=k-p}^{p} \alpha_r u_{j+r}^n \right) e^{-\mathbf{i}\theta j\Delta x} = \Delta x \sum_{j\in\mathbb{Z}} \left(\sum_{r=k-p}^{p} \alpha_r e^{i\theta r\Delta x} \right) u_{j+r}^n e^{-\mathbf{i}\theta(j+r)\Delta x}$$

On fait le changement d'indice $j' = j + r$. D'où

$$\widehat{u}_h^{n+1}(\theta) = \left(\sum_{r=k-p}^{k} \alpha_r e^{\mathbf{i}\theta r\Delta x} \right) \widehat{u}_h^n(\theta) = \lambda(\theta\Delta x)\widehat{u}_h^n(\theta).$$

Donc $\|U_h^n\|^2 = \frac{1}{2\pi} \int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} |\lambda(\theta\Delta x)|^{2n} \left|\widehat{u}_h^0(\theta)\right|^2 d\theta$ ce qui montre que la stabilité au sens de Von Neumann est une condition suffisante pour la stabilité uniforme en norme quadratique. Comme $\widehat{u}_h^0$ est quelconque, la condition est aussi nécessaire. □

Le symbole permet aussi de caractériser la stabilité d'un schéma en norme l^∞ ou en norme l^1.

Lemme 16. *Le schéma numérique (3.1) est stable en norme l^1 ou l^∞ ssi il existe une constante $C \in \mathbb{R}^+$ indépendante de h telle que*

$$\sup_{n\geq 0} \left(\sum_{j\in\mathbb{Z}} \left| \int_0^{2\pi} \lambda(\theta)^n e^{-\mathbf{i}j\theta} d\theta \right| \right) \leq C. \tag{3.8}$$

Démonstration. Une formule classique d'algèbre linéaire indique que les normes l^1 et l^∞ d'une matrice $M = (m_{ij})_{ij}$ sont donnés par $\|M\|_1 = \sup_j \left(\sum_i |m_{ij}|\right)$ et $\|M\|_\infty = \sup_i \left(\sum_j |m_{ij}|\right)$. La matrice M_h étant une matrice bande on obtient

$$\|M_h\|_1 = \|M_h\|_\infty = \sum_{r\in\mathbb{Z}} |\alpha_r| \tag{3.9}$$

où les coefficients α_r peuvent se déterminer à partir du symbole par

$$\alpha_r = \frac{1}{2\pi} \int_0^{2\pi} \lambda(\theta) e^{-\mathbf{i}j\theta} d\theta.$$

Or le symbole du produit de deux matrices bande est le produit des symboles, car les vecteurs propres sont communs. Donc le symbole de M_h^n est λ^n. L'utilisation de l'identité (3.9) termine la preuve. □

3.3 Convergence

La consistance et la convergence de la méthode numérique peuvent se caractériser en comparant les symboles à partir du critère (3.10), ce qui est aisé à vérifier pour un schéma donné.

Lemme 17 (Consistance et convergence)**.** *On suppose qu'il existe $p, q, r \in \mathbb{N}^*$ et une constante $C > 0$ indépendante de Δt, Δx et $\theta \in \mathbb{R}$ avec l'inégalité*

$$\left| e^{\mu(\theta)\Delta t} - \lambda(\theta\Delta x) \right| \leq C(1 + |\theta|^r) \left(\Delta x^p + \Delta t^q\right) \Delta t, \qquad -\pi \leq \theta\Delta x \leq \pi. \tag{3.10}$$

Supposons que les critères de stabilité unitaires sont vérifiées, tant continu (3.3) que discret (3.7).
Alors le schéma est convergent à l'ordre p en espace et q en temps en norme quadratique pour des solutions dans $H^r(\mathbb{R})$.

Démonstration. Une norme adaptée, évaluée en Fourier, est $\|u\|^2_{H^r(\mathbb{R})} = \int_{\mathbb{R}}(1+\theta^{2r})|\widehat{u}(\theta)|^2 d\theta$.
Pour les commodités de la preuve nous définissons un opérateur de projection Π_h^4 particulier sous la forme $\Pi_h^4 v = \Pi_h^1 F_h v$ où $F_h v$ est la fonction tronquée en Fourier

$$F_h v(x) = \frac{1}{2\pi}\int_{-\frac{\pi}{\Delta x}}^{\frac{\pi}{\Delta x}} e^{\mathbf{i}\theta x}\widehat{v}(\theta)d\theta.$$

On peut vérifier que $\Pi_h^1 F_h v$ est correctement défini car $F_h vestunefonctioncontinue$: c'est garanti au moins si $v \in H^2(\mathbb{R})$ avec $r \geq 2$. Donc cette condition ne pose pas de difficulté.
Soit la solution numérique u_h^n issue de la donnée initiale $u_h^0 = \Pi_h^4 u_0$. On pose $v_j^n = (\Pi_h^4 u)(n\Delta t, j\Delta x)$ c'est à dire

$$v_j^n = \frac{1}{2\pi}\int_{|\theta|<\frac{\pi}{\Delta x}} e^{\mu(\theta)n\Delta t}e^{\mathbf{i}\theta j\Delta x}\ \widehat{u}_0(\theta)d\theta.$$

Par ailleurs on a

$$u_j^n = \frac{1}{2\pi}\int_{|\theta|<\frac{\pi}{\Delta x}} \lambda(\theta\Delta x)^n e^{\mathbf{i}\theta j\Delta x}\ \widehat{u}_0(\theta)d\theta.$$

Grâce à la formule de Plancherel (3.6), on a

$$\left\|\Pi_h^4 u(t_n) - u_h^n\right\|_h = \frac{1}{\sqrt{2\pi}}\left\|\left(e^{\mu(\theta)n\Delta t} - \lambda(\theta\Delta x)^n\right)\widehat{u}_0(\theta)\right\|_{L^2(-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x})}.$$

Posant $\alpha = e^{\mu(\theta)\Delta t}$ et $\beta = \lambda(\theta\Delta x)$, on peut estimer la parenthèse par $\alpha^n - \beta^n = (\alpha-\beta)\sum_{p=0}^{n-1}\alpha^{n-1-p}\beta^p$. Compte tenu de la stabilité unitaire du problème continu, i.e. $|\alpha| \leq 1$, et de celle du problème discret, i.e. $|\beta| \leq 1$, on obtient $|\alpha^n - \beta^n| \leq n\,|\alpha-\beta|$. En conséquence on a

$$\left\|\Pi_h^4 u(t_n) - u_h^n\right\|_h \leq \frac{n}{\sqrt{2\pi}}\left\|\left(e^{\mu(\theta)\Delta t} - \lambda(\theta\Delta x)\right)\widehat{u}_0(\theta)\right\|_{L^2(-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x})} \leq \frac{Cn\Delta t}{\sqrt{2\pi}}\left\|1+|\theta|^r\widehat{u}_0(\theta)\right\|_{L^2(-\frac{\pi}{\Delta x},\frac{\pi}{\Delta x})}(\Delta x^p + \Delta t^q)$$

grâce à l'hypothèse de consistance sur les symboles (3.10). Pour une donnée initiale dans $H^r(\mathbb{R})$, on a (quitte à redéfinir la constante $C>0$)

$$\left\|\Pi_h^4 u(t_n) - u^n\right\|_h \leq CT\left\|u_0\right\|_{H^r(\mathbb{R})}(\Delta x^p + \Delta t^q), \qquad n\Delta t \leq T. \tag{3.11}$$

La preuve est terminée. □

On peut compléter la preuve en mesurant l'erreur entre l'interpolation ponctuelle classique $\Pi_h^1 v$ et l'interpolation ponctuelle de la fonction tronquée en Fourier $\Pi_h^4 v$. On a le résultat suivant pour une fonction un tout petit plus que $H^r(\mathbb{R})$.

Lemme 18. *Soit $v \in V^r(\mathbb{R}) \subset H^r(\mathbb{R})$ avec*

$$V^r(\mathbb{R})\left\{w \in H^r(\mathbb{R}),\ x(\partial_x)^{r-1}w \in L^2(\mathbb{R})\right\} \qquad r \geq 1.$$

Alors $\left\|\Pi_h^1 v - \Pi_h^4 v\right\|_h \leq C\left\|v\right\|_{V^r(\mathbb{R})}\Delta x^{r-1}$.

Démonstration. On peut vérifier qu'une norme adaptée évaluée en Fourier est

$$\|v\|^2_{V^r(\mathbb{R})} = \int_{\mathbb{R}}\left[(1+\theta^{2r})|\widehat{v}(\theta)|^2 + \theta^{2r-2}|\widehat{v}'(\theta)|^2\right]d\theta.$$

Soit $w = \Pi_h^1 v - \Pi_h^4 v$ avec

$$w_j = \frac{1}{2\pi}\int_{|\theta|>\frac{\pi}{\Delta x}}\widehat{v}(\theta)e^{\mathbf{i}\theta j\Delta x}d\theta = \underbrace{\frac{1}{2\pi}\int_{\frac{\pi}{\Delta x}}^{\infty}\widehat{v}(\theta)e^{\mathbf{i}\theta j\Delta x}d\theta}_{=a_j} + \underbrace{\frac{1}{2\pi}\int_{-\infty}^{-\frac{\pi}{\Delta x}}\widehat{v}(\theta)e^{\mathbf{i}\theta j\Delta x}d\theta}_{=b_j}.$$

Pour $j>0$ on commence par faire une intégration par partie, en intégrant $e^{\mathbf{i}\theta j\Delta x}$ par rapport à θ. D'où par exemple pour le premier terme $a_j = -\frac{1}{2\pi}\int_{\frac{\pi}{\Delta x}}^{\infty}\widehat{v}'(\theta)\frac{e^{\mathbf{i}\theta j\Delta x}-(-1)^j}{\mathbf{i}j\Delta x}d\theta$. On obtient

$$|a_j| \leq \frac{1}{\pi j\Delta x}\int_{\frac{\pi}{\Delta x}}^{\infty}\left|\widehat{v}'(\theta)\right|d\theta$$

puis

$$|a_j| \leq \frac{1}{\pi j\Delta x}\int_{\frac{\pi}{\Delta x}}^{\infty}\left(\theta^{r-1}\left|\widehat{v}'(\theta)\right|\right)\frac{1}{\theta^{r-1}}d\theta \leq \frac{1}{\pi j\Delta x}\left(\int_{\frac{\pi}{\Delta x}}^{\infty}\theta^{2r-2}\left|\widehat{v}'(\theta)\right|^2 d\theta\right)^{\frac{1}{2}}\left(\int_{\frac{\pi}{\Delta x}}^{\infty}\frac{d\theta}{\theta^{2r-2}}\right)^{\frac{1}{2}}$$
$$\leq \frac{C_r}{\pi j\Delta x}\left\|v\right\|_{V^r(\mathbb{R})}\Delta x^{r-\frac{1}{2}} \leq \frac{C_r'\Delta x^{r-\frac{3}{2}}}{j}\left\|u_0\right\|_{V^r(\mathbb{R})}.$$

Le terme a_0 se majore en utilisant que $v \in H^s(\mathbb{R})$. D'où par une nouvelle inégalité de Cauchy-Schwarz (et $1+\frac{1}{4}+\cdots+\frac{1}{j^2}+\cdots<\infty$) : $\|a\|_h \leq C''\left\|v\right\|_{V^r(\mathbb{R})}\Delta x^{r-1}$. De même pour $b=(b_j)$. La preuve est terminée. □

On peut alors mesurer l'erreur entre l'interpolation classique (opérateur Π_h^1) de la solution exacte et la solution numérique issue de l'interpolation classique $u_h^n = J_h^n \Pi_h^1 u(t_0)$. En notant J_h l'opérateur d'itération on a par exemple la décomposition télescopique

$$\Pi_h^1 u(t_n) - u_h^n = \left(\Pi_h^1 u(t_n) - \Pi_h^4 u(t_n)\right) + \left(\Pi_h^4 u(t_n) - J_h^n \Pi_h^4 u(t_0)\right) + J_h^n \left(\Pi_h^4 u(t_0) - \Pi_h^1 u(t_0)\right).$$

Par inégalité triangulaire on obtient une majoration de l'erreur numérique sous la forme $\mathcal{E}_{\text{num}}^n \leq \mathcal{E}_{\text{sch}}^n + \mathcal{E}_{\text{inter}}$. L'erreur du schéma est estimée par le Lemme (17)

$$\mathcal{E}_{\text{sch}}^n = \left\|\Pi_h^4 u(t_n) - J_h^n \Pi_h^4 u(t_0)\right\|_h \leq CT\left(\Delta x^p + \Delta t^q\right) \qquad n\Delta t \leq T, \tag{3.12}$$

et est a priori une fonction croissante de l'indice d'itération n. L'erreur d'interpolation

$$\mathcal{E}_{\text{inter}} \leq C\Delta x^{r-1}$$

est indépendante du temps (de n), et peut être aussi petite que souhaitée pour un r suffisamment grand, ce paramètre étant indépendant des paramètres du schéma caractérisé par p et q. Pour $r \geq p+1$ l'erreur d'interpolation est au moins du même ordre que l'erreur du schéma.

Principe 7. *Au final on retiendra que l'erreur (3.12) de consistance en norme quadratique peut se mesurer directement sur la différence (3.10) entre le symbole exact et le symbole discret. Ce principe est valable pour tout schéma de Différences Finis.*

3.4 Applications

Pour l'équation d'advection un cas couramment rencontré concerne $r = p+1$ avec $p = q$. On note $\nu = a\frac{\Delta t}{\Delta x}$.

Proposition 6 (Forme simplifiée de (3.10) pour l'advection)**.** *Considérons le cas de l'équation d'advection. Pour $r = p+1 = q+1$, le critère (3.10) est équivalent*

$$\left|e^{-\mathbf{i}\nu\theta} - \lambda(\theta)\right| \leq C\left|\theta\right|^p \nu, \qquad -\pi \leq \theta \leq \pi. \tag{3.13}$$

Démonstration. Evident avec le changement de variables $\theta \leftarrow \theta\Delta x$. □

Par exemple le symbole du schéma upwind est $\lambda^{up}(\theta) = (1-\nu) + \nu e^{-\mathbf{i}\theta}$. Comme on vérifie sans peine que

$$\left|(1-\nu) + \nu e^{-\mathbf{i}\theta} - e^{-\mathbf{i}\nu\theta}\right| \leq C\nu\theta,$$

on retrouve bien le fait que le schéma est d'ordre un en norme quadratique.
Considérons à présent le symbole du schéma à trois points (2.19) pour l'équation de diffusion

$$\lambda(\theta) = 1 - 2\nu(1-\cos\theta), \qquad \nu = \frac{\Delta t}{\Delta x^2}, \quad \theta \leftarrow \theta\Delta x^2.$$

Un développement local montre que

$$\left|e^{-\nu\theta^2} - \lambda(\theta)\right| \leq C\nu\theta^2$$

ce qui permet de retrouver la convergence à l'ordre deux en espace (et un en temps mais c'est la même chose pour un schéma explicite de ce type). Le cas général est présenté dans la propriété suivante.

Proposition 7 (Forme simplifiée de (3.10) pour la chaleur)**.** *Considérons l'équation de la chaleur. Pour $r = p+2$ et sans tenir compte de l'ordre en temps, le critère de consistance (3.10) est équivalent*

$$\left|e^{-\nu\theta^2} - \lambda(\theta)\right| \leq C\left|\theta\right|^p \nu, \qquad -\pi \leq \theta \leq \pi. \tag{3.14}$$

Démonstration. Laissée au lecteur. □

Chapitre 4

Analyse numérique abstraite : l'approche de Lax

L'analyse numérique des méthodes de Différences Finies se réalise à partir des notions fondamentales de **consistance** et de **stabilité**, ce qui permet d'évaluer et parfois de mesurer quantitativement la **précision numérique**. Cela pose par ailleurs les bases de l'analyse numérique de la plupart des méthodes numériques instationnaires. La présentation qui suit est tout à fait classique [33, 27, 11, 1, 20, 14], en veillant toutefois à permettre l'analyse numérique des méthodes de Volumes Finis avec les mêmes outils au chapitre suivant.

4.1 Consistance, stabilité et théorème de Lax

La présentation du cadre théorique sera développée à partir du problème linéaire modèle

$$\left\{ \begin{array}{ll} \frac{\partial}{\partial t} u = Au, & t > 0, \\ u(0) = u_0 \in V, & \end{array} \right. \tag{4.1}$$

où V un espace de Banach de norme $||\cdot||$. L'opérateur linéaire est $A : D(A) \to V$ de domaine dense $D(A) \subset V$. Nous supposerons [1] qu'il existe un unique $u(t) \in C^1([0,+\infty) : V) \cap C^0([0,+\infty) : D(A))$ solution de (4.1). Par convention, on représentera $u(t)$ sous la forme abstraite

$$u(t) = e^{tA} u_0. \tag{4.2}$$

Définition 9. *Nous considèrerons que le semi-groupe d'opérateur e^{tA} est* **borné**

$$\exists\ K,\ L \geq 0 \text{ tels que } \quad ||e^{tA}|| \leq K e^{Lt},\ \ t \in \mathbb{R}. \tag{4.3}$$

Nous dirons que e^{tA} est **uniformément borné** *si $L = 0$.*
Nous dirons que e^{tA} est **unitairement borné** *si de plus $K = 1$, auquel cas on a $||e^{tA}|| \leq 1$ pour tout temps.*

La plupart des exemples considérées dans ces notes correspond à des semi-groupes uniformément voire unitairement bornés.

1. Dans le cas plus restreint où V est un space de Hilbert, et sous la condition que A soit maximal monotone

$$\left\{ \begin{array}{l} \text{l'opérateur } A \text{ est monotone : } (Av, v) \geq 0 \quad \forall v \in D(A), \\ \text{l'opérateur } A \text{ est maximal monotone : } (I + A) \text{ est surjectif de } D(A) \text{ dans } V. \end{array} \right.$$

ce problème est bien posé (existence et unicité de la solution) dans le cadre du Théorème de Hille-Yosida [11, 6]. Pour tout $u_0 \in D(A)$, il existe un unique

$$u(t) \in C^1([0,+\infty) : V) \cap C^0([0,+\infty) : D(A))$$

solution de (4.1). L'hypothèse de monotonie implique que $\frac{d}{dt}\|u(t)\|^2 \leq 0$, ce qui implique alors que $\|u(t)\| \leq \|u_0\|$ pour tout $t \geq 0$.

Le problème modèle avec second membre s'écrit

$$\begin{cases} \frac{\partial}{\partial t}u = Au + f, \quad t > 0, \\ u(0) = u_0. \end{cases} \tag{4.4}$$

Sa solution est donnée par la formule de Duhamel

$$u(t) = e^{tA}u_0 + \int_0^t e^{(t-s)A} f(s) ds.$$

Cependant pour des raisons de simplicité de notations, nous ne considérerons que le problème homogène $f = 0$. Par ailleurs cela ne changerait pas fondamentalement les conclusions auxquelles nous arriveront.

4.1.1 Opérateur Π_h d'interpolation/projection

On suppose l'existence d'un sous-espace dense dans V noté $X \subset D(A) \subset V$

$$\forall u \in V, \quad \inf_{v \in X} \|u - v\| = 0.$$

Le sous espace X est typiquement constitué de fonctions régulières, par exemple de classe C^2 voire même C^∞. Ce qu'il faut c'est que X permette au moins de définir l'opérateur d'interpolation et de réaliser les différentes études de consistance nécessaires qui vont suivre.
Soit $V_h \subset V$ un sous-espace vectoriel de V et Π_h un **opérateur d'interpolation**,

$$\Pi_h : X \to V_h.$$

Ici interpolation fait référence au fait que $X \neq V$, ce qui est le cas pour l'exemple central (4.5). Si $\Pi_h\Pi_h = \Pi_h$ on dira que plus Π_h est un **opérateur de projection**. Dans la plupart des situations rencontrés dans ces notes et avec quelques adaptations dans les notations, Π_h est à la fois un opérateur d'interpolation et de projection.
L'exemple central qui permet d'illustrer ces définitions est le suivant. On peut prendre $V = L^p(\mathbb{R}^2)$ pour $1 \leq p \leq \infty$. Un sous-espace X qui convient naturellement est $X = C^0(\mathbb{R}^2)$. On peut aussi prendre $X = C^q(\Omega)$ pour $q \in \mathbb{N}$ assez grand ce qui se révèlera adapté pour l'étude de consistance. L'espace discret s'appuie sur un maillage c'est-à-dire une collection de points

$$\mathbf{x}_{ij} = (i\Delta x, j\Delta y), \quad (i,j) \in \mathbb{Z}^2.$$

Un **opérateur d'interpolation ponctuel naturel** est

$$\Pi_h(u) = (u(\mathbf{x}_{ij}))_{i,j \in \mathbb{Z}} \tag{4.5}$$

qui est défini pour des fonctions de classe C^∞. Le pas d'espace Δx dans la direction x est éventuellement différent du pas Δy dans la direction y. On aura naturellement $h = \max(\Delta x, \Delta y)$. L'espace discret V_h est constitué des fonctions discrètes dont les values sont spécifiées aux points du maillage

$$V_h = \left\{ v_h = (v_{ij})_{i,j \in \mathbb{Z}} \right\}.$$

On pourrait objecter que V_h n'est pas un sous-espace de V. Mais ce n'est en rien une restriction pour peu que l'on identifie (que l'on confonde) V_h et W_h qui est l'espace des fonctions constantes par morceaux sur des carrés $C_{ij} =](i - \frac{1}{2})\Delta x, (i + \frac{1}{2})\Delta x[\times](j - \frac{1}{2})\Delta y, (j + \frac{1}{2})\Delta y[$ de centre $\mathbf{x}_{ij}$

$$W_h = \{ v \in V, \ v \text{ est constant sur tous } C_{ij} \}.$$

On a alors $V_h \approx W_h \subset V$ et la norme dans V_h est la norme de V : $\|v_h\| = \left(\Delta x \Delta y \sum_{ij} |v_{ij}|^p \right)^{\frac{1}{p}}$. Dans ces conditions on a bien que $V_h \subset V$. Enfin il est aisé de donner un sens à la relation $\Pi_h\Pi_h = \Pi_h$ ce qui fait que Π_h est aussi un **opérateur de projection**.
Ces objets dépendent d'un paramètre $h > 0$ qui est destiné à converger vers zéro. Ce paramètre représente les paramètres numériques de la méthode. On peut identifier h à la plus grande longueur du maillage telle que définie dans (2.28).
Nous considérerons que Π_h est un bon opérateur d'approximation au sens où

$$\forall u \in X, \ \lim_{h \to 0} ||u - \Pi_h u|| = 0. \tag{4.6}$$

4.1.2 Opérateur discret A_h

Soit $A_h : V_h \to V_h$. un schéma numérique qui réalise une approximation de A.

Définition 10 (Consistance, ordre d'approximation)**.** *On dit que le schéma numérique A_h est une approximation consistante de A ssi*

$$\forall u \in X, \quad \lim_{h\to 0} ||A_h \Pi_h u - Au|| = 0. \tag{4.7}$$

On dira que l'approximation est d'ordre $p > 0$ ssi il existe une constante $C > 0$ indépendante de h (mais dépendant de u et de ses dérivées) telle que

$$||A_h \Pi_h u - Au|| \leq Ch^p, \qquad u \in X. \tag{4.8}$$

On note que l'ordre d'approximation peut dépendre de X et Π_h.

4.1.3 Cas instationnaire

Soit $\Delta t > 0$ un pas de temps et $t_n = n\Delta t$ pour $n \in \mathbb{N}$. Avec l'ensemble de ces notations, le schéma d'Euler explicite pour la discrétisation de (4.1) s'écrit

$$\begin{cases} \dfrac{u_h^{n+1} - u_h^n}{\Delta t} = A_h u_h^n, \ \ n \geq 0, \\ u_h^0 = \Pi_h u_0. \end{cases} \tag{4.9}$$

L'erreur de troncature de ce schéma est naturellement définie par $r_h^n \in V_h$ avec

$$r_h^n = \frac{1}{\Delta t} \left(\Pi_h u(t_{n+1}) - \Pi_h u(t_n)\right) - A_h \Pi_h u(t_n), \ \ \forall n, h. \tag{4.10}$$

Définition 11. *On dira que le schéma (4.9) est une approximation consistante de (4.1) ssi*

$$\forall u \in C^1([0,T] : X), \quad \lim_{h,\Delta t\to 0} \left(\max_{n \leq \frac{T}{\Delta t}} \|r_h^n - \left(\partial_t u - Au\right)(t_n)\| \right) = 0. \tag{4.11}$$

Comme on montre grâce au théorème de Heine [2] que $\partial_t u$ est uniformément continu, on peut vérifier que

$$\lim_{h,\Delta t\to 0} \left\| \frac{1}{\Delta t} \left(\Pi_h u(t_{n+1}) - \Pi_h u(t_n)\right) - \Pi_h \partial_t u \right\| = 0, \qquad \text{pour } u \in C^1([0,T] : X).$$

Aussi le critère de consistance (4.11) pour ce problème instationnaire se trouve être équivalent au critère de consistance (4.7) pour le problème stationnaire.
Une extension naturelle pour les schémas d'ordre élevé consiste à construire une discrétisation de A qui dépendent de tous les paramètres discrets, ce qui est le cas pour les schémas de Strang d'ordre élevé présentés à la fin de ce chapitre. On notera l'opérateur discret $A_{h,\Delta t} : V_h \to V_h$. Le schéma explicite devient alors

$$\begin{cases} \dfrac{u_h^{n+1} - u_h^n}{\Delta t} = A_{h,\Delta t} u_h^n, \ \ n \geq 0, \\ u_h^0 = \Pi_h u_0. \end{cases} \tag{4.12}$$

L'erreur de troncature devient $r_h^n \in V_h$

$$r_h^n = \frac{1}{\Delta t} \left(\Pi_h u(t_{n+1}) - \Pi_h u(t_n)\right) - A_{h,\Delta t} \Pi_h u(t_n), \ \ \forall n, h. \tag{4.13}$$

2. Toute fonction continue d'un espace métrique dans un espace métrique est uniformément continue si l'espace de départ est compact.

Définition 12 (Critère de consistance précisé et simplifié)**.** *Supposons que : pour toute solution suffisamment régulière $u \in C^r([0,T]:X)$ de $\partial_t u - Au = 0$, on a*

$$\max_{n \leq \frac{T}{\Delta t}} \|r_h^n\| \leq C(h^p + \Delta t^q), \quad p, q > 0. \tag{4.14}$$

Alors on dira que l'approximation (4.12) est d'ordre p en espace et q (pour des solutions $u \in C^r([0,T]:X)$).

Notons qu'on a augmenté la régularité en temps dans (4.14) par rapport à (4.11) car sinon il y a peu de chance d'obtenir un précision en temps à l'ordre q pour q assez grand.

4.1.4 Analyse du schéma d'Euler explicite

Pour la suite nous analyserons le schéma d'Euler explicite qui peut se récrire sous la forme

$$u_h^{n+1} = J_{h,\Delta t} u_h^n$$

où $J_{h,\Delta t} = I_h + \Delta t A_h$ est l'opérateur d'itération et I_h est l'opérateur identité dans V_h : $I_h v_h = v_h$ pour tout $v_h \in V_h$. On peut aussi considérer plus généralement $J_{h,\Delta t} = I_h + \Delta t A_{h,\Delta t}$. La question de la stabilité de cet opérateur d'itération est naturelle. On étend alors la définition 9 en introduisant une possible **restriction sur le pas de temps** [3] pour suivre ce que l'on a observé aux simulations numériques présentées dans les figures 2.2 et 2.5.

Définition 13 (Stabilité et condition CFL de Courant-Friedrichs-Levy)**.** *Nous supposons qu'il existe une fonction $\tau : \mathbb{R}^+ \to \mathbb{R}^+$ et deux constantes $K', L' \geq 0$ avec la propriété suivante : pour tous h et Δt satisfaisant la condition CFL de restriction sur le pas de temps*

$$\Delta t \leq \tau(h), \tag{4.15}$$

on a

$$\left\| J_{h,\Delta t}^n \right\| \leq K' e^{L' n \Delta t}. \tag{4.16}$$

Nous dirons alors que l'opérateur d'itération est **stable** *pour la condition CFL (4.15).*
Si $L' = 0$, on dira que l'opérateur d'itération est **uniformément stable** *pour la condition CFL (4.15).*
Si enfin $L' = 0$ et $K' = 1$, on se propose de dire que l'opérateur d'itération est **unitairement stable** *pour la condition CFL (4.15).*

En général pour un opérateur A_h qui discrétise un opérateur aux dérivées partielles donné A, le pas de temps maximal est tel que

$$\lim_{h \to 0} \tau(h) = 0. \tag{4.17}$$

Une conséquence de la condition CFL (4.17) est alors : plus le maillage est fin, plus le pas de temps est petit, ce qui accroit d'autant la charge de calcul de l'ordinateur.

Théorème 1 (Théorème de Lax : première version)**.** *Soit un schéma linéaire (4.9) consistant au sens de (4.14). Supposons que le pas de temps satisfasse à la condition CFL (4.17). Alors il est convergent au sens où*

$$\forall T > 0, \quad \lim_{h, \Delta t \to 0} \left(\max_{n \leq \frac{T}{\Delta t}} \|\Pi_h u(t_n) - u_h^n\| \right) = 0 \tag{4.18}$$

pour tout $u \in C^1([0,T]:X)$ solution de (4.1).

3. Cela fait référence au célèbre article de 1928 de Courant, Friedrichs et Levy [10].

Démonstration. On définit l'erreur numérique $e_h^n = u_h^n - \Pi_h u(t_n)$. On a la formule de récurrence $e_h^{n+1} = (I_h + \Delta t A_h)\, e_h^n + \Delta t\, r_h^n$ avec l'initialisation $e_h^0 = 0$. D'où la formule de représentation

$$e_h^n = (I_h + \Delta t A_h)^n\, e_h^0 + \Delta t \sum_{p=0}^{n-1} (I_h + \Delta t A_h)^{n-1-p}\, r_h^p,$$

ou plus précisément $e_h^n = \Delta t \sum_{p=0}^{n-1} (I_h + \Delta t A_h)^{n-1-p}\, r_h^p$. D'où

$$||e_h^n|| \leq \Delta t \sum_{p=0}^{n-1} ||\, (I_h + \Delta t A_h)^{n-1-p}\, ||\, ||r_h^p|| \leq \left(\Delta t \sum_{p=0}^{n-1} ||r_h^p|| \right) e^{L'T} \leq T e^{L'T} \max_{n \leq \frac{T}{\Delta t}} \|r_h^n\| \tag{4.19}$$

pour tout n tel que $n\Delta t \leq T$. Or $u \in C^1([0,T] : X)$ est solution de (4.1). Donc le critère de consistance (4.14) montre que l'erreur de troncature tend vers zéro, $\lim_{h,\Delta t \to 0} \left(\max_{n \leq \frac{T}{\Delta t}} \|r_h^n\| \right) = 0$. D'où le résultat recherché. □

La forme utile en pratique est plutôt la suivante.

Théorème 2 (Théorème de Lax : deuxième version)**.** *Soit un schéma linéaire (4.12) vérifiant le critère de consistance précisé (4.14) à l'ordre p en espace et q en temps. Supposons que le schéma est stable (4.16) sous CFL (4.17). Alors il est convergent à l'ordre p en espace et q en temps et*

$$\max_{n \leq \frac{T}{\Delta t}} \|\Pi_h u(t_n) - u_h^n\| \leq C T e^{L'T} \left(h^p + \Delta t^q \right).$$

Démonstration. Préciser (4.19). □

Ce théorème est central dans la compréhension des propriétés d'approximation des schémas numériques linéaires.

Exercice 10. *Enoncer et montrer un théorème de Lax pour le problème avec second membre (4.4).*

Schéma d'Euler implicite

On analyse ici un fait bien connu qui est que les schémas implicites ont souvent des propriétés de stabilité supérieures par rapport à celles des schémas explicites.
Soit par exemple le schéma d'Euler implicite pour résolution numérique de (4.1)

$$\begin{cases} \dfrac{u_h^{n+1} - u_h^n}{\Delta t} = A_h u_h^{n+1}, \quad n \geq 0, \\ u_h^0 = \Pi_h u_0. \end{cases} \tag{4.20}$$

On remarque que A_h est ici indépendant de Δt. La relation de récurrence est à présent

$$(I_h - \Delta t A_h) u_h^{n+1} = u_h^n.$$

Cela définit u_h^{n+1} à condition que l'opérateur linéaire $I_h - \Delta t A_h$ soit inversible.

Proposition 8 (Stabilité du schéma d'Euler implicite)**.** *Supposons que l'opérateur d'itération explicite $I_h + \Delta t A_h$ est uniformément stable*

$$\|(I_h + \Delta t A_h)^n\| \leq K', \quad n \in \mathbb{N}$$

pour un pas de temps $\Delta t \leq \tau(h)$ restreint par une condition CFL (4.15).
Alors pour tout $\Delta t > 0$, l'opérateur $I - \Delta t A_h$ est inversible et uniformément stable avec la même constante

$$\left\|(I_h - \Delta t A_h)^{-n}\right\| \leq K', \quad n \in \mathbb{N}.$$

Démonstration. Il est remarquable que le schéma implicite soit stable indépendamment de toute condition CFL sur le pas de temps.
On définit $T_h = I_h + \tau(h)A_h$, $\alpha = \frac{\Delta t}{\Delta t+\tau(h)}$ et $\beta = 1-\alpha = \frac{\tau(h)}{\Delta t+\tau(h)}$. Alors

$$I_h - \Delta t A_h = \frac{1}{\beta}\left(I_h - \alpha T_h\right).$$

On a la série de Neumann

$$\left(I_h - \Delta t A_h\right)^{-1} = \beta\left(I_h - \alpha T_h\right)^{-1} = \beta\sum_{p=0}^{\infty}\alpha^p T_h^p.$$

Cette série est bien convergente et

$$\left\|\sum_{p=0}^{\infty}\alpha^p T_h^p\right\| \le \sum_{p=0}^{\infty}\alpha^p\|T_h\|^p \le \sum_{p}\alpha^p K' = \frac{K'}{\beta}.$$

Cela montre que la majoration $\left\|(I_h - \Delta t A_h)^{-1}\right\| \le K'$, et implique l'inversibilité de $I_h - \Delta t A_h$.
Par ailleurs on a

$$\left(I_h - \Delta t A_h\right)^{-n} = \beta^n\left(\sum_{p=0}^{\infty}\alpha^p T_h^p\right)^n = \beta^n\sum_{q=0}^{\infty}C(q,n)\alpha^q T_h^q.$$

avec $C(q,n)$ qui désigne le nombre de combinaisons possibles pour que

$$p_1 + \cdots + p_n = q \text{ avec } 0 \le p_i \le n \text{ pour tout } 1 \le i \le n.$$

D'où

$$\left\|\left(I_h - \Delta t A_h\right)^{-n}\right\| \le \beta^n\sum_{q=0}^{\infty}C(q,n)\alpha^q K' = \beta^n\left(\sum_{p=0}^{\infty}\alpha^p\right)^n K' = \beta^n\frac{1}{\beta^n}K' = K'.$$

□

Remarque 5 (Calcul effectif de u_h^{n+1})**.** *En pratique, c'est à dire pour des calculs* **sur ordinateur**, *l'opérateur linéaire $M_h = I_h - \Delta t A_h$ est une* **matrice de dimension finie**. *Le calcul de u_h^{n+1} s'effectue en inversant un système linéaire, ce qui doit s'opérer par des méthodes efficaces d'algèbre linéaire qui ne sont pas évoquées dans ces notes.*

Proposition 9. *Soit un schéma numérique d'Euler explicite uniformément stable sous condition CFL, consistant et donc convergent. Alors le schéma numérique d'Euler implicite associé (4.20) est également convergent.*

Démonstration. La stabilité ayant déjà été montrée, il reste à vérifier la consistance ce qui permettra d'appliquer le théorème de Lax sous la forme générale 1. L'erreur de consistance ou de troncature du schéma implicite

$$\widehat{r}_h^n = \frac{1}{\Delta t}\left(\Pi_h u(t_{n+1}) - \Pi_h u(t_n)\right) - A_h\Pi_h u(t_{n+1}), \quad n \in \mathbb{N}. \tag{4.21}$$

On a immédiatement $\lim_{h,\Delta t\to 0}\left(\max_{n\le\frac{T}{\Delta t}}\left\|\widehat{r}_h^n - \Pi_h\left(\partial_t u - Au\right)(t_n)\right\|\right) = 0$ pour tout $u \in C^1([0,T]:X)$ solution de (4.1). Cela établit la consistance.
Pour finir la preuve de convergence, on peut récrire (4.21) sous la forme

$$\Pi_h u(t_{n+1}) = \left(I_h - \Delta t A_h\right)^{-1}\Pi_h u(t_n) + \Delta t s_h^n, \qquad s_h^n = \left(I_h - \Delta t A_h\right)^{-1} r_h^n.$$

De son côté le schéma se récrit

$$u_h^{n+1} = \left(I_h - \Delta t A_h\right)^{-1} u_h^n.$$

Donc la différence $e_h^n = u_h^n - \Pi_h u(t_n)$ est solution de $e_h^{n+1} = \left(I_h - \Delta t A_h\right)^{-1} e_h^n + \Delta t s_h^n$ avec une erreur initialement nulle $e_h^0 = 0$. On peut alors se contenter de reprendre la preuve des théorèmes 1 ou 2. La preuve est terminée. □

4.1.5 Schéma de Crank-Nicholson

Le terme $A_h u_h^n$ pour le schéma explicite, ou $A_h u_h^{n+1}$ pour le schéma implicite est un discrétisation du premier ordre de la dérivée en temps. La méthode de Cranck Nicholson est a priori plus précise car du deuxième ordre d'approximation pour la partie temporelle. Elle s'écrit

$$\begin{cases} \dfrac{u_h^{n+1}-u_h^n}{\Delta t} = A_h \dfrac{u_h^{n+1}+u_h^n}{2}, \ \ n \geq 0, \\ u_h^0 = \Pi_h u_0, \end{cases} \tag{4.22}$$

où A_h est indépendant de Δt. La relation de récurrence est

$$\left(I_h - \frac{1}{2}\Delta t A_h\right) u_h^{n+1} = \left(I_h + \frac{1}{2}\Delta t A_h\right) u_h^n.$$

Sous les hypothèses de la proposition 9, l'opérateur $I_h - \frac{1}{2}\Delta t A_h$ est inversible. Le schéma est également uniformément stable, consistant et donc est convergent.

4.1.6 Schéma semi-discret

A présent nous considérons le **schéma semi-discret** qui est la limite continue en temps du schéma explicite (ou du schéma implicite) c'est à dire pour $\Delta t \to 0$ et h fixe. Au contraire des précédents schémas, c'est un schéma purement théorique au sens où il n'est pas possible de le programmer sur ordinateur. Son intérêt est qu'il peut simplifier de manière importante l'étude des méthodes numériques.
Formellement on écrit que $v_h(t)$ est solution du système

$$\begin{cases} \frac{d}{dt} v_h(t) = A_h v_h(t), \\ v_h(0) = \Pi_h u_0. \end{cases} \tag{4.23}$$

Encore une fois A_h est ici indépendant de Δt.
Dans le cas où V_h est un espace de dimension fini, A_h est de fait une matrice carrée de taille finie. La solution est donnée par l'exponentielle de matrice

$$v_h(t) = e^{tA_h}\Pi_h u_0.$$

Il suffit que A_h soit un opérateur borné pour donner un sens à cette représentation de la solution. Or c'est bien le cas si le schéma explicite est stable sous condition CFL, car alors $\|(I_h + \tau(h)A_h)^n\| \leq K'$ d'où l'on tire que $\|A_h\| \leq \frac{1+K'}{\tau(h)} < \infty$ ce qui fait que A_h est bien un opérateur linéaire borné. On en déduit que $\left\|e^{tA_h}\right\| \leq e^{t\|A_h\|}$.
On a en fait mieux en supposant la stabilité uniforme du schéma explicite. En effet on a la formule $e^{tA_h} = e^{-\mu}e^{\mu(I_h+\tau(h)A_h)}$ pour $\mu = \frac{t}{\tau(h)}$. Cela montre que

$$e^{tA_h} = e^{-\mu}\sum_{n=0}^{\infty}\frac{\mu^n}{n!}\left(I_h + \tau(h)A_h\right)^n.$$

On obtient l'estimation

$$\left\|e^{tA_h}\right\| \leq e^{-\mu}\sum_{n=0}^{\infty}\frac{\mu^n}{n!}K = e^{-\mu}e^{\mu}K = K. \tag{4.24}$$

Cela montre que le schéma semi-discret bénéficie de la même propriété de stabilité que le schéma explicite. Ce résultat est en fait une extension de la proposition 9.

4.1.7 Principe de comparaison et supra-convergence

Nous montrons un principe de comparaison pour un opérateur A_h dont l'opérateur d'itération explicite est stable (4.16) sous un condition de type CFL telle que (4.15). Ce principe sera utilisé au chapitre 5 pour l'analyse numérique des schémas de Volumes Finis.
Soit u_h^n solution du schéma d'Euler explicite pour une certaine donnée initiale u_0

$$\begin{cases} \dfrac{u_h^{n+1}-u_h^n}{\Delta t} = A_h u_h^n, \ \ n \geq 0, \\ u_h^0 = \Pi_h u_0. \end{cases} \tag{4.25}$$

Soit v_h^n donné par

$$\begin{cases} \dfrac{v_h^{n+1}-v_h^n}{\Delta t} = A_h v_h^n + r_h^n, \ \ n \geq 0, \\ v_h^0 = \Pi_h u_0, \end{cases} \tag{4.26}$$

où r_h^n joue le rôle d'une erreur de troncature avec des propriétés particulières. On fait l'hypothèse que l'on peut écrire

$$r_h^n = \tau(h) A_h s_h^n \quad \text{avec } \|s_h^n\| \leq S < \infty \text{ pour tout } h, n. \tag{4.27}$$

Comme la condition de stabilité (4.16) permet simplement de borner $\|\tau(h)A_h\| \leq C$, on déduit à partir de (4.27) que $\left\|r_h^n\right\| \leq C$. Ce terme est $O(1)$ par rapport à h. Donc une stratégie basé sur le théorème de Lax pour estimer la différence entre u_h^n et v_h^n ne donnera que $\left\|v_h^n - u_h^n\right\| = O(1)$. L'intérêt du résultat suivant est qu'il indique que la structure (4.27) fait que la différence tend vers zéro, avec un taux de convergence explicite. Cette propriété abstraite développée dans [12] sera utile pour l'étude de certaines méthodes de Volumes Finis, et tente de correspondre à la notion de **supraconvergence** énoncée dans [38].

Lemme 19. *Il existe une constante $C > 0$ (qui dépend des estimations de stabilité) telle que*

$$\|v_h^n - u_h^n\| \leq CS\sqrt{\frac{T\tau(h)}{1-\nu}}, \qquad n\Delta t \leq T, \tag{4.28}$$

avec $\nu = \frac{\Delta t}{\tau(h)} < 1$ et S donné dans (4.27).

Dans les cas que nous considérons, on a $\tau(h) \to 0$ pour h tendant vers 0. Aussi cette inégalité est en fait un résultat de convergence de v_h^n vers u_h^n. Notons que la condition CFL est stricte, au sens où Δt doit être strictement inférieur au pas de temps maximal $\tau(h)$.

Démonstration. Soit $e_h^n = v_h^n - u_h^n$ avec $\frac{e_h^{n+1} - e_h^n}{\Delta t} = A_h e_h^n + r_h^n$ et $e_h^0 = 0$. Donc

$$e_h^n = \Delta t \sum_{p=0}^{n-1} (I_h + \Delta t A_h)^{n-1-p} r_h^p. \tag{4.29}$$

Posons $T_h = I_h + \tau(h)A_h$ dont les puissances sont bornées sous la forme $\left\|T_h^q\right\| \leq K' e^{L' q \Delta t}$ grâce à la stabilité (4.16). On posera $C = K' e^{L'T}$ un majorant uniforme des $\left\|T_h^q\right\|$ pour $q\Delta t \leq T$.
Posons $\nu = \frac{\Delta t}{\tau(h)}$ avec $\nu \leq 1$ du fait de l'hypothèse (4.15). On note également $q = n - 1 - p$ pour simplifier. Alors on peut écrire

$$(I_h + \Delta t A_h)^q r_h^p = ((1-\nu)I_h + \nu T_h)^q (T_h - I_h) s_h^p = \underbrace{\left(\sum_{j=0}^{q} \begin{pmatrix} q \\ j \end{pmatrix} (1-\nu)^{q-j} \nu^j T_h^j \right) (T_h - I_h)}_{=A_q} s_h^p.$$

On pose $a_j^q = \begin{pmatrix} q \\ j \end{pmatrix} (1-\nu)^{q-j}\nu^j$ ainsi que $a_j^q = 0$ pour $j < 0$ ou $j > q+1$. Alors $A_q = \sum_j (a_{j-1}^q - a_j^q) T_h^j$. Or

$$a_{j-1}^q - a_j^q = [j - (q+1)\nu] \times \frac{q!}{(q-j)!(j-1)!} (1-\nu)^{q-j}\nu^{j-1}$$

La fonction entre crochets $j \mapsto j - (q+1)\nu$ est croissante, négative pour $j = 0$ et positive pour $n = q$. Donc i$j - (q+1)\nu \leq 0$ pour $j \leq j_* \equiv [(q+1)\nu]$ et $0 \leq j - (q+1)\nu$ pour $j_* < j$. On a

$$\|A_q\| \leq \sum_{j \leq j_*} \left(a_j^q - a_{j-1}^q \right) C + \sum_{j \geq j_*+1} \left(-a_j^q + a_{j-1}^q \right) C = 2a_{j_*}^q C$$

Or une estimation basique [4] montre que $a_j^q \leq \min\left(\frac{2}{\sqrt{\nu(1-\nu)q}}, 1 \right)$ pour tous j et q. On est en mesure d'estimer l'erreur (4.29) par

$$\|e_h^n\| \leq \Delta t \left(2 + \sum_{p=2}^{n-1} \frac{2}{\sqrt{\nu(1-\nu)p}} \right) 2CS \leq \Delta t \left(2 + \frac{2}{\sqrt{\nu(1-\nu)}} \int_1^n \frac{dx}{\sqrt{x}} \right) 2CS \leq \Delta t \left(2 + \frac{2}{\sqrt{\nu(1-\nu)}} \left(2\sqrt{n} - 2 \right) \right) 2CS.$$

On peut vérifier que $2 - \frac{4}{\sqrt{\nu(1-\nu)}} \leq 0$ pour toute valeur de $\nu \in]0,1[$. Donc

$$\|e_h^n\| \leq \Delta t \frac{8n}{\sqrt{\nu(1-\nu)}} CS.$$

Par ailleurs $\Delta t \sqrt{\frac{n}{\nu}} = \sqrt{n \Delta t \tau(h)} \leq \sqrt{T\tau(h)}$ ce qui termine la preuve quitte à redéfinir la constante C. □

4. Par exemple on a par un calcul en Fourier en développant $a_j^n = \frac{1}{2\pi} \int_0^{2\pi} \left((1-\nu) + \nu e^{\mathbf{i}\theta} \right)^n e^{-\mathbf{i}j\theta} d\theta$. Or $\left| (1-\nu) + \nu e^{\mathbf{i}\theta} \right|^2 = 1 - 4\nu(1-\nu)\sin^2\frac{\theta}{2}$. D'où par des majorations élémentaires

$$|a_j^n| \leq \frac{1}{\pi} \int_0^\pi \left(1 - 4\nu(1-\nu)\sin^2\frac{\theta}{2} \right)^{\frac{n}{2}} d\theta \leq \frac{1}{\pi} \int_0^\pi \left(1 - 4\nu(1-\nu)\frac{\theta^2}{\pi^2} \right)^{\frac{n}{2}} d\theta$$
$$\leq \frac{1}{\pi} \int_0^\pi e^{-2\nu(1-\nu)n\frac{\theta^2}{\pi^2}} d\theta \leq \frac{1}{\pi} \int_0^\infty e^{-2\nu(1-\nu)n\frac{\theta^2}{\pi^2}} d\theta = \frac{1}{\sqrt{2\nu(1-\nu)n}} \int_0^\infty e^{-u^2} du.$$

Reconnaissant l'intégrale de Gauss $\int_{-\infty}^\infty e^{-u^2} du = \sqrt{\pi}$, on trouve $a_j^n \leq \frac{\pi}{\sqrt{8}} \frac{1}{\sqrt{\nu(1-\nu)n}} \leq \frac{2}{\sqrt{\nu(1-\nu)n}}$. Par ailleurs $|a_j^n| \leq 1$.

On peut utiliser ce principe pour estimer la différence entre le schéma explicite et le schéma implicite. Partons de u_h^n solution du schéma explicite (4.25) et de v_h^n solution du schéma implicite

$$\begin{cases} \dfrac{v_h^{n+1}-v_h^n}{\Delta t} = A_h v_h^{n+1}, \ \ n \geq 0, \\ v_h^0 = \Pi_h u_0, \end{cases} \tag{4.30}$$

que l'on récrit comme un schéma explicite avec un reste

$$\begin{cases} \dfrac{v_h^{n+1}-v_h^n}{\Delta t} = A_h v_h^n + r_h^n, \ \ n \geq 0, \\ v_h^0 = \Pi_h u_0, \end{cases} \tag{4.31}$$

où

$$r_h^n = A_h \left(v_h^{n+1} - v_h^n \right) = \tau(h) A_h s_h^n \tag{4.32}$$

et

$$s_h^n = \nu A_h v_h^{n+1}, \qquad \nu = \frac{\Delta t}{\tau(h)}. \tag{4.33}$$

Lemme 20. *Supposons la condition CFL vérifiée sous la forme $\nu < 1$. Alors il existe une constante C telle que*

$$\|v_h^n - u_h^n\| \leq C \|A_h \Pi_h u_0\| \sqrt{T\tau(h)}, \qquad n\Delta t \leq T.$$

Cela établit que la différence entre le schéma explicite et le schéma implicite tend vers 0 avec h. Il faut cependant s'assurer d'une estimation naturelle annexe $\|A_h \Pi_h u_0\| \leq C'$ qu'il faut en pratique vérifier en utilisant la condition initiale et les propriétés du schéma numérique.

Démonstration. On a $v_h^n = (I_h + \Delta t A_h)^{-n} v_h^0$ d'où $A_h v_h^n = (I_h + \Delta t A_h)^{-n} A_h v_h^0$. Le schéma implicite étant stable, on a immédiatement $\left\|A_h v_h^n\right\| \leq C'' \left\|A_h v_h^0\right\| = C'' \|A_h \Pi_h u_0\|$ où C'' est la constante de stabilité. Pour $\nu \leq 1$, on obtient

$$\|s_h^n\| = \left\|\nu A_h v_h^{n+1}\right\| \leq C'' \|A_h \Pi_h u_0\| \text{ ce qui définit } S = C'' \|A_h \Pi_h u_0\|.$$

La preuve est terminée par application du principe de comparaison (4.28). □

L'extension au schéma semi-discret est immédiate.

Lemme 21. *Supposons la condition CFL vérifiée sous la forme $\nu < 1$. Alors il existe une constante C telle que la différence entre le schéma semi-discret et le schéma est explicite est majorée par*

$$\|v_h(n\Delta t) - u_h^n\| \leq C \|A_h \Pi_h u_0\| \sqrt{T\tau(h)}, \qquad n\Delta t \leq T.$$

Démonstration. Posons $v_h^n = v_h(n\Delta t)$ de sorte que (4.31) est satisfait avec $r_h^n = \frac{v_h(t_{n+1}) - v_h(t_n)}{\Delta t} - A_h v_h(t_n) = \tau(h) A_h s_h^n$ et

$$s_h^n = \nu \frac{1}{\Delta t} \int_{t_n}^{t_{n+1}} \frac{v_h(s) - v_h(t_n)}{\Delta t} ds.$$

Sous la condition que $A_h \Pi_h u_0$ est borné indépendamment de h, on obtient une estimation uniforme de la dérivée $\frac{d}{dt} v_h(s)$ grâce à la définition (4.23) et à la stabilité (4.24). D'où $\left\| \frac{v_h(s) - v_h(t_n)}{\Delta t} \right\| \leq C''' \|A_h \Pi_h u_0\|$ ce qui implique une majoration uniforme de s_h^n. Cela termine la preuve. □

4.1.8 Caractérisation spectrale de la stabilité

Pour une méthode de Différences Finies l'opérateur d'interpolation est naturellement défini par les valeurs aux points de grille. Cela garantit également la consistance. Aussi la difficulté est souvent de montrer la stabilité. Une approche efficace quand elle peut être menée consiste à passer par l'étude du spectre (des valeurs propres) de l'opérateur d'itération. C'est la stabilité **au sens de von Neumann**, la référence initiale trouvant dans [7]. Le schéma général s'écrit

$$u_h^{n+1} = J_{h,\Delta t} u_h^n. \tag{4.34}$$

L'opérateur d'itération $J_{h,\Delta t}$ est égal par exemple à $I_h + \Delta t A_h$ pour le schéma d'Euler explicite, à $(I_h - \Delta t A_h)^{-1}$ pour le d'Euler schéma implicite et à $\left(I_h - \frac{1}{2}\Delta t A_h\right)^{-1} \left(I_h + \frac{1}{2}\Delta t A_h\right)$ pour le schéma Cranck-Nicholson.
Pour simplifier on suppose que V_h **est de dimension finie**. Les valeurs propres de l'opérateur d'itération sont notées $\lambda_h^p \in \mathbb{C}$ avec

$$J_{h,\Delta t} v_h^p = \lambda_{h,\Delta t}^p v_h^p, \qquad v_h^p \neq 0, \quad 1 \leq p \leq \dim(V_h).$$

Le rayon spectral de J_h est

$$\rho(J_{h,\Delta t}) = \max_p |\lambda_{h,\Delta t}^p|.$$

Lemme 22 (Condition nécessaire de stabilité en dimension finie)**.** *Soit $J_{h,\Delta t}$ un opérateur d'itération stable. Alors il existe une constante $C > 0$ telle que $\rho(J_{h,\Delta t}) \leq 1 + C\Delta t$ pour tout $\Delta t \in (0, 1]$ et tout h.*
Si $J_{h,\Delta t}$ est uniformément stable, alors $\rho(J_{h,\Delta t}) \leq 1$.

Démonstration. On sait que $\rho(J_{h,\Delta t}) \leq \left\|J^n_{h,\Delta t}\right\|^{\frac{1}{n}}$. Partant d'un opérateur stable au sens de (4.16) on a $\rho(J_{h,\Delta t}) \leq (K')^{\frac{1}{n}} e^{L'\Delta t}$. D'où

$$\rho(J_{h,\Delta t}) \leq \lim_{\infty} (K')^{\frac{1}{n}} e^{L'\Delta t} = e^{L'\Delta t} \leq 1 + C\Delta t$$

pour une constante $C > 0$ bien choisie. Si l'opérateur est uniformément stable, $L' = 0$ ce qui clôt la preuve. □

La définition en **dimension finie** d'un **opérateur normal** est qu'il commute avec son opérateur adjoint. Aussi l'opérateur $J_{h,\Delta t}$ est normal ssi

$$J_{h,\Delta t}\ J^*_{h,\Delta t} = J^*_{h,\Delta t}\ J_{h,\Delta t}.$$

Cette notion n'a de sens qu'au sein d'un espace de Hilbert car l'opérateur adjoint est défini grâce au produit scalaire par

$$(J_{h,\Delta t} u_h, v_h) = (u_h, J^*_{h,\Delta t} v_h), \qquad u_h, v_h \in V_h.$$

Pour une matrice $M \in \mathbb{R}^{n\times n}$, on dit que M est normale ssi $MM^t = M^t M$.

Lemme 23 (Condition suffisante pour les opérateurs normaux en dimension finie)**.** *Soit l'opérateur d'itération $J_{h,\Delta t}$ pour le schéma (4.34) posé dans un espace de Hilbert. Supposons que $J_{h,\Delta t}$ est normal, et supposons que $\rho(J_{h,\Delta t}) \leq 1$ pour tout $h, \Delta t$. Alors le schéma est unitairement stable.*

Démonstration. Pour un opérateur normal en dimension finie on sait que $\|J_{h,\Delta t}\| = \rho(J_{h,\Delta t})$. Voir [11]. D'où le résultat. □

4.1.9 Schéma de splitting

Les méthodes de Splitting se rencontrent lors de l'implémentation effective de méthodes numériques. Elles ont été évoquées pour la méthode de différences finies en dimension deux lors de l'énoncé du principe 6.
Nous considèrerons le problème abstrait

$$\partial_t u = Au + Bu \tag{4.35}$$

dont le second membre est splitté (i.e. décomposé) en la somme de deux termes.

Exemple 3 (Splitting directionnel)**.** *Cela correspond aux situations où A est un opérateur aux dérivées partielles dans la direction x, et B est un opérateur aux dérivées partielles dans la direction y. Par exemple*

$$\partial_t u = \underbrace{a\partial_x u - \partial_{xx} u}_{=Au} + \underbrace{b\partial_y u - \partial_y (D(x,y)\partial_y u}_{=Bu} \tag{4.36}$$

où $D \geq 0$ est un coefficient de diffusion a priori borné et régulier.

Nous considérons tout d'abord le schéma explicite

$$\begin{cases} \dfrac{u_h^{n+\frac{1}{2}} - u_h^n}{\Delta t} = A_h u_h^n, \\ \dfrac{u_h^{n+1} - u_h^{n+\frac{1}{2}}}{\Delta t} = B_h u_h^{n+\frac{1}{2}}. \end{cases} \tag{4.37}$$

Les deux étapes sont explicites par simplicité, mais peuvent être remplacées par des discrétisations implicites. La forme explicite est

$$u_h^{n+1} = J_h u_h^n \qquad \text{avec } J_h = (I_h + \Delta t B_h)(I_h + \Delta t A_h).$$

Que peut-on dire en terme de stabilité ?

Lemme 24. *Supposons que les opérateurs d'itération sont stables au sens où il existe K'', L'' tels que*

$$\|(I_h + \Delta t A_h)^n\| \leq K'' e^{L'' n\Delta t} \text{ et } \|(I_h + \Delta t B_h)^n\| \leq K'' e^{L'' n\Delta t}.$$

Alors

— *soit A_h et B_h commutent, auquel cas l'opérateur d'itération J_h est stable*

$$\|J_h^n\| \leq K'' e^{2L'' n\Delta t}.$$

— *soit A_h et B_h sont unitairement stables ($K''=1$ et $L''=0$), auquel cas l'opérateur d'itération J_h est aussi unitairement stable.*

Démonstration. Evident. □

La situation vraiment intéressante correspond au cas unitairement stable car elle se rencontre souvent dans les applications qui sont dominées par le transport et la diffusion.

Exercice 11. *Proposer pour l'exemple (4.36) un splitting directionnel par schéma explicite unitairement stable.*

On peut déterminer une condition CFL de stabilité unitaire pour l'opérateur non splitté $I_h + \Delta t\,(A_h + B_h)$.

Proposition 10. *Supposons que $I_h + \Delta t A_h$ et $I_h + \Delta t B_h$ sont chacun unitairement stable sous une condition CFL égale respectivement à $\tau_A(h)$ et $\tau_B(h)$. Alors le schéma non splitté est unitairement stable sous la condition CFL*

$$\Delta t \leq \tau_{A+B}(h) = \frac{\tau_A(h)\tau_B(h)}{\tau_A(h)+\tau_B(h)}.$$

Démonstration. On a la décomposition

$$I_h + \Delta t\,(A_h + B_h) = \alpha\left(I_h + \frac{\Delta t}{\alpha}A_h\right) + (1-\alpha)\left(I_h + \frac{\Delta t}{1-\alpha}B_h\right) \quad 0<\alpha<1.$$

Si $\frac{\Delta t}{\alpha} \leq \tau_A(h)$ et $\frac{\Delta t}{\alpha} \leq \tau_B(h)$, alors $\|I_h + \Delta t\,(A_h + B_h)\| \leq 1$. Cela fait apparaitre une condition de stabilité

$$\Delta t \leq \min\left(\alpha\tau_A(h), (1-\alpha)\tau_B(h)\right)$$

dans laquelle α est une valeur arbitraire que l'on peut choisir pour maximiser le terme à droite de l'inégalité, ce qui permettra de prendre un grand pas de temps. La valeur optimale correspond à $\alpha\tau_A(h) = (1-\alpha)\tau_B(h)$ dont la solution est $\alpha = \frac{\tau_B(h)}{\tau_A(h)+\tau_B(h)}$. On trouve $\tau_{A+B}(h) = \alpha\tau_A(h) = \frac{\tau_A(h)\tau_B(h)}{\tau_A(h)+\tau_B(h)}$ ce qui termine la preuve. □

Proposition 11. *Supposons qu'il existe un opérateur d'interpolation commun Π_h et un espace dense commun X tels que les opérateurs A_h et B_h sont tous deux consistants (avec A et B respectivement). Alors $A_h + B_h$ est consistant avec $A+B$.*

Démonstration. Considérons pour simplifier le critère de consistance stationnaire (4.7). On a pour $u \in X$

$$\|(A_h + B_h)\Pi_h u - (A+B)u\| \leq \|A_h\Pi_h u - Au\| + \|B_h\Pi_h u - Bu\|$$

grâce à l'inégalité triangulaire. D'où le résultat. □

4.2 Applications

On illustre l'utilisation des différents concepts de consistance et stabilité à partir de quelques exemples.

4.2.1 Schéma décentré en dimension un

Soit le schéma numérique décentré (2.3) pour l'advection en dimension $d=1$, la vitesse d'advection étant positive $a>0$. La condition initiale est $x \mapsto u_0(x)$. Soit $h = \Delta x > 0$ le pas constant du maillage en espace.

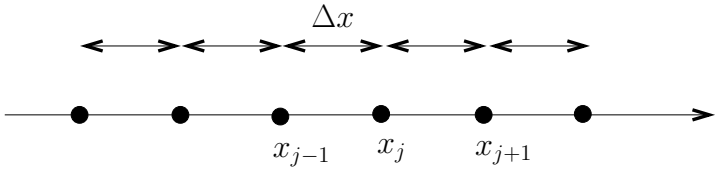


FIGURE 4.1 – Maillage Différences Finies à pas constant.

Soit $u_h^n = \left(u_j^n\right)_{j\in\mathbb{Z}} \in \mathbb{R}^{\mathbb{Z}}$ la solution numérique au temps $t_n = n\Delta t$. Le schéma (2.3) se récrit

$$u_h^{n+1} = \left(I_h + \Delta t A_h\right) u_h^n,$$

où I_h est l'identité de $\mathbb{R}^{\mathbb{Z}}$ et $A_h : \mathbb{R}^{\mathbb{Z}} \to \mathbb{R}^{\mathbb{Z}}$ est l'opérateur défini par

$$A_h u = (w_j) \text{ avec } w_j = -a\frac{u_j^n - u_{j-1}^n}{\Delta x}.$$

Montrer la convergence consiste *in fine* à comparer u_h^n et $v_h^n = \Pi_h u(t_n)$, mais aussi à choisir l'espace fonctionnel et la norme pour lesquels la convergence va être étudiée. L'approche la plus simple, quand elle est possible, consiste à mener cette étude dans un espace de fonctions bornée. Aussi nous prendrons ici

$$V = L^\infty(\mathbb{R}) \text{ et } V_h = \left\{ v_h = (v_j)_{j\in\mathbb{Z}}, \ \sup_j |v_j| < \infty \right\} = l_\infty.$$

On écrira indistinctement $\|v_h\| = \|v_h\|_\infty = \|v_h\|_{L^\infty(\mathbb{R})}$. L'opérateur d'interpolation $\Pi_h : C^\infty(\mathbb{R}) \to V_h$ est

$$\Pi_h(u) = \left(u(x_j)\right)_{j\in\mathbb{Z}}, \qquad x_j = j\Delta x.$$

On sait grâce à (2.4) que le schéma est stable sous CFL avec

$$\|I_h + \Delta t A_h\| \leq 1 \text{ pour } \Delta t \leq \tau(h) = \frac{h}{a} = \frac{\Delta x}{a}.$$

On note $v_h^n = \Pi_h u(t_n)$.
Soit l'erreur numérique $e_h^n = v_h^n - u_h^n$ qui est solution du processus itératif

$$\frac{e_h^{n+1} - e_h^n}{\Delta t} = A_h e_h^n + r_h^n \text{ avec la donnée initiale } e_h^0 = 0.$$

Le terme source est l'erreur de troncature $r_h^n = \left(r_j^n\right)$ avec

$$r_j^n = \frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_j^n - v_{j-1}^n}{\Delta x}.$$

Pour continuer l'analyse nous supposons ici que la donnée initiale est suffisamment régulière, $u_0 \in W^{2,\infty}(\mathbb{R})$. On a par un développement de Taylor

$$v_j^{n+1} = v_j^n + \Delta t \partial_t u\left(t_n, x_j\right) + \left(\Delta t^2 \|\partial_t^2 u\|_\infty\right) \alpha_j^n, \quad \left|\alpha_j^n\right| \leq \frac{1}{2}, \tag{4.38}$$

et

$$v_{j-1}^n = v_j^n - \Delta x \partial_x u\left(t_n, x_j\right) + \left(\Delta x^2 \|\partial_x^2 u\|_\infty\right) \beta_j^n, \quad \left|\beta_j^n\right| \leq \frac{1}{2}. \tag{4.39}$$

On obtient

$$\begin{aligned} r_j^n &= \partial_t u\left(n\Delta t, j\Delta x\right) + \left(\Delta t \|\partial_t u\|_\infty\right) \alpha_j^n + a\partial_x u\left(n\Delta t, j\Delta x\right) + a\left(\Delta x \|\partial_x^2 u\|_\infty\right) \beta_j^n \\ &= \left(\Delta t \|\partial_t u\|_\infty\right) \alpha_j^n + a\left(\Delta x \|\partial_x^2 u\|_\infty\right) \beta_j^n. \end{aligned}$$

Notons que $\|\partial_t^2 u\|_\infty = a^2 \|\partial_x^2 u_0\|_\infty$ et $\|\partial_x^2 u\|_\infty = \|\partial_x^2 u_0\|_\infty$. D'où

$$\left|r_j^n\right| \leq a\left(\Delta x \alpha_j^n + a\Delta t \beta_j^n\right) \|\partial_x^2 u_0\|_\infty.$$

Or la condition CFL implique que le pas de temps est borné par le pas d'espace sous la forme $\Delta t \leq \frac{\Delta x}{a}$. Cela implique que

$$\|r_h^n\|_\infty \leq a\Delta x \|\partial_x^2 u_0\|_\infty. \tag{4.40}$$

On dira que le schéma est consistant à l'ordre 1 en $O(\Delta x)$ dans L^∞.
On obtient le résultat de convergence.

Lemme 25. *Supposons que $u_0 \in W^{2,\infty}(\mathbb{R})$. Supposons la condition CFL satisfaite. Soit $T > 0$ donné. Alors pour tout n tel que $t_n = n\Delta t \leq T$, on a l'estimation d'erreur*

$$\|e_h^n\|_\infty \leq \|\partial_x^2 u_0\|_\infty (aT\Delta x). \tag{4.41}$$

Le schéma converge à l'ordre un en espace (et en temps).

Démonstration. La preuve est ne fait que reprendre la démonstration du théorème de Lax. On a $e_h^{n+1} = (I_h + \Delta t A_h)\, e_h^n + \Delta t r_h^n$. Donc $\|e_h^{n+1}\|_\infty \leq \|\,(I_h + \Delta t A_h)\, e_h^n\|_\infty + \Delta t \|r_h^n\|_\infty$. Or la stabilité fait que $\|I_h + \Delta t A_h\|_\infty \leq 1$. Donc $\|e_h^{n+1}\|_\infty \leq \|e_h^n\|_\infty + \Delta t \|r_h^n\|_\infty$. Comme $e^0 = 0$, on obtient finalement que $\|e^n\|_\infty \leq \Delta t \sum_{p=0}^{n-1} \|r^p\|_\infty$. Le résultat est démontré grâce à (4.40). □

4.2.2 Donnée moins régulière et ordre de convergence fractionnaire

Une question intéressante est de déterminer un ordre de convergence pour la solution numérique du schéma upwind (2.3) avec une donnée moins dérivable, par exemple $u_0 \in W^{1,\infty}(\mathbb{R})$. Nos verrons que le prix à payer sera que l'ordre de convergence est sous-linéaire.

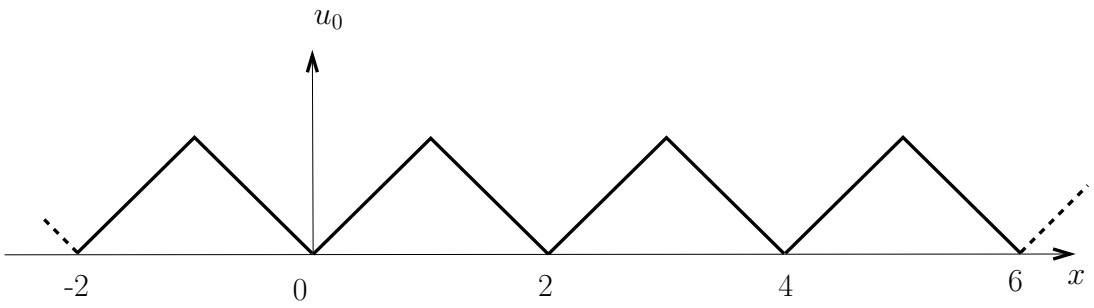


FIGURE 4.2 – Exemple d'une donnée initiale $u_0 \in W^{1,\infty}(\mathbb{R})$ pour laquelle le résultat de convergence d'ordre fractionnaire s'applique : $u_0(x) = \min_{k\in\mathbb{Z}} |x - 2k|$. Cette fonction est par ailleurs 2-périodique.

Définition 14 (Régularisation)**.** *Soit $\varphi \in W_0^{1,\infty}(\mathbb{R})$ une fonction positive ou nulle, de dérivée bornée, à support compact*[5] *et telle que*

$$\int_{\mathbb{R}} \varphi(z) dz = 1.$$

Pour une fonction donnée $w \in W^{1,\infty}(\mathbb{R})$, nous définissons la fonction régularisée par convolution

$$w_\varepsilon(x) = \frac{1}{\varepsilon} \int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) w(y) dy. \tag{4.43}$$

Lemme 26. *On a les inégalités*

$$\|w_\varepsilon\|_\infty \leq \|w\|_\infty, \quad \|w_\varepsilon'\|_\infty \leq \|w'\|_\infty, \quad \|w_\varepsilon''\|_\infty \leq \frac{\int |\varphi'(z)| dz}{\varepsilon} \|w'\|_\infty, \tag{4.44}$$

et

$$\|w_\varepsilon - w\|_\infty \leq \varepsilon \int \varphi(z)|z| dz \|\partial_x w\|_\infty. \tag{4.45}$$

5. On peut prendre par exemple

$$\varphi(z) = 1 - |z| \text{ pour } |z| \leq 1, \text{ et } \varphi(z) = 0 \text{ pour } |z| \geq 1. \tag{4.42}$$

Démonstration. Cela est standard [6]. A partir de la définition de w^ε on a

$$|w^\varepsilon(x)| \leq \left(\frac{1}{\varepsilon}\int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) dy\right) \|w\|_\infty.$$

Comme $\frac{1}{\varepsilon}\int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) dy = \int_{\mathbb{R}} \varphi(z)dz = 1$, cela montre immédiatement que $\|w^\varepsilon\|_\infty \leq \|w\|_\infty$.
On a aussi l'inégalité

$$w'_\varepsilon(x) = -\frac{1}{\varepsilon^2}\int_{\mathbb{R}} \varphi'\left(\frac{x-y}{\varepsilon}\right) w(y)dy = \frac{1}{\varepsilon}\int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) w'(y)dy$$

qui montre que $\|w'_\varepsilon\|_\infty \leq \|w'\|_\infty$.
Considérons ensuite

$$w''_\varepsilon(x) = -\frac{1}{\varepsilon^2}\int_{\mathbb{R}} \varphi'\left(\frac{x-y}{\varepsilon}\right) w'(y)dy$$

qui implique que

$$|w''_\varepsilon(x)| \leq \frac{1}{\varepsilon^2}\int_{\mathbb{R}} \left|\varphi'\left(\frac{x-y}{\varepsilon}\right)\right| dy \|w'\|_\infty = \frac{\int |\varphi'(z)|dz}{\varepsilon}\|w'\|_\infty.$$

Il reste à montrer (4.45). Or on a par construction

$$w_\varepsilon(x) - w(x) = \frac{1}{\varepsilon}\int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right)(w(y) - w(x))\, dy$$

d'où l'on tire

$$|w_\varepsilon(x) - w(x)| \leq \left(\frac{1}{\varepsilon}\int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right)|x-y|dy\right)\|w'\|_\infty = \left(\varepsilon\int \varphi(z)|z|dz\right)\|w'\|_\infty.$$

Cela termine la preuve. □

On peut alors montrer le résultat suivant pour le schéma (2.3).

Lemme 27 (Convergence à l'ordre $\frac{1}{2}$)**.** *Soit une donnée initiale $u_0 \in W^{1,\infty}(\mathbb{R})$. Supposons la condition CFL satisfaite. Soit $T > 0$ un temps final donné. Alors pour tout $t_n = n\Delta t \leq T$, on a l'estimation*

$$\|e_h^n\|_\infty \leq \frac{4}{\sqrt{3}}\|\partial_x u_0\|_\infty \sqrt{aT\Delta x}.$$

Démonstration. On commence par régulariser la donnée initiale

$$u_{0,\varepsilon}(x) = \frac{1}{\varepsilon}\int_{\mathbb{R}} \varphi\left(\frac{x-y}{\varepsilon}\right) u_0(y)dy.$$

La solution numérique découlant de cette donnée initiale est notée $u^n_{\varepsilon,h} = \left(u^n_{\varepsilon,j}\right)_{j\in\mathbb{Z}}$ avec

$$\begin{cases} \dfrac{u^{n+1}_{\varepsilon,h} - u^n_{\varepsilon,h}}{\Delta t} = A_h u^n_{\varepsilon,h}, \\ u^n_{\varepsilon,j} = u_{0,\varepsilon}(j\Delta x). \end{cases}$$

On a l'inégalité triangulaire

$$\|e_h^n\|_\infty = \|v_h^n - u_h^n\|_\infty \leq \|v_h^n - v^n_{\varepsilon,h}\|_\infty + \|v^n_{\varepsilon,h} - u^n_{\varepsilon,h}\|_\infty + \|u^n_{\varepsilon,h} - u_h^n\|_\infty, \tag{4.46}$$

où $v_h^n = (u(n\Delta t, j\Delta x))_{j\in\mathbb{Z}}$ et $v^n_{\varepsilon,h} = (u_\varepsilon(n\Delta t, j\Delta x))_{j\in\mathbb{Z}}$.
La stabilité du schéma montre que le troisième terme est borné par $\|u^n_{\varepsilon,h} - u^n_h\|_\infty \leq \|u^0_{\varepsilon,h} - u^0_h\|_\infty \leq \|u_{0,\varepsilon} - u_0\|_\infty$.
Comme la régularisation commute avec l'advection, on a pour le deuxième terme $\|v^n_h - v^n_{\varepsilon,h}\|_\infty \leq \|u_{0,\varepsilon} - u_0\|_\infty$.
Il reste à estimer le deuxième terme. Grâce (4.44) on obtient $\|u_{0,\varepsilon} - u_0\|_\infty \leq \varepsilon\int \varphi(z)|z|dz\|u'_0\|_\infty$. La dérivée seconde de la solution régularisée peut se contrôler grâce à (4.44). Aussi, utilisant (4.41) on trouve

$$\|v^n_{\varepsilon,h} - u^n_{\varepsilon,h}\|_\infty \leq \frac{\int |\varphi'(z)|dz}{\varepsilon}\|u'_0\|_\infty(aT\Delta x).$$

Après insertion dans (4.46) on obtient

$$\|e_h^n\|_\infty \leq \left(2\varepsilon\int \varphi(z)|z|dz + \frac{\int |\varphi'(z)|dz}{\varepsilon}aT\Delta x\right)\|u'_0\|_\infty.$$

Il reste à choisir la valeur optimale de ε qui est celle qui permet de minimiser le résultat : on prend $\varepsilon = \left(\frac{aT\Delta x\int |\varphi'(z)|dz}{2\int \varphi(z)|z|dz}\right)^{\frac{1}{2}}$.
Finalement

$$\|e_h^n\|_\infty \leq 2\left(2aT\Delta x\int |\varphi'(z)|dz \times \int \varphi(z)|z|dz\right)^{\frac{1}{2}}\|u'_0\|_\infty.$$

Pour le noyau (4.42) on a $\int |\varphi'(z)|dz \times \int \varphi(z)|z|dz = \frac{2}{3}$. Le reste de la preuve est évident. □

4.2.3 Maillage non uniforme

Enfin nous considérons l'équation d'advection $\partial_t u + a\partial_x u = 0$ discrétisée en dimension $d = 1$ avec le schéma upwind (2.3) sur un maillage non uniforme par le schéma (2.16). Cet exemple permet d'illustrer une difficulté spécifique de l'analyse numérique des schémas aux Différences Finies et aux Volumes Finis sur maillage non uniforme. La difficulté sera nettement plus conséquente en dimension supérieure, voir chapitre 5.

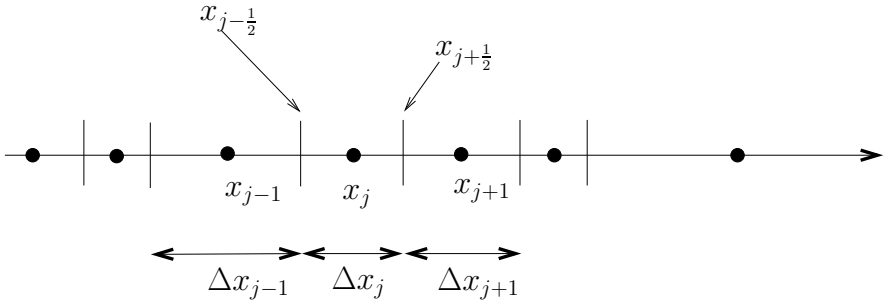


FIGURE 4.3 – Maillage non uniforme en 1D. Ici $\Delta x_{j-1} \neq \Delta x_j \neq \Delta x_{j+1}$. Les centres des mailles sont d'indice entier. Les bords de mailles sont d'indices demi-entier.

On commence par définir la finesse du maillage

$$h = \sup_j \Delta x_j$$

où $\Delta x_j = x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}$ est la longueur de la maille d'indice j. Il s'agit ensuite de définir l'opérateur de projection sur la maillage ce qui nécessite de définir préalablement les centres de mailles par

$$x_j = \frac{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}}{2}, \tag{4.47}$$

d'où une première définition naturelle de l'opérateur d'interpolation/projection $\Pi_h^1 : W^{2,\infty}(\mathbb{R}) \to V_h = l_\infty$ par

$$\Pi_h^1(v) = \left(v(x_j)\right)_{j\in\mathbb{Z}}.$$

Une deuxième définition possible de l'opérateur d'interpolation/projection, elle aussi naturelle, est fournie par les valeurs moyennes

$$\Pi_h^2(v) = \left(\frac{1}{\Delta x_j}\int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} v(x)dx\right)_{j\in\mathbb{Z}}.$$

Proposition 12. *L'erreur de consistance associée à Π_h^1 ou Π_h^2 ne tend pas vers zéro pour un maillage non uniforme.*

Démonstration. Commençons par évaluer l'erreur de consistance pour Π_h^1 à partir de l'un des deux critères (4.7) ou (4.11) au choix. Pour (4.11) on a $r_h^n = \left(r_j^n\right)_{j\in\mathbb{Z}}$ avec

$$r_j^n = \frac{v_j^{n+1} - v_j^n}{\Delta t} + a\frac{v_j^n - v_{j-1}^n}{\Delta x_j} - \partial_t u(t_n, x_j) - a\partial_x u(t_n, x_j).$$

Reprenant (4.38-4.39) pour une fonction dont les dérivées secondes sont bornées, on a

$$r_j^n = \left(\frac{x_j - x_{j-1}}{\Delta x_j} - 1\right) a\partial_x u(t_n, x_j) + O(\Delta t) + O(\Delta x). \tag{4.48}$$

Le terme principal disparait pour $\frac{x_j - x_{j-1}}{\Delta x_j} = 1$ pour tout j, ce qui revient *in fine* à considérer que le maillage est uniforme : $\Delta x_j = \Delta x_k = \Delta x$ pour tout j, k.
Cependant pour un maillage non uniforme on a uniquement $r_h^n = O(1)$ ce qui fait que cette erreur de consistance de tend pas vers zéro.
Pour $v \in W^{2,\infty}(\mathbb{R})$, on a $\left\|\Pi_h^1 v - \Pi_h^2 v\right\|_\infty \leq \Delta x^2 \|v''\|_{L^\infty(\mathbb{R})}$. Cette différence étant d'ordre deux en h, la résultat est le même en partant de Π_h^2. La preuve est terminée. □

Cette analyse montre d'une part que l'analyse numérique des schémas sur grille non uniforme est moins évident que pour des grilles uniformes, et d'autre part que le critère de consistance (4.7) ou (4.11) dépend bien du choix de l'opérateur d'interpolation Π_h. Cependant on a bien la convergence à partir d'un autre opérateur d'interpolation adapté au schéma. Soit $\Pi_h^3 : W^{2,\infty}(\mathbb{R}) \to V_h = l_\infty$ défini par

$$\Pi_h^3(v) = \left(v(x_{j+\frac{1}{2}})\right)_{j\in\mathbb{Z}}. \tag{4.49}$$

On observe que le point d'interpolation est décentré sur le bord droit des mailles.

Proposition 13. *L'erreur de consistance associée à Π_h^3 tend vers zéro à l'ordre un pour une donnée suffisamment régulière et pour tout maillage.*

Démonstration. On part de l'erreur de consistance définie par (4.14). Pour $r_h^n = \frac{\Pi_h^3 u(t_{n+1}) - \Pi_h^3 u(t_n)}{\Delta t} - A_h u_h^n - \Pi_h^3 \left(\partial_t u - Au(t_n)\right)$ on a

$$r_j^n = \frac{u(t_{n+1}, x_{j+\frac{1}{2}}) - u(t_n, x_{j+\frac{1}{2}})}{\Delta t} + a\frac{u(t_n, x_{j+\frac{1}{2}}) - u(t_n, x_{j-\frac{1}{2}})}{\Delta x_j} - \partial_t u(t_n, x_{j+\frac{1}{2}}) - a\partial_x u(t_n, x_{j+\frac{1}{2}}).$$

Reprenant (4.38-4.39) pour une fonction dont les dérivées secondes sont bornées, on a

$$r_j^n = \left(\frac{x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}}}{\Delta x_j} - 1\right) a\partial_x u(t_n, x_j) + O(\Delta t) + O(\Delta x_j) = O(\Delta t) + O(\Delta x)$$

car $x_{j+\frac{1}{2}} - x_{j-\frac{1}{2}} = \Delta x_j$. Cela termine la preuve. □

Lemme 28. *Soit le schéma (2.16) avec l'initialisation $u_h^0 = \Pi_h^3 u_0$ pour une donnée initiale $u_0 \in W^{2,\infty}(\mathbb{R})$. Supposons la condition CFL satisfaite. Alors*

$$\|\Pi_h^3 u(t_n) - u_h^n\|_\infty \leq aT\|u_0''\|_\infty h, \qquad n\Delta t \leq T. \tag{4.50}$$

Pour une donnée initiale moins régulière $u_0 \in W^{1,\infty}(\mathbb{R})$, on a l'ordre de convergence fractionnaire moitié

$$\|\Pi_h^3 u(t_n) - u_h^n\|_\infty \leq \frac{4}{\sqrt{3}}||u_0'||_\infty \times \sqrt{aTh}, \qquad n\Delta t \leq T. \tag{4.51}$$

Démonstration. Il s'agit de la même preuve que pour le lemme 25, à partir de l'erreur d'interpolation associée à Π_h^3.

□

Chapitre 5

Analyse numérique des Volumes Finis

Les méthodes de Volumes Finis sur maillage non structurés sont à la base des codes de CFD (Computational Fluid Dynamics) et de résolution de systèmes hyperboliques non linéaires pour lesquels l'objectif est le calcul précis de solutions très peu régulières voire même discontinues (les discontinuités et les ondes de chocs). Le calcul de transport et diffusion en milieux poreux sont eux aussi très demandeurs en méthodes de Volumes Finis.
L'analyse numérique des méthodes de Volumes Finis met en évidence deux propriétés fortes qui sont d'une part la stabilité et le principe du maximum et d'autre part une structure de données simple. Cela explique l'intérêt fort de ces méthodes en calcul scientifique et ingénierie numérique.
Cependant la convergence avec le pas du maillage apparaît nettement plus délicate à analyser. On verra qu'il est cependant possible de montrer par exemple la convergence à l'ordre $\frac{1}{2}$ pour le transport de données BV ce qui est représentatif de la convergence des méthodes de Volumes Finis pour des données peu régulières. La difficulté principale est qu'il faudra obtenir ces résultats sans passer par la consistance au sens des Différences Finies, c'est à dire sans utiliser la méthode de régularisation de la preuve du lemme 27.

5.1 Equation d'advection

Le problème modèle est

$$\left\{ \begin{array}{ll} \partial_t u + \mathbf{a} \cdot \nabla u = 0, & \mathbf{x} \in \Omega, \quad t > 0, \\ u(0, \mathbf{x}) = u_0(\mathbf{x}), & \mathbf{x} \in \Omega, \end{array} \right. \tag{5.1}$$

dans un domaine Ω que l'on prend **sans bord** pour simplifier les notations. Par exemple on pourra considérer soit que $\Omega = \mathbb{R}^2$ soit que $\Omega = \mathcal{T} = [0,1] \times [0,1]$ est le tore (carré académique périodique) : on peut identifier $x + 1 = x$ et $y + 1 = y$.
Nous considérons ici un champ de vitesse éventuellement non constant, mais régulier $\mathbf{a} \in C^1(\Omega)$ et à divergence nulle

$$\nabla \cdot \mathbf{a} = 0.$$

On utilise les notations générales de la section 2.2.2. On pose

$$a_{jk} = \frac{1}{l_{jk}} \int_{\Sigma_{jk}} \mathbf{a}(\mathbf{x}) \cdot \mathbf{n}_{jk}(\mathbf{x}) d\sigma$$

qui est la valeur moyenne de $\mathbf{a}$ après produit scalaire contre la normale extérieure. Pour simplifier un peu les notations, on définit

$$I^+(j) = \{k \text{ tels que } a_{jk} > 0\} \text{ et } I^-(j) = \{k \text{ tels que } a_{jk} < 0\}$$

et on utilisera la convention de notation

$$k^\pm \text{ au lieu et place de de } k \in I^\pm(j).$$

On utilise aussi la notation

$$m_{jk} = l_{jk}\,|a_{jk}| \quad \forall j,k..$$

Un schéma de Volumes Finis qui généralise (2.35) s'écrit

$$s_j \frac{u_j^{n+1} - u_j^n}{\Delta t} + \sum_{k^+} m_{jk} u_j^n - \sum_{k^-} m_{jk} u_k^n = 0. \tag{5.2}$$

En dimension $d = 2$ on peut évaluer la valeur numérique des a_{jk} sans difficulté.

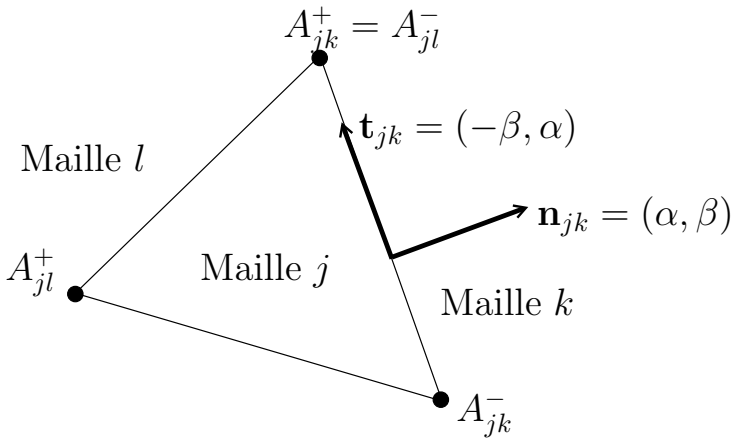


FIGURE 5.1 – Orientation des interfaces

En effet on peut supposer que $\mathbf{a}$ est le rotationnel d'un potentiel scalaire donné $q \in C^2(\Omega)$

$$\mathbf{a} = \nabla \wedge q = (-\partial_{x_2} q, \partial_{x_1} q)\,.$$

Par construction $\nabla \cdot \mathbf{a} = \partial_{x_1}(-\partial_{x_2} q) + \partial_{x_2}(\partial_{x_1} q) = 0$. Posons $\mathbf{n} = (\alpha, \beta)$ et $\mathbf{t} = (-\beta, \alpha)$: alors

$$a_{jk} = \frac{1}{l_{jk}} \int_{\Sigma_{jk}} \nabla \wedge q \cdot \mathbf{n}_{jk} d\sigma = \frac{1}{l_{jk}} \int_{\Sigma_{jk}} (-\partial_{x_2} q\ \alpha + \partial_{x_1} q\ \beta)\, d\sigma = -\frac{1}{l_{jk}} \int_{\Sigma_{jk}} \nabla q \cdot \mathbf{t}\ d\sigma = -\frac{1}{l_{jk}} \int_{\Sigma_{jk}} \frac{\partial q}{\partial \mathbf{t}}\ d\sigma,$$

ou encore

$$a_{jk} = \frac{q\left(A_{jk^-}\right) - q\left(A_{jk^+}\right)}{l_{jk}}.$$

Par convention (A_{jk}^-, A_{jk}^+) sont orientés dans le sens des aiguilles d'une montre sur le bord Σ_{jk}. On a par ailleurs que $A_{jk}^- = A_{kj}^+$.

Lemme 29. *On a l'égalité* $\sum_k l_{jk} a_{jk} = 0$.

Démonstration. En effet $\sum_k l_{jk} a_{jk} = \sum_k l_{jk} \left(\frac{q\left(A_{jk^-}\right) - q\left(A_{jk^+}\right)}{l_{jk}} \right) = \sum_k \left(A_{jk^-} - A_{jk^+}\right) = 0$ pour tout contour fermé. On peut aussi utiliser la condition de divergence nulle $\sum_k l_{jk} a_{jk} = \int_{\partial\Omega_j} \mathbf{a} \cdot \mathbf{n}_j\ d\sigma = \int_{\Omega_j} \nabla \cdot \mathbf{a}\ dx = 0$. □

Lemme 30. *On a* $\sum_{k^+} m_{jk} = \sum_{k^-} m_{jk}$.

Immédiat à partir de la définition de m_{jk} du lemme 29.

5.1.1 Analyse de la condition de stabilité

Le schéma (5.2) peut se mettre sous la forme explicite

$$u_j^{n+1} = \left(1 - \frac{\Delta t}{s_j}\sum_{k^-} m_{jk}\right) u_j^n + \frac{\Delta t}{s_j}\sum_{k^-} m_{jk} u_k^n. \tag{5.3}$$

Lemme 31. *Supposons que le pas de temps satisfasse à l'inégalité de stabilité (condition CFL)*

$$\frac{\Delta t}{s_j}\sum_{k^-} m_{jk} \leq 1, \qquad \forall j. \tag{5.4}$$

Alors la solution numérique vérifie le principe du maximum

$$\inf_k (u_k^n) \leq u_j^{n+1} \leq \sup_k (u_k^n). \tag{5.5}$$

Démonstration. Soit $m = \inf_k u_k^n$ le minimum de la solution numérique au temps t_n. Nous allons commencer par montrer que $m \leq u_j^{n+1}$ pour toute maille j. On a

$$u_j^{n+1} - m = \left(1 - \frac{\Delta t}{s_j}\sum_{k^-} m_{jk}\right)(u_j^n - m) + \frac{\Delta t}{s_j}\sum_{k^-} m_{jk}(u_k^n - m).$$

Les coefficients $1 - \frac{\Delta t}{s_j}\sum_{k^-} m_{jk}$ et $\frac{\Delta t}{s_j}\sum_{k^-} m_{jk}$ sont positifs ou nuls et leur somme fait 1. Donc $u_j^{n+1} - m$ est une combinaison convexe, c'est à dire une moyenne, des u_k^n. Donc $m \leq u_j^{n+1}$ pour tout j.
Une inégalité similaire se démontre pour la borne supérieure $M = \sup_k u_k^n$. Cela termine la preuve. □

Soit plus généralement une fonction convexe $u \mapsto \varphi(u)$

$$\varphi(\theta u_1 + (1-\theta)u_2) \leq \theta\varphi(u_1) + (1-\theta)\varphi(u_2)$$

pour tous u_1 et u_2 et pour tout $\theta \in [0,1]$.

Lemme 32. *Supposons la condition CFL (5.4) satisfaite. On a l'inégalité*

$$\sum_j s_j \varphi\left(u_j^{n+1}\right) \leq \sum_j s_j \varphi\left(u_j^n\right). \tag{5.6}$$

Démonstration. Comme φ est convexe, on a plus l'inégalité $\varphi(\sum \theta_i u_i) \leq \sum \theta_i \varphi(u_i)$ sous les conditions $\theta_i \geq 0$ pour tout i et $\sum \theta_i = 1$. Donc

$$\varphi\left(u_j^{n+1}\right) \leq \left(1 - \frac{\Delta t}{s_j}\sum_{k^-} m_{jk}\right)\varphi\left(u_j^n\right) + \frac{\Delta t}{s_j}\sum_{k^-} m_{jk}\varphi(u_k^n).$$

Sommons sur tout le maillage

$$\sum_j \varphi\left(u_j^{n+1}\right) \leq \sum_j \varphi\left(u_j^n\right) - \sum_j\sum_{k^-} \frac{\Delta t\, m_{jk}}{s_j}\varphi\left(u_j^n\right) + \sum_j\sum_{k^-} \frac{\Delta t\, m_{jk}}{s_j}\varphi(u_k^n).$$

On a

$$\sum_j\sum_{k^-} \frac{\Delta t\, m_{jk}}{s_j}\varphi\left(u_j^n\right) = \sum_j \sum_{k,\ a_{jk}>0} \frac{\Delta t\, m_{jk}}{s_j}\varphi\left(u_j^n\right)$$

et

$$\sum_j\sum_{k^-} \frac{\Delta t\, m_{jk}}{s_j}\varphi(u_k^n) = \sum_j \sum_{k,\ a_{jk}<0} \frac{\Delta t\, m_{jk}}{s_j}\varphi(u_k^n).$$

Or on a l'égalité

$$\sum_j \sum_{k,\, a_{jk}>0} \frac{\Delta t\, m_{jk}}{s_j} \varphi\left(u_j^n\right) = \sum_j \sum_{k,\, a_{jk}<0} \frac{\Delta t\, m_{jk}}{s_j} \varphi\left(u_k^n\right).$$

Le reste de la preuve est évident. □

Remarque 6. *Dans le cas où le maillage est infini ce qui correspond par exemple à $\Omega = \mathbb{R}^2$, il faut cependant justifier la convergence et la permutation des sommes infinies dans la preuve de (5.6). C'est bien le cas si $\varphi(0) = 0$ et la solution numérique est à support compact. Cela couvre les cas particuliers étudiés ci-dessous. Pour un maillage fini, cette difficulté n'a pas lieu.*

Lemme 33. *Le schéma de Volumes Finis (5.2) est stable dans tous les L^p sous la même condition (5.4). Plus précisément*

$$\|u^{n+1}\|_{L^p(\Omega)}^p \leq \|u^n\|_{L^p(\Omega)}^p \quad 1 \leq p \leq \infty. \tag{5.7}$$

Démonstration. Tout d'abord on considère $1 \leq p < \infty$. La fonction $\varphi(u) = |u|^p$ étant convexe, on peut appliquer l'inégalité précédente. D'où le résultat. Le cas $p = \infty$ est une conséquence du principe du maximum (5.5). □

Cependant l'inégalité 32 permet de dériver aussi le principe du maximum, ce qui fournit une deuxième démonstration de la stabilité dans L^∞.

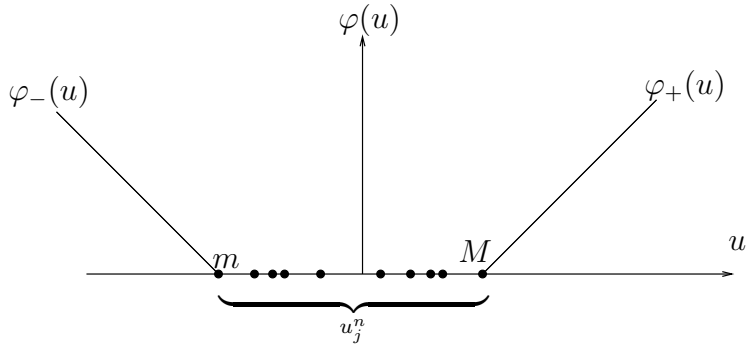


FIGURE 5.2 – φ_- et φ_+

On pose $m = \min_k u_k^n$ et $M = \max_k u_k^n$. Soit la fonction

$$\varphi_-(u) = \begin{cases} m - u & \text{pour } u \leq m, \\ 0 & \text{pour } m \leq u. \end{cases}$$

Cette fonction φ_- est continue, convexe et $\varphi_-(0) = 0$. L'inégalité (32) implique que

$$\sum_j s_j \varphi_-\left(u_j^{n+1}\right) \leq 0.$$

Or $\varphi_- \geq 0$. Donc $\varphi_-\left(u_j^{n+1}\right) = 0$ pour tout j, ce qui montre que $m \leq u_j^{n+1}$ pour tout j.
Pour montrer que $u_j^{n+1} \leq M$, nous considérons une deuxième fonction convexe

$$\varphi_+(u) = \begin{cases} 0 & \text{pour } u \leq M, \\ u - M & \text{pour } M \leq u. \end{cases}$$

Un raisonnement similaire montre que $u_j^{n+1} \leq M$ pour tout j.
A présent nous interprétons géométriquement la condition de CFL. Le membre de droite de

$$\Delta t \leq \frac{s_j}{\sum_{k^+} m_{jk}} \qquad \forall j, \tag{5.8}$$

dépend de la structure locale du maillage. Il importe de s'assurer que ce terme n'est pas excessivement petit, d'augmenter les temps de calcul dans des proportions excessives. Pour simplifier l'analyse on considère que

$$\mathbf{a} \in \mathbb{R}^2 \text{ est constant en espace.} \tag{5.9}$$

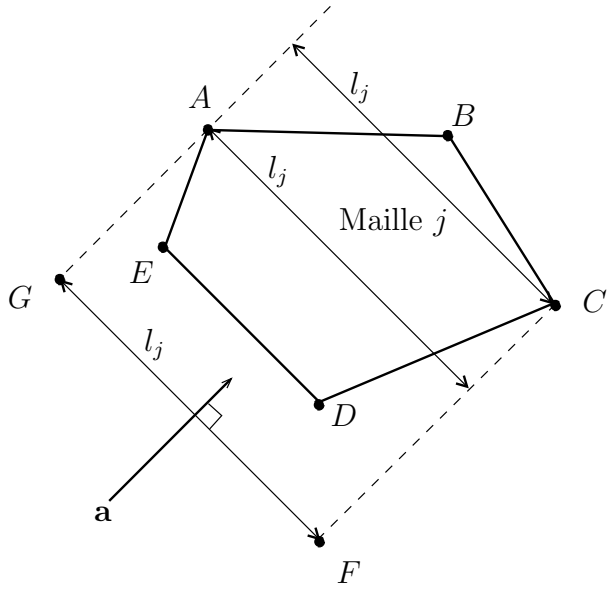


FIGURE 5.3 – Largeur apparente d'une maille

Nous définissons l_j la largeur apparente de la maille Ω_j comme la dimension de cette maille vue par un observateur à l'infini dans la direction $\mathbf{a}$.

Lemme 34. *Supposons les mailles convexes. Alors l'inégalité de stabilité se récrit*

$$\Delta t \leq \frac{s_j}{|\mathbf{a}| l_j} \tag{5.10}$$

où l_j est la longueur apparente comme sur la figure 5.3.

Démonstration. Nos montrons cette propriété sur l'exemple de la maille pentagonale Ω_j de sommets $ABCDE$ de la figure 5.3.
On construit une maille plus grande $\Omega_{j'}$ avec $\Omega_j \subset \Omega_{j'}$: ses sommets sont $ABCFG$, les segments AG et CF étant parallèles au vecteur $\mathbf{a}$. Comme Ω_j est convexe par hypothèse et que $\mathbf{a}$ est constant, les bords sortants $k \in I^+(j)$ (i.e. AB et BC sur la figure) forment une ligne brisée connexe. De la même manière les bords entrants $k \in I^-(j)$ (i.e. CD, DE et EA sur la figure) forment une ligne brisée connexe. Les bords sortants de $\Omega_{j'}$ sont les mêmes que ceux de Ω, ce que l'on peut noter par

$$I^+(j) = I^+(j').$$

Alors

$$\sum_{k \in I^+(j)} m_{jk} = \sum_{k \in I^+(j')} m_{jk} = \sum_{k \in I^-(j')} m_{jk} = |\mathbf{a}| l_j.$$

Or AG et CF sont parallèle à $\mathbf{a}$, donc ils ne contribuent pas. La preuve est terminée. □

Remarque 7. *En revanche $\sum_{k^+} m_{jk} = \sum_{k^-} m_{jk} > |\mathbf{a}| l_j$ est tout à fait possible pour une maille non convexe. Dans ce cas le pas de temps est plus restreint que pour (5.10).*

Soit à présent une maille Ω_j **convexe** : on définit r_j^- le plus grand rayon des cercles internes et r_j^+ le plus petit rayon des cercles externes. On a

$$\mathrm{diam}(\Omega_j) \leq 2r_j^+.$$

Pour un triangle r_j^- est le rayon du cercle inscrit, et r_j^+ est le rayon du cercle circonscrit.

Lemme 35. *Soit une maille convexe. Alors une condition suffisante pour obtenir (5.8) est que*

$$\Delta t \leq \frac{\pi(r_j^-)^2}{2|\mathbf{a}|r_j^+}. \tag{5.11}$$

Démonstration. Par définition $s_j \geq \pi(r_j^-)^2$ et $l_j \leq 2r_j^+$. Aussi (5.10) est une conséquence de (5.11). □

Définition 15. *On définit le facteur de qualité, aussi appelé rapport d'aspect, du maillage*

$$Q = \sup_j \left(\frac{r_j^+}{r_j^-} \right) \geq 1,$$

et la longueur caractéristique du maillage

$$h = \sup_j \left(diam(\Omega_j) \right).$$

La définition de h est une alternative possible à une définition similaire (2.28).
Avec ces notations, la condition sur le pas de temps (5.11) est vérifiée dès que

$$|\mathbf{a}| \left(\frac{Q^2}{\pi} \right) \frac{\Delta t}{h} \leq 1. \tag{5.12}$$

Pour un calcul sur ordinateur, on a toujours intérêt à utiliser le plus grand pas de temps possible. Le pas du maillage h est le plus souvent dicté par la précision souhaitée. En revanche Q est donné par la structure du maillage. De ce point de vue l'intérêt pratique dicte d'utiliser un maillage avec une constante Q la plus petite possible.

Définition 16. *Une suite de maillages indicés par n et de longueur caractéristique h_n avec $h_n \to 0$ pour $n \to \infty$ est dite régulière si*

$$1 \leq Q_n \leq C, \qquad \forall n.$$

Les preuves de convergence utilisent une telle hypothèse de régularité de maillage. Il faut noter que la situation est identique pour la théorie de convergence des méthodes d'éléments finis [8].

5.1.2 Approximation, erreur de projection initiale et inégalité de Poincaré-Wirtinger

On démontre quelques inégalités d'interpolation de base qui utiles pour l'analyse numérique générale des méthodes de Volumes Finis, et en particulier pour caractériser l'erreur d'approximation $\|u_0 - \Pi_h u_0\|_{L^p(\Omega)}$ de la donnée initiale à la toute première itération.
La première inégalité, lemme 36, est un résultat classique qui mesure l'erreur de projection en moyenne. La deuxième inégalité, lemme 38, est tout aussi classique. Elle mesure l'erreur entre la valeur moyenne dans les mailles par rapport à la valeur moyenne sur les segments aux interfaces des mailles.
Les mailles en dimension deux d'espace sont supposées convexes. La longueur caractéristique du maillage h est par définition plus grandes que tous les bords de mailles. Le maillage est pris régulier. Enfin le nombre de voisins est borné par une constante indépendante de h.

Lemme 36 (Inégalité de type Poincaré-Wirtinger)**.** *Soit Π_h l'opérateur de projection en moyenne. Alors il existe une constante $C > 0$ telle que*

$$\|u - \Pi_h u\|_{L^p(\Omega)} \leq Ch \|\nabla u\|_{L^p(\Omega)} \tag{5.13}$$

pour tout $u \in W^{1,p}(\Omega)$.

Démonstration. L'inégalité (5.13) ne fait que préciser la dépendance par rapport au maillage de la constante de l'inégalité de Poincaré-Wirtinger dans $L^p(\Omega)$.
Le cas $p=\infty$ est évident aussi on considère $1\leq p<\infty$. On a

$$\|u-\Pi_h u\|^p_{L^p(\Omega)}=\sum_j\int_{\mathbf{x}\in\Omega_j}\left|u(\mathbf{x})-\frac{1}{s_j}\int_{\mathbf{y}\in\Omega_j}u(\mathbf{y})dy\right|^p dx=\sum_j\frac{1}{s_j^p}\int_{\mathbf{x}\in\Omega_j}\left|\int_{\mathbf{y}\in\Omega_j}(u(\mathbf{x})-u(\mathbf{y}))dy\right|^p dx.$$

L'inégalité de Hölder pour $\frac{1}{p}+\frac{1}{q}=1$ implique $\left|\int_{\mathbf{y}\in\Omega_j}(u(\mathbf{x})-u(\mathbf{y}))dy\right|\leq\left(\int_{\mathbf{y}\in\Omega_j}|u(\mathbf{x})-u(\mathbf{y})|^p\,dy\right)^{\frac{1}{p}}s_j^{\frac{1}{q}}$. D'où

$$\|u-\Pi_h u\|^p_{L^p(\Omega)}\leq\sum_j\frac{1}{s_j}\int_{\mathbf{x}\in\Omega_j}\int_{\mathbf{y}\in\Omega_j}|u(\mathbf{x})-u(\mathbf{y})|^p dxdy. \tag{5.14}$$

Or $u(\mathbf{x})-u(\mathbf{y})=\int_0^1\nabla u\left(t\mathbf{x}+(1-t)\mathbf{y}\right)dt\cdot(\mathbf{x}-\mathbf{y})$ d'où l'on tire $|u(\mathbf{x})-u(\mathbf{y})|^p\leq h^p\int_0^1|\nabla u\left(t\mathbf{x}+(1-t)\mathbf{y}\right)|^p\,dt$. Il s'ensuit que

$$\int_{\mathbf{x}\in\Omega_j}\int_{\mathbf{y}\in\Omega_j}|u(\mathbf{x})-u(\mathbf{y})|^p dxdy\leq h^p\int_{\mathbf{x}\in\Omega_j}\int_{\mathbf{y}\in\Omega_j}\int_0^1|\nabla u\left(t\mathbf{x}+(1-t)\mathbf{y}\right)|^p\,dtdxdy$$

puis en utilisant un principe de symétrie

$$\int_{\mathbf{x}\in\Omega_j}\int_{\mathbf{y}\in\Omega_j}|u(\mathbf{x})-u(\mathbf{y})|^p dxdy\leq 2h^p\int_{\mathbf{y}\in\Omega_j}\int_{\frac{1}{2}}^1\left(\int_{\mathbf{x}\in\Omega_j}|\nabla u\left(t\mathbf{x}+(1-t)\mathbf{y}\right)|^p\,dx\right)dydt.$$

Le terme entre parenthèse s'évalue aisément grâce à un changement de variables. On pose $\mathbf{z}=t\mathbf{x}+(1-t)\mathbf{y}$, pour t et $\mathbf{y}$ donnés. On remarque que $\mathbf{z}\in\Omega_j$ et que $td\mathbf{x}=\mathbf{z}$ ce qui implique que $t^2dx=dz$. On remarque surtout que que la troncature en $\frac{1}{2}<t<1$ dans les intégrales a permis d'éviter une singularité en $\frac{1}{t^2}$ qui aurait été catastrophique. Cette idée semble remonter à la démonstration initiale de Poincaré lui-même.
On a alors

$$\int_{\mathbf{x}\in\Omega_j}|\nabla u\left(t\mathbf{x}+(1-t)\mathbf{y}\right)|^p\,dx\leq\frac{1}{t^2}\int_{\mathbf{z}\in\Omega_j}|\nabla u\left(\mathbf{z}\right)|^p\,dz\leq 4\int_{\mathbf{x}\in\Omega_j}|\nabla u\left(\mathbf{x}\right)|^p\,dx$$

qui est une majoration indépendant de $t\in[0,1]$ et $\mathbf{y}\in\Omega_j$. Cela implique que

$$\int_{\mathbf{x}\in\Omega_j}\int_{\mathbf{y}\in\Omega_j}|u(\mathbf{x})-u(\mathbf{y})|^p dxdy\leq 4s_jh^p\int_{\mathbf{x}\in\Omega_j}|\nabla u\left(\mathbf{x}\right)|^p.$$

Une insertion de cette inégalité dans (5.14) et une simplification donnent

$$\|u-\Pi_h u\|_{L^p(\Omega)}\leq 4^{\frac{1}{p}}h\|\nabla u\|_{L^p(\Omega)}.$$

La constante peut-être prise indépendant de p, soit $C=4$. La preuve est terminée. □

Soit $u\in W^{1,p}(\Omega_j)$, $p\in[1,\infty]$. On note u_j la valeur moyenne dans la maille Ω_j et u_{jk} la valeur moyenne sur le bord Σ_{jk}

$$u_j=\frac{1}{s_j}\int_{\Omega_j}u(x)dx,\qquad u_{jk}=\frac{1}{l_{jk}}\int_{\Sigma_{jk}}u(x)d\sigma,\quad\forall k.$$

Soit A_j une mesure de la différence dans une norme de type L^p

$$A_j=\left(h\sum_{k\in I(j)}l_{jk}\,|u_{jk}-u_j|^p\right)^{\frac{1}{p}}. \tag{5.15}$$

Lemme 37. *On a l'inégalité $A_j \leq Ch \|\nabla u\|_{L^p(\Omega)}$, où la constante C ne dépend pas de u ni des paramètres du maillage.*

Démonstration. On note $\Theta = \Omega_j$, puis enlève les indices j pour plus de lisibilité. Soit

$$v = u - \frac{1}{\Theta}\int_\Theta u(\mathbf{x})dx$$

dont la valeur moyenne est nulle dans Θ. En considérant que le bord de Θ est constitué d'un nombre fini de segments de droites de longueur l_k, on a (par exemple en utilisant la convexité de la fonction $v \mapsto |v|^p$)

$$\sum_k l_k |v_k|^p \leq \int_{\partial\Theta} |v(\mathbf{x})|^p d\sigma \tag{5.16}$$

où les v_k sont les valeurs moyennes de v sur chacun des segments.

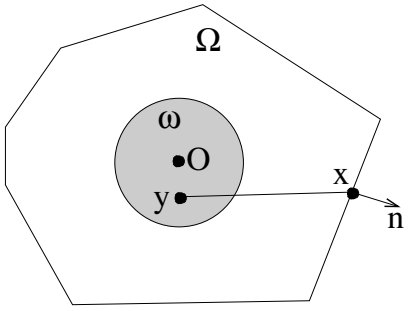


FIGURE 5.4 – Disque ω à l'intérieur d'une maille convexe Ω polygonale. On a bien $(\mathbf{x}-\mathbf{y}, \mathbf{n}(\mathbf{x})) \geq 0$ pour tout $\mathbf{x} \in \partial\Omega$ et tout $\mathbf{y} \in \omega$.

Un peu de géométrie : à une translation près, on peut toujours supposer que l'origine appartient à Ω qui est convexe par hypothèse, et que l'origine est centre d'un disque $\omega \subset \Omega$ de rayon r^-. Notons que pour des maillages réguliers, ce rayon minimum est borné inférieurement par $r^- \geq ch$, $c > 0$ indépendant de h. La maille Θ étant convexe, on a que

$$(\mathbf{x}-\mathbf{y}, \mathbf{n}(\mathbf{x})) \geq 0, \quad \forall \mathbf{x} \in \partial\Theta \text{ et } \forall \mathbf{y} \in \Theta.$$

Soit $\mathbf{y} = ch\mathbf{n}(\mathbf{x}) \in \omega$. Alors on a une inégalité géométrique

$$(\mathbf{x}, \mathbf{n}(\mathbf{x})) \geq ch, \quad \forall \mathbf{x} \in \partial\Theta. \tag{5.17}$$

Soit alors le champ de vecteurs $\mathbf{y}(\mathbf{x}) = |v(\mathbf{x})|^p \mathbf{x}$ pour lequel on peut utiliser la formule de Stokes $\int_{\partial\Theta} (\mathbf{y}, \mathbf{n})\, d\sigma = \int_\Theta \nabla \cdot \mathbf{y} dx$, ou encore

$$\int_{\partial\Theta} (\mathbf{x}, \mathbf{n})\, |v(\mathbf{x})|^p d\sigma = \int_\Theta \left(2|v(\mathbf{x})|^p + p|v(\mathbf{x})|^{p-1} \text{signe}(v(\mathbf{x}))\, (\nabla v(\mathbf{x}), \mathbf{x})\right) dx$$

Donc

$$ch \int_{\partial\Theta} |v(\mathbf{x})|^p d\sigma \leq 2\, \|v\|^p_{L^p(\Theta)} + hp \int_\Theta |v(\mathbf{x})|^{p-1} |\nabla v| dx.$$

Or l'inégalité de Hölder pour $|v(\mathbf{x})|^{p-1} \in L^q(\Theta)$ et $|\nabla v| \in L^p(\Theta)$ indique que

$$\int_\Theta |v|^{p-1} |\nabla v| dx \leq \|v\|^{\frac{p}{q}}_{L^p(\Theta)}\, \| |\nabla v| \|_{L^p(\Theta)}.$$

L'inégalité (5.13) appliqué à v dans Θ montre que $\|v\|_{L^p(\Omega)} \leq Ch \|\nabla v\|_{L^p(\Omega)}$. D'où

$$ch \int_{\partial\Theta} |v(\mathbf{x})|^p d\sigma \leq \left(2C^p h^p + phh^{\frac{p}{q}}\right) \|\nabla v\|^p_{L^p(\Omega)} = (2C^p + p)\, h^p \|\nabla v\|^p_{L^p(\Omega)}$$

puis en reprenant (5.13)

$$\left(h \sum_k l_k |v_k|^p\right)^{\frac{1}{p}} \leq \left(\frac{2C^p + p}{c}\right)^{\frac{1}{p}} h \|\nabla v\|_{L^p(\Omega)}.$$

Le résultat est démontré pour une constante $\widehat{C} = \left(\frac{2C^p+p}{c}\right)^{\frac{1}{p}}$ que l'on peut majorer indépendamment de p. □

Soit

$$A = \left(\sum_j A_j^p\right)^p. \tag{5.18}$$

Lemme 38. *On a l'inégalité $A \leq Ch \|\nabla u\|_{L^p(\Omega)}$ (évident).*

5.1.3 Consistence des schémas de Volumes Finis pour l'advection

Nous allons plutôt montrer que la **non consistance au sens des Différences Finies** est la règle générale pour l'équation d'advection. Cette non consistance formelle est la raison des difficultés d'analyse numérique générées pas ces méthodes.
Nous partons du schéma de Volumes Finis pour un maillage général (5.2) ou (5.3). Pour analyser la consistance au sens des Différences Finis, il faut définir un opérateur de projection Π_h à partir de points $\mathbf{x}_j$. Ces points peuvent être les centres de masses des mailles mais ce n'est pas obligatoire. Il apparait raisonnable de demander que $\mathbf{x}_j \in \Omega_j$, mais ce n'est pas obligatoire non plus.
Soit $u = u_0(\mathbf{x} - \mathbf{a}t)$ une solution exacte pour la donnée initiale $u_0 \in W^{2,\infty}(\mathbb{R}^2)$. L'erreur de troncature est alors

$$r_j^n = \frac{v_j^{n+1} - v_j^n}{\Delta t} + \frac{1}{s_j}\sum_{k^+} m_{jk} v_j^n - \frac{1}{s_j}\sum_{k^-} m_{jk} v_k^n = \frac{v_j^{n+1} - v_j^n}{\Delta t} - \frac{1}{s_j}\sum_{k^-} m_{jk}\left(v_k^n - v_j^n\right).$$

où v_j^n, v_j^{n+1} et les v_k^n sont les valeurs ponctuelles associées à des points $\mathbf{x}_j$ que l'on a choisi préliminairement : $v_j^n = u(n\Delta t, \mathbf{x}_j)$ pour tout j et tout n. Pour des fonctions régulières un développement de Taylor montre que

$$v_j^{n+1} = v_j^n + \partial_t u(n\Delta t, \mathbf{x}_j)\Delta t + O(\Delta t^2) = v_j^n - \mathbf{a}\cdot\nabla u(n\Delta t, \mathbf{x}_j)\Delta t + O(\Delta t^2)$$

et

$$v_k^n = v_j^n + \nabla u(n\Delta t, \mathbf{x}_j)\cdot(\mathbf{x}_k - \mathbf{x}_j) + O(h^2).$$

On supposera que les points sont tels que

$$\sup_{k\in I(j)} |\mathbf{x}_j - \mathbf{x}_k| \leq Ch \text{ pour } C \text{ indépendant de } h.$$

On obtient

$$r_j^n = -\mathbf{a}\cdot\nabla u(n\Delta t, \mathbf{x}_j)\Delta t - \frac{1}{s_j}\sum_{k^-} l_{jk}\mathbf{a}\cdot\mathbf{n}_{jk}\nabla u(n\Delta t, \mathbf{x}_j)\cdot(\mathbf{x}_k - \mathbf{x}_j) + O(\Delta t) + O(h),$$

ou encore

$$r_j^n = \left(\mathbf{M}_{jk}^t\mathbf{a}\right)\cdot\nabla u(n\Delta t, \mathbf{x}_j) + O(\Delta t) + O(h). \tag{5.19}$$

avec une matrice $\mathbf{M}_j = -\mathbf{I} + \frac{1}{s_j}\sum_{k^-} l_{jk}\mathbf{n}_{jk}\otimes(\mathbf{x}_k - \mathbf{x}_j) \in \mathbb{R}^{2\times 2}$. Donc pour avoir $r_j^n = O(\Delta t + h)$ il faut et il suffit que $\mathbf{M}_j^t\mathbf{a}$ s'annule. On note que l'on retrouve exactement le critère déjà étudié (4.48) en dimension un d'espace. Comme il apparait raisonnable que les points $\mathbf{x}_j$ soit indépendants autant que possible de l'équation, on retiendra la définition suivante.

Définition 17 (Consistence au sens des Différences Finies : première version)**.** *On dira que le schéma est consistent au sens des Différences Finies si il existe des points* $(\mathbf{x}_j)$ *solution de l'équation* $\mathbf{M}_j = 0$*, c'est à dire*

$$\sum_{k^-} l_{jk}\mathbf{n}_{jk}\otimes(\mathbf{x}_k - \mathbf{x}_j) = s_j\mathbf{I}, \quad \forall j. \tag{5.20}$$

Si de plus $\mathbf{x}_j \in \overline{\Omega_j}$ *la solution est locale.*

La somme étant sur les k^-, il subsiste une dépendance par rapport à $\mathbf{a}$ dans cette définition.
Comme nous le savons déjà , les maillages cartésiens bénéficient de la consistance au sens des Différences Finies, ce que l'on retrouve rapidement de la façon suivante. Soit en effet un maillage cartésien (en dimension $d = 2$). Considérons que $\mathbf{x}_j$ est le centre de masse (aussi centre de gravité ou barycentre) de la maille d'indice j. Alors la condition de consistance (5.20) est vraie pour tout $\mathbf{a}$. En effet $s_j = \Delta x^2$, $l_{jk} = \Delta x$ et $\mathbf{x}_k - \mathbf{x}_j = \Delta x\ \mathbf{n}_{jk}$. Deux bords au plus contribuent dans (5.20). Le reste est affaire de calcul évident.
Un résultat négatif est le suivant.

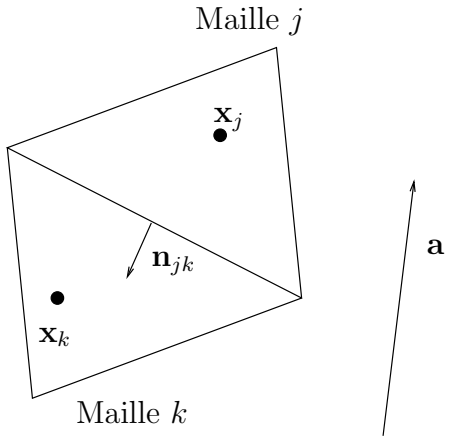


FIGURE 5.5 – Un cas particulier sans solution au critère de consistance (5.20)

Lemme 39. *Il existe des maillages pour lesquels il n'y a aucune solution au critère de consistance (5.20).*

Démonstration. Considérons le maillage en triangles de la figure 5.5.
Une seule maille est dans $I^-(j)$. La somme (5.20) se réduit à une seule contribution $l_{jk}\mathbf{n}_{jk} \otimes (\mathbf{x}_k - \mathbf{x}_j)$. Cette matrice est au plus de rang un, et ce pour tout $\mathbf{x}_j$ et tout $\mathbf{x}_k$. Elle ne peut donc pas être égale à $s_j\mathbf{I}$ qui est de rang deux.

□

Cela incite à étudier une version affaiblie de la relation de consistance (5.20).

Définition 18 (Consistance au sens des Différences Finies : deuxième version)**.** *Un deuxième critère de consistance au sens des Différences Finis s'écrit :* $\mathbf{M}_{jk}^t\mathbf{a} = 0$*. Ou encore*

$$\mathbf{x}_j = \sum_{k^-} \left(\frac{l_{jk}\ \mathbf{a}\cdot\mathbf{n}_{jk}}{\sum_{r^-} l_{jr}\ \mathbf{a}\cdot\mathbf{n}_{jr}} \right) \mathbf{x}_k - \frac{s_j}{\sum_{k^-} l_{jk}\ \mathbf{a}\cdot\mathbf{n}_{jk}}\mathbf{a}. \tag{5.21}$$

Si (5.21) est vrai alors l'erreur de troncature (5.19) est en $O(\Delta t + h)$, ce qui est exactement la définition de la consistance au sens des Différences Finies.
Il est possible a priori de résoudre (5.21) de proche en proche en considérant que les $\mathbf{x}_k$ ont déjà été calculés. Cela détermine le point $\mathbf{x}_j$ comme une moyenne des $\mathbf{x}_k$ plus une correction géométrique, ce qui propage la connaissance de $\mathbf{x}_j$ de mailles en mailles. Cependant l'étude de ce système, même élémentaire, n'a pas évidente. Par exemple la solution peut ne pas être locale, $\mathbf{x}_j \notin \Omega_j$. Cela rend l'interprétation de la solution délicate. On pourra consulter [3].

5.2 Convergence dans L^2

Nous montrons la convergence dans L^2 du schéma de Volumes Finis pour l'advection, en utilisant une combinaison de techniques adaptées. Le domaine d'étude est le tore $\mathcal{T}$. Le champ de vitesse $\mathbf{a} \in \mathbb{R}^2$ est constant en temps et en espace. La donnée initiale est une fois dérivable dans L^2, soit $u_0 \in H^1(\mathcal{T})$.
Le maillage est régulier avec un nombre de voisins par mailles qui est borné indépendant de h. Cela est assuré pour un maillage dont les mailles sont des polygones avec un nombre donné maximal de côtés. Les mailles sont toutes convexes.

Théorème 3. *Supposons la condition CFL vérifiée. Alors le schéma de Volumes Finis est convergent avec l'estimation d'ordre fractionnaire*

$$\|u_h^n - \Pi_h u(n\Delta t)\|_{L^2(\mathcal{T})} \leq C\|\nabla u\|_{L^2(\mathcal{T})}(Th)^{\frac{1}{2}}, \quad n\Delta t \leq T. \tag{5.22}$$

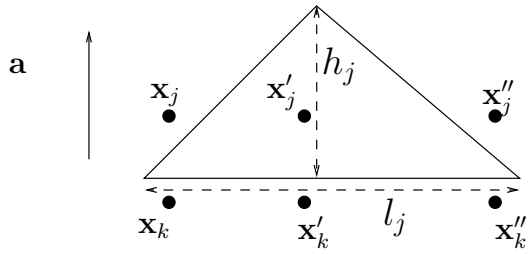


FIGURE 5.6 – Ici l'équation (5.21) se simplifie en $\mathbf{x}_j = \frac{h_j}{2|\mathbf{a}|}\mathbf{a} + \mathbf{x}_k$. La hauteur du triangle est $h_j = \frac{s_j}{l_j}$. Si le second membre de (5.21) est $\mathbf{x}_k$ alors $\mathbf{x}_j \in \Omega_j$. Cependant si $\mathbf{x}'_k$ ou $\mathbf{x}''_k$ sont près des coins, alors $\mathbf{x}_j \notin \Omega_j$.

Remarque 8. *La comparaison avec les résultats en dimension un d'espace de la section 4.2.3 montre que ce résultat est optimal car il retrouve exactement l'ordre de convergence moitié pour une donnée une fois dérivable (dans L^2).*

Pour simplifier un peu, la preuve est décomposé en deux étapes. La première étape est plus générale car dans L^p.

5.2.1 Première étape : estimation en temps dans L^p

On s'appuie sur le lemme 21 pour remplacer le schéma explicite

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{1}{s_j}\left(\sum_{k^+} m_{jk} u_j^n - \sum_{k^-} m_{jk} u_k^n\right) = 0,$$

par le schéma semi-discret

$$v_j'(t) + \frac{1}{s_j}\left(\sum_{k^+} m_{jk} v_j(t) - \sum_{k^-} m_{jk} v_k(t)\right) = 0.$$

La condition initiale est commune

$$u_j^0 = v_j(0) = \frac{1}{s_j}\int_{\Omega_j} u_0(\mathbf{x})dx.$$

Lemme 40. *Soit une donnée initiale $u_0 \in W^{1,p}(\mathcal{T})$. Soit une suite de maillages réguliers de pas $h \to 0$. Alors il existe une constante universelle $C > 0$ telle que*

$$\|u_h^n - v_h(n\Delta t)\|_{L^p(\mathcal{T})} \leq C\|\nabla u_0\|_{L^p(\mathcal{T})}(Th)^{\frac{1}{2}}, \quad n\Delta t \leq T. \tag{5.23}$$

Démonstration. On applique l'inégalité de comparaison du lemme 21. Tout d'abord les hypothèses sur le maillage et l'étude de la condition CFL montrent que $\tau(h) \leq Ch$ pour une constante $C > 0$ bornée indépendamment de h. Il reste à obtenir une bonne estimation sur $A_h\Pi_h u_0 = (w_j)$ avec

$$w_j = \frac{1}{s_j}\sum_{k^+} l_{jk}\left(u_k^0 - u_j^0\right)$$

où $u_j^0 = \frac{1}{s_j}\int_{\Omega_j} u(\mathbf{x})dx$ et $u_k^0 = \frac{1}{s_k}\int_{\Omega_k} u(\mathbf{x})dx$ sont les valeurs moyennes obtenues par projection de la donnée initiale. On a

$$w_j = \frac{1}{s_j}\sum_{k^+} l_{jk}\left(u_{jk}^0 - u_j^0\right) + \frac{1}{s_j}\sum_{k^+} l_{jk}\left(u_k^0 - u_{jk}^0\right)$$

où $u_{jk}^0 = \frac{1}{l_{jk}}\int_{\partial\Omega_j\cap\partial\Omega_k} u(\mathbf{x})dx$ est la valeur moyenne sur l'interface commune de la donnée initiale. L'inégalité de Hölder montre que

$$\left|\sum_{k^+} l_{jk}\left(u_{jk}^0 - u_j^0\right)\right| \leq \left(\sum_{k^+} l_{jk}\left|u_{jk}^0 - u_j^0\right|^p\right)^{\frac{1}{p}}\left(\sum_{k^+} l_{jk}\right)^{\frac{1}{q}} \leq \left(Ch^{-\frac{1}{p}}A_j\right)h^{\frac{1}{q}}$$

où on a repris la définition (5.15) de A_j. De même pour les autres termes. D'où grâce à la minoration uniforme $s_j \geq ch^2$:

$$|w_j| \leq Ch^{\frac{1}{q}-\frac{1}{p}-2}\left(A_j + \sum_{k^+} A_k\right).$$

Puis en utilisant le fait que le nombre de voisins est borné indépendamment de h, on obtient $\|A_h\Pi_h u_0\| = \left(\sum_j s_j |w_j|^p\right)^{\frac{1}{p}} \leq Ch^{\frac{1}{q}-\frac{1}{p}-2+\frac{2}{p}} A$, avec A défini par (5.18) : le terme $h^{\frac{2}{p}}$ vient des contributions de la forme $s_j^{\frac{1}{p}}$. Or $A \leq ch \|\nabla u_0\|_{L^p(\mathcal{T})}$ par le lemme 38. Comme $\frac{1}{q} - \frac{1}{p} - 2 + \frac{2}{p} + 1 = 0$, cela établit le résultat. □

5.2.2 Deuxième étape : estimation en espace dans L^2

Nous étudions à présent la différence entre la solution du schéma semi-discret et la projection de la solution exacte, en norme L^2.

Lemme 41. *Supposons : $u_0 \in H^1(\mathcal{T})$; la condition CFL réalisée ; et les maillages réguliers. On a l'estimation d'erreur*

$$||u(n\Delta t) - v_h(n\Delta t)||_{L^2} \leq C \|\nabla u_0\|_{L^2(\mathcal{T})} \left(h + (Th)^{\frac{1}{2}}\right), \qquad n\Delta t \leq T. \tag{5.24}$$

Démonstration. La fonction $v_h(t) \in L^2(\Omega)$ est constante par mailles. On étudie

$$E(t) = \frac{1}{2}\int_{\mathcal{T}} (u(t) - v_h(t))^2. \tag{5.25}$$

avec $E(0) \leq C(\nabla u_0)h^2$ au temps initial en utilisant le résultat du lemme 36. Le résultat final (5.24) sera démontré si nous pouvons montrer que $E'(t) \leq C||\nabla u_0||_{L^2}^2 h$. Or nous allons voir que c'est affaire de calculs élémentaires.
On a $E(t) = \frac{1}{2}\int_\Omega u(t)^2 + \frac{1}{2}\int_\Omega v_h(t)^2 - \int_\Omega v_h(t)u(t)$. Donc

$$E'(t) = \underbrace{\frac{d}{dt}\left(\frac{1}{2}\int_{\mathcal{T}} u(t)^2\right)}_{=A_1} + \underbrace{\left(-\frac{1}{2}\sum_j \sum_{k\in I^+(j)} m_{jk}(v_j - v_k)^2\right)}_{=A_2}$$
$$+ \underbrace{\sum_j \left(\frac{\sum_{k^+} m_{jk}v_j - \sum_{k^-} m_{jk}v_k}{s_j}\right)\int_{\Omega_j} u(t)}_{=A_3} + \underbrace{\sum_j u_j(-\sum_{k^+} m_{jk}u_{jk} + \sum_{k^-} m_{jk}u_{jk})}_{=A_4}.$$

où u_{jk} dénote la valeur moyenne de la solution exacte au temps t sur l'interface $\partial\Omega_j \cap \partial\Omega_k$. Le premier terme A_1 est nul car la norme L^2 de la solution de l'équation d'advection est constante pour un domaine sans bord

$$A_1 = \frac{d}{dt}\int_\Omega \frac{u^2}{2} = \int_\Omega u\partial_t u = -\int_\Omega u\mathbf{a}\cdot\nabla u = -\int_\Omega \mathbf{a}\cdot\nabla\frac{u}{2} = 0.$$

Le deuxième terme A_2 est négatif ou nul. Les termes suivants A_3 et A_4 sont *a priori* tels que leur somme est homogène à $\approx \int_\Omega (\mathbf{a}.\nabla(uw_h) = 0$. On peut alors anticiper qu'une réécriture adaptée permet de mettre en évidence que leur somme est petite en un sens à définir. Vérifions.
Comme $\sum_{k^+} m_{jk} = \sum_{k^-} m_{jk}$, alors

$$A_3 = \underbrace{-\sum_j\sum_{k^-} m_{jk}(v_j - v_k)u_{jk}}_{=A_5} + \underbrace{\sum_j\sum_{k^-} m_{jk}(v_j - v_k)\left(u_{jk} - \frac{\int_{\Omega_j} u}{s_j}\right)}_{=A_6}. \tag{5.26}$$

Une intégration par partie discrète, c'est à dire une permutation des indices de sommation, montre que $A_5 = -A_4$. Par ailleurs une inégalité de la forme $\alpha\beta \leq \frac{1}{4}\alpha^2 + \beta^2$ montre que $A_6 \leq -\frac{1}{2}A_2 + \sum_j\sum_{k^-} m_{jk}\left(u_{jk} - \frac{\int_{\Omega_j} u}{s_j}\right)^2$. One obtient alors

$$E'(t) + \frac{1}{2}\sum_j \sum_{k\in I^+(j)} m_{jk}(v_j - v_k)^2 \leq \frac{1}{2}\sum_j\sum_{k^-} m_{jk}\left(u_{jk} - \frac{\int_{\Omega_j} u}{s_j}\right)^2.$$

Le résultat du lemme 38 pour $p = 2$ montre que $E'(t) + \frac{1}{2}\sum_j\sum_{k\in I^+(j)} m_{jk}(v_j - v_k)^2 \leq Ch||\nabla u_0||_{L^2}^2$. Il s'ensuit que

$$E(t) + \frac{1}{2}\int_0^t \sum_j \sum_{k\in I^+(j)} m_{jk}(v_j(s) - v_k(s))^2 ds \leq Ch^2 \|\nabla u_0\|_{L^2}^2 + Ch\|\nabla u_0\|_{L^2}^2 T. \tag{5.27}$$

Le résultat est démontré avec de plus une estimation sur les différences de la solution semi-discrète qui sera utilisée dans ce qui suit. □

Remarque 9. *La structure de la preuve de l'inégalité (5.24) s'appuie d'une part sur la dissipation de l'énergie L^2 ce qui est une propriété courante pour une méthode numérique et d'autre part sur la structure d'un schéma de Volumes Finis qui peut se caractériser par la relation $\sum_{k^+} m_{jk} = \sum_{k^-} m_{jk}$ qui est fondamentale dans les méthodes de Volumes Finis. En résumé le point clé de la preuve est la transformation (5.26).*

Preuve final du théorème 3. Le théorème de convergence 3 s'obtient par inégalité triangulaire à partir de l'inégalité (5.23) et de l'inégalité (5.24). □

5.3 Convergence dans L^1

On fera l'hypothèse que

$$u_0 \in \mathrm{BV}(\mathcal{T})$$

ce qui permet de traiter les cas des fonctions indicatrices, lesquelles sont liées au calcul numérique de la propagation d'interfaces par des schémas de Volumes Finis.
La stratégie générale de preuve de convergence est identique au cas précédent. D'abord se ramener au schéma semi-discret ce qui ne pose pas de difficultés à partir du résultat du lemme 40 et du fait que $W^{1,1}$ est dense dans BV. D'où une première estimation

$$\|u_h^n - v_h(n\Delta t)\|_{L^1(\mathcal{T})} \leq C\,|u_0|_{\mathrm{BV}(\mathcal{T})}\,(Th)^{\frac{1}{2}}, \quad n\Delta t \leq T. \tag{5.28}$$

L'estimation en espace va être montrée pour des fonctions indicatrices.

5.3.1 Cas des fonctions indicatrices

La donnée initiale $u_0 = \mathbf{1}_\omega$ est prise comme la fonction indicatrice d'une partie $\omega \subset \mathcal{T}$

$$u_0(x) = 1 \text{ pour } x \in \omega, \quad \text{ et } u_0(x) = 1 \text{ pour } x \notin \Theta.$$

On supposera le périmètre ω borné, auquel cas

$$|u_0|_{\mathrm{BV}} = |\omega| < \infty.$$

On commence par régulariser/convoluer la donner initiale u_0 à l'aide d'un noyau positif ou nul, borné, de masse unité et à support compact

$$u_0^\varepsilon(\mathbf{x}) = (\varphi_\varepsilon * u_0)(\mathbf{x}) = \frac{1}{\varepsilon}\int_{\mathbf{y}\in\mathcal{T}} \varphi\left(\frac{\mathbf{x}-\mathbf{y}}{\varepsilon}\right) u_0(\mathbf{y}) d\mathbf{y}.$$

Un résultat classique [19] montre que

$$\|u_0^\varepsilon - u_0\|_{L^1(\mathcal{T})} \leq C\varepsilon|\omega|. \tag{5.29}$$

On a également que

$$\nabla u_0^\varepsilon = \frac{1}{\varepsilon^2}\int_{\mathbf{y}\in\mathcal{T}} \nabla\varphi\left(\frac{\mathbf{x}-\mathbf{y}}{\varepsilon}\right) u_0(\mathbf{y}) dy$$

d'où l'on tire à partir de la définition d'une fonction BV que

$$\|\nabla u_0^\varepsilon\|_{L^\infty(\mathcal{T})} \leq C\frac{|u_0|_{BV(\mathcal{T})}}{\varepsilon} \text{ et } \|\nabla u_0^\varepsilon\|_{L^1(\mathcal{T})} \leq C\,|u_0|_{BV(\mathcal{T})}\,.$$

Il s'ensuit une inégalité qui va jouer un rôle dans la suite

$$\|\nabla u_0^\varepsilon\|_{L^2(\mathcal{T})} \leq C\frac{|u_0|_{BV(\mathcal{T})}}{\varepsilon^{\frac{1}{2}}}. \tag{5.30}$$

La solution du schéma semi-discret issu de u_0^ε est $v_h^\varepsilon(t) = e^{A_h t}\Pi_h u_0^\varepsilon$. La solution du schéma semi-discret issu de u_0 est $v_h(t) = e^{A_h t}\Pi_h u_0$.

Lemme 42. *L'erreur entre la solution numérique semi-discrète et la solution exacte peut se majorer par l'erreur entre les solutions régularisées plus un reste*

$$\|v_h(t) - u(t)\|_{L^1(\mathcal{T})} \leq \|v_h^\varepsilon(t) - u^\varepsilon(t)\|_{L^1(\mathcal{T})} + C\varepsilon|\omega|. \tag{5.31}$$

Démonstration. On a l'inégalité triangulaire

$$\|v_h(t) - u(t)\|_{L^1(\mathcal{T})} \leq \|v_h(t) - v_h^\varepsilon(t)\|_{L^1(\mathcal{T})} + \|v_h^\varepsilon(t) - u^\varepsilon(t)\|_{L^1(\mathcal{T})} + \|u^\varepsilon(t) - u(t)\|_{L^1(\mathcal{T})}\,.$$

On a $\|u^\varepsilon(t) - u(t)\|_{L^1(\mathcal{T})} \leq \|u_0^\varepsilon - u_0\|_{L^1(\mathcal{T})}$. Or on a aussi $\|v_h(t) - v_h^\varepsilon(t)\|_{L^1(\mathcal{T})} \leq \|v_h(0) - v_h^\varepsilon(0)\|_{L^1(\mathcal{T})}$ en utilisant la stabilité unitaire dans L^1 du schéma semi-discret, laquelle peut soit se voir comme une conséquence de la propriété générale (4.24) qui étend au semi-discret les propriétés de stabilité des schémas explicites, soit se re-démontrer directement. Puis $\|v_h(t) - v_h^\varepsilon(t)\|_{L^1(\mathcal{T})} \leq \|u_0^\varepsilon - u_0\|_{L^1(\mathcal{T})}$. Donc

$$\|v_h(t) - u(t)\|_{L^1(\mathcal{T})} \leq \|v_h^\varepsilon(t) - u^\varepsilon(t)\|_{L^1(\mathcal{T})} + 2\,\|u_0^\varepsilon - u_0\|_{L^1(\mathcal{T})}$$

La preuve est terminée grâce à (5.29). □

Lemme 43. *On a la formule*

$$\|u^\varepsilon(t) - v_h^\varepsilon(t)\|_{L^1(\mathcal{T})} = \|u^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 - \|v_h^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 + \|u^\varepsilon(t) - v_h^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 + O(\varepsilon|\omega|)$$

où le terme $O(\varepsilon|\omega|)$ est indépendant du temps.

Démonstration. Le support du noyau de convolution φ est compact, aussi $u_0^\varepsilon(x) = u_0(x)$ sauf éventuellement dans une région dont l'aire peut se majorer en $\mathcal{A}_\varepsilon = \mathrm{Per}(\omega)O(\varepsilon)$. Cela étant vrai pour ω de la forme d'un disque ou d'un carré, nous l'admettons sans démonstration pour le cas général. Après advection on a la même propriété entre $u^\varepsilon(t)$ et $u(t)$.
Dans les régions où $u^\varepsilon(t) = 1$

$$|u^\varepsilon(t) - w_h^\varepsilon(t)| = 1 - w_h^\varepsilon(t).$$

En effet $w_h^\varepsilon(t) \leq 1$ car le schéma semi-discret vérifie aussi le principe du maximum : cela qui peut se voir comme une conséquence de la propriété générale (4.24) qui étend au semi-discret les propriétés de stabilité des schémas explicites, soit se re-démontrer directement.
Dans les régions où $u^\varepsilon(t) = 0$ on a par un principe similaire

$$|u^\varepsilon(t) - w_h^\varepsilon(t)| = w_h^\varepsilon(t).$$

Ces deux situations peuvent se résumer par

$$|u^\varepsilon(t) - w_h^\varepsilon(t)| = (u^\varepsilon(t) - w_h^\varepsilon(t)) \times (2u^\varepsilon(t) - 1)\,,$$

qui est valide presque partout excepté dans un domaine d'aire $\mathcal{A}_\varepsilon = O(\varepsilon|\omega|)$. On a donc

$$\|u^\varepsilon(t) - v_h^\varepsilon(t)\|_{L^1(\mathcal{T})} = \int_\Omega (u^\varepsilon(t) - w_h^\varepsilon(t)) \times (2u^\varepsilon(t) - 1) + O(\varepsilon|\omega|).$$

Or l'initialisation de la donnée initiale en valeur moyenne et la conservativité (lemme 12) du schéma font que $\int_\Omega \left(u^\varepsilon(t) - w_h^\varepsilon(t)\right) = 0$. Il reste alors les termes $\left(u^\varepsilon - w_h^\varepsilon\right) 2u^\varepsilon = |u^\varepsilon|^2 - \left|w_h^\varepsilon\right|^2 + \left|u^\varepsilon - w_h^\varepsilon\right|^2$ que l'on retrouve directement dans le résultat. La preuve est terminée. □

Théorème 4. *Soient $T > 0$ et $h \leq 1$. Il existe une constante $C > 0$ telle que*

$$\|u^\varepsilon(t) - v_h^\varepsilon(t)\|_{L^1(\mathcal{T})} \leq C\,|u_0|_{BV(\mathcal{T})}^{\frac{1}{2}}\, h^{\frac{1}{2}}, \qquad t \leq T.$$

Démonstration. En effet l'inégalité (5.23) en norme L^2 combinée avec (5.30) l'estimation sur le gradient également en norme L^2 implique

$$\|u^\varepsilon(t) - v_h^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 \leq C\,\|\nabla u_0^\varepsilon\|_{L^2(\mathcal{T})}^2 \left(h^2 + th\right) \leq C\frac{|u_0|_{BV(\mathcal{T})}^2}{\varepsilon}\left(h^2 + th\right).$$

D'autre part

$$\frac{d}{dt}\left(\|u^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 - \|v_h^\varepsilon(t)\|_{L^2(\mathcal{T})}^2\right) = \frac{1}{2}\sum_j \sum_{k\in I^+(j)} m_{jk}(v_j^\varepsilon - v_k^\varepsilon)^2$$

car la norme L^2 de u^ε est constante, et le schéma est dissipatif. Le terme dissipatif est exactement le terme A_2 dans la preuve du lemme 41, et dont l'intégrale en temps peut se majorer par (5.27). Donc on peut écrire

$$\|u^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 - \|v_h^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 \leq \left(\|u_0^\varepsilon\|_{L^2(\mathcal{T})}^2 - \|v_h^\varepsilon(0)\|_{L^2(\mathcal{T})}^2\right) + Ch^2\,\|\nabla u_0^\varepsilon\|_{L^2}^2 + Cth\,\|\nabla u_0^\varepsilon\|_{L^2}^2\,.$$

On a immédiatement que

$$\|u_0^\varepsilon\|_{L^2(\mathcal{T})}^2 - \|v_h^\varepsilon(0)\|_{L^2(\mathcal{T})}^2 = (u_0^\varepsilon - v_h^\varepsilon(0), u_0^\varepsilon + v_h^\varepsilon(0))_{L^2(\mathcal{T})} \leq \|u_0^\varepsilon - v_h^\varepsilon(0)\|_{L^1(\mathcal{T})} \times \|u_0^\varepsilon + v_h^\varepsilon(0)\|_{L^\infty(\mathcal{T})}\,.$$

D'où grâce à (5.13) et au fait que les données sont bornées dans L^∞ : $\left\|u_0^\varepsilon\right\|_{L^2(\mathcal{T})}^2 - \left\|v_h^\varepsilon(0)\right\|_{L^2(\mathcal{T})}^2 \leq Ch\left\|\nabla u_0^\varepsilon\right\|_{L^1(\mathcal{T})}$ puis $\left\|u_0^\varepsilon\right\|_{L^2(\mathcal{T})}^2 - \left\|v_h^\varepsilon(0)\right\|_{L^2(\mathcal{T})}^2 \leq Ch\,|u_0|_{BV(\mathcal{T})}$. Donc on peut écrire

$$\|u^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 - \|v_h^\varepsilon(t)\|_{L^2(\mathcal{T})}^2 \leq Ch\,|u_0|_{BV(\mathcal{T})} + C\frac{h^2 + th}{\varepsilon}\,|u_0|_{BV(\mathcal{T})}^2\,.$$

En regroupant ces diverses expressions on obtient

$$\|u(t) - v_h(t)\|_{L^1(\mathcal{T})} \leq Ch\,|u_0|_{BV(\mathcal{T})} + C\frac{h^2 + th}{\varepsilon}\,|u_0|_{BV(\mathcal{T})}^2 + C\varepsilon\,|u_0|_{BV(\mathcal{T})}\,.$$

Une valeur optimale du paramètre de convolution est $\varepsilon = \sqrt{h}$. On obtient

$$\|u(t) - v_h(t)\|_{L^1(\mathcal{T})} \leq C\left(h + h^{\frac{3}{2}} + th^{\frac{1}{2}} + h^{\frac{1}{2}}\right)|u_0|_{BV(\mathcal{T})}\,.$$

On obtient alors le résultat pour $h \leq 1$ et $t \leq T$ donné. □

5.3.2 Données générales

Le résultat du théorème de convergence dans L^1 pour une fonction indicatrice peut s'étendre aux fonctions de BV à partir de l'inégalité de la co-aire (1.1).

5.4 Convergence du schéma de diffusion

On analyse à présent la version implicite du schéma (2.49) pour l'équation de la chaleur. Il s'écrit pour tout $n \geq 0$

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \frac{1}{s_j} \sum_k l_{jk} \frac{u_k^{n+1} - u_j^{n+1}}{d_{jk}} = 0, \quad \forall j. \tag{5.32}$$

Les éléments caractéristiques du maillage du tore $\mathcal{T}$ sont l'aire de la maille courante notée $s_j > 0$, la longueur de l'interface entre les mailles voisines notée $l_{jk} > 0$: la distance entre les centres de gravité $\mathbf{x}_j$ et $\mathbf{x}_k$ de deux mailles voisines, initialement dénotée $\widehat{d}_{jk}$, sera noté $d_{jk} > 0$ pour alléger la notation. On envisage l'initialisation ponctuelle

$$u_j^0 = u_0(\mathbf{x}_j), \quad \forall j. \tag{5.33}$$

On pourrait tout aussi bien étudier les variantes explicites ou semi-discrètes avec des résultats similaires.
Comme noté précédemment, la matrice du système linéaire qui permet de calculer u_h^{n+1} en fonction de u_h^n est inversible. On peut le retrouver comme conséquence de la décroissance de la norme L^2.

Lemme 44 (Stabilité inconditionnelle en norme quadratique)**.** *Soit $v_h = (v_j)$ donné. Soit $u_h = (u_j)$ une solution de $\frac{u_j - v_j}{\Delta t} + \frac{1}{s_j}\sum_k l_{jk}\frac{u_k - u_j}{d_{jk}} = 0, \quad \forall j$. Alors pour tout $\Delta t > 0$*

$$\|u_h\|_{L^2(\mathcal{T})} \leq \|v_h\|_{L^2(\mathcal{T})}.$$

Démonstration. Le schéma se récrit

$$u_j - \frac{\Delta t}{s_j} \sum_k l_{jk} \frac{u_k - u_j}{d_{jk}} = v_j.$$

Multipliant par u_j et sommant sur toutes les mailles on trouve

$$\sum_j s_j u_j^2 - \Delta t \sum_j \left(u_j \sum_k l_{jk} \frac{u_k - u_j}{d_{jk}} \right) = \sum_j s_j u_j v_j$$

ce qui implique après quelques manipulations

$$\sum_j s_j u_j^2 + \Delta t \sum_{j<k} \frac{l_{jk}}{d_{jk}} (u_j - u_k)^2 = \frac{1}{2} \sum_j s_j u_j^2 + \frac{1}{2} \sum_j s_j v_j^2 - \sum_j s_j (u_j - v_j)^2$$

où $\sum_{j<k} = \sum_j \sum_{k,\ j<k}$ est une somme double sur toutes les interfaces entre maille d'indice j et mailles d'indices k. D'où

$$\frac{1}{2} \sum_j s_j u_j^2 + \Delta t \sum_{j<k} \frac{l_{jk}}{d_{jk}} (u_j - u_k)^2 + \sum_j s_j (u_j - v_j)^2 = \frac{1}{2} \sum_j s_j v_j^2 \tag{5.34}$$

ce qui termine la preuve. □

On retrouve bien que si $v_h \equiv 0$, alors l'unique solution du système linéaire est $u_h \equiv 0$. Or c'est un des critères possibles pour caractériser l'invisibilité du système linéaire qui permet de déterminer u_h en fonction de v_h. Donc le système linéaire est inversible.
Pour poursuivre l'analyse numérique on empreinte une idée très courante dans les formulations variationnelles pour les problèmes elliptiques qui est de récrire (5.32) sous une **forme mixte**, c'est à dire en faisant apparaitre explicitement des discrétisations d'opérateurs différentiels du premier ordre. On obtient

$$\begin{cases} \dfrac{u_j^{n+1} - u_j^n}{\Delta t} - \dfrac{1}{s_j} \displaystyle\sum_k l_{jk} p_{jk}^{n+1} = 0, & \forall j, \\ p_{jk}^{n+1} - \dfrac{u_k^{n+1} - u_j^{n+1}}{d_{jk}} = 0, & \forall (j,k). \end{cases}$$

On note bien sûr que $p_{kj}^{n+1} = -p_{jk}^{n+1}$. Puis on reprend l'idée de la consistance au sens des Différences Finies, qui est d'introduire la solution exacte dans le schéma est d'évaluer l'erreur de troncature. Le point important est que l'on effectue cette étude de consistance pour les deux équations discrètes séparément.
On prend $v_j^n = u(\mathbf{x}_j, t_n)dx$ qui est la valeur au point $\mathbf{x}_j$ de la solution exacte et

$$q_{jk}^n = \frac{1}{l_{jk}} \int_{\partial\Omega_j \cap \partial\Omega_k} \nabla u(\mathbf{x}, t_n) \cdot \mathbf{n}_{jk} d\sigma$$

qui est la projection en moyenne sur les segments d'interface. La valeur ponctuelle est correctement définie pour une fonction continue ce qui sera le cas pour la régularité envisagée dans le lemme qui suit. On définit alors **deux erreurs de troncature**

$$\begin{cases} r_j^n = \dfrac{v_j^{n+1} - v_j^n}{\Delta t} - \dfrac{1}{s_j} \displaystyle\sum_k l_{jk} q_{jk}^{n+1}, & \forall j, \\ t_{jk}^n = q_{jk}^{n+1} - \dfrac{v_k^{n+1} - v_j^{n+1}}{d_{jk}}, & \forall (j,k). \end{cases}$$

Le terme t_{jk}^n évalue la consistance du flux numérique. Ces deux erreurs de troncature $r_h^n = \left(r_j^n\right)_j$ et $t_h^n = \left(t_{jk}^n\right)_{jk}$ ne vivent pas dans les mêmes espaces, mais peuvent toutes deux s'estimer dans des normes quadratiques adaptées. Comme auparavant on prendra $\|r_h\|_{L^2(\mathcal{T})}^2 = \sum_j s_j r_j^2$. On définit

$$|||t_h|||_{L^2(\mathcal{T})}^2 = \sum_{jk} l_{jk} d_{jk} t_{jk}^2.$$

On remarque $s_j = O(h^2)$ et $l_{jk} d_{jk} = O(h^2)$ ce qui fait que ce sont deux normes de type L^2.

Proposition 14 (Consistance des erreurs de troncature)**.** *Supposons que la solution soit $u \in \mathcal{W}^{2,\infty}([0,T] \times \mathcal{T})$. Supposons le maillage triangulaire et satisfaisant la condition du lemme 14.*
Alors il existe C qui dépend de u et de ses dérivées telle que

$$\|r_h^n\|_{L^2(\mathcal{T})} \leq C(\Delta t + h) \;\; et \;\; |||t_h^n|||_{L^2(\mathcal{T})} \leq Ch, \qquad n\Delta t \leq T.$$

Cette preuve est loin d'être optimale, ne serait-ce que parce que la régularité de la solution est évaluée dans des espaces de type L^∞ et que l'erreur est mesurée dans L^2. Mais la structure de la preuve est intéressante en elle-même car la dépendance des estimations par rapport aux éléments caractéristiques du maillage apparait clairement. Les conditions sur le maillage sont elles-aussi restrictives.
On consultera [15] pour des développements complémentaires.

Démonstration. On a

$$\begin{aligned} r_j^n &= \frac{u(\mathbf{x}_j, t_{n+1}) - u(\mathbf{x}_j, t_n)}{\Delta t} - \frac{1}{s_j} \int_{\partial\Omega_j} \nabla u(\mathbf{x}, t_{n+1}) \cdot \mathbf{n}_j d\sigma \\ &= \frac{u(\mathbf{x}_j, t_{n+1}) - u(\mathbf{x}_j, t_n)}{\Delta t} - \frac{1}{s_j} \int_{\Omega_j} \Delta u(\mathbf{x}, t_{n+1}) dx \\ &= \frac{u(\mathbf{x}_j, t_{n+1}) - u_j(\mathbf{x}, t_n)}{\Delta t} - \frac{1}{s_j} \int_{\Omega_j} \partial_t u(\mathbf{x}, t_{n+1}) dx \\ = \left(\frac{u(\mathbf{x}_j, t_{n+1}) - u(\mathbf{x}_j, t_n)}{\Delta t} - \partial_t u(\mathbf{x}_j, t_{n+1}) \right) &+ \frac{1}{s_j} \int_{\Omega_j} \left(\partial_t u(\mathbf{x}_j, t_{n+1}) - \partial_t u(\mathbf{x}, t_{n+1}) \right) dx. \end{aligned}$$

Or $\partial_{tt} u \in L^\infty([0,T] \times \mathcal{T})$. On a alors pour le terme entre parenthèse

$$\left| \frac{u(\mathbf{x}_j, t_{n+1}) - u(\mathbf{x}_j, t_n)}{\Delta t} - \partial_t u(\mathbf{x}_j, t_{n+1}) \right| \leq \frac{1}{2} \Delta t \, \|\partial_{tt} u(t_{n+1})\|_{L^\infty([0,T] \times \mathcal{T})}.$$

Comme on a aussi $\nabla\partial_t u \in L^\infty([0,T]\times\mathcal{T})$ le terme sous l'intégrale s'estime par

$$|\partial_t u(\mathbf{x}_j, t_{n+1}) - \partial_t u(\mathbf{x}, t_{n+1})| \leq h \left\|\nabla\partial_t u\right\|_{L^\infty([0,T]\times\mathcal{T})}$$

avec $\mathrm{diam}(\Omega_j) \leq h$. Donc

$$\left|r_j^n\right| \leq \frac{1}{2}\Delta t \left\|\partial_{tt} u(t_{n+1})\right\|_{L^\infty(\mathcal{T})} + h \left\|\nabla\partial_t u\right\|_{L^\infty([0,T]\times\mathcal{T})}.$$

Comme $\|r_h^n\|_{L^2(\mathcal{T})} \leq \|r_h^n\|_{L^\infty(\mathcal{T})} |\mathcal{T}|^{\frac{1}{2}}$, on obtient une première estimation

$$\|r_h^n\|_{L^2(\mathcal{T})} \leq \left(\underbrace{\frac{1}{2}\left\|\partial_{tt} u(t_{n+1})\right\|_{L^\infty(\mathcal{T})}}_{=c_1} \Delta t + \underbrace{\left\|\nabla\partial_t u\right\|_{L^\infty([0,T]\times\mathcal{T})}}_{=c_2} h \right) |\mathcal{T}|^{\frac{1}{2}} \leq C(\Delta t + h). \tag{5.35}$$

avec $C = \max(c_1, c_2)\,|\mathcal{T}|^{\frac{1}{2}}$.
On évalue à présent le deuxième terme. On a

$$t_{jk}^n = \left(q_{jk}^{n+1} - \nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk}\right) + \left(\nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk} - \frac{v_k^{n+1} - v_j^{n+1}}{d_{jk}}\right).$$

Or

$$q_{jk}^n - \nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk} = \frac{1}{l_{jk}} \int_{\partial\Omega_j \cap \partial\Omega_k} (\nabla u(\mathbf{x}, t_{n+1}) - \nabla u(\mathbf{x}_{jk}, t_{n+1}))\cdot\mathbf{n}_{jk} d\sigma$$

Comme la matrice Hessienne des dérivées secondes de u est bornée, $\nabla^2 u \in L^\infty([0,T]\times\mathcal{T})^4$, on en déduit que

$$\left|q_{jk}^n - \nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk}\right| \leq \left\|\nabla^2 u\right\|_{L^\infty([0,T]\times\mathcal{T})^4} h.$$

Le dernier terme à estimer est

$$\nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk} - \frac{v_k^{n+1} - v_j^{n+1}}{d_{jk}} = \nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk} - \frac{u(\mathbf{x}_k, t_{n+1}) - u(\mathbf{x}_j, t_{n+1})}{d_{jk}}$$

où le point $\mathbf{x}_{jk}$ est situé entre $\mathbf{x}_j$ et $\mathbf{x}_k$, et où d_{jk} est précisément la distance entre en $\mathbf{x}_j$ et $\mathbf{x}_k$. Bien que bidimensionnelle, la situation est identique à celle de la figure 2.6 en dimension un d'espace. Il s'ensuit que

$$\left|\nabla u(\mathbf{x}_{jk}, t_{n+1})\cdot\mathbf{n}_{jk} - \frac{v_k^{n+1} - v_j^{n+1}}{d_{jk}}\right| \leq \left\|\nabla^2 u\right\|_{L^\infty([0,T]\times\mathcal{T})^4} h.$$

Cela implique que $\left|t_{jk}^n\right| \leq 2\left\|\nabla^2 u\right\|_{L^\infty([0,T]\times\mathcal{T})^4} h$. Or

$$\|t^n\|_{L^2(\mathcal{T})} \leq \max_{jk}\left(\left|t_{jk}^n\right|\right) \sqrt{\sum_{jk} l_{jk} d_{jk}} \leq \max_{jk}\left(\left|t_{jk}^n\right|\right) \sqrt{\sum_j s_j} \max_j \left(\frac{\sum_k l_{jk} d_{jk}}{s_j}\right).$$

Avec les hypothèses usuelles sur le maillage, on obtient

$$\|t^n\|_{L^2(\mathcal{T})} \leq K \left\|\nabla^2 u\right\|_{L^\infty([0,T]\times\mathcal{T})^4} h \tag{5.36}$$

où $K > 0$ ne dépend que du maillage.
La preuve est terminée. □

A présent que la consistance est établie, il reste à utiliser une nouvelle fois la stabilité pour obtenir la convergence.

Théorème 5. *Soit $T > 0$ un temps final donné. Sous les hypothèses précédentes, il existe une constante $C > 0$ telle que*

$$\|u_h^n - v_h^n\|_{L^2(\mathcal{T})} \leq C(\Delta t + h). \tag{5.37}$$

Démonstration. On définit les différences $e_j^n = v_j^n - u_j^n$ et $f_{jk}^n = q_{jk^n} - p_{jk}^n$ qui vérifient

$$\begin{cases} \dfrac{e_j^{n+1} - e_j^n}{\Delta t} - \dfrac{1}{s_j} \displaystyle\sum_k l_{jk} f_{jk}^{n+1} = r_j^n, & \forall j, \\ f_{jk}^{n+1} - \dfrac{e_k^{n+1} - e_j^{n+1}}{d_{jk}} = t_{jk}^n, & \forall (j,k). \end{cases} \tag{5.38}$$

La condition initiale (5.33) devient $e_j^0 = 0$ pour tout j. Il n'y a pas de condition initiale pour f_{jk}^0. On peut alors reprendre l'analyse de la stabilité qui donne lieu à (5.34) sous la forme suivante : on multiplie la première équation de (5.38) par $\Delta t s_j e_j^{n+1}$ et on somme ; dans le même temps multiplie la deuxième équation de (5.38) par $\Delta t l_{jk} d_{jk} f_{jk}^{n+1}$ et on somme. On obtient

$$\begin{aligned} &\frac{1}{2}\sum_j s_j \left|e_j^{n+1}\right|^2 + \Delta t \sum_{j<k} \frac{l_{jk}}{d_{jk}} \left|e_j^{n+1} - e_k^{n+1}\right|^2 + \sum_j s_j \left|e_j^{n+1} - e_j^n\right|^2 \\ &= \frac{1}{2}\sum_j s_j \left|e_j^n\right|^2 + \Delta t \sum_j s_j r_j^n e_j^{n+1} + \Delta t \sum_{jk} l_{jk} d_{jk} t_{jk}^n f_{jk}^{n+1} \\ &= \frac{1}{2}\sum_j s_j \left|e_j^n\right|^2 + \Delta t \sum_j s_j r_j^n \left(e_j^{n+1} - e_j^n\right) + \Delta t \sum_j s_j r_j^n e_j^n + \Delta t \sum_{jk} l_{jk} d_{jk} \left|t_{jk}^n\right|^2 + \Delta t \sum_{jk} l_{jk} t_{jk}^n \left(e_k^{n+1} - e_j^{n+1}\right) \end{aligned} \tag{5.39}$$

après élimination des f_{jk}^{n+1} par la deuxième équation de (5.38). On a les différentes inégalités de type Minkovski [1]

$$\Delta t \sum_j s_j r_j^n \left(e_j^{n+1} - e_j^n\right) \leq \frac{1}{2}\sum_j s_j \left|e_j^{n+1} - e_j^n\right|^2 + \frac{1}{2}\Delta t^2 \sum_j s_j \left|r_j^n\right|^2,$$

$$\Delta t \sum_j s_j r_j^n e_j^n \leq \frac{1}{2}\Delta t \sum_j s_j \left|r_j^n\right|^2 + \frac{1}{2}\Delta t \sum_j s_j \left|e_j^n\right|^2,$$

et

$$\Delta t \sum_{jk} l_{jk} t_{jk}^n \left(e_k^{n+1} - e_j^{n+1}\right) \leq \frac{1}{2}\Delta t \sum_{j<k} \frac{l_{jk}}{d_{jk}} \left|e_j^{n+1} - e_k^{n+1}\right|^2 + \frac{1}{2}\Delta t \sum_{jk} l_{jk} d_{jk} \left|t_{jk}^n\right|^2.$$

En insérant ces inégalités dans l'expression précédente (5.39) on obtient après quelques simplifications évidentes

$$\frac{1}{2}\sum_j s_j \left|e_j^{n+1}\right|^2 \leq \frac{1}{2}(1 + \Delta t)\sum_j s_j \left|e_j^n\right|^2 + \frac{1}{2}\left(\Delta t + \Delta t^2\right)\Delta t \sum_j s_j \left|r_j^n\right|^2 + \frac{1}{2}\Delta t \sum_{jk} l_{jk} d_{jk} \left|t_{jk}^n\right|^2$$

ou plus simplement

$$\left\|e_h^{n+1}\right\|_{L^2(\mathcal{T})}^2 \leq (1 + \Delta t)\left\|e_h^n\right\|_{L^2(\mathcal{T})}^2 + \Delta t \left((1 + \Delta t)\left\|r_h^{n+1}\right\|_{L^2(\mathcal{T})}^2 + |||t_h^{n+1}|||_{L^2(\mathcal{T})}^2\right).$$

Utilisant à présent les estimations de consistance, on obtient $\left\|e_h^{n+1}\right\|_{L^2(\mathcal{T})}^2 \leq e^{\Delta t}\left\|e_h^n\right\|_{L^2(\mathcal{T})}^2 + K\Delta t\left(\Delta t + h\right)^2$ pour une constante $K > 0$ qui dépend de u, et pour $0 < \Delta t < 1$. D'où $\|e_h^n\|_{L^2(\mathcal{T})}^2 \leq \Delta t K \sum_{p=0}^{n-1} e^{p\Delta t}\left(\Delta t + h\right)^2$. Soit un temps final donné $T > 0$. Pour $n\Delta t \leq T$ on peut écrire $\|e_h^n\|_{L^2(\mathcal{T})}^2 \leq Q\left(\Delta t + h\right)^2$. La preuve est terminée. □

Remarque 10. *Le résultat de convergence (5.37) est encore vrai pour $u \in H^2([0,T]\times\mathcal{T})$. On peut se référer au théorème 3.4 page 55 de [15] pour les idées principales. C'est un peu plus technique en ce qui concerne l'étude des erreurs de troncature, mais est strictement identique en ce qui concerne le schéma lui-même.*

1. On entend par là toute inégalité de la forme $ab \leq \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2$ pour $\varepsilon > 0$ bien choisi, a et b étant quelconques.

Bibliographie

[1] G. Allaire, Analyse numérique et optimisation, Editions de l'Ecole Polytechnique, 2012.

[2] G. Allaire, X. Blanc, F. Golse et B. Després, Transport et diffusion, cours de l'Ecole polytechnique, 2013.

[3] D. Bouche, J.-M. Ghidaglia, F. Pascal, Error estimate and the geometric corrector for the upwind finite volume method applied to the linear advection equation, SIAM J. Numer. Anal., 43(2), p. 578-603, 2005.

[4] F. Boyer, Aspects théoriques et numériques de l'équation de transport, Université de Aix-Marseille, en ligne http ://www.cmi.univ-mrs.fr/ fboyer/en/accueil.

[5] Brenner, V. Thomée et L. Wahlbin, Besov Spaces and Applications to Difference Methods for Initial Value Problems, Lecture Notes in Math. 434, Springer-Verlag, Berlin, New York, 1975.

[6] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Springer Verlag, (2010).

[7] J.G. Charney, R. Fjortoft et J. von Neumann, Numerical integration of the barotropic vorticity equation, vol. 2, 4, 237-254, Tellus, 1950.

[8] P. G. Ciarlet, The Finite Element Method for Elliptic Problems, North-Holland, Amsterdam, 1978

[9] A. Cohen, Approximation variationnelle des fonctions, Master de la modélisation, LJLL-UPMC.

[10] R. Courant, K. O. Fredrichs et H. Lewy (1928), "Uber die Differenzengleichungen der Mathematischen Physik", Math. Ann, vol.100, p.32, 1928.

[11] R. Dautray et J.L. Lions, Analyse mathématique et calcul numérique pour les sciences et les techniques. Vol. 9. (French) [Mathematical analysis and computing for science and technology. Vol. 9] Évolution : numérique, transport, 1985.

[12] Despres, Bruno Lax theorem and finite volume schemes. Math. Comp. 73 (2004), no. 247, 1203-1234.

[13] B. Després, Uniform asymptotic stability of Strang's explicit compact schemes for linear advection, Siam J. Numer. Anal., Vol. 47, No. 5, pp. 3956-3976

[14] A. Ern and J.-L. Guermond, Theory and Practice of Finite Elements, vol. 159 of Applied Mathematical Series, Springer, New York, 2004.

[15] R. Eymard, T. Gallouët, et R. Herbin, Finite volume methods, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 2000, pp. 713-1020.

[16] P. Frey et P.L. George, Mesh generation. Application to finite elements. Second edition. ISTE, London ; John Wiley & Sons, Inc., Hoboken, NJ, 2008.

[17] V. Girault et P.A. Raviart, Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms. Berlin-Heidelberg-New York-Tokyo, Springer-Verlag 1986

[18] E. Giusti, Minimal surfaces and functions of bounded variation. Birkhäuser Verlag, Basel, 1984.

[19] E. Godlewski et P. A. Raviart, Numerical Approximation of Hyperbolic Systems of Conservation Laws, Appl. Math. Sci. 118, Springer-Verlag, New York, 1996

[20] S. Godunov et Ryaben'kii, Introduction to the theory of difference schemes, Fizmatgiz, 1962.

[21] Harten, Ami On a class of high resolution total-variation-stable finite-difference schemes. With an appendix by Peter D. Lax. SIAM J. Numer. Anal. 21 (1984), no. 1, 123.

[22] F. Hermeline, Two Coupled Particle-Finite Volume Methods Using Delaunay-Voronoi Meshes for the Approximation of Vlasov-Poisson and Vlasov-Maxwell Equations, Journal of Computational Physics, Volume 106, Issue 1, May 1993, Pages 1-18

[23] A. Iserles et G. Strang, The optimal accuracy of difference schemes, Trans. of the AMS, Vol. 277, 2, 198, 779–803, 1983.

[24] S. Jaouen et F. Lagoutière, Numerical transport of an arbitrary number of components. Comput. Methods Appl. Mech. Engrg. 196 (2007), no. 33-34, 3127-3140.

[25] T. Kato, Perturbation theory for linear operators, Springer, 1995.

[26] Flux-corrected transport. Principles, algorithms, and applications. Edited by D. Kuzmin, R. Löhner et S. Turek. Scientific Computation. Springer-Verlag, Berlin, 2005.

[27] P.D. Lax et B. Wendroff, On the stability of difference schemes, Comm. Pure and Appl. Math., 15 1962, 363–371.

[28] P. Lesaint et P. A. Raviart, On a finite element method for solving the neutron transport equation, in Mathematical Aspects of Finite Elements in Partial Differential Equations, Academic Press, New York, 1974, pp. 89-123

[29] R.J. LeVeque, Numerical methods for conservation laws, (ETHZ Zurich, Birkhauser, Basel 1992).

[30] Pietro Perona and Jitendra Malik (November 1987). "Scale-space and edge detection using anisotropic diffusion". Proceedings of IEEE Computer Society Workshop on Computer Vision, pp. 16-22.

[31] S. Osher et P.K. Sweby, Recent developments in the numerical solution of nonlinear conservation laws. The state of the art in numerical analysis (Birmingham, 1986), 681-701, Inst. Math. Appl. Conf. Ser. New Ser., 9, Oxford Univ. Press, New York, 1987.

[32] W. Reed et T. Hill. Triangular mesh methods for the neutron transport equation. Technical Report Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.

[33] Richtmyer R. D. Richtmyer et K. W. Morton, Difference methods for initial-value problems, Interscience Publishers, 1957.

[34] G. Strang, Trigonometric polynomials and difference methods of maximum accuracy, J. Math. Phys, 41, 147–520, 1962.

[35] V. Thomée, Stability of difference schemes in the maximum-norm, J. Differential Equations, 1 (1965), pp. 273-292.

[36] B. van Leer, Towards the ultimate conservative difference scheme. V. A second-order sequel to Godunov's method [J. Comput. Phys. 32 (1979), no. 1, 101-136].

[37] R.F. Warming et R.M. Beam, Recent advances in the development of implicit schemes for the equations of gas dynamics, Seventh International Conf. on Numerical Methods in Fluid Dynamics, 429–433, Lecture Notes in Physics, Springer, 1981.

[38] B. Wendroff et A. B. White, A supraconvergent scheme for nonlinear hyperbolic systems, Comput. Math. Appl., 18, pp 761-767 (1989).