Équations aux dérivées partielles et leurs approximations.

Sommaire

Table des figures

Chapitre 1

Introduction générale

Soit d, m, n, s des entiers naturels non nuls. Soit Ω un ouvert de $\mathbb{R}^d$.

On appelle système d'*équations aux dérivées partielles* (EDP) à coefficients réels dans $\mathbb{R}^d$ de taille n d'ordre m une relation de la forme

$$\begin{aligned}\varphi\Big(x, u(x), (\partial_i u(x))_{i\in\{1,\dots,d\}}, \left(\partial^2_{i,j}u(x)\right)_{(i,j)\in\{1,\dots,d\}^2},\\ \dots, \left(\partial^m_{i_1,i_2,\dots,i_m}u(x)\right)_{(i_1,i_2,\dots,i_m)\in\{1,\dots,d\}^m}\Big) = 0\end{aligned} \tag{1.1}$$

où

$$\begin{cases} u : \Omega \subset \mathbb{R}^d \longrightarrow \mathbb{R}^n \text{ est la solution de l'EDP (du système d'EDP),}\\ \varphi : \Omega \times \mathbb{R}^n \times \mathbb{R}^n \times \cdots \times \mathbb{R}^n \longrightarrow \mathbb{R}^s, \end{cases}$$

Habituellement, on aura $s = n$ (et on aura le même nombre d'équations et d'inconnues).

On dit que l'EDP est *linéaire* si et seulement si $\varphi(x, \cdot)$ l'est pour tout $x \in \Omega$. On dit que l'EDP est *homogène* si et seulement si 0 en est solution. On dit que l'équation est *scalaire* si et seulement si $s = n = 1$.

Nous n'étudierons dans ce cours que des EDP d'ordres 1 et 2.

1.1 Classification des EDP scalaires linéaires d'ordre 2

On considère une EDP scalaire linéaire d'ordre 2 :

$$a(x)u(x) + \sum_{i=1}^{d} b_i(x)\partial_i u(x) + \sum_{i=1}^{d}\sum_{j=1}^{d} c_{i,j}(x)\partial^2_{i,j}u(x) = f(x). \tag{1.2}$$

Notons $C(x)$ la matrice $(c_{i,j}(x))_{i,j=1}^{d}$. À une modification (qui n'a pas d'influence sur l'EDP si la solution en est de classe $\mathcal{C}^2$) près, c'est une matrice symétrique ; elle est donc diagonalisable et ses

valeurs propres sont réelles. Notons-les $(\lambda_i(x))_{i=1}^d$, et notons $d_+(x)$ le nombre de valeurs propres strictement positives, $d_-(x)$ le nombre de valeurs propres strictement négatives (en tenant compte de leur multiplicité) et $d_0(x)$ la multiplicité de la valeur propre 0 : $d = d_0(x) + d_-(x) + d_+(x)$ $\forall x \in \Omega$.
On dit que l'EDP (1.2) est *elliptique en* $x \in \Omega$ si et seulement si

$$\begin{aligned} & d_+(x) = d \\ \text{ou} \quad & d_-(x) = d \end{aligned}$$

(la forme $(y_i)_{i=1}^d \longmapsto \sum_{i=1}^d b_i(x)y_i + \sum_{i=1}^d \sum_{j=1}^d y_j c_{i,j}(x) y_i$ définit une quadrique elliptique).
On dit que l'EDP (1.2) est *hyperbolique en* $x \in \Omega$ si et seulement si

$$\begin{aligned} & d_+(x) = d-1 \quad \text{et} \quad d_-(x) = 1 \\ \text{ou} \quad & d_+(x) = 1 \quad \text{et} \quad d_-(x) = d-1. \end{aligned}$$

On dit que l'EDP (1.2) est *parabolique en* $x \in \Omega$ si et seulement si

$$d_0(x) > 0.$$

Exercice 1
Déterminer le type de l'équation de Tchaplyguin sur $\mathbb{R}^2$:

$$\partial^2_{1,1} u(x_1, x_2) + x_1 \partial^2_{2,2} u(x_1, x_2) = f(x_1, x_2).$$

1.2 Exemples d'EDP tirés de la physique

1.2.1 Déformation d'un fil élastique

Considérons un fil élastique mono-dimensionnel dans le segment $[0, 1]$ maintenu en $x = 0$ et en $x = 1$ à l'altitude 0 et soumis à un chargement $f(x)$ perpendiculaire au segment, à l'équilibre. Notons $u(x)$ l'altitude du fil à l'abscisse x. L'altitude du fil est solution du problème

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

où $c(x)$ est donné par les caractéristiques du matériau qui constitue le fil (c'est le coefficient d'élasticité). Il s'agit d'un problème de nature elliptique. On se posera dans ce cours les questions de l'existence d'une solution, de son unicité, et de son calcul (ou calcul approché). Cette EDP est en fait une équation différentielle ordinaire (EDO), mais ce n'est pas un problème de Cauchy (donc, le théorème de Cauchy-Lipschitz ne permet pas de conclure directement à l'existence d'une solution).

1.2.2 Déformation d'une membrane élastique

On considère cette fois une membrane élastique horizontale à l'équilibre soumise à un chargement vertical dans un ouvert (borné) Ω de $\mathbb{R}^2$ et maintenue dans une position fixe (à l'altitude 0) sur le bord de Ω. L'altitude de la membrane est alors solution de

$$\begin{cases} -\Delta u(x) + c(x)u(x) = f(x) \quad \forall x \in \Omega, \\ u(x) = 0 \quad \forall x \in \partial\Omega. \end{cases}$$

C'est un problème elliptique. Nous nous poserons à propos de cette équation les mêmes questions que pour le fil élastique. La quantité Δu est appelé laplacien de u et vaut $\partial^2_{1,1}u + \partial^2_{2,2}u$, que nous noterons dans la suite $\partial^2_{x,x}u + \partial^2_{y,y}u$.

1.2.3 Vibration d'une corde

Le fil de la sous-section 1.2.1 n'est plus ici supposé à l'équilibre : on veut précisément étudier les phénomènes instationnaires, en se donnant des conditions initiales pour le dispositif. En faisant l'« hypothèse des petites déformations [1] », l'équation mathématique associée à ce problème est

$$\begin{cases} \partial^2_{t,t}u(t,x) - \partial^2_{x,x}u(t,x) = f(t,x) \quad \forall t \in \mathbb{R}^*_+, \forall x \in]0,1[, \\ u(t,0) = u(t,1) = 0 \quad \forall t \in \mathbb{R}_+, \\ u(0,x) = u^0(x) \quad \forall x \in]0,1[, \\ \partial_t u(0,x) = u^1(x) \quad \forall x \in]0,1[\end{cases}$$

où t est la variable de temps (et le chargement dépend à la fois du temps et de l'espace). C'est un problème hyperbolique. On se posera à son sujet les mêmes questions que dans les exemples précédents, et l'on se demandera aussi si la solution mathématique est stable au cours du temps.

Ce problème admet bien entendu des généralisations en dimension 2 (vibration d'une membrane) et en dimension 3, au même titre que le problème de la déformation d'un fil élastique.

1.2.4 Diffusion de la chaleur

On considère encore un fil sur le segment $[0,1]$, et l'on s'intéresse cette fois non pas à son déplacement mais à sa température. On note $u(t,x)$ la température du fil à l'instant t et au point $x \in [0,1]$. Ce problème physique est modélisé (dans un cadre simplifié, en utilisant la loi de Fourier) par

$$\begin{cases} \partial_t u(t,x) - \kappa\partial^2_{x,x}u(t,x) = f(t,x) \quad \forall t \in \mathbb{R}^*_+, \forall x \in]0,1[, \\ u(t,0) = u(t,1) = 0 \quad \forall t \in \mathbb{R}_+, \\ u(0,x) = u^0(x) \quad \forall x \in]0,1[. \end{cases}$$

1. C'est-à-dire en supposant que u, $\partial_x u$ et $\partial^2_{x,x}u$ sont *petits*.

$f(t,x)$ représente un terme de chauffage. Le coefficient κ est le coefficient de conductivité thermique, supposé constant [2] et positif. La donnée u^0 est la température initiale en tout point de $]0,1[$. La condition $u(t,0) = u(t,1) = 0\ \forall t \in \mathbb{R}_+$ indique que la température est fixée à 0 pour tout temps sur les bords du domaine. C'est un problème parabolique. Il se généralise lui aussi en dimensions supérieures. Les questions que nous nous poserons et auxquelles nous tâcherons de répondre sont encore celles de l'existence, de l'unicité, de la stabilité. Il faut remarquer que l'équation de la chaleur est en rapport avec une équation elliptique par le biais suivant : si la solution $u(t,x)$ converge en temps infini vers une fonction qui ne dépend pas du temps (on peut supposer que f ne dépend que de x, pour fixer les idées), la limite en temps infini de cette solution, notons-la $v(x)$, vérifie

$$\begin{cases} -\kappa\partial^2_{x,x} v(x) = f(x), \\ v(0) = v(1) = 0, \end{cases}$$

qui est l'archétype du problème elliptique. Pour une démonstration rigoureuse de ce résultat, voir l'examen de juin 2004 et celui de mai 2006 (et leur corrigé).

1.2.5 Évolution du trafic routier sur une autoroute

Assimilons (!) la répartition des automobiles sur une autoroute de longueur infinie à une densité de répartition sur $\mathbb{R}$. Notons $\rho(t,x)$ cette densité, a priori fonction du temps et de la position. Notons encore $v(t,x)$ la vitesse locale des automobiles. L'équation de transport des automobiles est

$$\partial_t \rho(t,x) + \partial_x(\rho v)(t,x) = 0.$$

Supposons (pour simplifier...) que les conducteurs adaptent leur vitesse à la densité locale d'automobiles : $v(t,x) = V(\rho(t,x))$. Il est logique de choisir une fonction V décroissante [3]. On peut de plus modéliser l'apparition d'un bouchon lorsque la densité de voitures est trop importante par l'hypothèse mathématique qu'il existe une densité de saturation ρ_s telle que $V(\rho_s) = 0$. Ceci conduit à un *flux* $q = \rho v = \rho V(\rho)$ d'allure représentée sur la figure 1.1.

2. Si l'on ne suppose pas que ce coefficient est constant mais qu'il dépend de u et x, l'équation de la chaleur s'écrit $\partial_t u - \partial_x(\kappa(u,x)\partial_x u) = f$.

3. De plus, si les distances sont mesurées en kilomètres et les temps en heures, si l'autoroute est française et les automobilistes disciplinés (ne sont pas français), on aura $\lim_{\rho\to 0} V(\rho) = 130$.

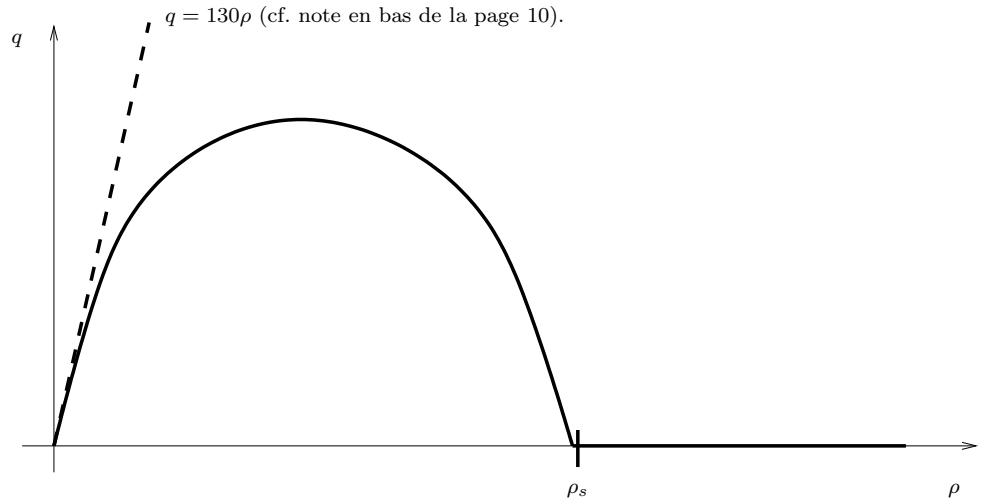

FIGURE 1.1 – Allure générale d'un flux autoroutier simplifié.

1.2.6 Hydrodynamique compressible

Soit $\rho(t,x) \in \mathbb{R}$, $u(t,x) \in \mathbb{R}^3$ et $e(t,x) \in \mathbb{R}$ les densité, vitesse et densité massique d'énergie totale d'un fluide compressible (dans $\mathbb{R}^3$). Les équations de conservation de la masse, de la quantité de mouvement et de l'énergie totale forment le système d'EDP d'Euler

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho u) = 0, \\ \partial_t(\rho u) + \operatorname{div}(\rho u \otimes u + pI) = 0, \\ \partial_t(\rho e) + \operatorname{div}(\rho u e + pu) = 0, \end{cases}$$

p étant la pression dans le fluide (supposé ici newtonien) : par exemple, pour un gaz parfait, $p = (\gamma - 1)\rho(e - u^2/2)$. Lorsque $\gamma p/\rho > 0$ [4], ce système est hyperbolique (comme il ne s'agit pas d'une EDP d'ordre 2, la définition de l'hyperbolicité est ici différente de celle que nous avons introduite : voir le chapitre 3 pour plus de précisions).

1.2.7 Évolution du prix d'une option

Notons $c(t,s)$ le prix d'achat d'une option pour la somme s au temps t (avant un temps final). Une équation régissant l'évolution de ce prix est proposée par le modèle de Black et Scholes :

$$\partial_t c(t,s) + \frac{s^2\sigma^2(t,s)}{2}\partial^2_{s,s} c(t,s) + r(t)s\partial_s c(t,s) - r(t)c(t,s) = 0.$$

Les paramètres σ et r représentent respectivement la volatilité du marché et le taux d'intérêt de l'actif sans risque.

En supposant les coefficients σ et r constants, on peut montrer que cette EDP est équivalente à l'équation de la chaleur.

4. C'est le cas par exemple pour $\gamma > 1$ lorsque $\rho > 0$ et $e - u^2/2 > 0$.

1.3 Encore quelques généralités

Ce cours a pour but d'apporter quelques réponses aux questions posées dans cette introduction, qui concernent le caractère *bien posé* de problèmes d'EDP, l'existence de solutions, leur unicité, leur stabilité en temps, stabilité par rapport aux données du problème (étude qualitative). On insistera aussi sur le calcul (approché le plus souvent) de ces solutions ; ce sera l'occasion d'introduire successivement les méthodes de différences finies, de volumes finis et d'éléments finis.

Une particularité de l'étude des EDP est qu'il n'existe pas de résultat général utilisable permettant de prédire l'existence ou l'unicité d'une solution à (1.1). Voici à titre indicatif LE théorème général relatif à cette question.

Théorème 1 (Cauchy-Kowalewskaya)
Considérons le système d'EDP

$$\partial_t u_j = \sum_{i=1}^{d} \sum_{k=1}^{n} \alpha_{i,k}^{j}(x, u_1, u_2, \ldots, u_n)\partial_i u_k + \beta^j(x, u_1, u_2, \ldots, u_n), \quad j = 1, \ldots, n$$

avec les données initiales $u_j(0, x) = 0\ \forall x, \forall j = 1, \ldots, n$. Supposons que les coefficients $\alpha_{i,k}^{j}$ et β^j sont analytiques réels [5] au voisinage de $(x, u_1, u_2, \ldots, u_n) = (0, 0, 0, \ldots, 0)$.
Alors, le problème de Cauchy considéré a une unique solution $u_j(t, x)$ analytique réelle au voisinage de $(0, 0)$. □

Il s'agit d'un résultat beaucoup plus faible que celui qui concerne les EDO que nous rappelons pour comparaison.

Théorème 2 (Cauchy-Lipschitz)
Considérons l'EDO

$$x'(t) = f(t, x(t))$$

posée dans $I \times \Omega$ où I est un intervalle ouvert de $\mathbb{R}$ contenant 0 et Ω est un ouvert de $\mathbb{R}^d$, avec la donnée initiale $x(0) = x^0 \in \Omega$. Supposons que f est continue et qu'elle est localement lipschitzienne par rapport à sa seconde variable sur $I \times \Omega$. [6]
Alors, il existe une unique solution maximale à l'EDO, définie sur l'intervalle maximal I_M, intervalle ouvert tel que $0 \in I_M \subset I$. □

5. Rappel : notations de Schwartz. Soit $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, soit $n = (n_1, n_2, \ldots, n_d) \in \mathbb{N}^d$. On note $x^n = x_1^{n_1} x_2^{n_2} \ldots x_d^{n_d}$. Soit Ω un ouvert de $\mathbb{R}^d$ qui contient 0. Soit $f : \Omega \longrightarrow \mathbb{R}$. On dit que f est *analytique réelle* au voisinage de 0 si et seulement s'il existe un voisinage $\mathcal{O}$ de 0 et $(c_n)_{n\in\mathbb{N}^d}$ tels que $\forall x \in \mathcal{O}$,

$$f(x) = \sum_{n\in\mathbb{N}^d} c_n x^n.$$

6. On entend par ceci : f est localement en (t, x) lipschitzienne par rapport à x, c'est-à-dire que $\forall (t_0, x_0) \in I\times\Omega$, $\exists \mathcal{O}(t_0, x_0)$ voisinage de (t_0, x_0) et $\kappa(t_0, x_0) \in \mathbb{R}$ tels que $|f(t, y) - f(t, x)| \leq \kappa(t_0, x_0)\,|y - x|\ \forall t, x, y$ tels que $(t, x) \in \mathcal{O}(t_0, x_0)$ et $(t, y) \in \mathcal{O}(t_0, x_0)$.

Une étude des EDP dans toute leur généralité est impossible. Nous nous intéresserons donc à certaines catégories d'EDP, séparément. Nous analyserons premièrement les EDP paraboliques (en focalisant sur l'équation de la chaleur), puis les EDP hyperboliques, et enfin les EDP elliptiques. Nous aborderons dans le même ordre les méthodes de différences finies, de volumes finis et d'éléments finis.

Chapitre 2

Équations paraboliques

L'essentiel de ce chapitre concerne l'équation de la chaleur à coefficient de conductivité constant dans un intervalle borné en dimension 1 avec des conditions de bord de Dirichlet homogènes. Le problème mathématique considéré est

$$\begin{cases} \partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = f(t,x) \text{ dans }]0,T]\times]0,1[, \\ u(t,0) = 0 \quad \forall t \in]0,T], \\ u(t,1) = 0 \quad \forall t \in]0,T], \\ u(0,x) = u^0(x) \quad \forall x \in]0,1[\end{cases} \tag{2.1}$$

où T est un réel supposé strictement positif.

Des conditions de bord de type $u(t,0) = u_0(t)\ \forall t \in]0,T]$, $u(t,1) = u_1(t)\ \forall t \in]0,T]$ sont appelées *conditions de Dirichlet*; si $u_0(t) = 0$ et $u_1(t) = 0\ \forall t \in]0,T]$, ce qui est le cas dans le problème modèle (2.1), ces conditions sont dites de *Dirichlet homogènes*.

Remarque 1
A priori, le temps final T fait partie du problème : il s'agit de trouver le plus grand $T \in \mathbb{R}$ tel qu'il existe une solution sur $]0,T]$, ou $]0,T[$. Nous verrons que pour cette EDP (moyennant des hypothèses *ad hoc* sur les données), il existe une solution pour tout $T \in \mathbb{R}$. Une autre remarque, importante, concerne le domaine $]0,T]\times]0,1[$ sur lequel on cherche une solution de l'EDP. L'on pourrait aussi chercher une solution sur $[0,T]\times[0,1]$ (en précisant que sur les bords de ce domaine, les dérivées partielles seraient à comprendre « à gauche » ou « à droite »), mais, nous allons le voir, ce serait très réducteur en nous privant de toute une famille de solutions intéressantes, celles issues de conditions initiales non régulières. Ce choix nous imposera de préciser ultérieurement ce que nous entendons par condition limite, puisque la régularité au bord (en temps...) n'est pas garantie. □

Notons en premier lieu que l'EDP $\partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = f(t,x)$ est une EDP scalaire linéaire (non homogène sauf si f est identiquement nulle). En conséquence, si u vérifie $\partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = f(t,x)$ et si v vérifie $\partial_t v(t,x) - \kappa \partial^2_{x,x} v(t,x) = 0$, $w = u + \lambda v$ vérifie $\partial_t w(t,x) - \kappa \partial^2_{x,x} w(t,x) = f(t,x)$. Autrement dit, l'ensemble des solutions de $\partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = f(t,x)$ est un espace affine (vectoriel si $f = 0$). Cette remarque servira par la suite.

2.1 Existence et unicité d'une solution

La méthode que nous allons utiliser ici est, à peu de choses près, celle qu'a développée Fourier en 1822 dans sa Théorie analytique de la chaleur. Cette méthode est basée sur la remarque suivante : quel que soit $k \in \mathbb{N}$,

$$e^{-\kappa k^2 \pi^2 t} \sin(k\pi x)$$

est solution de

$$\begin{cases} \partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = 0 \text{ dans }]0,T]\times]0,1[, \\ u(t,0) = 0 \quad \forall t \in]0,T], \\ u(t,1) = 0 \quad \forall t \in]0,T]. \end{cases}$$

2.1.1 Une base hilbertienne de $L^2(]0,1[)$

Proposition 1
$(\sqrt{2}\sin(k\pi\cdot))_{k\in\mathbb{N}^*}$ est une base hilbertienne de $L^2(]0,1[)$. □

Démonstration
Soit $f \in L^2(]0,1[)$ et soit $\tilde{f}$ le prolongement par imparité de f sur $]-1,1[$: $\tilde{f} \in L^2(]-1,1[)$. Le théorème de Féjer assure que $((1/\sqrt{2})e^{ik\pi\cdot})_{k\in\mathbb{Z}}$ est une base hilbertienne de $L^2(]-1,1[)$. Donc, en posant, pour tout $k \in \mathbb{Z}$, $\widehat{\tilde{f}}(k) = \left(1/\sqrt{2}\right)\int_{-1}^{1} \tilde{f}(x)e^{-ik\pi x}\,dx$, on a

$$\tilde{f} = \frac{1}{\sqrt{2}} \sum_{k\in\mathbb{Z}} \widehat{\tilde{f}}(k) e^{ik\pi\cdot} \quad \text{dans } L^2(]-1,1[).$$

D'autre part,

$$\begin{aligned} \widehat{\tilde{f}}(k) = \frac{1}{\sqrt{2}}\int_{-1}^{1} \tilde{f}(x)e^{-ik\pi x}\,dx &= \frac{1}{\sqrt{2}}\int_{-1}^{0} \tilde{f}(x)e^{-ik\pi x}\,dx + \frac{1}{\sqrt{2}}\int_{0}^{1} \tilde{f}(x)e^{-ik\pi x}\,dx \\ &= \frac{1}{\sqrt{2}}\int_{-1}^{0} -f(-x)e^{-ik\pi x}\,dx + \frac{1}{\sqrt{2}}\int_{0}^{1} f(x)e^{-ik\pi x}\,dx \\ &= \frac{1}{\sqrt{2}}\int_{0}^{1} f(x)\left(e^{-ik\pi x} - e^{ik\pi x}\right)\,dx = -\frac{2i}{\sqrt{2}}\int_{0}^{1} f(x)\sin(k\pi x)\,dx. \end{aligned}$$

Ainsi, $\widehat{\tilde{f}}(-k) = -\widehat{\tilde{f}}(k)$ et

$$f = \sqrt{2} \sum_{k\in\mathbb{N}^*} \widehat{f}(k)\sin(k\pi\cdot)$$

dans $L^2(]0,1[)$ si l'on pose $\widehat{f}(k) = \sqrt{2}\int_0^1 f(x)\sin(k\pi x)\,dx$ pour $k \in \mathbb{N}^*$. □

Maintenant, un petit lemme :

Lemme 1
Soit $g \in \mathcal{C}^1([0,1])$ une fonction deux fois dérivable sur $]0,1[$ et à dérivée seconde dans $L^2(]0,1[)$ telle que $g(0) = g(1) = 0$. On a

$$\widehat{g''}(k) = -k^2\pi^2\widehat{g}(k) \quad \forall k \in \mathbb{N}^*.$$

□

Démonstration
Soit $g \in \mathcal{C}^0([0,1])$ une fonction dérivable sur $]0,1[$ et à dérivée dans $L^2(]0,1[)$. Pour $k \in \mathbb{N}^*$, on a

$$\begin{aligned}\widehat{g'}(k) &= \sqrt{2}\int_0^1 g'(x)\sin(k\pi x)\,dx \\ &= \sqrt{2}\left[g(x)\sin(k\pi x)\right]_0^1 - \sqrt{2}\int_0^1 g(x)k\pi\cos(k\pi x)\,dx \\ &= -\sqrt{2}k\pi\int_0^1 g(x)\cos(k\pi x)\,dx.\end{aligned}$$

Soit $g \in \mathcal{C}^1([0,1])$ une fonction deux fois dérivable sur $]0,1[$ et à dérivée seconde dans $L^2(]0,1[)$:

$$\begin{aligned}\widehat{g''}(k) &= -\sqrt{2}k\pi\int_0^1 g'(x)\cos(k\pi x)\,dx \\ &= -\sqrt{2}k\pi\left[g(x)\cos(k\pi x)\right]_0^1 - \sqrt{2}k^2\pi^2\int_0^1 g(x)\sin(k\pi x)\,dx \\ &= -\sqrt{2}k\pi\left[g(x)\cos(k\pi x)\right]_0^1 - k^2\pi^2\widehat{g}(k).\end{aligned}$$

Si g est de plus nulle en 0 et en 1, on a $\widehat{g''}(k) = -k^2\pi^2\widehat{g}(k)\ \forall k \in \mathbb{N}^*$. □

2.1.2 Unicité de la solution. Stabilité

Revenons maintenant au problème (2.1). Supposons que u est solution de ce problème : u est une fonction de $[0,T]\times[0,1]$ dans $\mathbb{R}$ dérivable une fois par rapport à sa première variable t et deux fois par rapport à sa seconde variable x sur $]0,T]\times[0,1]$ vérifiant l'EDP ainsi que les conditions aux limites. On fait de plus l'hypothèse que

$$\begin{aligned}&\partial_t u \in \mathcal{C}^0\left(]0,T]\times[0,1]\right), \\ &\partial^2_{x,x} u \in \mathcal{C}^0\left(]0,T]\times[0,1]\right).\end{aligned}$$

L'EDP donne $f = \partial_t u - \kappa\partial^2_{x,x}u \in \mathcal{C}^0(]0,T]\times[0,1])$, mais on demande de plus que $f \in \mathcal{C}^0([0,T]\times[0,1])$. Alors, $f(t,\cdot)$, $u(t,\cdot)$, $\partial_t u(t,\cdot)$ et $\partial^2_{x,x}u(t,\cdot)$ sont développables sur la base hilbertienne $(\sqrt{2}\sin(k\pi x))_{k\in\mathbb{N}^*}$, quel que soit $t \in]0,T]$. De plus, d'après le théorème de convergence dominée

de Lebesgue (ou plutôt, sa conséquence concernant la dérivation sous un signe d'intégrale), pour tout $k \in \mathbb{N}^*$, $\widehat{u(t,\cdot)}(k)$ est dérivable par rapport à t si $t > 0$, et

$$\partial_t \widehat{u(t,\cdot)}(k) = \sqrt{2} \int_0^1 \partial_t u(t,x) \sin(k\pi x)\, dx = \widehat{\partial_t u(t,\cdot)}(k)$$

et

$$\partial_t u(t,\cdot) = \sqrt{2} \sum_{k\in\mathbb{N}^*} \widehat{\partial_t u(t,\cdot)}(k) \sin(k\pi\cdot) = \sqrt{2} \sum_{k\in\mathbb{N}^*} \partial_t \widehat{u(t,\cdot)}(k) \sin(k\pi\cdot).$$

En effet :

- pour tout $t > 0$, $u(t,\cdot)\sin(k\pi\cdot)$ est mesurable ;
- pour tout $t > 0$, $u(t,\cdot)\sin(k\pi\cdot)$ est intégrable, donc... Il existe t tel que $u(t,\cdot)\sin(k\pi\cdot)$ est intégrable ;
- pour tout x, $u(\cdot,x)\sin(k\pi x)$ est dérivable : sa dérivée vaut $\partial_t u(\cdot,x)\sin(k\pi x)$;
- quel que soit $\epsilon \in]0,T[$, pour tout $t \in [\epsilon, T]$, $|\partial_t u(t,x)\sin(k\pi x)| \leq ||\partial_t u||_{L^\infty([\epsilon,T]\times[0,1])}$ pour tout x, et cette constante est intégrable sur $]0,1[$.

Par la suite, nous noterons plutôt $\widehat{u}(k)(t)$ les termes $\widehat{u(t,\cdot)}(k)$. On a donc

$$\partial_t \widehat{u}(k)(t) = \partial_t \widehat{u(t,\cdot)}(k) = \widehat{\partial_t u(t,\cdot)}(k),$$

la première égalité étant due à la nouvelle notation, la seconde étant conséquence du théorème de Lebesgue.

Par ailleurs, puisque $u(t,0) = u(t,1) = 0$ (condition de Dirichlet homogène), on a, d'après le lemme 1, $\widehat{\partial^2_{x,x} u}(k)(t) = -k^2\pi^2 \widehat{u}(k)(t)$ $\forall k \in \mathbb{N}^*$, $\forall t \in]0,T]$ sous l'hypothèse

$$u(t,\cdot) \in \mathcal{C}^1([0,1])\ \forall t \text{ et est pour tout } t \text{ une fonction deux fois dérivable sur }]0,1[\text{ et à dérivée seconde dans } L^2(]0,1[).$$

Ces propriétés sont bien assurées par les hypothèses faites en tête de cette section. L'EDP vérifiée par u se réécrit au moyen de sa série de Fourier

$$\sum_{k\in\mathbb{N}^*} \left[\widehat{u}(k)'(t) - \kappa \widehat{\partial^2_{x,x} u}(k)(t)\right] \sin(k\pi\cdot) = \sum_{k\in\mathbb{N}^*} \widehat{f}(k)(t) \sin(k\pi\cdot),$$

qui devient, compte tenu des remarques précédentes,

$$\sum_{k\in\mathbb{N}^*} \left[\widehat{u}(k)'(t) + \kappa k^2\pi^2 \widehat{u}(k)(t) - \widehat{f}(k)(t)\right] \sin(k\pi\cdot) = 0.$$

Comme $(\sqrt{2}\sin(k\pi\cdot))_{k\in\mathbb{N}^*}$ est une base hilbertienne, ceci signifie que chacun des coefficients de la somme ci-dessus est nul :

$$\widehat{u}(k)'(t) + \kappa k^2\pi^2 \widehat{u}(k)(t) = \widehat{f}(k)(t) \quad \forall k \in \mathbb{N}^*, \forall t \in]0,T].$$

Ces équations différentielles ordinaires admettent chacune une unique solution (sous réserve que l'on fournisse une donnée initiale) sous l'hypothèse que $\widehat{f}(k) \in \mathcal{C}^0([0,T])$ $\forall k \in \mathbb{N}^*$. Cette condition

est remplie dès que $f \in \mathcal{C}^0([0,T]\times[0,1])$: c'est ici qu'intervient cette hypothèse faite au début de la section. On a transformé le problème aux dérivées partielles en une famille d'équations différentielles ordinaires. Chacune de ces EDO se résout facilement, la solution, *unique à la donnée initiale* $\widehat{u}(k)(0)$ *près*, en est donnée par

$$\widehat{u}(k)(t) = \left[\widehat{u}(k)(0) + \int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right] e^{-\kappa k^2\pi^2 t}$$

(la fonction $\widehat{u}(k)(t)$ étant ainsi convenablement définie, bien sûr grâce à l'hypothèse de régularité sur f). Donc, sous les hypothèses faites sur u et f, on a

$$u(t,\cdot) = \sqrt{2}\sum_{k\in\mathbb{N}^*} \widehat{u}(k)(t)\sin(k\pi\cdot) \quad \text{dans } L^2(]0,1[)$$

avec

$$\widehat{u}(k)(t) = \left[\widehat{u}(k)(0) + \int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right] e^{-\kappa k^2\pi^2 t},$$
$$\widehat{f}(k)(t) = \sqrt{2}\int_0^1 f(t,x)\sin(k\pi x)\,dx,$$

les coefficients $\widehat{u}(k)(0)$ étant encore à définir. Le fait que la solution $u(t,\cdot)$ soit décrite dans $L^2(]0,1[)$ nous incite à donner à la prescription de la condition initiale le sens (à peine plus) faible d'égalité (forte) dans $L^2(]0,1[)$, exprimée sous la forme

$$\lim_{t\to 0^+} u(t,\cdot) = u^0 \quad \text{dans } L^2(]0,1[),$$

en supposant que $u^0 \in L^2(]0,1[)$. Ceci se réécrit

$$\lim_{t\to 0^+} \sqrt{2}\sum_{k\in\mathbb{N}^*} \widehat{u}(k)(t)\sin(k\pi\cdot) = \sqrt{2}\sum_{k\in\mathbb{N}^*} \widehat{u^0}(k)\sin(k\pi\cdot) \quad \text{dans } L^2(]0,1[),$$

dont une conséquence est que

$$\lim_{t\to 0^+} \widehat{u}(k)(t) = \widehat{u^0}(k) \quad \forall k \in \mathbb{N}^*$$

puisque les $\sin(k\pi\cdot)$ forment une base hilbertienne (ceci peut aussi se voir rapidement grâce au théorème de Bessel-Parseval). On a donc $\widehat{u^0}(k) = \widehat{u}(k)(0)$ et

$$\widehat{u}(k)(t) = \left[\widehat{u^0}(k) + \int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right] e^{-\kappa k^2\pi^2 t} \quad \forall k \in \mathbb{N}^*.$$

Regroupons maintenant les résultats que nous avons obtenus.

Proposition 2
Avec $f \in \mathcal{C}^0([0,T]\times[0,1])$ et $u^0 \in L^2(]0,1[)$ le problème (2.1) [1] admet au plus une solution u dérivable par rapport à t, 2 fois dérivable par rapport à x sur $]0,T]\times[0,1]$ et telle que $\partial_t u$, $\partial_x u$ et $\partial^2_{x,x} u$ sont dans $\mathcal{C}^0(]0,T]\times[0,1])$. □

1. La condition initiale étant vérifiée au sens $L^2(]0,1[)$.

On montre de plus le résultat de stabilité suivant.

Proposition 3
Soit $f \in \mathcal{C}^0([0,T]\times[0,1])$ et soit $u^0 \in L^2(]0,1[)$. Soit u une solution de (2.1) dérivable par rapport à t, 2 fois dérivable par rapport à x sur $]0,T]\times[0,1]$ et telle que $\partial_t u(t,\cdot)$, $\partial_x u(t,\cdot)$ et $\partial^2_{x,x} u(t,\cdot)$ sont dans $\mathcal{C}^0(]0,T]\times[0,1])$. Cette solution u vérifie

$$||u(t,\cdot)||_{L^2(]0,1[)} \leq \left|\left|u^0\right|\right|_{L^2(]0,1[)} + \frac{1}{\sqrt{2\kappa\pi}} ||f||_{L^2(]0,T[\times]0,1[)} \quad \forall t \in]0,T].$$

□

Démonstration
Le théorème de Bessel-Parseval affirme que, $\forall t$,

$$||u(t,\cdot)||_{L^2(]0,1[)} = \left|\left| \left(\widehat{u}(k)(t)\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} = \sqrt{\sum_{k\in\mathbb{N}^*} |\widehat{u}(k)(t)|^2}.$$

Or

$$\widehat{u}(k)(t) = \left[\widehat{u}(k)(0) + \int_0^t \widehat{f}(k)(s) e^{\kappa k^2\pi^2 s}\, ds\right] e^{-\kappa k^2\pi^2 t}.$$

Donc

$$\begin{aligned}
&\left|\left| \left(\widehat{u}(k)(t)\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} \\
&\qquad \leq \left|\left| \left(\widehat{u^0}(k) e^{-\kappa k^2\pi^2 t}\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} + \left|\left| \left(\int_0^t \widehat{f}(k)(s) e^{\kappa k^2\pi^2 s}\, ds e^{-\kappa k^2\pi^2 t}\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} \\
&\qquad\qquad \leq \left|\left| \left(\widehat{u^0}(k)\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} + \sqrt{\sum_{k\in\mathbb{N}^*} \int_0^t |\widehat{f}(k)(s)|^2\, ds \int_0^t e^{2\kappa k^2\pi^2(s-t)}\, ds}
\end{aligned}$$

d'après l'inégalité de Cauchy-Schwarz. Comme

$$\begin{aligned}
&\int_0^t |\widehat{f}(k)(s)|^2\, ds \int_0^t e^{2\kappa k^2\pi^2(s-t)}\, ds \\
&\qquad = \int_0^t |\widehat{f}(k)(s)|^2\, ds \times \frac{1}{2\kappa k^2\pi^2} \left[e^{2\kappa k^2\pi^2(s-t)}\right]_0^t \\
&\qquad\qquad = \int_0^t |\widehat{f}(k)(s)|^2\, ds \times \frac{1}{2\kappa k^2\pi^2}(1 - e^{-2\kappa k^2\pi^2 t}),
\end{aligned}$$

on a la majoration

$$\int_0^t |\widehat{f}(k)(s)|^2\, ds \int_0^t e^{2\kappa k^2\pi^2(s-t)}\, ds \leq \int_0^t |\widehat{f}(k)(s)|^2\, ds \times \frac{1}{2\kappa\pi^2}$$

pour tout $k \in \mathbb{N}^*$, et finalement

$$\left|\left| \left(\widehat{u}(k)(t)\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} \leq \left|\left| \left(\widehat{u^0}(k)\right)_{k\in\mathbb{N}^*} \right|\right|_{l^2} + \frac{1}{\sqrt{2\kappa\pi}} \sqrt{\sum_{k\in\mathbb{N}^*} \int_0^t |\widehat{f}(k)(s)|^2\, ds},$$

soit

$$\left|\left|\left(\widehat{u}(k)(t)\right)_{k\in\mathbb{N}^*}\right|\right|_{l^2} \leq \left|\left|\left(\widehat{u^0}(k)\right)_{k\in\mathbb{N}^*}\right|\right|_{l^2} + \frac{1}{\sqrt{2\kappa}\pi}\sqrt{\int_0^t \sum_{k\in\mathbb{N}^*} |\widehat{f}(k)(s)|^2\, ds}$$

(grâce au théorème de convergence monotone de Lebesgue). À nouveau grâce à l'égalité de Bessel-Parseval,

$$\left|\left|\left(\widehat{u}(k)(t)\right)_{k\in\mathbb{N}^*}\right|\right|_{l^2} \leq \left|\left|\left(\widehat{u^0}(k)\right)_{k\in\mathbb{N}^*}\right|\right|_{l^2} + \frac{1}{\sqrt{2\kappa}\pi}\sqrt{\int_0^t ||f(s,\cdot)||^2_{L^2(]0,1[)}\, ds},$$

ce qui donne enfin

$$\left|\left|\left(\widehat{u}(k)(t)\right)_{k\in\mathbb{N}^*}\right|\right|_{l^2} \leq \left|\left|\left(\widehat{u^0}(k)\right)_{k\in\mathbb{N}^*}\right|\right|_{l^2} + \frac{1}{\sqrt{2\kappa}\pi}\,||f||_{L^2(]0,T[\times]0,1[)}\,.$$

□

Remarque 2

On pourra montrer (exercice !) que sous les mêmes hypothèses on a même

$$||u(t,\cdot)||_{L^2(]0,1[)} \leq \left|\left|u^0\right|\right|_{L^2(]0,1[)} e^{-\kappa\pi^2 t} + ||f||_{L^2(]0,T[\times]0,1[)} \frac{1}{\sqrt{2\kappa}\pi}\left(1-e^{-2\kappa\pi^2 t}\right)^{1/2} \quad \forall t\in]0,T].$$

□

Remarque 3

— La précédente inégalité est valable pour tout t, donc on a

$$\sup_{t\in[0,T]} ||u(t,\cdot)||_{L^2(]0,1[)} \leq \left|\left|u^0\right|\right|_{L^2(]0,1[)} + 1/(\sqrt{2\kappa}\pi)\,||f||_{L^2(]0,T[\times]0,1[)}\,,$$

que l'on réécrit

$$||u(t,\cdot)||_{L^\infty([0,T],L^2(]0,1[))} \leq \left|\left|u^0\right|\right|_{L^2(]0,1[)} + 1/(\sqrt{2\kappa}\pi)\,||f||_{L^2([0,T]\times]0,1[)}\,.$$

— D'un point de vue physique, la quantité $E(t) = \sqrt{\int_0^1 u^2(t,x)\,dx}$ est l'énergie thermique contenue dans le fil au temps t. Parallèlement, $\sqrt{\int_0^1 f^2(t,x)\,dx}$ est l'énergie fournie au fil par unité de temps au temps t. L'inégalité obtenue signifie que l'énergie contenue dans le fil en un instant t est inférieure à la somme de l'énergie initialement contenue dans le fil et de l'énergie fournie (par chauffage) au fil entre l'instant initial et l'instant t.

— D'un point de vue mathématique, l'inégalité exprime la continuité de la solution par rapport aux données du système (qui sont u^0 et f) et garantit la stabilité de cette solution. Soit en effet $\overline{u^0}$ et $\overline{f}$ d'autres données, supposées proches de u^0 et f : soit $\varepsilon > 0$ tel que

$$\left|\left|u^0 - \overline{u^0}\right|\right|_{L^2(]0,1[)} \leq \varepsilon, \quad \left|\left|f-\overline{f}\right|\right|_{L^2(]0,T[\times]0,1[)} \leq \varepsilon.$$

La linéarité de l'EDP permet de voir que $u - \overline{u}$ est solution de l'équation de la chaleur avec les mêmes conditions de bord (Dirichlet homogènes), avec donnée initiale $u^0 - \overline{u^0}$ et terme source $f - \overline{f}$. L'inégalité que donne la proposition 3 pour $u - \overline{u}$ est

$$\begin{aligned} ||u(t,\cdot) - \overline{u}(t,\cdot)||_{L^2(]0,1[)} &\leq \left|\left|u^0 - \overline{u^0}\right|\right|_{L^2(]0,1[)} + \frac{1}{\sqrt{2\kappa\pi}} \left|\left|f - \overline{f}\right|\right|_{L^2(]0,T[\times]0,1[)} \\ &\leq \varepsilon(1 + 1/(\sqrt{2\kappa\pi})) \quad \forall t \in]0,T]. \end{aligned}$$

Ceci signifie que $\overline{u}(t,\cdot)$ reste proche de $u(t,\cdot)$ en norme $L^2(]0,1[)$: c'est de la stabilité L^2. □

2.1.3 Existence d'une solution. Régularité

Les calculs que nous avons effectués, utilisant les séries de Fourier, permettent aussi de prouver l'existence d'une solution. Il suffit pour cela de considérer la série de Fourier où les coefficients sont solutions des EDO mises en évidence et de montrer qu'elle définit une fonction régulière qui vérifie l'EDP de la chaleur ainsi que les conditions aux limites prescrites. C'est ce que nous allons faire dans cette section.

Définition 1
Soit I et J des intervalles de $\mathbb{R}$. On note $\mathcal{C}^{l,m}(I,J)$ l'ensemble des fonctions u de $I \times J$ dans $\mathbb{R}$ telles que

$$\begin{aligned} &u(\cdot,x) \in \mathcal{C}^l(I) \quad \forall x \in J \quad \text{et} \\ &u(t,\cdot) \in \mathcal{C}^m(J) \quad \forall t \in I. \end{aligned}$$

On note $\mathcal{C}_b^{l,m}(I,J)$ l'ensemble des fonctions $u \in \mathcal{C}^{l,m}(I,J)$ dont les l premières dérivées partielles par rapport à la première variable et les m premières dérivées partielles par rapport à la seconde variable sont uniformément bornées sur $I \times J$. □

Théorème 3
Considérons le problème (2.1) où T est un réel strictement positif quelconque, avec une donnée initiale $u^0 \in L^2(]0,1[)$ et un terme source $f \in \mathcal{C}^2([0,T] \times [0,1])$ vérifiant $f(t,0) = f(t,1) = 0$ $\forall t \in [0,T]$. Il a une unique solution [2] $u \in \mathcal{C}^{1,2}(]0,T],[0,1])$. Cette solution vérifie la propriété de stabilité de la proposition 3. □

La démonstration de ce résultat est assez fastidieuse (surtout à cause du terme de chauffage : d'ailleurs, allez, oui, une lecture rapide pourra être faite en annulant par la pensée tous les termes venant de f).

Démonstration
Elle consiste à vérifier que la série de Fourier dont on a calculé les coefficients précédemment est solution de l'EDP et des conditions aux limites. Il faut vérifier que cette série définit une fonction

2. La condition initiale étant vérifiée au sens $L^2(]0,1[)$.

$u(t,x)$ dérivable une fois par rapport à t, deux fois par rapport à x et telle que ces dérivées partielles sont dans $\mathcal{C}^0(]0,T]\times[0,1])$, puis que $u(t,0)=u(t,1)=0$ et enfin que $u(0,\cdot)=u^0(\cdot)$ (en un sens à préciser).

Commençons par montrer que cette série est sommable pour tout $t>0$. On considère, comme promis, la somme de la série

$$\sqrt{2}\sum_{1}^{\infty}\left(\widehat{u^0}(k)+\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right)e^{-\kappa k^2\pi^2 t}\sin(k\pi x), \tag{2.2}$$

qui est candidate à être solution de (2.1). Puisque $u^0\in L^2(]0,1[)$, la suite $\left(\widehat{u^0}(k)\right)_{k\in\mathbb{N}^*}$ est bornée (disons, par $B\in\mathbb{R}$) et l'on a

$$\sum_{k\in\mathbb{N}^*}\left|\widehat{u^0}(k)e^{-\kappa k^2\pi^2 t}\sin(k\pi x)\right|\leq\sum_{k\in\mathbb{N}^*}Be^{-\kappa k^2\pi^2 t}<+\infty$$

si $t>0$. La série $\sum_{k\in\mathbb{N}^*}\widehat{u^0}(k)e^{-\kappa k^2\pi^2 t}\sin(k\pi x)$ est donc convergente (normalement convergente en espace).

Occupons-nous maintenant de $\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x)$. Notons d'abord que $\forall k\in\mathbb{N}^*$, $\widehat{f}(k)(\cdot)\in L^2(]0,T[)$:

$$\begin{aligned}\int_0^T|\widehat{f}(k)(s)|^2\,ds&\leq\int_0^T\sum_{k\in\mathbb{N}^*}\left|\widehat{f}(k)(s)\right|^2\,ds\\&=\int_0^T\int_0^1|f(s,x)|^2\,dx\,ds=||f||^2_{L^2(]0,T[\times]0,1[)}<+\infty.\end{aligned}$$

D'autre part, $e^{\kappa k^2\pi^2(\cdot-t)}\in L^2(]0,t[)\ \forall t\in\mathbb{R}_+^*$. On peut donc écrire, en utilisant l'inégalité de Cauchy-Schwarz,

$$\left|\int_0^t\widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x)\right|\leq\sqrt{\int_0^t|\widehat{f}(k)(s)|^2\,ds}\sqrt{\int_0^t e^{2\kappa k^2\pi^2(s-t)}\,ds}$$

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

$$\int_0^t e^{2\kappa k^2\pi^2(s-t)}\,ds=\left[\frac{1}{2\kappa k^2\pi^2}e^{2\kappa k^2\pi^2(s-t)}\right]_0^t=\frac{1}{2\kappa k^2\pi^2}(1-e^{-2\kappa k^2\pi^2 t})\leq\frac{1}{2\kappa k^2\pi^2}.$$

Ainsi

$$\left|\int_0^t\widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x)\right|\leq\frac{1}{\sqrt{2\kappa}k\pi}\sqrt{\int_0^t|\widehat{f}(k)(s)|^2\,ds}\leq\frac{1}{\sqrt{2\kappa}k\pi}\sqrt{\int_0^T|\widehat{f}(k)(s)|^2\,ds}.$$

De plus,

$$\sum_{k\in\mathbb{N}^*}\frac{1}{\sqrt{2\kappa}k\pi}\sqrt{\int_0^T|\widehat{f}(k)(s)|^2\,ds}\leq\sqrt{\sum_{k\in\mathbb{N}^*}\frac{1}{2\kappa k^2\pi^2}}\sqrt{\sum_{k\in\mathbb{N}^*}\int_0^T|\widehat{f}(k)(s)|^2\,ds}$$

car les deux séries sont dans l^2. On a ainsi montré que

$$\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x)$$

est pour tout $t\in]0,T]$ le terme général d'une série convergente (normalement convergente) : la série que nous avons définie en (2.2) est convergente pour $(t,x)\in]0,T]\times[0,1]$. Notons $u(t,x)$ la somme de cette série. Nous allons montrer que u est de classe $\mathcal{C}^1$ par rapport à sa première variable sur $]0,T]\times[0,1]$. Chaque terme de la série est dérivable par rapport à t, et la dérivée en vaut

$$\sqrt{2}\left[-\kappa k^2\pi^2\left(\widehat{u^0}(k)+\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right)e^{-\kappa k^2\pi^2 t}+\widehat{f}(k)(t)\right]\sin(k\pi x).$$

Ici, nous supposons que $\left(\sup_{t\in[0,T]}\left|\widehat{f}(k)(t)\right|\right)_{k\in\mathbb{N}^*}$ est une suite de l^1. Ceci sera ensuite vérifié (grâce aux hypothèses faites sur f).

Soit $\varepsilon\in]0,T[$; $\left(-\kappa k^2\pi^2\widehat{u^0}(k)e^{-\kappa k^2\pi^2 t}+\widehat{f}(k)(t)\right)\sin(k\pi x)$ est le terme général d'une série normalement convergente sur $[\varepsilon,T]\times[0,1]$ car

- le terme $\left|-\kappa k^2\pi^2\widehat{u^0}(k)e^{-\kappa k^2\pi^2 t}\sin(k\pi x)\right|$ y est majoré par $\kappa Bk^2\pi^2e^{-\kappa k^2\pi^2\varepsilon}$ où $B\in\mathbb{R}$ est une constante telle que $\left|\widehat{u^0}(k)\right|\le B\ \forall k\in\mathbb{N}^*$;
- $\left|\widehat{f}(k)(t)\sin(k\pi x)\right|\le\sup_{[\varepsilon,T]}\left|\widehat{f}(k)(t)\sin(k\pi x)\right|\le\sup_{[\varepsilon,T]}\left|\widehat{f}(k)(t)\right|$ qui est le terme général d'une suite de l^1.

Il reste à étudier le terme

$$-\kappa k^2\pi^2\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x).$$

Puisque $\left(\sup_{t\in[0,T]}\left|\widehat{f}(k)(t)\right|\right)_{k\in\mathbb{N}^*}$ est une suite de l^1,

$$\begin{aligned}\left|\kappa k^2\pi^2\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x)\right| &\le \sup_{t\in[0,T]}|\widehat{f}(k)(t)|\kappa k^2\pi^2\int_0^t e^{\kappa k^2\pi^2(s-t)}\,ds\\ &= \sup_{t\in[0,T]}|\widehat{f}(k)(t)|\left(1-e^{-\kappa k^2\pi^2 t}\right)\le\sup_{t\in[0,T]}|\widehat{f}(k)(t)|\end{aligned}$$

qui est le terme général d'une série convergente. Récapitulons : sous l'hypothèse que

$$\left(\sup_{t\in[0,T]}\left|\widehat{f}(k)(t)\right|\right)_{k\in\mathbb{N}^*}$$

est une suite de l^1, la série des dérivées par rapport à t est normalement convergente sur $[\varepsilon,T]\times[0,1]$; il en découle[3] que u est dérivable par rapport à sa première variable et que $\partial_t u(t,x)$ est

3. Il faut aussi vérifier que la série converge en un point, ce que nous avons montré à l'étape précédente en montrant qu'elle convergeait pour tout $(t,x)\in]0,T]\times[0,1]$.

de classe $\mathcal{C}^0$ par rapport à (t,x) sur $[\varepsilon,T]\times[0,1]$ pour tout $\varepsilon>0$ [4], donc sur $]0,T]\times[0,1]$, et enfin que

$$\partial_t u(t,x)=\sqrt{2}\sum_{k\in\mathbb{N}^*}\left[-\kappa k^2\pi^2\left(\widehat{u^0}(k)+\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right)e^{-\kappa k^2\pi^2 t}+\widehat{f}(k)(t)\right]\sin(k\pi x).$$

La démonstration que nous avons faite permet aussi de voir que, sous les mêmes hypothèses, u est deux fois dérivable par rapport à sa seconde variable et que $\partial^2_{x,x}$ est de classe $\mathcal{C}^0$ par rapport à (t,x) sur $]0,T]\times[0,1]$, et enfin que

$$\partial^2_{x,x}u(t,x)=\sqrt{2}\sum_{k\in\mathbb{N}^*}\left[k^2\pi^2\left(\widehat{u^0}(k)+\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right)e^{-\kappa k^2\pi^2 t}\right]\sin(k\pi x).$$

Il reste à trouver une condition sur f pour que $\left(\sup_{t\in[0,T]}\left|\widehat{f}(k)(t)\right|\right)_{k\in\mathbb{N}^*}$ soit une suite de l^1. Le lemme 2 ci-après assure que si $f\in\mathcal{C}^2([0,T]\times[0,1])$ et $f(t,0)=f(t,1)=0\ \forall t\in[0,T]$, cette condition est vérifiée. Nous avons montré que la somme u de la série de Fourier était dans $\mathcal{C}^{1,2}(]0,T],[0,1])$ et qu'elle vérifiait l'EDP. Il faut maintenant montrer qu'elle vérifie les conditions aux limites demandées.
Tout d'abord : les conditions aux limites en espace,

$$u(t,0)=u(t,1)=0\quad\forall t\in]0,T].$$

Ces conditions sont trivialement vérifiées puisque $\sin(0)=\sin(k\pi)=0\ \forall k\in\mathbb{N}^*$.
Maintenant, intéressons-nous à la condition initiale. Ce point est un peu plus délicat : on note en particulier que sans hypothèse supplémentaire sur la régularité de la condition initiale, il n'y a aucune raison pour que $u\in\mathcal{C}^{1,2}([0,T],[0,1])$. On va cependant montrer que $\lim_{t\to0^+}u(t,\cdot)=u(0,\cdot)$ dans $L^2(]0,1[)$. Pour ceci, appliquons l'égalité de Bessel-Parseval à $u^0-u(t,\cdot)$:

$$\int_0^1\left|u^0(x)-u(t,x)\right|^2dx=\sum_{k\in\mathbb{N}^*}\left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 t}\right)-\int_0^t\widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\right|^2,$$

cette série convergeant évidemment [5]. Comme suite, et puisque $(a+b)^2\le 2a^2+2b^2\ \forall a,b\in\mathbb{R}$,

$$\int_0^1\left|u^0(x)-u(t,x)\right|^2dx\le 2\sum_{k\in\mathbb{N}^*}\left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 t}\right)\right|^2+2\sum_{k\in\mathbb{N}^*}\left|\int_0^t\widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\right|^2.$$

Or pour tout $\epsilon>0$, $\exists N\in\mathbb{N}$ tel que $\forall t\in[0,T]$

$$\sum_{k=N}^{\infty}\left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 t}\right)\right|^2\le\sum_{k=N}^{\infty}\left|\widehat{u^0}(k)\right|^2\le\frac{\epsilon}{2}.$$

4. Car chaque terme de la série des dérivées est de classe $\mathcal{C}^0$.
5. Ce n'est pas une nouvelle notion inconnue (?) de convergence.

D'autre part, $\lim_{t\to 0^+} e^{-\kappa k^2\pi^2 t} = 1\ \forall k \in \mathbb{N}^*$, donc $\exists \eta \in \mathbb{R}$ (dépendant de N) tel que

$$\sum_{k=1}^{N-1} \left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 t}\right)\right|^2 \leq \sum_{k=1}^{N-1} B^2 \left|1-e^{-\kappa k^2\pi^2 t}\right|^2 \leq \frac{\epsilon}{2} \text{ si } 0 \leq t \leq \eta.$$

Donc [6], $\forall \epsilon > 0$, $\exists \eta \in \mathbb{R}$ tel que si $0 \leq t \leq \eta$,

$$\sum_{k\in\mathbb{N}^*} \left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 t}\right)\right|^2 \leq \epsilon.$$

Et le terme $\sum_{k\in\mathbb{N}^*} \left|\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\right|^2$? Il suffit de lui appliquer les majorations déjà utilisées dans la cette démonstration : $\forall k \in \mathbb{N}^*$,

$$\left|\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\right|^2 \leq \sup_{t\in[0,T]} \left|\widehat{f}(k)(t)\right|^2 \left(\int_0^t e^{\kappa k^2\pi^2(s-t)}\,ds\right)^2$$
$$\leq \sup_{t\in[0,T]} \left|\widehat{f}(k)(t)\right|^2 \left(\frac{1}{\kappa k^2\pi^2}\left(1-e^{-\kappa k^2\pi^2 t}\right)\right)^2,$$

d'où

$$\left|\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\right|^2 \leq \sup_{t\in[0,T]} \left|\widehat{f}(k)(t)\right|^2 \frac{1}{\kappa^2\pi^4}$$

pour tout $k \in \mathbb{N}^*$ et finalement

$$\sum_{k\in\mathbb{N}^*} \left|\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\right|^2 \leq \frac{1}{\kappa^2\pi^4} \sum_{k\in\mathbb{N}^*} \sup_{t\in[0,T]} \left|\widehat{f}(k)(t)\right|^2.$$

Or $\left(\sup_{t\in[0,T]} \left|\widehat{f}(k)(t)\right|\right)_{k\in\mathbb{N}^*}$ est une suite de l^1, donc une suite de l^2. Donc le membre de droite de la dernière inégalité converge : la série en question converge normalement sur $[0,T]$, donc sa somme est continue sur $[0,T]$ et vaut 0 en $t=0$.

Nous avons donc montré que $\lim_{t\to 0^+} u(t,\cdot) = u^0$ dans $L^2(]0,1[)$. La condition initiale est vérifiée dans $L^2(]0,1[)$. □

Remarque 4

On a en fait montré un résultat plus fort que l'énoncé : non seulement $\partial_t u(\cdot,x)$ est de classe $\mathcal{C}^0(]0,T])$ pour tout $x \in [0,1]$, mais encore $\partial_t u$ est de classe $\mathcal{C}^0(]0,T]\times[0,1])$. De même, $\partial^2_{x,x} u$ est de classe $\mathcal{C}^0(]0,T]\times[0,1])$. Cette remarque est la base de la compréhension du théorème 4. □

Remarque 5

Il est tout à fait naturel que la condition initiale ne soit pas vérifiée en un sens plus fort que L^2 (par exemple, ponctuellement en x), puisque cette condition initiale n'est pas supposée régulière. Ceci amène d'ailleurs à faire les deux commentaires suivants.

6. Hum, démonstration plus simple de ce résultat : la série $\sum_{k\in\mathbb{N}^*} \left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 t}\right)\right|^2$ converge normalement sur $[0,T]$, donc sa limite lorsque t tend vers 0 est $\sum_{k\in\mathbb{N}^*} \left|\widehat{u^0}(k)\left(1-e^{-\kappa k^2\pi^2 0}\right)\right|^2 = 0$.

D'une part, remarquons que l'EDP proprement dite n'est vérifiée par la solution *que* sur $]0,T] \times [0,1]$, et pas en $t = 0$. Il est hors de question de montrer qu'elle est vérifiée en $t = 0$ puisque, u^0 n'étant pas supposée de classe $\mathcal{C}^2$, $\partial^2_{x,x} u^0$ n'a pas de sens. On pourrait néanmoins trouver une solution du problème sur le pavé fermé (*i.e.* $[0,T] \times [0,1]$) en supposant la condition initiale de classe $\mathcal{C}^2$. Ceci aurait masqué une propriété essentielle de l'opérateur $\partial^2_{x,x}$, évoquée dans le commentaire qui suit.

Nous voyons que même si la condition initiale n'est pas régulière, la solution est de classe $\mathcal{C}^2$ en espace, pour tout t strictement positif. On peut même pousser le raisonnement plus loin : on remarque que le terme général de la série de Fourier de la solution est de classe $\mathcal{C}^\infty$ par rapport à sa seconde variable et que la série des dérivées n^{es} est normalement convergente sur $[\varepsilon, T] \times [0,1]$ pour tout $n \in \mathbb{N}$, la solution est donc dans $\mathcal{C}^{1,\infty}(]0,T], [0,1])$. On montrerait de même que si f est de classe $\mathcal{C}^{2n}$ et telle que $f^{(2l)}(t,0) = f^{(2l)}(t,1) = 0 \ \forall t \in [0,T]$, pour $l = 0, 1, \ldots, n-1$, la solution u est dans $\mathcal{C}^{n,\infty}(]0,T], [0,1])$. En particulier, si $f = 0$ (équation sans terme source), $u \in \mathcal{C}^{\infty,\infty}(]0,T][0,1])$. Comme de plus $\partial_t u(t,x) = \kappa \partial^2_{x,x} u(t,x) \ \forall (t,x) \in]0,T] \times [0,1]$, on en déduit que $u \in \mathcal{C}^\infty(]0,T] \times [0,1])$. C'est une propriété de régularisation de l'opérateur $\partial^2_{x,x}$. On peut montrer que la solution de l'équation de la chaleur sans terme source est même analytique réelle en espace sur $[0,1]$ pour tout $t > 0$: voir pour ceci le sujet du partiel du 19 avril 2005 (et son corrigé). En exercice, montrer grâce à cela une propriété de non-localité de cette EDP : $\forall x \in]0,1[, \ \forall t > 0$, il n'existe pas de voisinage de x $V(x) \subsetneq]0,1[$ tel que $u(t,x)$ ne dépende que de $\{u^0(y) \text{ t. q. } y \in V(x)\}$. □

Le lecteur attentif considérerait comme une arnaque d'en rester là : il nous reste à démontrer le lemme suivant, utile à la démonstration du précédent théorème.

Lemme 2
Soit $g \in \mathcal{C}^2([0,1])$ vérifiant $g(0) = g(1) = 0$. Il existe $C \in \mathbb{R}$ tel que

$$|\widehat{g}(k)| \leq \frac{C}{k^2} \quad \forall k \in \mathbb{N}^*.$$

□

Démonstration
L'essentiel de cette démonstration a déjà été fait dans la démonstration du lemme 1. N'hésitons cependant pas à répéter...

Soit $k \in \mathbb{N}^*$.

$$\widehat{g}(k) = \sqrt{2} \int_0^1 g(x) \sin(k\pi x)\, dx.$$

Donc

$$\widehat{g}(k) = -\frac{\sqrt{2}}{k\pi} \left[g(x) \cos(k\pi x) \right]_0^1 + \frac{\sqrt{2}}{k\pi} \int_0^1 g'(x) \cos(k\pi x)\, dx = \frac{\sqrt{2}}{k\pi} \int_0^1 g'(x) \cos(k\pi x)\, dx$$

car $g(0) = g(1) = 0$. Et on continue :

$$\widehat{g}(k) = \frac{\sqrt{2}}{k^2\pi^2}\left[g'(x)\sin(k\pi x)\right]_0^1 - \frac{\sqrt{2}}{k^2\pi^2}\int_0^1 g''(x)\sin(k\pi x)\,dx$$
$$= -\frac{\sqrt{2}}{k^2\pi^2}\int_0^1 g''(x)\sin(k\pi x)\,dx$$

car $\sin(0) = \sin(k\pi) = 0$. Donc

$$|\widehat{g}(k)| \leq \frac{\sqrt{2}}{k^2\pi^2}\max_{[0,1]}|g''|\int_0^1 |\sin(k\pi x)|\,dx = \frac{2\sqrt{2}}{k^2\pi^3}\max_{[0,1]}|g''|.$$

Il suffit donc de poser

$$C = \frac{2\sqrt{2}}{\pi^3}\max_{[0,1]}|g''|.$$

□

On laisse le soin au lecteur de démontrer le résultat nécessaire pour achever la démonstration du théorème précédent (dans le cas où f dépend aussi de t).

Théorème 4
Considérons le problème (2.1) où T est un réel strictement positif quelconque et avec une donnée initiale $u^0 \in \mathcal{C}^4([0,1])$ vérifiant $u^0(0) = u^0(1) = u^{0\prime\prime}(0) = u^{0\prime\prime}(1) = 0$ et un terme source $f \in \mathcal{C}^2([0,T]\times[0,1])$ vérifiant $f(t,0) = f(t,1) = 0\ \forall t \in [0,T]$. Il a une unique solution $u \in \mathcal{C}_b^{1,2}([0,T],[0,1])$. □

À la différence du théorème 3, le théorème 4 fait appel à une hypothèse de régularité sur la donnée initiale.

Démonstration
On sait déjà qu'il existe une solution $u \in \mathcal{C}^{1,2}(]0,T],[0,1])$ et qu'elle est la somme de la série

$$\sqrt{2}\sum_{k=1}^{\infty}\left(\widehat{u^0}(k) + \int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right)e^{-\kappa k^2\pi^2 t}\sin(k\pi x).$$

On s'intéresse au comportement de la dérivée par rapport à t au voisinage de $t = 0$ de cette fonction. La série des dérivées par rapport à t, déjà étudiée, est la série de terme général

$$\sqrt{2}\left[-\kappa k^2\pi^2\left(\widehat{u^0}(k) + \int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2 s}\,ds\right)e^{-\kappa k^2\pi^2 t} + \widehat{f}(k)(t)\right]\sin(k\pi x).$$

Nous allons reprendre le schéma de la démonstration de la dérivabilité (par rapport à t) de la série, en y incorporant ici les éléments nouveaux. Rappelons que la régularité de f implique que $\left(\sup_{t\in[0,T]}\left|\widehat{f}(k)(t)\right|\right)_{k\in\mathbb{N}^*}$ est une suite de l^1. D'autre part, puisque $u^0 \in \mathcal{C}^4([0,1])$ et $u^0(0) = u^0(1) = u^{0\prime\prime}(0) = u^{0\prime\prime}(1) = 0$, le lemme 2 « dopé » à l'ordre 4 assure qu'il existe $C \in \mathbb{R}$ tel que

$$\left|\widehat{u^0}(k)\right| \leq \frac{C}{k^4}.$$

Nous avons maintenant : $\left(-\kappa k^2\pi^2\widehat{u^0}(k)e^{-\kappa k^2\pi^2 t} + \widehat{f}(k)(t)\right)\sin(k\pi x)$ est le terme général d'une série normalement convergente sur $[0,T]\times[0,1]$ car

— le terme $\left|-\kappa k^2\pi^2\widehat{u^0}(k)e^{-\kappa k^2\pi^2 t}\sin(k\pi x)\right|$ y est majoré par $\kappa\pi^2\frac{C}{k^2}$;

— $\left|\widehat{f}(k)(t)\sin(k\pi x)\right| \leq \sup_{[0,T]}\left|\widehat{f}(k)(t)\sin(k\pi x)\right| \leq \sup_{[0,T]}\left|\widehat{f}(k)(t)\right|$ qui est le terme général d'une suite de l^1.

L'étude du terme

$$-\kappa k^2\pi^2\int_0^t \widehat{f}(k)(s)e^{\kappa k^2\pi^2(s-t)}\,ds\sin(k\pi x)$$

a été faite lors de la démonstration du théorème 3 et a permis de montrer que ce terme est celui d'une série normalement convergente sur $[0,T]\times[0,1]$. En guise de récapitulation : la série des dérivées par rapport à t des coefficients de Fourier de u est normalement convergente sur $[0,T]\times[0,1]$; en conséquence, u est dérivable par rapport à sa première variable et $\partial_t u(t,x)$ est de classe $\mathcal{C}^0$ par rapport à (t,x) sur $[0,T]\times[0,1]$. La dérivée de u par rapport à t est donc uniformément bornée sur $[0,T]\times[0,1]$.

Un raisonnement similaire pour la dérivée seconde de u par rapport à x est effectuable... Et à effectuer à titre d'exercice. □

Remarque 6

Il peut paraître stérile de faire une hypothèse aussi forte sur la condition initiale, vue la faiblesse du résultat obtenu. En fait, nous montrerons dans le chapitre concernant l'approximation numérique la convergence des solutions approchées vers la solution exacte *pourvu que cette solution exacte ait ses dérivées partielles uniformément bornées*. Le théorème 4 a pour seule ambition de montrer qu'il existe de telles solutions, et que les algorithmes discrets que nous allons étudier convergent dans de « vrais cas ». Remarquer par ailleurs que, comme nous l'avons déjà observé (remarque 5), moyennant des hypothèses suffisamment fortes sur u^0 et f, on est en mesure de produire des solutions de (2.1) dans $\mathcal{C}_b^{l,m}([0,T],[0,1])$ pour tout (l,m). □

Remarque 7 (Conditions aux limites de Dirichlet non homogènes)

Le problème de la chaleur dans $[0,1]$ avec conditions aux limites de Dirichlet non homogènes s'écrit

$$\begin{cases} \partial_t u(t,x) - \kappa\partial^2_{x,x}u(t,x) = f(t,x) \text{ dans }]0,T]\times]0,1[, \\ u(t,0) = u_0 \quad \forall t \in]0,T], \\ u(t,1) = u_1 \quad \forall t \in]0,T], \\ u(0,x) = u^0(x) \quad \forall x \in]0,1[\end{cases} \tag{2.3}$$

(on suppose ici pour simplifier que u_0 ni u_1 ne dépendent de t). Pour résoudre ce problème, il suffit de remarquer que la fonction $\Delta(t,x)$ définie par $\Delta(t,x) = u_0 + (u_1 - u_0)x$ vérifie l'EDP de la chaleur sans terme source ainsi que les conditions aux limites de Dirichlet non homogènes en

0 et en 1. Soit donc $v(t,x)$ la solution de

$$\begin{cases} \partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = f(t,x) \text{ dans }]0,T]\times]0,1[, \\ u(t,0) = 0 \quad \forall t \in]0,T], \\ u(t,1) = 0 \quad \forall t \in]0,T], \\ u(0,x) = u^0(x) - \Delta(0,x) \quad \forall x \in]0,1[\end{cases}$$

(on sait que la solution de ce problème existe puisque c'est le problème de Dirichlet homogène). Posons maintenant $u(t,x) = v(t,x) + \Delta(t,x)$: u est la solution cherchée. □

Remarque 8 (Conditions aux limites de Neumann homogènes)
Un autre type de conditions aux limites très souvent pris en compte est celui de Neumann, où la *dérivée normale* de la solution est imposée sur le bord. Le problème en dimension 1 devient alors, dans le cas homogène,

$$\begin{cases} \partial_t u(t,x) - \kappa \partial^2_{x,x} u(t,x) = f(t,x) \text{ dans }]0,T]\times]0,1[, \\ \partial_x u(t,0) = 0 \quad \forall t \in]0,T], \\ \partial_x u(t,1) = 0 \quad \forall t \in]0,T], \\ u(0,x) = u^0(x) \quad \forall x \in]0,1[. \end{cases} \tag{2.4}$$

On peut pour ce problème faire la même étude, en utilisant exactement la même technique, celle des séries de Fourier. Il faut pour ceci faire usage de la base hilbertienne $\left\{1, \left(\sqrt{2}\cos(k\pi x)\right)_{k\in\mathbb{N}^*}\right\}$ de $L^2(]0,1[)$. C'est en effet avec cette base-ci que $u \in \mathcal{C}^2([0,1])$ vérifiant $\partial_x u(0) = \partial_x u(1) = 0$ sera telle que

$$\widehat{u''}(k) = -k^2\pi^2 \widehat{u}(k) \quad \forall k \in \mathbb{N}.$$

□

2.2 Principes du maximum

Nous avons déjà évoqué la question de la stabilité de la solution de (2.1) et nous avons répondu par l'affirmative en norme L^2 : voir la proposition 3. La présente section est consacrée à l'étude de la stabilité en norme L^∞. La stabilité que nous allons démontrer porte le nom de « principe du maximum » et sera à nouveau étudiée dans les chapitres concernant les équations hyperboliques et les équations elliptiques.

2.2.1 Entropies

On appelle *entropie* pour le problème (2.1) toute fonction S de classe $\mathcal{C}^2(\mathbb{R})$ telle que, si u désigne la solution de (2.1),

$$\partial_t S(u)(t,x) - \kappa \partial^2_{x,x} S(u)(t,x) \leq S'(u)(t,x) f(t,x).$$

Remarque 9
Le terme « entropie » est généralement réservé aux équations hyperboliques, mais il désigne dans ce cadre exactement la même chose... □

Proposition 4
Supposons que les hypothèses du théorème 3 sont vérifiées. Toute fonction de classe $\mathcal{C}^2(\mathbb{R})$ et convexe sur $\mathbb{R}$ est une entropie pour (2.1). □

Dans cette proposition, les hypothèses du théorème 3 sont faites afin d'assurer l'existence d'une solution au problème (2.1) et sa régularité.

Démonstration
On a $\partial_t S(u) = S'(u)\partial_t u$ et $\partial_x S(u) = S'(u)\partial_x u$, d'où $\partial^2_{x,x} S(u) = S''(u)(\partial_x u)^2 + S'(u)\partial^2_{x,x} u$. Donc,

$$\partial_t S(u) - \kappa \partial^2_{x,x} S(u) = S'(u) f(t,x) - \kappa S''(u)(\partial_x u)^2 \leq S'(u) f(t,x).$$

□

Pour simplifier, on suppose à partir d'ici et pour le reste de la section 2.2 que le terme source est nul : $f(t,x) = 0\ \forall t, \forall x$.

2.2.2 Décroissance en temps de la norme $L^\infty(]0,1[)$

Théorème 5
Supposons que les hypothèses du théorème 3 sont vérifiées. Soit $u(t,x)$ l'unique solution de (2.1) avec terme source nul et donnée initiale $u^0 \in L^\infty(]0,1[)$. Cette solution vérifie

$$||u(t,\cdot)||_{L^\infty(]0,1[)} \leq \left|\left|u^0\right|\right|_{L^\infty(]0,1[)} \quad \forall t \in \mathbb{R}_+.$$

□

Démonstration
Toute fonction $S \in \mathcal{C}^2(\mathbb{R})$ telle que $S''(x) \geq 0\ \forall x \in \mathbb{R}$ est une entropie pour (2.1) et vérifie

$$\partial_t S(u) - \kappa \partial^2_{x,x} S(u) \leq 0.$$

Notons $M = \max(0, \text{supess } u^0)$. Choisissons une entropie particulière (voir la figure) :

$$S(u) = \begin{cases} (u-M)^3 \text{ si } u \geq M, \\ 0 \text{ si } u > M \end{cases}$$

(exercice : vérifier que c'est une entropie !).

Intégrons l'inégalité d'entropie en espace entre 0 et 1 :

$$\int_0^1 \partial_t S(u)(t,x)\, dx - \kappa \left[\partial_x S(u)(t,x)\right]_0^1 \leq 0.$$

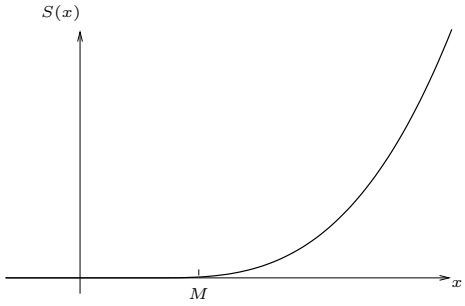

FIGURE 2.1 – Graphe de l'entropie.

Or (pour $t > 0$) $u(t,0) = u(t,1) = 0 \leq M$, donc $\partial_x S(u(t,0)) = S'(u(t,0))\partial_x u(t,0) = 0 = S'(u(t,1))\partial_x u(t,1) = \partial_x S(u(t,1))$ grâce à la forme particulière de l'entropie choisie, donc

$$\int_0^1 \partial_t S(u)(t,x)\,dx \leq 0.$$

De plus, $\int_0^1 \partial_t S(u)(t,x)\,dx = \partial_t \int_0^1 S(u)(t,x)\,dx$ (à vérifier en exercice), d'où, en intégrant en temps entre 0 et t,

$$\int_0^1 S(u)(t,x)\,dx - \int_0^1 S(u)(0,x)\,dx \leq 0$$

et, puisque $S(u(0,x)) = 0$ pour presque tout x,

$$\int_0^1 S(u)(t,x)\,dx \leq 0.$$

Puisque $S(u) \geq 0$ pour tout $u \in \mathbb{R}$, cela implique que $S(u)(t,x) = 0$ pour presque tout x, pour tout $t > 0$. Ceci signifie que $u(t,x) \leq M$ pour presque tout x, pour tout t, à nouveau grâce à la forme particulière de l'entropie. En prenant maintenant l'entropie

$$\tilde{S}(u) = \begin{cases} -(u-m)^3 \text{ si } u \leq m, \\ 0 \text{ si } u > m \end{cases}$$

avec $m = \min(0, \operatorname{infess}\, u^0)$, on démontre de la même manière que $u(t,x) \geq m$ pour presque tout x, pour tout t. Ceci termine la démonstration du théorème. □

Remarque 10

— Nous avons démontré en fait un résultat plus fort que l'énoncé (lequel ?).

— Avec des conditions de bord de Dirichlet non homogènes

$$u(t,0) = u_0,$$
$$u(t,1) = u_1,$$

on démontrerait que

$$u(t,x) \leq \max(u_0, u_1, \text{supess } u^0) \quad p.p.(x), \forall \text{t},$$
$$u(t,x) \geq \min(u_0, u_1, \text{infess } u^0) \quad p.p.(x), \forall \text{t}$$

(le faire à titre d'exercice).

— La même méthode permet de montrer le même résultat avec des conditions de Neumann homogènes. On pose $M = \text{supess } u^0$. En intégrant en espace l'inégalité d'entropie, il vient

$$\partial_t \int_0^1 S(u)\, dx - \kappa \left[\partial_x S(u)(t,x)\right]_0^1 \leq 0.$$

Or $\partial_x S(u)(t,0) = S'(u)(t,0)\partial_x u(t,0) = 0$ et $\partial_x S(u)(t,1) = S'(u)(t,1)\partial_x u(t,1) = 0$ grâce aux conditions de Neumann homogènes, d'où

$$\partial_t \int_0^1 S(u)\, dx \leq 0,$$

donc $\int_0^1 S(u)(t,x)\, dx \leq \int_0^1 S(u)(0,x)\, dx = 0$. Mais puisque $S(u) \geq 0$ quelque soit u, on a nécessairement $S(u(t,x)) = 0$ pour presque tout x, pour tout t. D'où enfin $u(t,x) \leq M$ pour presque tout x, pour tout t...

— La méthode que nous avons utilisée est une adaptation de la méthode des troncatures de Stampacchia pour montrer le principe du maximum pour des équations elliptiques, que nous verrons dans le chapitre qui y sera consacré.

— Une méthode plus classique pour montrer ce principe du maximum consiste à poser $v(t,x) = u(t,x)e^{-\lambda t}$, pour $\lambda > 0$, et à étudier v comme solution d'un problème aux limites.

□

Remarque 11

Le problème (2.1) est *mal posé* pour $t < 0$. En effet, la série de Fourier de la solution diverge pour $t < 0$, sauf si u^0 ne « contient qu'un nombre fini de modes de Fourier ». La stabilité en norme L^2 (norme de l'énergie) est alors fausse. La stabilité en norme L^∞ est perdue aussi (les entropies augmentent en temps négatif). Il en est bien entendu de même pour $t > 0$ avec $\kappa < 0$.
□

Remarque 12

Les résultats que nous avons obtenus ne sont pas optimaux. Des résultats plus forts, assurant l'existence de *solutions faibles* sous des hypothèses plus faibles peuvent être démontrés. Ils font cependant appel à des techniques moins classiques que nous n'introduirons que dans la suite (chapitres sur les équations hyperboliques et les équations elliptiques). □

2.3 Résolution approchée par la méthode des différences finies

La méthode que nous avons employée dans la section 2.1 pour montrer l'existence et l'unicité d'une solution au problème (2.1) est *constructive* : elle permet un *calcul* de la solution. Cela demande néanmoins le calcul des coefficients de Fourier de la condition initiale, de ceux du terme source, le calcul de leur évolution, et enfin le calcul de la transformation de Fourier inverse. Ceci est long et coûteux ; mais rendu réalisable grâce à la transformée de Fourier rapide (Fast Fourier Transform en anglais). Nous n'utiliserons pas cette méthode. Nous préférons en effet proposer l'étude d'une méthode beaucoup plus générale. Il s'agit de la méthode des *différences finies*. Son principe général repose sur la définition de la dérivée d'une fonction f :

$$f'(x) = \lim_{h\to 0} \frac{f(x+h/2) - f(x-h/2)}{h}.$$

La méthode consiste à remplacer, dans l'EDP, $\partial_x u(t,x)$ par $(u(t,x+h/2)-u(t,x-h/2))/h$ pour un h assez petit... C'est-à-dire des dérivées par des différences *finies*. Cette méthode permet un calcul approché de la solution de (2.1) sur une grille de $[0,T]\times[0,1]$. Pour simplifier, on suppose que cette grille est régulière en temps et en espace : on note Δt le pas de temps et Δx le pas d'espace. On note N le nombre de pas de temps entre 0 et T et $t^n = n\Delta t$ pour $n \in \{0,\dots,N\}$, avec $t^0 = 0$ et $t^N = T$, soit $\Delta t = T/N$. On note J le nombre de points du maillage en espace situés à l'intérieur du segment $[0,1]$, et $x_j = j\Delta x$ pour $j \in \{0,\dots,J+1\}$, avec $x_0 = 0$ et $x_{J+1} = 1$, soit $\Delta x = 1/(J+1)$. Le but de la méthode des différences finies est le calcul approché de la solution du problème (2.1) en les points du maillage temps-espace, c'est-à-dire le calcul approché des valeurs $u(t^n,x_j)$ pour tout $n \in \{0,\dots,N\}$ et tout $j \in \{0,\dots,J+1\}$, où u est la solution exacte. Les valeurs approchées que nous calculons sont notées u_j^n.

Les conditions aux limites du problème ne posent a priori aucune difficulté : il est naturel d'imposer

$$\begin{aligned} &u_0^n = 0 \quad \forall n \in \{0,\dots,N\},\\ &u_{J+1}^n = 0 \quad \forall n \in \{0,\dots,N\},\\ &u_j^0 = u^0(x_j) \quad \forall j \in \{0,\dots,J+1\}, \end{aligned}$$

en supposant pour simplifier que la donnée initiale vérifie les conditions aux limites : $u^0(0) = u^0(1) = 0$.

Il nous reste maintenant à faire le calcul des autres valeurs u_j^n, qui doivent être proches des $u(t^n,x_j)$. Or, ces valeurs exactes vérifient

$$\partial_t u(t^n,x_j) - \kappa\partial_{x,x}^2 u(t^n,x_j) = f(t^n,x_j).$$

Bien entendu, la méthode que nous allons développer doit permettre un calcul approché des $u(t^n,x_j)$ sans l'aide des valeurs exactes $\partial_t u(t^n,x_j)$ et $\partial_{x,x}^2 u(t^n,x_j)$, c'est pourquoi nous allons remplacer ces dérivées par des différences finies.

Discrétisation de $\partial_{x,x}^2$

En utilisant

$$\partial_x u(t^n, x_j) \approx \frac{u(t^n, x_j + \Delta x/2) - u(t^n, x_j - \Delta x/2)}{\Delta x},$$

on écrit

$$\partial^2_{x,x} u(t^n, x_j) \approx \frac{\partial_x u(t^n, x_j + \Delta x/2) - \partial_x u(t^n, x_j - \Delta x/2)}{\Delta x},$$

et en réitérant cette approximation,

$$\partial^2_{x,x} u(t^n, x_j) \approx \frac{\frac{u(t^n, x_j+\Delta x) - u(t^n, x_j)}{\Delta x} - \frac{u(t^n, x_j) - u(t^n, x_j - \Delta x)}{\Delta x}}{\Delta x},$$

soit

$$\partial^2_{x,x} u(t^n, x_j) \approx \frac{u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})}{\Delta x^2}. \tag{2.5}$$

Discrétisation de ∂_t

La discrétisation naturelle serait ici

$$\partial_t u(t^n, x_j) \approx \frac{u(t^n + \Delta t/2, x_j) - u(t^n - \Delta t/2, x_j)}{\Delta t},$$

mais le problème de cette discrétisation est qu'elle fait intervenir les valeurs de la solution en des points qui ne sont pas sur la grille. Plusieurs modifications sont possibles, dont, par exemple,

$$\partial_t u(t^n, x_j) \approx \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t}, \tag{2.6}$$

$$\partial_t u(t^n, x_j) \approx \frac{u(t^n, x_j) - u(t^{n-1}, x_j)}{\Delta t}, \tag{2.7}$$

ou encore

$$\partial_t u(t^n, x_j) \approx \frac{u(t^{n+1}, x_j) - u(t^{n-1}, x_j)}{2\Delta t}. \tag{2.8}$$

Ces trois choix, qui paraissent très proches, conduisent en fait à des algorithmes aux comportements extrêmement différents :

- le premier choix conduit à un algorithme *convergent sous une condition sur les pas de temps et d'espace* ; cet algorithme est appelé schéma explicite ;
- le deuxième choix conduit à un algorithme *inconditionnellement convergent* ; cet algorithme est appelé schéma implicite ;
- le troisième choix conduit à un algorithme *non convergent* ; cet algorithme porte le nom de schéma saute-mouton.

Par *convergence*, on entend *convergence de la solution approchée vers la solution exacte lorsque les pas de discrétisation en temps et en espace tendent vers 0.*

La suite de cette section est consacrée, entre autres, à l'étude de ces trois discrétisations.

2.3.1 Étude du schéma explicite

Par *étude du schéma*, on comprend

— étude de la stabilité de la solution approchée, propriété qualitative ;

— étude de la convergence de la solution approchée (ou *numérique*, ou encore *discrète*) vers la solution exacte lorsque Δt et Δx tendent vers 0 ;

— étude de la vitesse de convergence...

Le schéma explicite, qui résulte de l'approximation proposée en (2.5) pour l'opérateur du second ordre, et de la première discrétisation de l'opérateur du premier ordre, (2.6), est

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \kappa \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = f(t^n, x_j). \tag{2.9}$$

Définition 2 (erreur de consistance)
On appelle erreur de consistance d'un schéma l'erreur que l'on commet en remplaçant l'équation exacte $\partial_t u - \kappa \partial_{x,x}^2 u - f = 0$ par l'équation aux différences finies. □

Un exemple vaut mieux qu'un long discours : pour le schéma explicite, l'erreur de consistance au temps t^n au point x_j, notée $\varepsilon_j^n(u)$, est donnée par

$$\varepsilon_j^n(u) = \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} - \kappa \frac{u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})}{\Delta x^2} - f(t^n, x_j).$$

On désignera par la suite par $\varepsilon^n(u)$ le vecteur des erreurs de consistance à l'instant t^n,

$$\varepsilon^n(u) = \left(\varepsilon_j^n(u)\right)_{j \in \{1,\dots,J\}}.$$

Remarque 13
Le terme « consistance » provient d'une erreur de traduction de l'anglais « consistency » et est à comprendre comme « cohérence ». □

Proposition 5
Supposons que $u \in \mathcal{C}_b^{2,4}([0,T],[0,1])$ est solution de (2.1). Alors, il existe $C \in \mathbb{R}$ tel que

$$\max_{n \in \{0,\dots,N-1\}} ||\varepsilon^n(u)||_\infty \leq C(\Delta t + \Delta x^2).$$

□

Remarque 14

— Il en résulte que

$$\lim_{\Delta t, \Delta x \to 0} \max_{n \in \{0,\dots,N-1\}} ||\varepsilon^n(u)||_\infty = 0,$$

on dit alors que le schéma (2.9) est *consistant* avec (2.1).

— On dit que le schéma est d'ordre 1 en temps et d'ordre 2 en espace.

□

Démonstration

Évaluons $\frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} - \partial_t u(t^n,x_j)$**.**

On a supposé que u est de classe $\mathcal{C}^2$ en temps. Donc il existe $\tau \in [t^n, t^{n+1}]$ tel que

$$u(t^{n+1},x_j) = u(t^n,x_j) + \Delta t \partial_t u(t^n,x_j) + \frac{\Delta t^2}{2}\partial^2_{t,t}u(\tau,x_j).$$

Ainsi,

$$\frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} - \partial_t u(t^n,x_j) = \frac{\Delta t}{2}\partial^2_{t,t}u(\tau,x_j).$$

Évaluons $\frac{u(t^n,x_{j+1})-2u(t^n,x_j)+u(t^n,x_{j-1})}{\Delta x^2} - \partial^2_{x,x}u(t^n,x_j)$**.**

Comme on a supposé u de classe $\mathcal{C}^4$ en espace, il existe $y \in [x_j, x_{j+1}]$ tel que

$$u(t^n,x_{j+1}) = u(t^n,x_j) + \Delta x \partial_x u(t^n,x_j) + \frac{\Delta x^2}{2}\partial^2_{x,x}u(t^n,x_j) + \frac{\Delta x^3}{6}\partial^3_{x,x,x}u(t^n,x_j) + \frac{\Delta x^4}{24}\partial^4_{x,x,x,x}u(t^n,y)$$

et il existe $z \in [x_{j-1}, x_j]$ tel que

$$u(t^n,x_{j-1}) = u(t^n,x_j) - \Delta x \partial_x u(t^n,x_j) + \frac{\Delta x^2}{2}\partial^2_{x,x}u(t^n,x_j) - \frac{\Delta x^3}{6}\partial^3_{x,x,x}u(t^n,x_j) + \frac{\Delta x^4}{24}\partial^4_{x,x,x,x}u(t^n,z).$$

Ainsi,

$$\frac{u(t^n,x_{j+1})-2u(t^n,x_j)+u(t^n,x_{j-1})}{\Delta x^2} - \partial^2_{x,x}u(t^n,x_j) = \frac{\Delta x^2}{24}\left[\partial^4_{x,x,x,x}u(t^n,y) + \partial^4_{x,x,x,x}u(t^n,z)\right].$$

Pour finir, on rappelle que $\partial_t u(t^n,x_j) - \kappa\partial^2_{x,x}u(t^n,x_j) - f(t^n,x_j) = 0$, donc

$$\begin{aligned}\varepsilon^n_j(u) = \frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} &- \partial_t u(t^n,x_j) \\ &- \kappa\frac{u(t^n,x_{j+1})-2u(t^n,x_j)+u(t^n,x_{j-1})}{\Delta x^2} + \kappa\partial^2_{x,x}u(t^n,x_j) \\ &- f(t^n,x_j) + f(t^n,x_j) \\ &= \frac{\Delta t}{2}\partial^2_{t,t}u(\tau,x_j) - \kappa\frac{\Delta x^2}{24}\left[\partial^4_{x,x,x,x}u(t^n,y) + \partial^4_{x,x,x,x}u(t^n,z)\right]\end{aligned}$$

On a le résultat annoncé en posant

$$C = 1/2 \max\left(\max_{[0,T]\times[0,1]}\left|\partial^2_{t,t}u\right|, \kappa/6 \max_{[0,T]\times[0,1]}\left|\partial^4_{x,x,x,x}u\right|\right),$$

car alors

$$|\varepsilon_j^n(u)| \leq C(\Delta t + \Delta x^2) \quad \forall n \in \{0, \ldots, N-1\}, \forall j \in \{1, \ldots, J\}.$$

□

Théorème 6

Notons $e^n(u)$ le vecteur de l'*erreur* au temps t^n : $e^n(u) = \left(e_j^n(u)\right)_{j\in\{1,\ldots,J}$ avec

$$e_j^n(u) = u_j^n - u(t^n, x_j) \quad \forall n \in \{0, \ldots, N\}, \forall j \in \{0, \ldots, J+1\}.$$

Supposons que $\kappa\Delta t/\Delta x^2 \leq 1/2$. Alors, sous les hypothèses de la proposition 5, il existe $C \in \mathbb{R}$ tel que

$$||e^n(u)||_\infty \leq C(\Delta t + \Delta x^2) \quad \forall n \in \{0, \ldots, N\}.$$

□

Ce théorème exprime une majoration de l'erreur entre la solution approchée et la solution exacte, en norme l^∞ discrète, c'est-à-dire en un certain nombre de points. On peut bien entendu en déduire une majoration de l'erreur en norme $L^\infty([0,T]\times[0,1])$, c'est ce qu'exprime le corollaire suivant, dont la démonstration, qui utilise une inégalité d'interpolation classique, est laissée au lecteur.

Corollaire 1

Soit $\overline{u}(t,x)$ la fonction affine en t (à x fixé) et affine en x (à t fixé) sur chaque pavé $[t^n, t^{n+1}] \times [x_j, x_{j+1}]$ telle que $\overline{u}(t^n, x_j) = u_j^n$ pour tout $n \in \{0, \ldots, N\}$ et tout $j \in \{0, \ldots, J+1\}$. Sous les hypothèses du théorème 6, il existe $C \in \mathbb{R}$ et $D \in \mathbb{R}$ tels que

$$||\overline{u} - u||_{L^\infty([0,T]\times[0,1])} \leq C(\Delta t + \Delta x^2) + D(\Delta t^2 + \Delta x^2).$$

□

Démonstration (du théorème 6)

Tout d'abord, $e_0^n(u) = e_{J+1}^n(u) = 0\ \forall n$. Puis, si $j \in \{1, \ldots J\}$,

$$\begin{aligned}
e_j^{n+1}(u) &= u_j^{n+1} - u(t^{n+1}, x_j) \\
&= u_j^n + \kappa\frac{\Delta t}{\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t f(t^n, x_j) \\
&\quad - u(t^n, x_j) - (u(t^{n+1}, x_j) - u(t^n, x_j)) \\
&= e_j^n(u) + \kappa\frac{\Delta t}{\Delta x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + \Delta t f(t^n, x_j) \\
&\quad - \left(\Delta t\varepsilon_j^n(u) + \kappa\frac{\Delta t}{\Delta x^2}\left(u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})\right) + \Delta t f(t^n, x_j)\right) \\
&= e_j^n(u) + \kappa\frac{\Delta t}{\Delta x^2}(e_{j+1}^n(u) - 2e_j^n(u) + e_{j-1}^n(u)) - \Delta t\varepsilon_j^n(u).
\end{aligned}$$

Il est important de remarquer que l'erreur est solution de l'équation discrète

$$\frac{e_j^{n+1}(u) - e_j^n(u)}{\Delta t} - \kappa\frac{e_{j+1}^n(u) - 2e_j^n(u) + e_{j-1}^n(u)}{\Delta x^2} = -\varepsilon_j^n(u)$$

qui est l'équation discrète de la chaleur avec le schéma explicite et comme second membre l'erreur de consistance.

On a donc, pour tout $n \in \{0, \ldots, N-1\}$ et tout $j \in \{1, \ldots, J\}$,

$$e_j^{n+1}(u) = \left(1 - 2\kappa\frac{\Delta t}{\Delta x^2}\right) e_j^n(u) + \kappa\frac{\Delta t}{\Delta x^2} e_{j+1}^n(u) + \kappa\frac{\Delta t}{\Delta x^2} e_{j-1}^n(u) - \Delta t \varepsilon_j^n(u).$$

On suppose que $0 \leq 2\kappa\Delta t/\Delta x^2 \leq 1$. Donc

$$|e_j^{n+1}(u)| \leq \left(1 - 2\kappa\frac{\Delta t}{\Delta x^2}\right) ||e^n(u)||_\infty + 2\kappa\frac{\Delta t}{\Delta x^2} ||e^n(u)||_\infty + \Delta t \, ||\varepsilon^n(u)||_\infty$$
$$\forall j \in \{0, \ldots, J+1\},$$

soit

$$|e_j^{n+1}(u)| \leq ||e^n(u)||_\infty + \Delta t \, ||\varepsilon^n(u)||_\infty \, \forall n \in \{0, \ldots, N-1\}, \forall j \in \{0, \ldots, J+1\}.$$

On applique maintenant le résultat obtenu à la proposition 5 et on a

$$\left|\left|e^{n+1}(u)\right|\right|_\infty \leq ||e^n(u)||_\infty + C\Delta t(\Delta t + \Delta x^2).$$

Ainsi,

$$\left|\left|e^{n+1}(u)\right|\right|_\infty \leq \left|\left|e^0(u)\right|\right|_\infty + C(n+1)\Delta t(\Delta t + \Delta x^2).$$

Or $e_j^0(u) = 0 \ \forall j \in \{0, \ldots, J+1\}$, d'où finalement

$$\left|\left|e^{n+1}(u)\right|\right|_\infty \leq C(n+1)\Delta t(\Delta t + \Delta x^2),$$

ceci étant vrai pour tout $n \in \{0, \ldots, N-1\}$. Comme $n\Delta t \leq T \ \forall n \in \{0, \ldots, N\}$, il en découle que

$$\left|\left|e^{n+1}(u)\right|\right|_\infty \leq CT(\Delta t + \Delta x^2).$$

□

Le théorème 6 montre que la solution numérique converge vers la solution exacte lorsque l'on fait tendre les pas de temps et d'espace vers 0, *si le pas de temps est choisi suffisamment petit pour que* $\kappa\Delta t/\Delta x^2 \leq 1/2$. On dit que le schéma explicite (2.9) est *(conditionnellement) convergent* dans L^∞.

Remarque 15

Par analogie avec le cas des équations de transport (voir le chapitre 3), la condition $\kappa\Delta t/\Delta x^2 \leq 1/2$, qui lie le pas de temps au pas d'espace, est parfois appelée condition de Courant-Friedrichs-Lewy. Cette condition est très contraignante car elle impose que $\Delta t \leq \Delta x^2/(2\kappa)$. Nous allons tâcher d'élaborer des algorithmes ne nécessitant pas cette condition, très coûteuse en temps de calcul sur les ordinateurs. □

Remarque 16 (principe du maximum discret)
Supposons que le terme source est nul. Le schéma explicite s'écrit alors

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

soit

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

La condition $\kappa \Delta t/\Delta x^2 \leq 1/2$ apparaît ici naturellement comme une condition sous laquelle u_j^{n+1} est une combinaison convexe de u_{j-1}^n, u_j^n et u_{j+1}^n. c'est une condition de stabilité l^∞ du schéma. Si cette condition est vérifiée, on a

$$\min_{j\in\{0,\ldots,J+1\}} u_j^n \leq u_j^{n+1} \leq \max_{j\in\{0,\ldots,J+1\}} u_j^n \quad \forall j \in \{0,\ldots,J+1\},$$

et donc

$$\min_{j\in\{0,\ldots,J+1\}} u_j^0 \leq u_j^{n+1} \leq \max_{j\in\{0,\ldots,J+1\}} u_j^0 \quad \forall j \in \{0,\ldots,J+1\}.$$

Cette estimation permet, grâce à la convergence du schéma, de retrouver le principe du maximum du théorème 5. □

Nous allons maintenant étudier une classe plus générale d'algorithmes de résolution approchée de l'équation de la chaleur. Afin de gagner en généralité, commençons par observer que le schéma explicite peut être écrit sous une forme matricielle :

$$\frac{U^{n+1} - U^n}{\Delta t} + \frac{\kappa}{\Delta x^2} A U^n = F^n$$

avec

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix} \in \mathcal{M}_J(\mathbb{R})$$

et

$$U^n = \begin{pmatrix} u_1^n \\ \vdots \\ u_J^n \end{pmatrix}, \quad F^n = \begin{pmatrix} f(t^n, x_1) \\ \vdots \\ f(t^n, x_J) \end{pmatrix}.$$

Au moyen de ces notations, nous pouvons réécrire le schéma obtenu en remplaçant $\partial_t u(t^n, x_j)$ par $\frac{u(t^n,x_j)-u(t^{n-1},x_j)}{\Delta t}$ (une des autres discrétisations de $\partial_t u$ proposées) sous la forme matricielle

$$\left(\frac{1}{\Delta t} I + \frac{\kappa}{\Delta x^2} A\right) U^{n+1} + \left(-\frac{1}{\Delta t} I\right) U^n = F^n.$$

Nous regroupons maintenant ces deux schémas différents sous une même formulation, tout en mettant au jour un grand ensemble de nouveaux algorithmes. Pour tout $\theta \in \mathbb{R}$, nous considérons l'algorithme défini par

$$\left(\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A\right)U^{n+1} + \left(-\frac{1}{\Delta t}I + \frac{(1-\theta)\kappa}{\Delta x^2}A\right)U^n = F^{n+1/2} \tag{2.10}$$

où le second membre est donné par

$$F^{n+1/2} = \begin{pmatrix} f(t^n + \Delta t/2, x_1) \\ \vdots \\ f(t^n + \Delta t/2, x_J) \end{pmatrix}.$$

La modification de ce second membre, par rapport au schéma explicite, va permettre d'obtenir un schéma d'ordre 2 en temps. Il est clair qu'elle ne change pas les résultats déjà obtenus pour le schéma explicite. Le schéma décrit par (2.10) est appelé *θ-schéma*. Nous allons l'étudier selon les valeurs du paramètre θ. Lorsque $\theta = 0$, on retrouve le schéma explicite que nous avons déjà étudié. Lorsque $\theta = 1$, le schéma est dit *totalement implicite*. La difficulté posée par les cas où $\theta > 0$ est l'inversion de la matrice $\left(\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A\right)$, mais nous allons voir que cela peut se révéler payant en termes d'efficacité.

2.3.2 Étude du θ-schéma

Il s'écrit sous forme scalaire

$$\begin{aligned}\frac{u_j^{n+1} - u_j^n}{\Delta t} &- (1-\theta)\kappa\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} \\ &- \theta\kappa\frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{\Delta x^2} = f(t^n + \Delta t/2, x_j)\end{aligned}$$

($\forall n \in \{0, \ldots, N-1\}$, $\forall j \in \{1, \ldots, J\}$). On définit l'erreur de consistance au temps t^n et au point x_j par

$$\begin{aligned}\varepsilon_j^n(u) = \quad & \frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} \\ & - \kappa\left((1-\theta)\frac{u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})}{\Delta x^2}\right. \\ & \qquad \left. + \theta\frac{u(t^{n+1}, x_{j+1}) - 2u(t^{n+1}, x_j) + u(t^{n+1}, x_{j-1})}{\Delta x^2}\right) \\ & - f(t^n + \Delta t/2, x_j)\end{aligned}$$

où u est la solution de (2.1).

Proposition 6
Supposons que $u \in \mathcal{C}_b^{3,4}([0,T],[0,1])$ est solution de (2.1) où $f \in \mathcal{C}^2([0,T]\times[0,1])$. Alors, il existe $C \in \mathbb{R}$ et $D \in \mathbb{R}$ tels que

$$\max_{n\in\{0,\ldots,N-1\}} ||\varepsilon^n(u)||_\infty \leq C\,|1-2\theta|\,\Delta t + D(\Delta t^2 + \Delta x^2)$$

(le θ-schéma est consistant pour tout $\theta \in \mathbb{R}$). □

Démonstration

Elle repose sur le même principe que celle de la proposition 5, en comparant cette fois les différences finies à $\partial_t u(t^n + \Delta t/2, x_j)$ et à $\partial^2_{x,x} u(t^n + \Delta t/2, x_j)$.

Évaluons $\frac{u(t^{n+1},x_j)-u(t^n,x_j)}{\Delta t} - \partial_t u(t^n + \Delta t/2, x_j)$.

On a supposé que $u(\cdot, x_j) \in \mathcal{C}^3([0,T])$, donc

$$u(t^{n+1}, x_j) = u(t^n + \Delta t/2, x_j) + \frac{\Delta t}{2}\partial_t u(t^n + \Delta t/2, x_j) + \frac{\Delta t^2}{8}\partial^2_{t,t} u(t^n + \Delta t/2, x_j) + \mathcal{O}(\Delta t^3)$$

et

$$u(t^n, x_j) = u(t^n + \Delta t/2, x_j) - \frac{\Delta t}{2}\partial_t u(t^n + \Delta t/2, x_j) + \frac{\Delta t^2}{8}\partial^2_{t,t} u(t^n + \Delta t/2, x_j) + \mathcal{O}(\Delta t^3),$$

ce qui donne

$$\frac{u(t^{n+1}, x_j) - u(t^n, x_j)}{\Delta t} = \partial_t u(t^n + \Delta t/2, x_j) + \mathcal{O}(\Delta t^2).$$

Évaluons $(1-\theta)\frac{u(t^n,x_{j+1})-2u(t^n,x_j)+u(t^n,x_{j-1})}{\Delta x^2} + \theta\frac{u(t^{n+1},x_{j+1})-2u(t^{n+1},x_j)+u(t^{n+1},x_{j-1})}{\Delta x^2} - \partial^2_{x,x} u(t^n + \Delta t/2, x_j)$.

En faisant la même opération que lors de la démonstration de la proposition 5, on obtient

$$\begin{aligned}
(1-\theta)&\frac{u(t^n, x_{j+1}) - 2u(t^n, x_j) + u(t^n, x_{j-1})}{\Delta x^2} \\
&\quad + \theta\frac{u(t^{n+1}, x_{j+1}) - 2u(t^{n+1}, x_j) + u(t^{n+1}, x_{j-1})}{\Delta x^2} \\
&\quad = (1-\theta)\partial^2_{x,x} u(t^n, x_j) + \theta\partial^2_{x,x} u(t^{n+1}, x_j) + \mathcal{O}(\Delta x^2) \\
&= \partial^2_{x,x} u(t^n + \Delta t/2, x_j) + (2\theta - 1)\frac{\Delta t}{2}\partial^3_{t,x,x} u(t^n + \Delta t/2, x_j) + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2)
\end{aligned}$$

si $\partial^2_{x,x} u(\cdot, x_j) \in \mathcal{C}^2([0,T])$, ce qui est le cas puisque $\partial^2_{x,x} u(\cdot, x_j) = 1/\kappa\,(\partial_t u(\cdot, x_j) - f(\cdot, x_j))$. Ainsi,

$$\begin{aligned}
\varepsilon^n_j(u) = \quad & \partial_t u(t^n + \Delta t/2, x_j) + \mathcal{O}(\Delta t^2) \\
& - \kappa\partial^2_{x,x} u(t^n + \Delta t/2, x_j) + (1-2\theta)\frac{\kappa\Delta t}{2}\partial^3_{t,x,x} u(t^n + \Delta t/2, x_j) \\
& + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2) \\
& - f(t^n + \Delta t/2, x_j),
\end{aligned}$$

soit, puisque u est solution de l'équation de la chaleur,

$$\varepsilon^n_j(u) = (1-2\theta)\frac{\kappa\Delta t}{2}\partial^3_{t,x,x} u(t^n + \Delta t/2, x_j) + \mathcal{O}(\Delta t^2) + \mathcal{O}(\Delta x^2),$$

ce qui prouve bien le résultat annoncé. □

Le θ-schéma est donc d'ordre 1 en temps et 2 en espace, sauf si $\theta = 1/2$, auquel cas il est d'ordre 2 en temps et en espace. Le 1/2-schéma est appelé *schéma de Crank-Nicolson*.

Remarque 17
Le remplacement du second membre F^n par $F^{n+1/2}$ a permis d'éliminer un terme d'erreur d'ordre 1 en temps qui ne s'annulerait pour aucune valeur de θ. On peut remplacer le second membre $f(t^n+\Delta t/2, x_j)$ par $(1-\theta)f(t^n, x_j)+\theta f(t^{n+1}, x_j)$ ou par $f((1-\theta)t^n+\theta t^{n+1}, x_j) = f(t^n+\theta\Delta t, x_j)$ pour obtenir le même résultat. □

Il nous faut maintenant nous pencher sur la convergence du θ-schéma. Pour ceci, nous notons, comme pour le schéma explicite (à ceci près que nous ne prenons pas cette fois les valeurs de l'erreur en $x = 0$ ni en $x = 1$), $e^n(u) \in \mathbb{R}^J$ le vecteur de l'erreur au temps t^n, de composantes

$$e_j^n(u) = u_j^n - u(t^n, x_j) \quad \forall n \in \{0, \dots, N\}, \forall j \in \{1, \dots, J\}.$$

La linéarité du schéma implique que ce vecteur de l'erreur vérifie l'équation

$$\left(\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A\right) e^{n+1}(u) + \left(-\frac{1}{\Delta t}I + \frac{(1-\theta)\kappa}{\Delta x^2}A\right) e^n(u) = -\varepsilon^n(u).$$

La présence d'une matrice non diagonale (lorsque $\theta \neq 0$) devant e^{n+1} rend difficile l'étude de la norme l^∞ du vecteur de l'erreur. L'étude que nous proposons ici est celle de la norme « l^2 » : on définit la norme $|||\cdot|||_2$ par

$$|||e^n|||_2 = \left(\frac{1}{J}\sum_{j=1}^{J} |e_j^n|^2\right)^{1/2}.$$

La division par J sert à normaliser (en fonction du nombre de points du maillage) : ainsi, si g est une fonction constante sur $[0, 1]$,

$$||g||_{L^2(]0,1[)} = \left|\left|\left| (g(x_j))_{j=1}^J \right|\right|\right|_2.$$

Étude de la matrice $\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A$.

La matrice $-A$ est appelée *matrice du laplacien discret en dimension 1 sur un maillage à J points.*

Lemme 3
La matrice A, de taille $J \times J$, a pour valeurs propres

$$\alpha_j = 4\sin^2\left(\frac{j\pi}{2(J+1)}\right) \text{ pour } j = 1, \dots, J,$$

et pour vecteurs propres associés, respectivement,

$$V_j = \begin{pmatrix} \sin(\frac{j\pi}{J+1}) \\ \sin(2\frac{j\pi}{J+1}) \\ \vdots \\ \sin(J\frac{j\pi}{J+1}) \end{pmatrix} \text{ pour } j = 1, \dots, J.$$

□

La démonstration de ce lemme est laissée au lecteur (elle ne fait appel qu'à des formules de trigonométrie classiques).

Remarque 18
Noter que les vecteurs propres ont pour composantes les valeurs aux points du maillage des J premiers vecteurs propres de l'opérateur exact $-\partial^2_{x,x}$. Un calcul simple (développement limité de la fonction sin à l'ordre 2) montre de plus que $4(J+1)^2\sin^2(j\pi/(2(J+1))) = j^2\pi^2 + \mathcal{O}((j\pi/(2(J+1)))^4)$, donc les valeurs propres de $A/\Delta x^2$ sont des approximations des valeurs propres de $-\partial^2_{x,x}$.

Avant de continuer l'étude de la matrice $\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A$, nous pouvons faire quelques remarques.

— Les valeurs propres de A sont toutes strictement positives.

— La matrice $\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A$ est symétrique réelle et donc diagonalisable sur $\mathbb{R}$ (ce que l'on constate aussi dans le lemme 3).

— La matrice $\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A$ a pour valeurs propres les réels

$$\mu_j = \frac{1}{\Delta t} + \frac{\theta\kappa}{\Delta x^2}\alpha_j$$

et pour vecteurs propres associés les vecteurs V_j (pour $j = 1, \ldots, J$). Pour $\theta \geq 0$, les μ_j étant strictement positifs, cette matrice est définie positive (on fait dorénavant l'hypothèse que $\theta \geq 0$).

— La matrice $\frac{1}{\Delta t}I + \frac{\theta\kappa}{\Delta x^2}A$ est diagonalisable et à valeurs propres non nulles, elle est donc inversible (ce qui assure l'existence d'une solution à l'équation approchée donnée par le θ-schéma).

Notons, pour tout $\theta \in \mathbb{R}$, B_θ la matrice $I + \frac{\theta\kappa\Delta t}{\Delta x^2}A$. Le vecteur de l'erreur satisfait à l'équation

$$e^{n+1}(u) = {B_\theta}^{-1}\left[B_{\theta-1}e^n(u) - \Delta t\varepsilon^n(u)\right]$$

et la solution numérique satisfait à l'équation

$$U^{n+1} = {B_\theta}^{-1}\left[B_{\theta-1}U^n + \Delta t F^{n+1/2}\right].$$

B_θ est bien sûr elle-même symétrique réelle, diagonalisable, définie positive, inversible.

Posons maintenant

$$L_\theta = {B_\theta}^{-1}B_{\theta-1}.$$

La formule donnant le vecteur de l'erreur est

$$e^{n+1}(u) = L_\theta e^n(u) - \Delta t {B_\theta}^{-1}\varepsilon^n(u).$$

La matrice L_θ est appelée *matrice d'amplification* du schéma.

On obtient aisément une majoration de la norme $|||\cdot|||_2$ du vecteur de l'erreur :

$$\begin{aligned}\left|\left|\left|e^{n+1}(u)\right|\right|\right|_2 &\leq |||L_\theta e^n(u)|||_2 + \Delta t\left|\left|\left|{B_\theta}^{-1}\varepsilon^n(u)\right|\right|\right|_2 \\ &\leq |||L_\theta|||_2\,|||e^n(u)|||_2 + \Delta t\left|\left|\left|{B_\theta}^{-1}\right|\right|\right|_2|||\varepsilon^n(u)|||_2\end{aligned}$$

où la norme matricielle $|||\cdot|||_2$ est la norme induite par la norme vectorielle $|||\cdot|||_2$:

$$|||M|||_2 = \sup_{v\in\mathbb{R}^{J^*}} \frac{|||Mv|||_2}{|||v|||_2}.$$

Remarquer que $|||M|||_2 = ||M||_2$.

On sait que $||M||_2 = (\rho(M^*M))^{1/2}$ où $\rho(C)$ désigne le rayon spectral de la matrice C, et que $||M||_2 = \rho(M)$ si M est *normale*, c'est-à-dire vérifie $M^*M = MM^*$ (voir [13] ou [2] par exemple). La matrice B_θ a pour valeurs propres les

$$\beta_j = 1 + \frac{\theta\kappa\Delta t}{\Delta x^2}\alpha_j = \Delta t\mu_j \text{ pour } j = 1,\dots,J.$$

Toutes les valeurs propres de B_θ sont donc supérieures ou égales à 1 (car $\theta \geq 0$). Cette matrice est inversible et toutes les valeurs propres de ${B_\theta}^{-1}$ sont dans l'intervalle $]0,1]$. D'autre part, B_θ est symétrique réelle, donc ${B_\theta}^{-1}$ est symétrique réelle. Ainsi,

$$\left|\left|\left|{B_\theta}^{-1}\right|\right|\right|_2 = \left|\left|{B_\theta}^{-1}\right|\right|_2 = \rho({B_\theta}^{-1}) \leq 1$$

car $||M||_2 = \rho(M)$ pour M symétrique réelle. Donc

$$\left|\left|\left|e^{n+1}(u)\right|\right|\right|_2 \leq |||L_\theta|||_2 \, |||e^n(u)|||_2 + \Delta t \, |||\varepsilon^n(u)|||_2 .$$

Un raisonnement par récurrence nous permet d'en déduire que

$$\left|\left|\left|e^{n+1}(u)\right|\right|\right|_2 \leq {|||L_\theta|||_2}^{n+1} \left|\left|\left|e^0(u)\right|\right|\right|_2 + \Delta t \sum_{m=0}^{n} {|||L_\theta|||_2}^{n-m} \, |||\varepsilon^m(u)|||_2 .$$

Comme de plus $\left|\left|\left|e^0(u)\right|\right|\right|_2 = 0$,

$$\left|\left|\left|e^{n+1}(u)\right|\right|\right|_2 \leq \Delta t \sum_{m=0}^{n} {|||L_\theta|||_2}^{n-m} \, |||\varepsilon^m(u)|||_2 .$$

Or $\left(\sum_{j=1}^{J} |\varepsilon_j^m(u)|^2\right)^{1/2} \leq \left(J \, ||\varepsilon^m(u)||_\infty^2\right)^{1/2}$, donc $|||\varepsilon^m(u)|||_2 \leq ||\varepsilon^m(u)||_\infty$ et

$$\begin{aligned}\left|\left|\left|e^{n+1}(u)\right|\right|\right|_2 &\leq \Delta t \sum_{m=0}^{n} {|||L_\theta|||_2}^{n-m} \, ||\varepsilon^m(u)||_\infty \\ &\leq \Delta t \sum_{m=0}^{n} {|||L_\theta|||_2}^{n-m} \left[C\,|1-2\theta|\,\Delta t + D(\Delta t^2 + \Delta x^2)\right].\end{aligned}$$

Supposons maintenant que $||L_\theta||_2 \leq 1$. Alors

$$\left|\left|\left|e^{n+1}(u)\right|\right|\right|_2 \leq T\left[C(1-2\theta)\Delta t + D(\Delta t^2 + \Delta x^2)\right] \quad \forall n \in \{0,\dots,N-1\}.$$

Si $||L_\theta||_2 \leq 1$ (et sous les hypothèses de la proposition 6), le schéma (2.10) est « convergent pour la norme $|||\cdot|||_2$ » (bien que cela n'ait pas le sens habituel, car $|||\cdot|||_2$ dépend de J).

Cherchons donc des conditions (simples) sous lesquelles $||L_\theta||_2 \leq 1$. Rappelons que

$$L_\theta = {B_\theta}^{-1} B_{\theta-1} = \left(I + \frac{\theta\kappa\Delta t}{\Delta x^2} A\right)^{-1} \left(I - \frac{(1-\theta)\kappa\Delta t}{\Delta x^2} A\right)$$

et que

- B_θ et ${B_\theta}^{-1}$ ont les mêmes vecteurs propres et des valeurs propres inverses ;
- $\left(I - \frac{(1-\theta)\kappa\Delta t}{\Delta x^2} A\right)$ a les mêmes vecteurs propres que B_θ, les vecteurs V_j, qui sont les vecteurs propres de A, et a pour valeurs propres les réels

$$\sigma_j = 1 - (1-\theta)\kappa \frac{\Delta t}{\Delta x^2} \alpha_j \text{ pour } j = 1, \ldots, J.$$

Les vecteurs propres de L_θ sont donc les vecteurs V_j. Les valeurs propres associées sont les réels

$$\lambda_j = \frac{\sigma_j}{\beta_j} = \frac{1 - 4(1-\theta)\kappa \frac{\Delta t}{\Delta x^2} \sin^2\left(\frac{j\pi}{2(J+1)}\right)}{1 + 4\theta\kappa \frac{\Delta t}{\Delta x^2} \sin^2\left(\frac{j\pi}{2(J+1)}\right)},$$

pour $j = 1, \ldots, J$. Or $||L_\theta||_2 = \rho(L_\theta)$, puisque L_θ est une matrice symétrique réelle[7]. Donc $||L_\theta||_2 \leq 1$ si et seulement si $\rho(L_\theta) \leq 1$, c'est-à-dire si et seulement si $\lambda_j \leq 1 \ \forall j \in \{1, \ldots, J\}$.

Remarque 19

La condition $\rho(L_\theta) \leq 1$ est appelée *condition de stabilité de Von Neumann*. Puisque L_θ est symétrique réelle, cette condition de stabilité est équivalente à la condition de stabilité l^2, $||L_\theta||_2 \leq 1$, selon laquelle la norme de la matrice d'amplification du schéma est inférieure à 1[8]. □

Proposition 7

Si $\theta \geq 1/2$, $||L_\theta||_2 \leq 1$.
Si $\theta \in [0, 1/2[$, $||L_\theta||_2 \leq 1$ pour tout J si et seulement si $\kappa \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\theta)}$. □

Démonstration

Étudions la fonction g définie par

$$g(s) = \frac{1 - 4(1-\theta)\kappa \frac{\Delta t}{\Delta x^2} s}{1 + 4\theta\kappa \frac{\Delta t}{\Delta x^2} s}$$

sur le segment $[0, 1]$. Elle est décroissante, donc elle atteint son maximum et son minimum en 1 et en 0 respectivement. Donc $|g(s)|$ est maximal lorsque $s = 0$ ou $s = 1$. Or $|g(0)| = 1$. Étudions $|g(1)|$ (en fonction des valeurs de $\theta \geq 0$).

$$g(1) = \frac{1 - 4(1-\theta)\kappa \frac{\Delta t}{\Delta x^2}}{1 + 4\theta\kappa \frac{\Delta t}{\Delta x^2}},$$

7. L_θ est en effet le produit de deux matrices, ${B_\theta}^{-1}$ et $I - \frac{(1-\theta)\kappa\Delta t}{\Delta x^2} A$, qui sont symétriques et qui commutent, ayant les mêmes vecteurs propres.

8. Voir la proposition 8 pour plus de détails.

donc $|g(1)| \leq 1$ si et seulement si

$$-1 - 4\theta\kappa\frac{\Delta t}{\Delta x^2} \leq 1 - 4(1-\theta)\kappa\frac{\Delta t}{\Delta x^2} \leq 1 + 4\theta\kappa\frac{\Delta t}{\Delta x^2},$$

ce qui est équivalent à

$$\begin{cases} \kappa\frac{\Delta t}{\Delta x^2}(4 - 8\theta) \leq 2 \\ \text{et } \kappa\frac{\Delta t}{\Delta x^2} \geq 0 \end{cases}$$

soit (puisque $\theta \geq 0$, $\kappa \geq 0$ et $\Delta t \geq 0$)

$$\begin{cases} \theta \geq 1/2 \\ \text{ou } \kappa\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\theta)}. \end{cases}$$

On a démontré que si $\theta \geq 1/2$ ou $\kappa\frac{\Delta t}{\Delta x^2} \leq \frac{1}{2(1-2\theta)}$, $||L_\theta||_2 \leq 1$ $(\forall J \in \mathbb{N}^*)$. Pour montrer la réciproque, il suffit de remarquer que

$$\lim_{J\to\infty} \sin^2\left(\frac{J\pi}{2(J+1)}\right) = 1,$$

et que donc

$$\lim_{J\to\infty} \lambda_J = \frac{1 - 4(1-\theta)\kappa\frac{\Delta t}{\Delta x^2}}{1 + 4\theta\kappa\frac{\Delta t}{\Delta x^2}} = g(1).$$

□

Une conséquence directe est le

Théorème 7
Supposons vérifiées les hypothèses de la proposition 6. Soit $\theta \in \mathbb{R}_+$. Supposons que $\theta \geq 1/2$ ou $\kappa\Delta t/\Delta x^2 \leq 1/2(1-2\theta)$. Alors, le schéma (2.10) vérifie : $\exists C \in \mathbb{R}$ et $D \in \mathbb{R}$ tels que

$$|||e^n(u)|||_2 \leq T\left[C|1-2\theta|\Delta t + D(\Delta t^2 + \Delta x^2)\right] \quad \forall n \in \{0,\ldots,N\}.$$

□

Remarque 20
Le schéma de Crank-Nicolson, obtenu pour $\theta = 1/2$, est à la fois d'ordre 2 et *inconditionnellement convergent.* □

Le corollaire qui suit (dont la démonstration est laissée au lecteur), qui est au théorème 7 ce que le corollaire 1 est au théorème 6, précise la convergence de l'interpolée de la solution numérique vers la solution exacte.

Corollaire 2
Soit $\overline{u}(t,x)$ la fonction affine en t (à x fixé) et affine en x (à t fixé) sur chaque pavé $[t^n, t^{n+1}] \times [x_j, x_{j+1}]$ telle que $\overline{u}(t^n, x_j) = u_j^n$ pour tout $n \in \{0,\ldots,N\}$ et tout $j \in \{0,\ldots,J+1\}$ (où les u_j^n

sont donnés par le schéma (2.10)). Sous les hypothèses du théorème 7, il existe $C \in \mathbb{R}$, $D \in \mathbb{R}$ et $E \in \mathbb{R}$ tels que

$$||\overline{u}(t,\cdot) - u(t,\cdot)||_{L^2(]0,1[)} \leq 2T\left[C|1-2\theta|\Delta t + D(\Delta t^2 + \Delta x^2)\right] + E(\Delta t^2 + \Delta x^2) \quad \forall t \in [0,T].$$

□

Remarque 21
C'est en fin de compte ce corollaire qui justifie le choix de la norme $|||\cdot|||_2$ dépendant du nombre de points du maillage J. □

La convergence du θ-schéma en norme $|||\cdot|||_2$ repose sur deux propriétés :
— sa consistance ($\lim_{N,J\to+\infty} \sup_{n\in\{0,\dots,N\}} ||\varepsilon^n(u)||_\infty = 0$) ;
— la majoration $||L_\theta||_2 \leq 1$. Cette majoration assure la stabilité au sens l^2 de la solution numérique : supposons pour simplifier que le terme source est nul, on a alors $U^{n+1} = L_\theta U^n$ et $\left|\left|\left|U^{n+1}\right|\right|\right|_2 \leq ||L_\theta||_2 \, |||U^n|||_2 \leq |||U^n|||_2$.

Rappelons que la consistance en norme l^∞ implique la consistance en norme $|||\cdot|||_2$. La convergence semble donc être une conséquence de la consistance et de la stabilité. Nous allons maintenant formaliser ce résultat et le généraliser à une norme quelconque pour démontrer le théorème de Lax. On considère à partir de maintenant un algorithme de la forme générale

$$B_1 U^{n+1} + B_0 U^n = F^n \quad \forall n \in \{0,\dots,N\} \tag{2.11}$$

où $U^0 \in \mathbb{R}^J$, $F^n \in \mathbb{R}^J$, $B_0, B_1 \in \mathcal{M}_J(\mathbb{R})$ sont donnés et B_1 est une matrice inversible, et cela pour tout $J \in \mathbb{N}$. Il est implicite dans cette définition que les données (les deux matrices et le terme source) dépendent de paramètres qui sont Δt et Δx : donc la solution U aussi.
On supposera pour simplifier, comme dans tout ce qui précède, que U^0 est constitué des valeurs ponctuelles de u^0, donnée initiale exacte.

Définition 3
1) On appelle *erreur de consistance* du schéma (2.11) à l'itération n le vecteur $\varepsilon^n(u) \in \mathbb{R}^J$ défini par

$$\varepsilon^n(u) = B_1\left(u(t^{n+1},x_j)\right)_{j=1}^J + B_0\left(u(t^n,x_j)\right)_{j=1}^J - F^n$$

où u est la solution de (2.1).
2) On dit que (2.11) est *consistant* avec (2.1) dans l'espace vectoriel de fonctions $\mathcal{U}$ pour la norme $||\cdot||$ si et seulement si $\forall u \in \mathcal{U}$ solution de (2.1),

$$\lim_{N,J\to+\infty} \sup_{n\in\{0,\dots,N\}} ||\varepsilon^n(u)|| = 0.$$

3) On dit que (2.11) est d'ordre p en temps et q en espace dans $\mathcal{U}$ si et seulement si $\forall u \in \mathcal{U}$ solution de (2.1), $\exists C_0(u) \in \mathbb{R}$ tel que

$$\sup_{n\in\{0,\dots,N\}} ||\varepsilon^n(u)|| \leq C_0(u)\left(\Delta t^p + \Delta x^q\right) \quad \text{pour tous } \Delta t, \Delta x \in \mathbb{R}_+^*.$$

4) On dit que (2.11) est *stable* pour la norme $||\cdot||$ si et seulement si pout tout $T \in \mathbb{R}_+$ il existe $C_1(T), C_2(T) \in \mathbb{R}$ tels que

$$\sup_{n\in\{0,\dots,N\}} ||U^n|| \leq C_1(T) \left|\left|U^0\right|\right| + C_2(T) \sup_{n\in\{0,\dots,N-1\}} ||F^n||$$

pour tous $\Delta t, \Delta x \in \mathbb{R}_+^*,$ pour tout $U^0 \in \mathbb{R}^J,$ pour tout N tel que $N\Delta t \leq T.$

5) On dit que (2.11) est convergent dans $\mathcal{U}$ pour la norme $||\cdot||$ si et seulement si

$$\sup_{n\in\mathbb{N} \text{ tel que } n\Delta t\leq T} \left|\left| \left(u_j^n - u(t^n, x_j)\right)_{j=1}^J \right|\right| = \sup_{n\in\mathbb{N} \text{ tel que } n\Delta t\leq T} ||e^n|| \longrightarrow_{\Delta t,\Delta x\to 0} 0$$

pour tout $u \in \mathcal{U}$ solution de (2.1). □

Dans cette définition, la norme $||\cdot||$ peut dépendre de J. On peut aussi ajouter des conditions à la stabilité et à la convergence : par exemple, liant Δt et Δx.

Théorème 8 (théorème de Lax)
On considère un schéma de la forme (2.11) dont on suppose qu'il est consistant dans $\mathcal{U}$ et stable. Il est convergent dans $\mathcal{U}$. □

Remarque 22
Ce théorème est parfois appelé « théorème d'équivalence de Lax » mais la réciproque demanderait de nombreuses hypothèses supplémentaires sur la forme du schéma. Le meilleur moyen d'énoncer (et de montrer rigoureusement) ce résultat d'équivalence est d'écrire le schéma en variable de Fourier. On peut consulter [15]... □

Démonstration
Supposons (2.11) consistant et stable. La linéarité du schéma implique que le vecteur de l'erreur vérifie

$$B_1 e^{n+1}(u) + B_0 e^n(u) = -\varepsilon^n(u).$$

Puisque le schéma est stable pour toute donnée initiale et tout second membre, il l'est pour la donnée initiale $e^0(u)$ et pour le second membre $-(\varepsilon^n(u))_{n=0}^{N-1}$: $\exists C_1(u) \in \mathbb{R}, C_2(u) \in \mathbb{R}$ tels que

$$\sup_{n\in\{0,\dots,N\}} ||e^n(u)|| \leq C_1(u) \left|\left|e^0(u)\right|\right| + C_2(u) \sup_{n\in\{0,\dots,N-1\}} ||\varepsilon^n(u)||.$$

Or $e^0(u) = 0$, donc

$$\sup_{n\in\{0,\dots,N\}} ||e^n(u)|| \leq C_2(u) \sup_{n\in\{0,\dots,N-1\}} ||\varepsilon^n(u)||.$$

Le schéma (2.11) étant de plus consistant, $\sup_{n\in\{0,\dots,N\}} ||e^n(u)|| \longrightarrow_{N,J\to\infty} 0$: le schéma est convergent. □

Remarque 23
Le fait que la norme $||\cdot||$ dépende de J ou que les vecteurs U^n et e^n soient de taille variable avec J ne modifie bien entendu pas la démonstration que nous venons de faire. Ce détail a été omis afin de simplifier les définitions et la démonstration. □

Remarque 24
En réalité, aucune hypothèse concernant l'EDP à résoudre n'a été nécessaire : peu importe que ce soit (2.1). La propriété importante qui a été utilisée est la linéarité du schéma (noter qu'un schéma linéaire ne peut pas être consistant avec une EDP non linéaire ; remarquer qu'en revanche un schéma non linéaire peut être consistant avec une EDP linéaire). □

Applications.
1) Le schéma explicite est convergent en norme $||\cdot||_\infty$ si $\kappa\Delta t/\Delta x^2 \leq 1/2$.
2) Le θ-schéma est convergent en norme $|||\cdot|||_2$ sous les conditions évoquées dans le théorème 7.

Remarque 25
Pour le θ-schéma avec $\theta < 1/2$, nous n'avons pas montré que la condition dite de CFL était nécessaire, nous avons seulement vu qu'elle était suffisante. Il est cependant possible de montrer qu'elle est effectivement nécessaire en choisissant une condition initiale dont la projection sur le vecteur propre V_J de la matrice L_θ associé à la valeur propre de plus grand module ne tend pas vers 0 lorsque J tend vers l'infini. Le phénomène de non-convergence lorsque le pas de temps est trop grand s'observe facilement. □

Une indication du mauvais comportement du schéma lorsque la condition sur le pas de temps est violée est donnée par le fait que ce schéma n'est alors pas stable. En effet, il s'écrit sous forme matricielle

$$U^{n+1} = L_\theta U^n + \Delta t {B_\theta}^{-1} F^n.$$

La proposition suivante va nous permettre de conclure.

Proposition 8
Un schéma de la forme

$$U^{n+1} = BU^n + \Delta t D F^n$$

est stable si et seulement si $\exists C \in \mathbb{R}$ indépendant de N et J tel que

$$\sup_{n\in\mathbb{N}} ||B^n|| \leq C.$$

□

Démonstration
Supposons que le schéma est stable : $\forall U^0$, $\forall \left(F^n\right)_{n\in\{1,\dots,N\}}$,

$$\sup_{n\in\{0,\dots,N\}} ||U^n|| \leq C_1 \left|\left|U^0\right|\right| + C_2 \sup_{n\in\{0,\dots,N\}} ||F^n||.$$

Prenons $F^n = 0\ \forall n \in \{0, \dots, N\}$. Alors, $U^n = B^n U^0$ et $\sup_{n\in\{0,\dots,N\}} ||U^n|| \leq C_1 \left||U^0|\right|$, d'où $\sup_{n\in\{0,\dots,N\}} \left||B^n U^0|\right| \leq C_1 \left||U^0|\right|$. Donc

$$||B^n|| = \sup_{U\in\mathbb{R}^J\setminus\{0\}} \frac{||B^n U||}{||U||} \leq C_1 \quad \forall n \in \{0, \dots, N\},$$

et cela est vrai pour tout N.
Montrons la réciproque : supposons qu'il existe $C \in \mathbb{R}$ tel que $\sup_{n\in\mathbb{N}} ||B^n|| \leq C$.

$$U^n = B^n U^0 + \Delta t \sum_{m=0}^{n-1} B^{n-m-1} DF^m,$$

donc

$$\begin{aligned} ||U^n|| &\leq ||B^n|| \left||U^0|\right| + \Delta t \sum_{m=0}^{n-1} \left||B^{n-m-1}|\right| ||DF^m|| \\ &\leq C \left||U^0|\right| + n\Delta t C ||D|| \sup_{m\in\{0,\dots,n-1\}} ||F^m|| \\ &\leq C \left||U^0|\right| + CT ||D|| \sup_{m\in\{0,\dots,N\}} ||F^m|| , \end{aligned}$$

ce qui termine la démonstration. □

Dans le cas du θ-schéma,

$$B = L_\theta = \left(I + \frac{\theta\kappa\Delta t}{\Delta x^2} A\right)^{-1} \left(I - \frac{(1-\theta)\kappa\Delta t}{\Delta x^2} A\right).$$

C'est une matrice symétrique réelle, donc $||B||_2 = \rho(B)$ et $||B^n||_2 = ||B||_2^n$. Or nous avons montré (proposition 7) que $||B||_2 > 1$ si la condition sur le pas de temps n'est pas vérifiée. Cela prouve que le schéma n'est pas stable.

Exercice 2
Étudier (consistance, ordre, stabilité ?) le *schéma saute-mouton* défini par

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} - \kappa \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = f_j^n$$

(cf. partiel du 6 avril 2003).

2.3.3 Quelques résultats numériques

On présente ici quelques résultats numériques obtenus avec le θ-schéma, pour différentes valeurs de $\theta \in [0, 1]$, différentes conditions initiales, différentes valeurs du pas d'espace et différentes valeurs du pas de temps. Le coefficient de diffusion est $\kappa = 1$.

Les premiers résultats numériques, figures 2.2 à 2.5, ont été obtenus pour la condition initiale $u^0(x) = \sin(2\pi x)$. La solution exacte est alors donnée par $u(t, x) = \sin(2\pi x)e^{-4\pi t}$. Le temps final

(pour lequel les solutions approchées ont été calculées) est $T = 0,02$. Le rapport $\Delta t/\Delta x^2$ est fixé à $0,5$ pour le schéma explicite, ce qui conduit à effectuer 209 pas de temps. Pour les schémas de Crank-Nicolson et implicite, on fixe le nombre de pas de temps à 50 (c'est arbitraire).

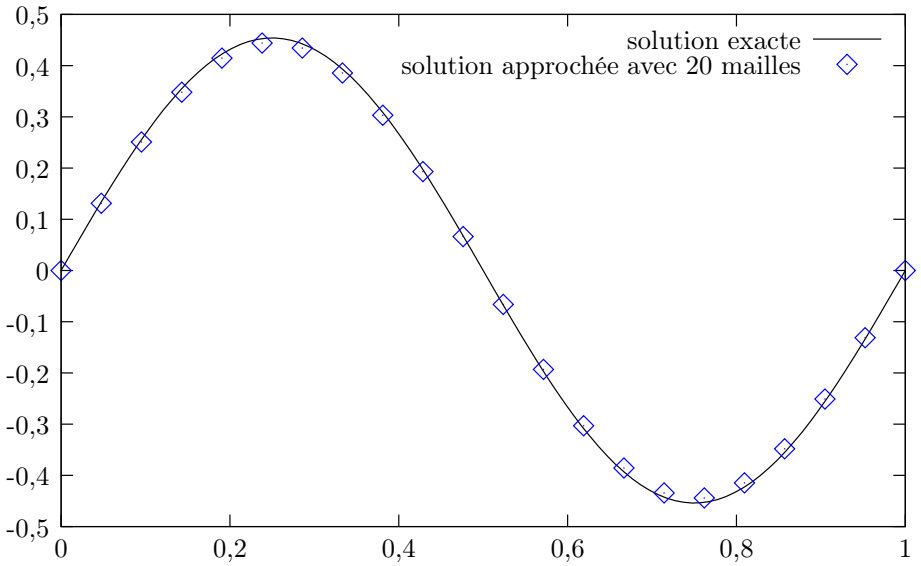

FIGURE 2.2 – Condition initiale sinusoïdale, 20 mailles avec le schéma explicite.

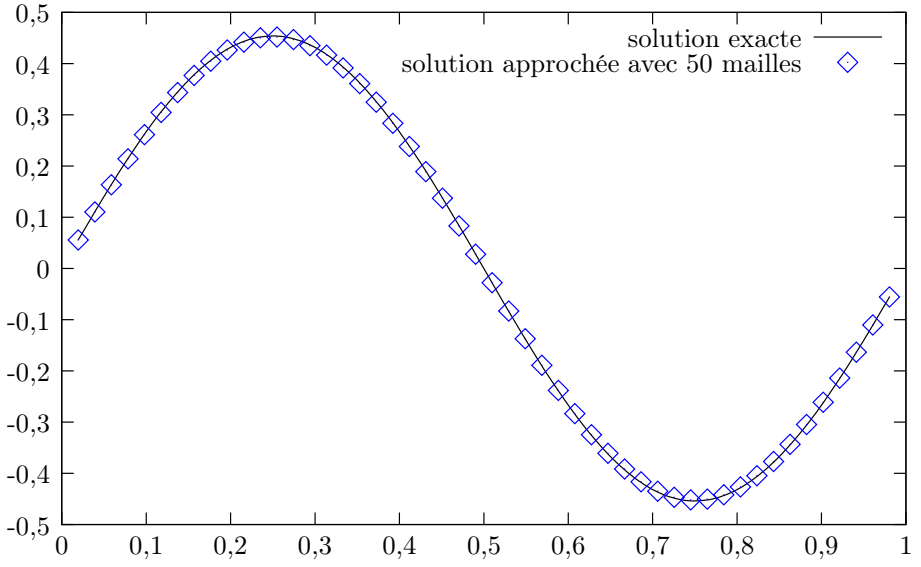

FIGURE 2.3 – Condition initiale sinusoïdale, 50 mailles avec le schéma explicite.

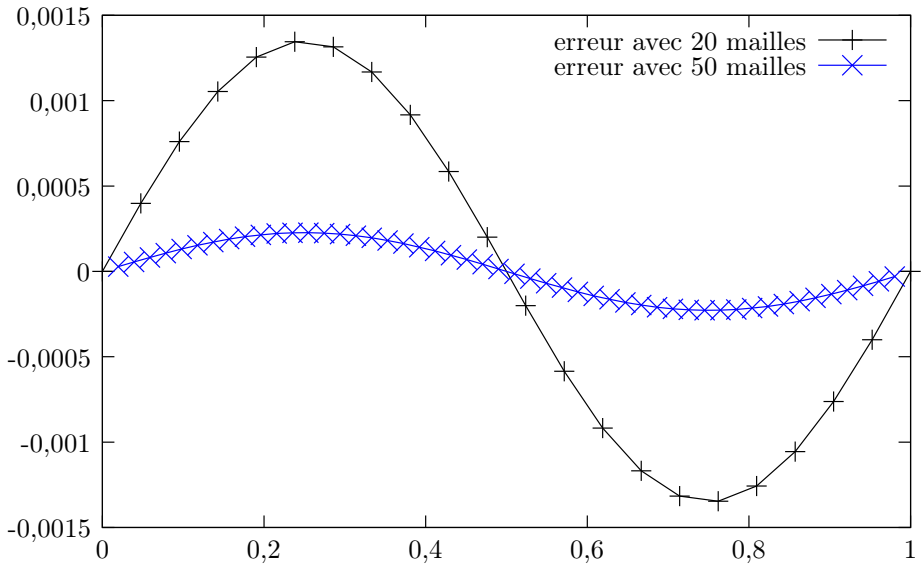

FIGURE 2.4 – Condition initiale sinusoïdale, comparaison des erreurs (schéma explicite).

On constate sur cette figure que l'erreur sur le maillage fin est inférieure à l'erreur sur le maillage grossier.

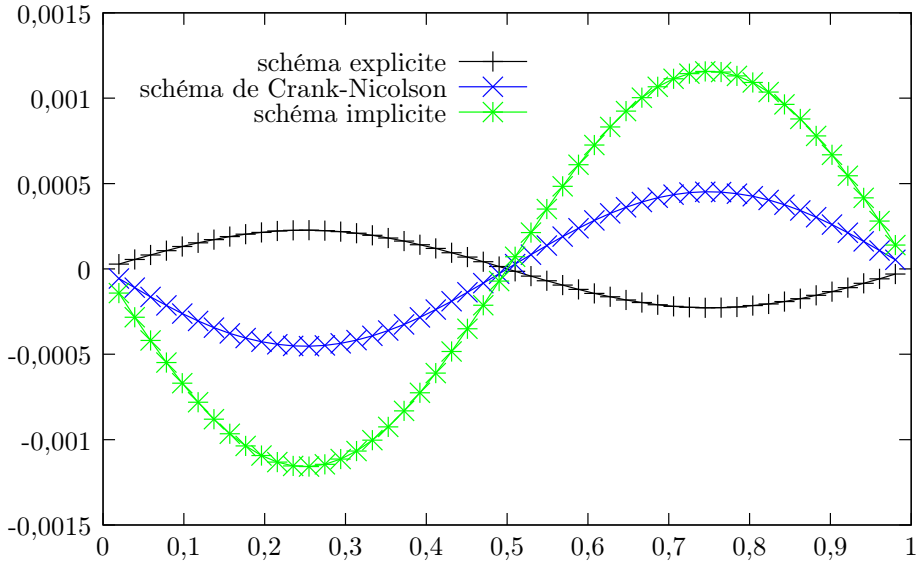

FIGURE 2.5 – Comparaison des erreurs avec différentes valeurs de θ, 50 mailles.

Pour cette condition initiale particulière et pour ce maillage, l'erreur du schéma explicite est inférieure à celle des schémas de Crank-Nicolson et du schéma implicite. Cependant, il faut se rappeler que ces derniers autorisent à faire le calcul en un nombre de pas de temps beaucoup plus petit puisqu'ils sont stables sans aucune condition sur ce pas de temps. Voici d'ailleurs le résultat obtenu avec le schéma explicite lorsque le rapport $\Delta t/\Delta x^2$ vaut $1,1$.

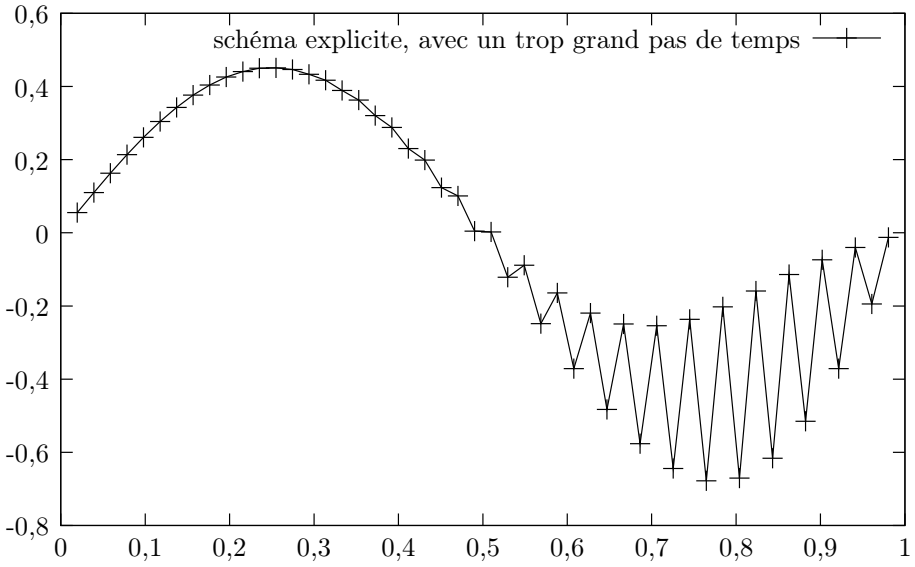

FIGURE 2.6 – Calcul où la condition de stabilité n'est pas vérifiée.

Ce résultat oscillant est dû à l'« explosion » des petites erreurs numériques, les calculs n'étant pas faits en arithmétique exacte. En effet, dans le cas précis de la condition initiale sinusoïdale, la condition exhibée pour le pas de temps n'est *théoriquement* pas nécessaire à la stabilité du résultat [9]. Un résultat plus convainquant est proposé pour le cas-test suivant.

Maintenant observons les résultats obtenus avec pour condition initiale la fonction de classe $\mathcal{C}^\infty([0,1])$ à support compact dans $]0,1[$

$$u^0(x) = \begin{cases} 0 \text{ si } x \leq 0,4 \\ e^{1-\frac{1}{1-100(x-1/2)^2}} \text{ si } 0,4 < x \leq 0,6 \\ 0 \text{ si } x > 0,6. \end{cases}$$

9. Nous ne nous étendrons pas sur ce phénomène, très marginal pour notre propos. Il faut cependant savoir que ce type de problèmes d'arithmétique non exacte peu être crucial dans certaines applications en analyse numérique.

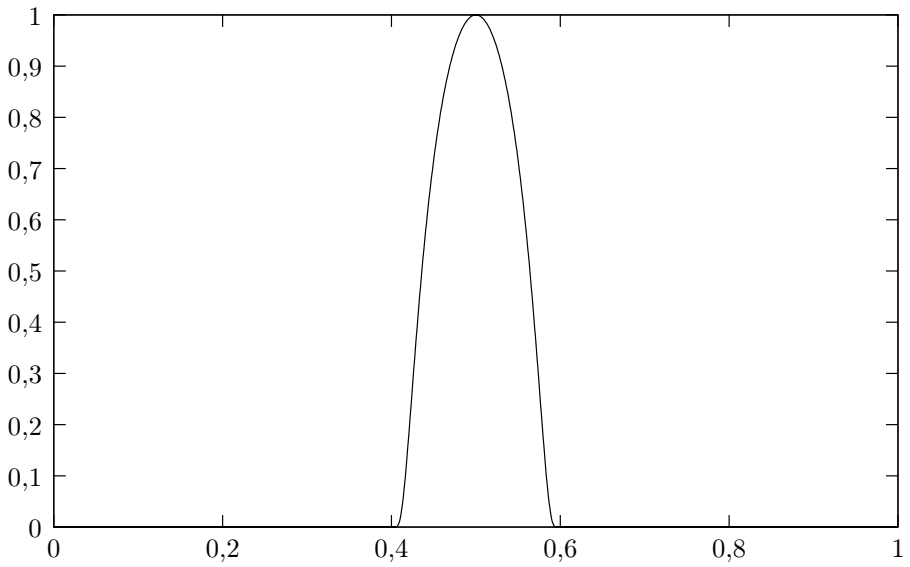

FIGURE 2.7 – Condition initiale régulière à support compact.

Le temps final choisi est $T = 0,01$. Remarquer que pour tout temps strictement positif, la solution n'est plus à support compact car elle est analytique réelle (cf. partiel du 19 avril 2005) et non identiquement nulle[10]. Comme nous n'avons pas cette fois de solution explicite à notre disposition, nous comparons les résultats obtenus pour les schémas explicite, de Crank-Nicolson et implicite avec 50 mailles à une *solution de référence* calculée avec le schéma explicite et 1000 mailles.

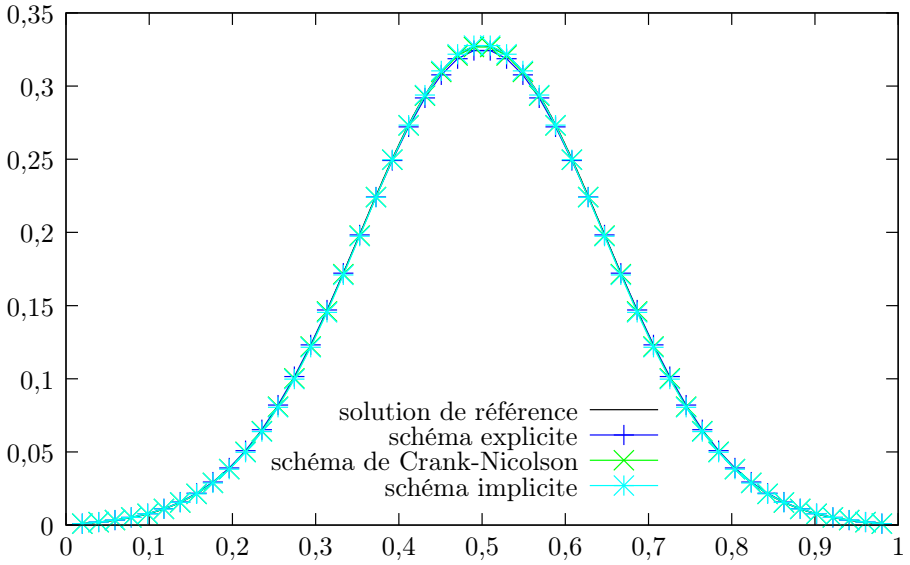

FIGURE 2.8 – Solutions avec 50 mailles.

10. La seule fonction analytique réelle et identiquement nulle sur un intervalle est la fonction nulle.

Les différences entre les résultats sont plus facilement appréciables sur le zoom proposé par la figure 2.9. On y remarque que le schéma de Crank-Nicolson, d'ordre 2 en temps et en espace, offre avec seulement 50 mailles une solution déjà très proche de la solution de référence.

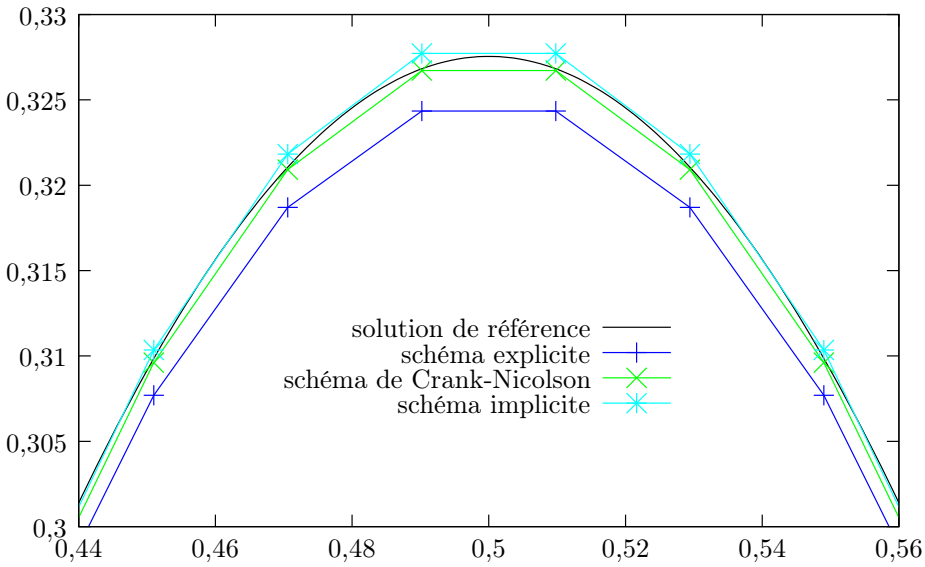

FIGURE 2.9 – Solutions avec 50 mailles, zoom.

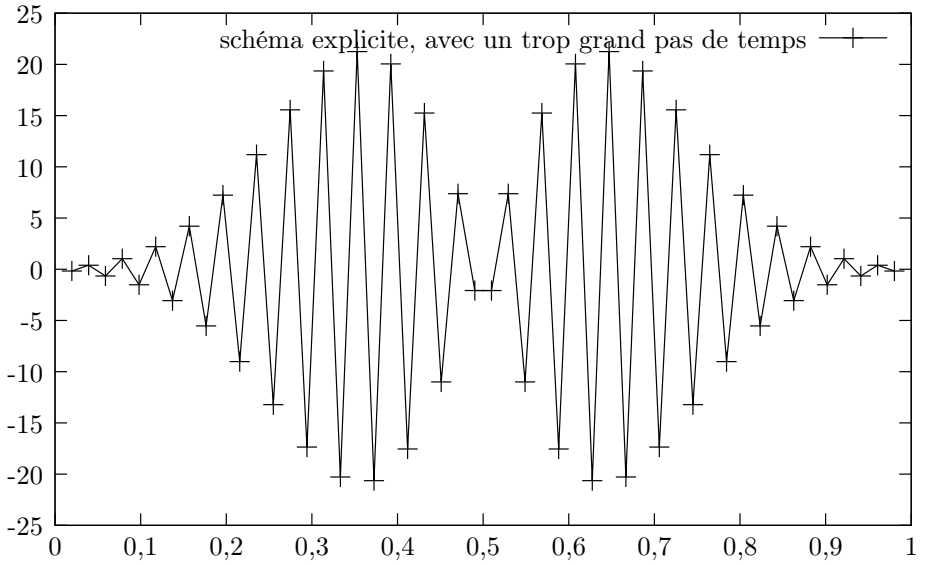

FIGURE 2.10 – La condition de stabilité n'étant pas vérifiée...

```
          ////////  θ-schéma pour l'équation de la chaleur  ////////
          ////////  en dimension 1 avec conditions de bord  ////////
          ////////       de Dirichlet (0 ≤ θ ≤ 1).       ////////


clear();
stacksize(100000000);

////////////////////////////// paramètres ////////////////////////////

kappa = 1.;             // Coefficient de diffusion thermique.
T = 0.01;               // Longueur de l'intervalle en temps.
Xmax = 1;               // Longueur de l'intervalle en espace.
M = 500;                // Nombre de mailles en espace.
dx = Xmax/(M+1);        // Pas en espace.
theta = .5;             // Paramètre du θ-schema :
                        // choisir une valeur theta ∈ [0, 1].

if theta < 0.5 then
  dtmax = 1/(2*kappa*(1-2*theta))*dx*dx
                        // Borne supérieure pour le pas de temps.
  rapport = 0.5;        // Rapport (à choisir) entre le pas de temps
                        // maximal autorisé et le pas de temps effectif.
                        // Ce réel doit être inférieur à 1
                        // pour que le schéma soit stable en norme L^2.
  dt = min(rapport*dtmax,T)   // Pas de temps.
  N = ceil(T/dt)        // Nombre de pas de temps.
else
  N = 500;              // Nombre de pas de temps.
  dt = T/N;             // Pas de temps.
end;

nb = kappa*dt/dx/dx;     // Nombre << de Courant >>.
x=(1:M)'*dx;             // Points de la grille en x.
xb=(0:M+1)'*dx;          // Grille incluant les bords du domaine.

function A = lap1D(n) // Matrice du laplacien sur une grille à n points.
```

```
  A = 2*eye(n,n) - diag(ones(n-1,1),1) - diag(ones(n-1,1),-1);
endfunction

function u=CI(x,xmax)    // Condition initiale.
  n = size(x,1);
  for i=1:n,
    if (x(i)-0.5)^2. < 0.1 then
      u(i) = 1.;
    else
      u(i) = 0.;
    end;
    //u(i) = sin(2.*%pi*x(i)));
    //u(i) = 0.;
  end;
endfunction;

function g=CLg(t)        // Condition limite à gauche.
  //g = sin(2.*%pi*t);
  g = 0.;
endfunction

function d=CLd(t)        // Condition limite à droite.
  d = 0.;
endfunction

function s=source(t,x)   // Terme source.
  n = size(x,1);
  for i=1:n;
    //s(i) = 100.*sin(2.*%pi*x(i));
    s(i) = 0.;
  end;
endfunction

//  Remplissage des matrices creuses B et R (conditions de Dirichlet).  //

A = lap1D(M);
B = eye(M,M) + theta*nb*A;
R = -eye(M,M) + (1. - theta)*nb*A;
```

```
B = sparse(B);          // Stockage sous forme creuse.
R = sparse(R);
[B,rk] = lufact(B);     // Factorisation de B afin de résoudre
                        // les systèmes linéaires plus rapidement.

u=CI(x,Xmax);           // Initialisation de la solution numérique.
f1 = zeros(M,1);        // Source à t^n.
f2 = source(0,x);       // Source à t^n + 1.
t=0;

for i=2:M+1             // w est la solution sur
    w(i) = u(i-1);      // le maillage incluant les
end;                    // points du bord.
w(1) = CLg(t);
w(M+2) = CLd(t);

xbasc();

xset("pixmap",1);       // Options d'affichage.
isoview;
hotcolormap;
plot2d(xb,w);           // Affichage de la condition initiale à l'écran.
xset("wshow")

print(%io(2),'Appuyer sur << Enter >> pour continuer');
halt();

// Boucle en temps.

for n=0:N-1,
  t = t + dt;               // Mise à jour du temps.
  f1 = f2;                  // Mise à jour de f1.
  f2 = source(t,x);         // Mise à jour de f2.
                            // Mise à jour du second membre :
  rhs = -R*u + dt*(theta*f2 + (1-theta)*f1);
                            // On tient compte des conditions de Dirichlet
                            // non homogènes en modifiant le terme source :
  rhs(1) = rhs(1) + nb*(theta*CLg(t) + (1-theta)*CLg(t-dt));
```

```
  rhs(M) = rhs(M) + nb*(theta*CLd(t) + (1-theta)*CLd(t-dt));

  u = lusolve(B,rhs);          // Résolution du système linéaire.

  for i=2:M+1                  // Calcul de la solution sur le
    w(i) = u(i-1);             // maillage incluant les points
  end;                         // du bord du domaine.
  w(1) = CLg(t);
  w(M+2) = CLd(t);

  xset("wwpc");               // Options d'affichage.
  plot2d(xb,w);               // Affichage de la solution à l'écran.
  xset("wshow");
end; // Fin de la boucle en temps.

unix('rm -f resultat');      // Écriture du résultat dans le fichier
plouf = file('open','resultat','unknown'); // << résultat >>.
for i=1:M+2,
  fprintf(plouf,'%f   %f\n',xb(i),w(i));
end;
file('close',plouf);


                              // Fin du programme.
```

Voici maintenant un programme permettant de résoudre l'équation de la chaleur sur un segment en dimension 1 avec des conditions aux limites de Dirichlet (homogènes ou non) ou de Neumann homogènes. La seule différence pour ces dernières conditions réside dans la matrice du laplacien, dont la première et la dernière lignes sont modifiées. . ..

```
////////      θ-schéma pour l'équation de la chaleur      ////////
////////    en dimension 1 avec conditions de bord de     ////////
//////// Dirichlet ou de Neumann homogènes (0 ≤ θ ≤ 1). ////////


clear();
stacksize(100000000);

////////////////////////// paramètres //////////////////////////

kappa = 1.;             // Coefficient de diffusion thermique.
T = 0.01;               // Longueur de l'intervalle en temps.
Xmax = 1;               // Longueur de l'intervalle en espace.
M = 100;                // Nombre de points en espace, de 0 à Xmax (compris).
dx = Xmax/(M-1);        // Pas en espace.
tcg = 'n';              // Type de condition limite à gauche : 'n' ou 'd'.
tcd = 'n';              // Type de condition limite à droite : 'n' ou 'd'.
                        // Si le type est 'n', cela correspond à une condition
                        // de Neumann homogène, sinon à une condition de
                        // Dirichlet, pas forcément homogène.
if(tcg=='d')
  if(tcd=='d')
    x = (1:M-2)'*dx;    // Points de la grille en espace où
  else                  // le calcul sera effectué.
    x = (1:M-1)'*dx;
  end;
else
  if(tcd=='d')
    x = (0:M-2)'*dx;
  else
    x = (0:M-1)'*dx;
  end;
end;
```

```
m = size(x,1);          // Nombre de points de calcul.
xb=(0:M-1)'*dx;         // Grille incluant
                        // les bords du domaine
                        // (pour la représentation graphique).

theta = .5;             // Paramètre du θ-schéma :
                        // choisir une valeur theta ∈ [0, 1].
if theta < 0.5 then
  dtmax = 1/(2*kappa*(1-2*theta))*dx*dx
                        // Borne supérieure pour le pas de temps.
  rapport = 0.5;        // Rapport (à choisir) entre le pas de temps
                        // maximal autorisé et le pas de temps effectif.
                        // Ce réel doit être inférieur à 1
                        // pour que le schéma soit stable en norme L^2.
  dt = min(rapport*dtmax,T)   // Pas de temps.
  N = ceil(T/dt)        // Nombre de pas de temps.
else
  N = 500;              // Nombre de pas de temps.
  dt = T/N;             // Pas de temps.
end;

//N = 1.;

nb = kappa*dt/dx/dx;    // Nombre << de Courant >>.

function A = lap1D(n,tcg,tcd) // Matrice du laplacien sur
                              // une grille à n points.
  A = 2*eye(n,n) - diag(ones(n-1,1),1) - diag(ones(n-1,1),-1);
  if(tcg=='n') then
    A(1,1) = 1;
  end;
  if(tcd=='n') then
    A(n,n) = 1;
  end;
endfunction

function u=CI(x,xmax)   // Condition initiale.
```

```
  n = size(x,1);
  for i=1:n,
    if (x(i)-0.5)^2. < 0.1 then
      u(i) = 1.;
    else
      u(i) = 0.;
    end;
    //u(i) = sin(2.*%pi*x(i)));
    //u(i) = 0.;
  end;
endfunction;

function g=CLg(t)        // Condition limite à gauche
  //g = sin(2.*%pi*t);   // (en cas de condition de Dirichlet).
  g = 1.;
endfunction

function d=CLd(t)        // Condition limite à droite
  d = 0.;                // (en cas de condition de Dirichlet).
endfunction

function s=source(t,x)   // Terme source.
  n = size(x,1);
  for i=1:n;
    s(i) = 100.*sin(2.*%pi*x(i));
  end;
endfunction

/////////       Remplissage des matrices creuses B et R    ///////////////

A = lap1D(m,tcg,tcd);
B = eye(m,m) + theta*nb*A;
R = -eye(m,m) + (1. - theta)*nb*A;
B = sparse(B);         // Stockage sous forme creuse.
R = sparse(R);
[B,rk] = lufact(B);    // Factorisation de B afin de résoudre
                       // les systèmes linéaires plus rapidement.
```

```
u=CI(x,Xmax);          // Initialisation de la solution numérique.
f1 = zeros(m,1);       // Source à t^n.
f2 = source(0,x);      // Source à t^n + 1.
t=0;

if(tcg=='d')           // w est la solution sur le maillage
  if(tcd=='d')         // incluant les points du bord.
    w = [CLg(t);u;CLd(t)];
  else
    w = [CLg(t);u];
  end;
else
  if(tcd=='d')
    w = [u;CLd(t)];
  else
    w = u;
  end;
end;

xbasc();

xset("pixmap",1);      // Options d'affichage.
isoview;
hotcolormap;
plot2d(xb,w);          // Affichage de la condition initiale à l'écran.
xset("wshow")

print(%io(2),'Appuyer sur << Enter >> pour continuer');
halt();

// Boucle en temps.

for n=0:N-1,
  t = t + dt;              // Mise à jour du temps.
  f1 = f2;                 // Mise à jour de f1.
  f2 = source(t,x);        // Mise à jour de f2.
                           // Mise à jour du second membre :
  rhs = -R*u + dt*(theta*f2 + (1-theta)*f1);
```

```
                          // On tient compte des conditions de Dirichlet
                          // non homogènes en modifiant le terme source :
  if(tcg=='d') then
    rhs(1) = rhs(1) + nb*(theta*CLg(t) + (1-theta)*CLg(t-dt));
  end;
  if(tcd=='d') then
    rhs(m) = rhs(m) + nb*(theta*CLd(t) + (1-theta)*CLd(t-dt));
  end;

  u = lusolve(B,rhs);     // Résolution du système linéaire.

if(tcg=='d')              // w est la solution sur le maillage
  if(tcd=='d')            // incluant les points du bord.
    w = [CLg(t);u;CLd(t)];
  else
    w = [CLg(t);u];
  end;
else
  if(tcd=='d')
    w = [u;CLd(t)];
  else
    w = u;
  end;
end;

  xset("wwpc");              // Options d'affichage.
  plot2d(xb,w);              // Affichage de la solution à l'écran.
  xset("wshow");
end; // Fin de la boucle en temps.

unix('rm -f resultat');      // Écriture du résultat dans le fichier
plouf = file('open','resultat','unknown'); // << résultat >>.
for i=1:m,
  fprintf(plouf,'%f   %f\n',xb(i),w(i));
end;
file('close',plouf);
```

Enfin, on propose maintenant un algorithme de résolution de l'équation de la chaleur sur un pavé P en dimension 2. Il va de soi que la démonstration d'existence, d'unicité, de régularité de solutions est la même qu'en dimension 1 puisqu'une base hilbertienne de $L^2(P)$ est disponible (formée par les produits des fonctions de base dans chaque direction). L'algorithme programmé ici est encore le θ-schéma, dont l'étude peut se faire de la même façon qu'en dimension 1. Signalons que la matrice du laplacien (avec conditions de bord de Dirichlet homogènes) en dimension 2 sur un maillage à $J \times J$ points est

$$A_2 = \begin{pmatrix} A_1 & -I_J & 0 & \cdots & \cdots & 0 \\ -I_J & A_1 & -I_J & 0 & \cdots & 0 \\ 0 & -I_J & A_1 & -I_J & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -I_J & A_1 & -I_J \\ 0 & \cdots & \cdots & 0 & -I_J & A_1 \end{pmatrix} \in \mathcal{M}_{J\times J}(\mathbb{R})$$

où I_J est la matrice identité en dimension J et

$$A_1 = \begin{pmatrix} 4 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 4 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 4 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & -1 & 4 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 4 \end{pmatrix} \in \mathcal{M}_J(\mathbb{R})$$

si l'on utilise la numérotation « lexicographique » selon laquelle les valeurs $u(t, i\Delta x, j\Delta y)$ sont représentées par le vecteur

$$U(t) = \begin{pmatrix} u(t,\Delta x,\Delta y) \\ u(t,\Delta x,2\Delta y) \\ \vdots \\ u(t,\Delta x,J\Delta y) \\ u(t,2\Delta x,\Delta y) \\ \vdots \\ u(t,\Delta x,J\Delta y) \\ \vdots \\ u(t,J\Delta x,\Delta y) \\ \vdots \\ u(t,J\Delta x,J\Delta y) \end{pmatrix},$$

c'est-à-dire que l'on y fait varier j plus vite que i...

```
//////          θ-schéma pour l'équation de la chaleur          //////
//////        en dimension 2 avec conditions de bord de         //////
//////                 Dirichlet (0 ≤ θ ≤ 1).                   //////


clear();
stacksize(10000000);

////////////////////////// paramètres //////////////////////////

kappa = 1.;             // Coefficient de diffusion thermique.
T = 0.1;                // Longueur de l'intervalle en temps.
Xmax = 1;               // Longueur de l'intervalle en espace en x et en y.
M = 30;                 // Nombre de mailles en espace en x et en y.
dx = Xmax/(M+1);        // Pas en espace, en x et en y.
theta = 0.6;            // Paramètre du θ-schéma :
                        // choisir une valeur theta ∈ [0, 1].
if theta < 0.5 then
  dtmax = 1/(4*kappa*(1-2*theta))*dx*dx
                        // Borne supérieure pour le pas de temps.
  rapport = 0.5;        // Rapport (à choisir) entre le pas de
                        // temps maximal autorisé et le pas de
                        // temps effectif. Ce réel doit être
                        // inférieur à 1 pour que le schéma
                        // soit stable en norme L^2.
  dt = min(rapport*dtmax,T)   // Pas de temps.
  N = ceil(T/dt)        // Nombre de pas de temps.
else
  N = 100;              // Nombre de pas de temps.
  dt = T/N;             // Pas de temps.
end;

nb = kappa*dt/dx/dx;     // Nombre << de Courant >>.

x=(1:M)'*dx;             // Points de la grille en x.
y=(1:M)'*dx;             // Points de la grille en y.
```

```
xb=(0:M+1)'*dx;          // Grille incluant
yb=(0:M+1)'*dx;          // les bords du domaine.

function L2 = lap2D(n) // Matrice du laplacien sur une grille nxn.
  L2 = 4*eye(n*n,n*n) - diag(ones(n*n-1,1),1) - diag(ones(n*n-1,1),-1);
  L2 = L2 - diag(ones((n-1)*n,1),n) - diag(ones((n-1)*n,1),-n)
  for i=1:n-1
    L2(n*i,n*i+1) = 0.;
    L2(n*i+1,n*i) = 0.;
  end;
endfunction

function u=CI(x,y,xmax)    // Condition initiale.
  n = size(x,1);
  for i=1:n,
    for j=1:n
      if (x(i)-0.5)^2. + (y(j)-0.5)^2/2. < 0.1 then
        u(n*(i-1)+j) = 1.;
      else
      u(n*(i-1)+j) = 0.;
      end;
      //u(n*(i-1)+j) = sin(2.*pi*x(i))*sin(2.*pi*y(j));
      //u(n*(i-1)+j) = 0.;
    end;
  end;
endfunction;

function g=CLg(t,x)        // Condition limite à gauche (x = 0).
  //g = sin(2.*%pi*x);
  g = 0.;
endfunction

function d=CLd(t,x)        // Condition limite à droite (x = Xmax).
  d = 0.;
endfunction

function b=CLb(t,x)        // Condition limite en bas (y = 0).
  b = 0.;
```

```
endfunction

function h=CLh(t,x)        // Condition limite en haut (y = Xmax).
  h = 0.;
endfunction

function s=source(t,x,y)   // Terme source.
  n = size(x,1);
  m = size(y,1);
  for i=1:n;
    for j=1:m;
      s(M*(i-1)+j) = 50.*sin(2.*%pi*x(i));
      //s(M*(i-1)+j) = 0.;
    end;
  end;
endfunction

//  Remplissage des matrices creuses B et R (conditions de Dirichlet).  //

A = lap2D(M);
B = eye(M*M,M*M) + theta*nb*A;
R = -eye(M*M,M*M) + (1. - theta)*nb*A;
B = sparse(B);         // Stockage sous forme creuse.
R = sparse(R);
[B,rk] = lufact(B);    // Factorisation de B afin de résoudre
                       // les systèmes linéaires plus rapidement.

u=CI(x,y,Xmax);        // Initialisation de la solution numérique.
f1=zeros(M*M,1);       // Source à t^n.
f2=source(0,x,y);      // Source à t^n+1.
t=0;

for i=1:M              // Mise de la solution sour forme matricielle
  for j=1:M            // pour représentation graphique.
    v(i,j) = u(M*(i-1)+j);
  end;
end;
for i=2:M+1
```

```
  for j=2:M+1
    w(i,j) = v(i-1,j-1);
  end;                  // w est la solution sur la maillage
end;                    // incluant les points du bord.
for i=2:M+1
    w(i,1) = CLb(t,xb(i));
    w(i,M+2) = CLh(t,xb(i));
end;
for j=1:M+2
  w(1,j) = CLg(t,yb(j));
  w(M+2,j) = CLd(t,yb(j));
end;                    // Fin de la mise sous forme matricielle.

xbasc();

xset("pixmap",1);       // Options d'affichage.
isoview;
hotcolormap;
plot3d(xb,yb,w);        // Affichage de la condition initiale à l'écran.
xset("wshow")

print(%io(2),'Appuyer sur << Enter >> pour continuer');
halt();

// Boucle en temps.

for n=0:N-1,
  t = t + dt;             // Mise à jour du temps.
  f1 = f2;                // Mise à jour de f1.
  f2 = source(t,x,y);     // Mise à jour de f2.
                          // Mise à jour du second membre :
  rhs = -R*u + dt*(theta*f2 + (1-theta)*f1);

  for i=1:M
    rhs(M*(i-1)+1) = rhs(M*(i-1)+1)...
      + nb*(theta*CLb(t,x(i)) + (1-theta)*CLb(t-dt,x(i)));
    rhs(M*i) = rhs(M*i)...
      + nb*(theta*CLh(t,x(i)) + (1-theta)*CLh(t-dt,x(i)));
```

```
  end;
  for j=1:M
    rhs(j) = rhs(j) + nb*(theta*CLg(t,y(j)) + (1-theta)*CLg(t-dt,y(j)));
    rhs(M*(M-1)+j) = rhs(M*(M-1)+j)...
      + nb*(theta*CLd(t,y(j)) + (1-theta)*CLd(t-dt,y(j)));
  end;

  u = lusolve(B,rhs);          // Résolution du système linéaire.

  for i=1:M              // Mise de la solution sous forme matricielle
    for j=1:M            // pour représentation graphique.
      v(i,j) = u(M*(i-1)+j);
    end;
  end;
  for i=2:M+1
    for j=2:M+1
      w(i,j) = v(i-1,j-1);
    end;                 // w est la solution sur la maillage
  end;                   // incluant les points du bord.

  for i=2:M+1
    w(i,1) = CLb(t,xb(i));
    w(i,M+2) = CLh(t,xb(i));
  end;
  for j=1:M+2
    w(1,j) = CLg(t,yb(j));
    w(M+2,j) = CLd(t,yb(j));
  end;                   // Fin de la mise sous forme matricielle.

  xset("wwpc");          // Options d'affichage.
  plot3d(xb,yb,w);       // Affichage de la solution à l'écran.
  xset("wshow");
end;                     // Fin de la boucle en temps.

unix('rm -f resultat');  // Écriture du résultat dans le fichier
plouf = file('open','resultat','unknown'); // << résultat >>.
for i=1:M+2,
  for j=1:M+2,
```

```
    fprintf(plouf,'%f   %f   %f\n',xb(i),yb,w(i));
  end;
end;
file('close',plouf);


                              // fin du programme.
```

Chapitre 3

Équations hyperboliques

Ce chapitre a pour objet l'étude de certaines EDP de la forme

$$\partial_t u(t,x) + A(u(t,x))\partial_x u(t,x) = 0 \tag{3.1}$$

où $u : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}^n$ et $A(u) \in \mathcal{M}_n(\mathbb{R})$ dépend « régulièrement » de u. Un système de ce type est dit *quasi-linéaire*.

3.1 Introduction, définitions, exemples

Définition 4
Le système (3.1) est dit *hyperbolique dans* $\mathcal{U} \subset \mathbb{R}^n$ si et seulement si $A(u)$ est diagonalisable et à valeurs propres réelles pour tout $u \in \mathcal{U}$.
Il est dit *strictement hyperbolique dans* $\mathcal{U}$ si et seulement s'il est hyperbolique et toutes ses valeurs propres sont distinctes, pour tout $u \in \mathcal{U}$. □

Remarque 26 (en dimension supérieure à 1)
Dans un espace de dimension $d > 1$, le système considéré est

$$\partial_t u + \sum_{j=1}^{d} A_j(u)\partial_j u = 0. \tag{3.2}$$

Il est dit *(strictement) hyperbolique dans* $\mathcal{U}$ si et seulement si la matrice $A(u,\xi) = \sum_{j=1}^{d} \xi_j A_j(u)$ est diagonalisable et à valeurs propres réelles (distinctes) pour tout $(u,\xi) \in \mathcal{U} \times \mathbb{R}^d$ (avec $\xi = (\xi_j)_{j=1}^d$). □

Remarque 27 (hyperbolicité des EDP linéaires d'ordre 2)
L'hyperbolicité au sens de la définition ci-dessus peut bien entendu être rapprochée de la notion d'hyperbolicité donnée dans le chapitre d'introduction de ce cours. Le rapprochement proposé ici est une analogie avec l'équation des ondes. Pour simplifier, on considère cette équation en dimension 1 seulement :

$$\partial^2_{t,t} u - c^2 \partial^2_{x,x} u = 0 \text{ avec } c \in \mathbb{R}^*.$$

Elle est hyperbolique au sens de l'hyperbolicité des EDP linéaires d'ordre 2 puisque la matrice formée par ses coefficients situés devant les termes d'ordre 2,

$$\begin{pmatrix} 1 & 0 \\ 0 & -c^2 \end{pmatrix},$$

a une valeur propre positive et une valeur propre négative (car $c \in \mathbb{R}^*$). Soit u une solution de cette EDP, posons $v_1 = \partial_t u$ et $v_2 = \partial_x u$. On a alors naturellement (si u est suffisamment régulière...) $\partial_t v_2 - \partial_x v_1 = 0$. D'autre part, l'EDP vérifiée par u donne $\partial_t v_1 - c^2 \partial_x v_2 = 0$, de sorte que l'EDP des ondes peut s'écrire sous forme du système du premier ordre

$$\partial_t \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} + A \partial_x \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0$$

avec

$$A = \begin{pmatrix} 0 & -c^2 \\ -1 & 0 \end{pmatrix}.$$

Les valeurs propres de A sont c et $-c$, elles sont réelles si et seulement si... $c \in \mathbb{R}$, et d'autre part A n'est diagonalisable que si $c \neq 0$: les deux notions d'hyperbolicité que nous avons définies coïncident dans le cas linéaire de l'équation des ondes. □

Ce chapitre ne traitera pas uniquement des EDP quasi-linéaires sous la forme (3.1). Nous serons aussi amenés à étudier des EDP sous la forme

$$\partial_t u(t,x) + \partial_x \left(F(u) \right)(t,x) = 0. \tag{3.3}$$

Définition 5
Le système (3.3) est dit *(strictement) hyperbolique* si et seulement si sa forme quasi-linéaire $\partial_t u + \nabla F(u) \partial_x u = 0$ est (strictement) hyperbolique. □

3.1.1 Exemples

Équation des ondes

Cette EDP linéaire d'ordre deux hyperbolique peut être ramenée à un système linéaire d'EDP d'ordre 1 hyperbolique, comme on l'a vu dans l'introduction ci-dessus.

Équations d'Euler

Le système des équations d'Euler,

$$\begin{cases} \partial_t \rho + \operatorname{div}(\rho u) = 0, \\ \partial_t(\rho u) + \operatorname{div}(\rho u \otimes u + pI) = 0, \\ \partial_t(\rho e) + \operatorname{div}(\rho u e + pu) = 0, \end{cases}$$

déjà introduit dans la section 1.2.6, est hyperbolique sous certaines conditions sur la pression $p(\rho, u, e)$.

Équations scalaires réelles

Les équation scalaires réelles d'ordre 1, du type

$$\partial_t u + \partial_x f(u) = 0$$

où $f : \mathbb{R} \longrightarrow \mathbb{R}$ est suffisamment régulière sont hyperboliques. La forme quasi-linéaire de l'équation ci-dessus,

$$\partial_t u + f'(u)\partial_x u = 0$$

est de la forme (3.1) avec $A(u)$ scalaire réel.

Ces équations seront l'objet principal de l'étude dans ce cours.

3.2 Méthode des caractéristiques pour l'advection

L'étude de l'équation

$$\partial_t u + a(t, x)\partial_x u = 0$$

où a est donnée est préliminaire à celle de

$$\partial_t u + f'(u)\partial_x u = 0.$$

Nous l'effectuons ici, au moyen de la méthode des caractéristiques. Plus précisément, nous allons étudier le problème

$$\begin{cases} \partial_t u + a(t, x)\partial_x u = 0 \text{ dans } \mathbb{R} \times \mathbb{R}, \\ u(0, x) = u^0(x) \text{ dans } \mathbb{R} \end{cases} \tag{3.4}$$

où u^0, condition initiale, est donnée, et a, *vitesse de transport*, est donnée.

Remarque 28 (terminologie)
Si $a(t, x) = a \in \mathbb{R}\ \forall (t, x) \in \mathbb{R}^2$, une solution évidente du problème est donnée par

$$u(t, x) = u^0(x - at).$$

Ceci signifie en particulier que la solution u est constante sur les courbes $x = at + x^0$ du plan (t, x). Cela permet de comprendre les termes « transport » et « vitesse ». La méthode des *caractéristiques* est une généralisation de cette remarque au cas où a n'est pas une constante. □

Remarque 29 (systèmes linéaires hyperboliques)
La remarque précédente permet aussi de comprendre l'importance de la notion d'hyperbolicité pour les systèmes. Pour simplifier, on considère le système hyperbolique linéaire à coefficients constants

$$\partial_t u + A\partial_x u = 0 \text{ dans } \mathbb{R}^2$$

associé à la donnée initiale $u(0,x) = u^0(x)$. A étant diagonalisable à valeurs propres réelles, il existe une matrice $P \in \mathcal{M}_n(\mathbb{R})$ inversible et une matrice $D \in \mathcal{M}_n(\mathbb{R})$ diagonale telles que $D = P^{-1}AP$. Une fonction u est donc solution de l'EDP si et seulement si

$$P^{-1}\partial_t u + P^{-1}APP^{-1}\partial_x u = 0,$$

soit

$$\partial_t v + D\partial_x v = 0,$$

où v est l'expression de u dans la base qui diagonalise A : $v = P^{-1}u$. Le système vérifié par v est diagonal, ce qui signifie que les n EDP vérifiées par les composantes de v sont indépendantes :

$$\partial_t v_i + d_i\partial_x v_i = 0 \quad i = \{1, \dots, n\}$$

où les d_i sont les coefficients diagonaux de D. La remarque précédente permet de voir qu'une (la ?) solution de chacune de ces EDP est

$$v_i(t,x) = v_i(0, x - d_i t).$$

D'autre part, on connaît $v_i(0,\cdot)$, qui nous est donné par la condition initiale

$$v(0,\cdot) = P^{-1}u^0(\cdot).$$

□

Le principe de la méthode des caractéristiques repose sur la recherche (et la découverte) de courbes $x = X(t)$ du plan (t,x) sur lesquelles la solution u reste constante [1]. Ceci peut s'exprimer ainsi : on cherche une fonction $X(t)$ telle que $u(t, X(t))$ ne dépend pas de t.

On suppose que $u(t,x)$ est une solution de classe $\mathcal{C}^1(\mathbb{R}^2)$ du problème (3.4). Soit $X(t)$ une fonction de classe $\mathcal{C}^1(\mathbb{R})$. Posons $\tilde{u}(t) = u(t, X(t))\ \forall t \in \mathbb{R}$. Le but est de trouver $X(t)$ tel que $\tilde{u}'(t) = 0$. Pour cela, on remarque que, sous les hypothèses de régularité faites sur u et X,

$$\tilde{u}'(t) = \partial_t u(t, X(t)) + X'(t)\partial_x u(t, X(t)).$$

Puisque u est solution du problème,

$$\partial_t u(t, X(t)) = -a(t, X(t))\partial_x u(t, X(t)),$$

1. Ces courbes sont appelées courbes caractéristiques, ou encore caractéristiques.

et la nullité de $\tilde{u}'(t)$ s'écrit

$$X'(t)\partial_x u(t, X(t)) = a(t, X(t))\partial_x u(t, X(t)).$$

Il *suffit* pour cela que X vérifie

$$X'(t) = a(t, X(t)) \quad \forall t \in \mathbb{R}.$$

Si X vérifie cette EDO, on a

$$u(t, X(t)) = u(0, X(0)) = u^0(X(0)) \quad \forall t \in \mathbb{R}.$$

Nous sommes donc amenés à considérer le problème de Cauchy

$$\begin{cases} X'(t) = a(t, X(t)) \quad \forall t \in \mathbb{R}, \\ X(0) = X^0. \end{cases}$$

D'après le théorème de Cauchy-Lipschitz (voir, dans l'introduction du cours, le théorème 2), ce problème a une unique solution maximale si a est continue sur $\mathbb{R}^2$ et localement (en (t, x)) lipschitzienne par rapport à sa seconde variable. Si de plus a est globalement (en (t, x)) lipschitzienne par rapport à sa seconde variable, l'unique solution maximale est *globale* (c'est une conséquence du théorème des bouts). De manière plus complète on a, sous ces hypothèses qui conduisent à l'existence et à l'unicité d'une solution globale,

pour tout $t^0 \in \mathbb{R}$ et tout $X^0 \in \mathbb{R}$, il existe une unique fonction $X(t, t^0, X^0)$ solution de
$\begin{cases} \partial_1 X(t, t^0, X^0) = a(t, X(t, t^0, X^0)) \quad \forall t \in \mathbb{R}, \\ X(t^0, t^0, X^0) = X^0 \end{cases}$

(X est le *flot* de l'équation différentielle). Cette fonction vérifie la propriété de semi-groupe

$$X(t + \Delta t, t^0, X^0) = X(t + \Delta t, t, X(t, t^0, X^0)) \quad \forall (t, t^0, \Delta t, X^0) \in \mathbb{R}^4.$$

En écrivant cette équation avec $t = 0$, $t^0 = t$, $\Delta t = t$ et $x = X^0$, on obtient

$$X(t, 0, X(0, t, x)) = X(t, t, x) = x.$$

On en déduit que $\forall t \in \mathbb{R}$, tout $x \in \mathbb{R}$ est atteint par une solution de l'EDO pour une certaine donnée initiale, en l'occurrence $X(\cdot, 0, X(0, t, x))$.

On appelle *courbe caractéristique* ou *caractéristique* toute courbe $(t, X(t, t^0, X^0))$. On appelle *pied* de la caractéristique $(t, X(t, t^0, X^0))$ le point $X(0, t^0, X^0)$. On a montré que par tout point $(t, x) \in \mathbb{R}^2$ passe une unique caractéristique du problème [2]. Le pied d'une caractéristique est unique (par unicité de la solution).

Revenons à notre EDP.

2. Sous l'hypothèse que a est continue, et globalement lipschitzienne par rapport à sa seconde variable.

Proposition 9
On considère le problème (3.4) et l'on suppose que $u^0 \in \mathcal{C}^1(\mathbb{R})$ et que $a \in \mathcal{C}^1(\mathbb{R}^2)$ est globalement lipschitzienne par rapport à sa seconde variable : $\exists K \in \mathbb{R}$ tel que

$$|a(t,x) - a(t,y)| \leq K\,|x-y| \quad \forall (t,x,y) \in \mathbb{R}^3.$$

Alors, le problème (3.4) admet une unique solution de classe $\mathcal{C}^1(\mathbb{R}^2)$: la fonction u définie par $u(t,x) = u^0(X(0,t,x))$. □

Démonstration
On sait (voir [9]) que $X \in \mathcal{C}^1(\mathbb{R}^3)$ (grâce à l'hypothèse $a \in \mathcal{C}^1(\mathbb{R}^2)$). En posant $u(t,x) = u^0(X(0,t,x))$, on a

$$\partial_t u(t,x) = u^{0\prime}(X(0,t,x))\partial_2 X(0,t,x)$$

et

$$\partial_x u(t,x) = u^{0\prime}(X(0,t,x))\partial_3 X(0,t,x)$$

Calculons $\partial_2 X(0,t,x)$ (ou presque).

Posons $g(t,x) = X(t,t,x)\ \forall (t,x) \in \mathbb{R}^2$. Puisque $X(t,t,x) = x\ \forall (t,x) \in \mathbb{R}^2$, $\partial_1 g = 0$. Or

$$g(t,x) = X(t,0,X(0,t,x)),$$

donc

$$\partial_1 g(t,x) = \partial_1 X(t,0,X(0,t,x)) + \partial_3 X(t,0,X(0,t,x))\partial_2 X(0,t,x)$$

et ainsi

$$\partial_3 X(t,0,X(0,t,x))\partial_2 X(0,t,x) = -\partial_1 X(t,0,X(0,t,x)).$$

Puisque $\partial_1 X(t,0,X(0,t,x)) = a(t,X(t,0,X(0,t,x)) = a(t,x)$, on a

$$\partial_3 X(t,0,X(0,t,x))\partial_2 X(0,t,x) = -a(t,x). \tag{3.5}$$

Calculons $\partial_3 X(0,t,x)$, ou presque.

Avec la même définition de g, $\partial_2 g = 1$ et, puisque $g(t,x) = X(t,0,X(0,t,x))$,

$$\partial_2 g(t,x) = \partial_3 X(t,0,X(0,t,x))\partial_3 X(0,t,x),$$

d'où

$$\partial_3 X(t,0,X(0,t,x))\partial_3 X(0,t,x) = 1.$$

Ainsi

$$a(t,x)\partial_3 X(t,0,X(0,t,x))\partial_3 X(0,t,x) = a(t,x). \tag{3.6}$$

En sommant les équations (3.5) et (3.6), on obtient

$$\partial_3 X(t,0,X(0,t,x))\,(\partial_2 X(0,t,x) + a(t,x)\partial_3 X(0,t,x)) = 0.$$

D'autre part, puisque $\partial_3 X(t,0,X(0,t,x))\partial_3 X(0,t,x) = 1$, on a $\partial_3 X(t,0,X(0,t,x)) \neq 0$ et

$$\partial_2 X(0,t,x) + a(t,x)\partial_3 X(0,t,x) = 0.$$

Ainsi

$$\partial_t u(t,x) + a(t,x)\partial_x u(t,x) = u^{0\prime}(X(0,t,x)\,(\partial_2 X(0,t,x) + a(t,x)\partial_3 X(0,t,x)) = 0.$$

L'unicité de la solution (régulière) résulte du fait que toute solution régulière est constante sur les caractéristiques, et est donc donnée par $u(t,x) = u^0(X(0,t,x))$. □

Munis de ce premier résultat, nous allons pouvoir aborder le sujet des équations scalaires non linéaires.

3.3 Équations scalaires conservatives

On s'intéresse ici à l'équation (non linéaire)

$$\partial_t u(t,x) + \partial_x f(u)(t,x) = 0 \text{ dans } \mathbb{R}^2 \tag{3.7}$$

où $f : \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction donnée, suffisamment régulière (soyons prêts à faire des concessions). Cette équation est appelée *conservative* car si u en est une solution (bornée), pour tout intervalle $[a,b] \subset \mathbb{R}$, on a

$$\partial_t \int_a^b u(t,x)\,dx = f(u(t,a)) - f(u(t,b)) \tag{3.8}$$

et, si u est intégrable et tend vers 0 en $\pm\infty$ (et est suffisamment régulière),

$$\partial_t \int_{\mathbb{R}} u(t,x)\,dx = 0$$

pour tout $t \in \mathbb{R}_+$: l'intégrale de u est conservée. L'équation (3.8) nous apprend que l'intégrale de u entre a et b varie en temps selon ce qui « passe » en b et a. Pour cette raison, la fonction f est appelée *flux*.

Nous allons dans un premier temps chercher à résoudre cette équation au moyen de la méthode des caractéristiques. Cette tentative va échouer (du moins en ce qui concerne les solutions globales en temps). Nous introduirons alors la notion de *solution faible* (solution au sens des distributions). Cette notion nous permettra de prendre en compte une classe beaucoup plus grande de solutions des EDP, notamment celle des solutions discontinues. Nous étudierons en détail le cas $f(u) = u^2/2$ (équation de Burgers). Nous passerons ensuite à la résolution approchée de (3.7) au moyen d'algorithmes de volumes finis, avec l'étude précise du schéma de Lax-Friedrichs. Cette étude nous permettra de généraliser le résultat d'existence et d'unicité précédemment obtenu dans le cas de l'équation de Burgers au cas où f est strictement convexe générale.

3.3.1 Méthode des caractéristiques pour les équations non linéaires

Le but de cette section est d'observer que la méthode des caractéristiques s'applique bien à l'étude de (3.7) tant que u est régulière, mais que ceci n'est pas vrai pour tout t en général.

Explosion en temps fini de la dérivée en espace

Nous allons montrer que les solutions régulières de (3.7) ont en général un temps de vie fini.

Proposition 10
On considère l'EDP (3.7) avec $f \in \mathcal{C}^2(\mathbb{R})$ et une donnée initiale $u(0,\cdot) = u^0(\cdot) \in \mathcal{C}^1(\mathbb{R}) \cap L^\infty(\mathbb{R})$ telle que $u^{0\prime} \in L^\infty(\mathbb{R})$. Alors

— si $f''(u^0(x))u^{0\prime}(x) \geq 0 \ \forall x \in \mathbb{R}$, le problème a une unique solution $u \in \mathcal{C}^1([0,+\infty[\times\mathbb{R})$;

— sinon, le problème a une unique solution $u \in \mathcal{C}^1([0, \frac{-1}{\inf_{x\in\mathbb{R}} f''(u^0(x))u^{0\prime}(x)}[\times\mathbb{R})$ qui ne peut être prolongée en temps supérieur.

Dans tous les cas, la solution vérifie $u(t, x + tf'(u^0(x))) = u^0(x)$. □

Démonstration
Si u est une solution de classe $\mathcal{C}^1$, elle vérifie

$$\partial_t u(t,x) + f'(u(t,x))\partial_x u(t,x) = 0.$$

Soit $X(t,x)$ la solution maximale de

$$\begin{cases} \partial_t X(t,x) = f'(u(t,X(t,x))) \\ X(0,x) = x \end{cases}$$

(comme il n'y a plus maintenant qu'une variable de temps pour X, nous abandonnons les notations ∂_1 et ∂_3 au profit de ∂_t et ∂_x, respectivement). Il existe en effet une telle solution maximale, d'après le théorème de Cauchy-Lipschitz, car le champ $g(t,x) = f'(u(t,x))$ est continu par rapport à (t,x) et localement (en (t,x)) lipschitzien par rapport à x sous l'hypothèse que u et f' sont de classe $\mathcal{C}^1$. Posons $\tilde{u}(t,x) = u(t,X(t,x))$. On a

$$\begin{aligned} \partial_t \tilde{u}(t,x) &= \partial_t u(t,X(t,x)) + \partial_x u(t,X(t,x))\partial_t X(t,x) \\ &= \partial_t u(t,X(t,x)) + f'(u(t,X(t,x)))\partial_x u(t,X(t,x)) \\ &= 0 \end{aligned}$$

(nous avons déjà fait ce calcul). Donc $u(t,X(t,x))$ ne dépend pas de t : une solution régulière de (3.7) est constante sur les caractéristiques. Le fait que la pente de la caractéristique au point t ne dépende que de $u(t,X(t,x))$ simplifie encore le problème :

$$\begin{aligned} f'(u(t,X(t,x))) &= f'(u(0,X(0,x))) \\ &= f'(u(0,x)) \\ &= f'(u^0(x)). \end{aligned}$$

L'équation vérifiée par les caractéristiques est donc

$$\begin{cases} \partial_t X(t,x) = f'(u^0(x)) \\ X(0,x) = x. \end{cases}$$

La solution $X(t,x)$ en est donnée par la formule

$$X(t,x) = x + f'(u^0(x))t.$$

La solution $u(t,x)$ vérifie donc

$$u(t, x + f'(u^0(x))t) = u^0(x).$$

Cela prouve la dernière assertion de la proposition. Mais bien entendu, tout cela n'est vrai que tant que u est de classe $\mathcal{C}^1$. Pour simplifier la suite, notons

$$X_t : \begin{cases} \mathbb{R} \longrightarrow \mathbb{R} \\ x \longmapsto x + f'(u^0(x))t \end{cases} \qquad \forall t \in \mathbb{R}.$$

On a $X_t'(x) = 1 + f''(u^0(x))u^{0\prime}(x)t$. Donc, en particulier, $X_0'(x) = 1 > 0$. Posons

$$m = \inf_{x \in \mathbb{R}} f''(u^0(x))u^{0\prime}(x).$$

On sait que $m \in \mathbb{R}$ car $u^0 \in L^\infty(\mathbb{R})$ donc $f''(u^0(x))$ est borné et $u^{0\prime} \in L^\infty(\mathbb{R})$. Si $m \geq 0$, $X_t'(x) > 1$ pour tout x et X_t est une bijection [3] de $\mathbb{R}$ dans $\mathbb{R}$, pour tout $t \geq 0$. Si ce n'est pas le cas, posons

$$\tau = \frac{-1}{m}.$$

On a : $\forall t \in [0, \tau[$, $X_t'(x) \geq m(t - \tau) > 0$ pour tout $x \in \mathbb{R}$; X_t est donc une bijection de $\mathbb{R}$ dans $\mathbb{R}$ pour tout t dans $[0, \tau[$ (voir la précédente note de bas de page). On pose maintenant, pour résumer,

$$T = \begin{cases} +\infty \text{ si } m \geq 0 \\ \tau \text{ si } m < 0. \end{cases}$$

Si u est une solution de classe $\mathcal{C}^1$ de (3.7), tant que X_t est une bijection, u vérifie $u(t,x) = u^0(X_t^{-1}(x))$. Nous allons montrer que la fonction u définie par $u(t,x) = u^0(X_t^{-1}(x))$ est de classe $\mathcal{C}^1$ et vérifie (3.7) tant que $X_t'(x) > 0$ $\forall x \in \mathbb{R}$ (donc pour $t \in [0, T[$) [4]. X_t est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et, si $X_t'(x) > 0$ $\forall x \in \mathbb{R}$, X_t est inversible sur $\mathbb{R}$, X_t^{-1} est de classe $\mathcal{C}^1$ sur $\mathbb{R}$ et

$$X_t^{-1\prime}(x) = \frac{1}{X_t'(X_t^{-1}(x))}.$$

3. En effet, l'application X_t est bijective de $\mathbb{R}$ dans $X_t(\mathbb{R})$ et $X_t(\mathbb{R}) = \mathbb{R}$.
4. Cette démonstration a en fait déjà été effectuée dans le paragraphe sur l'équation de transport.

Donc

$$\partial_x u(t,x) = \frac{u^{0\prime}(X_t^{-1}(x))}{X_t'(X_t^{-1}(x))} = \frac{u^{0\prime}(X_t^{-1}(x))}{1 + tf''(u^0(X_t^{-1}(x)))u^{0\prime}(X_t^{-1}(x))}. \tag{3.9}$$

D'autre part,

$$X_t(x) = x + tf'(u^0(x)) = x + tf'(u(t, X(t,x))) = x + tf'(u(t, X_t(x))),$$

d'où

$$X_t^{-1}(x) = x - tf'(u(t,x)).$$

Donc la dérivée par rapport à t de $X_t^{-1}(x)$ (considérée à nouveau comme fonction de (t,x)) est, au point (t,x), $-f'(u(t,x)) - tf''(u(t,x))\partial_t u(t,x)$. On en déduit que

$$\partial_t u(t,x) = u^{0\prime}(X_t^{-1}(x)) \left[-f'(u(t,x)) - tf''(u(t,x))\partial_t u(t,x)\right]$$

et

$$\partial_t u(t,x) \left[1 + tf''(u(t,x))u^{0\prime}(X_t^{-1}(x))\right] = -u^{0\prime}(X_t^{-1}(x))f'(u(t,x)).$$

Or

$$1 + tf''(u(t,x))u^{0\prime}(X_t^{-1}(x)) = 1 + tf''(u^0(X_t^{-1}(x)))u^{0\prime}(X_t^{-1}(x))$$

et pour tout $t \in [0,T[$, ce terme est strictement positif, donc

$$\partial_t u(t,x) = \frac{-u^{0\prime}(X_t^{-1}(x))f'(u^0(X_t^{-1}(x)))}{1 + tf''(u^0(X_t^{-1}(x)))u^{0\prime}(X_t^{-1}(x))}.$$

On en déduit en comparant à l'expression de $\partial_x u(t,x)$ donnée par (3.9) que l'on a bien

$$\partial_t u(t,x) + f'(u(t,x))\partial_x u(t,x) = 0$$

et, bien sûr, $u(0,\cdot) = u^0(\cdot)$. Supposons que $T < +\infty$, c'est-à-dire que

$$m = \inf_{x\in\mathbb{R}} f''(u^0(x))u^{0\prime}(x) < 0.$$

On va montrer qu'il ne peut exister une solution de (3.7) de classe $\mathcal{C}^1$ au delà du temps T. La seule chose à remarquer pour cela est que des caractéristiques se croisent pour les temps trop grands, et que, puisque la valeur de la solution u est transportée le long des caractéristiques, cela conduit à une *solution multivaluée*[5]. Soit $\overline{x} \in \mathbb{R}$ tel que $f''(u^0(\overline{x}))u^{0\prime}(\overline{x}) < 0$. Posons

$$\overline{t} = \frac{-1}{f''(u^0(\overline{x}))u^{0\prime}(\overline{x})}.$$

Nous allons montrer qu'il ne peut y avoir de solution de (3.7) de classe $\mathcal{C}^1([0,\tilde{t}]\times\mathbb{R})$ pour $\tilde{t} > \overline{t}$ (noter que $\overline{t} \geq T$). Soit $\tilde{t} > \overline{t}$. Soit $u \in \mathcal{C}^1([0,\tilde{t}]\times\mathbb{R})$ une solution de (3.7) associée à la donnée initiale u^0. Nous savons qu'elle vérifie $u(t, X(t,x)) = u^0(x)$ avec

$$X(t,x) = x + tf'(u^0(x)).$$

5. C'est-à-dire... Pas une solution.

On a

$$\overline{t} f''(u^0(\overline{x}))u^{0'}(\overline{x}) = -1,$$

donc

$$\tilde{t} f''(u^0(\overline{x}))u^{0'}(\overline{x}) < -1$$

puisque $\tilde{t} > \overline{t}$. De plus, d'après les hypothèses faites sur u^0 et f, $f'' \circ u^0(\cdot)u^{0'}(\cdot)$ est continue, donc il existe $\varepsilon > 0$ tel que pour tout $x \in [\overline{x} - \varepsilon, \overline{x}]$,

$$\tilde{t} f''(u^0(x))u^{0'}(x) < -1.$$

La caractéristique $X(t,x)$ est de classe $\mathcal{C}^1$ par rapport à x, donc il existe $y \in [\overline{x} - \varepsilon, \overline{x}]$ tel que

$$\begin{aligned} X(\tilde{t}, \overline{x} - \varepsilon) &= X(\tilde{t}, \overline{x}) - \varepsilon \partial_x X(\tilde{t}, y) \\ &= X(\tilde{t}, \overline{x}) - \varepsilon \Big(1 + \tilde{t} f''(u^0(y))u^{0'}(y)\Big) > X(\tilde{t}, \overline{x}). \end{aligned}$$

Les caractéristiques issues de $\overline{x} - \varepsilon$ et de $\overline{x}$ se sont croisées. Par continuité, $\exists t \in [0, \tilde{t}]$ tel que $X(t, \overline{x} - \varepsilon) = X(t, \overline{x})$. Or $u^0(\overline{x} - \varepsilon) \neq u^0(\overline{x})$ puisque $f''(u^0(x))u^{0'}(x) < 0$ sur $[\overline{x} - \varepsilon, \overline{x}]$. La contradiction est là :

$$u^0(\overline{x} - \varepsilon) = u(t, X(t, \overline{x} - \varepsilon)) = u(t, X(t, \overline{x})) = u^0(\overline{x}) \neq u^0(\overline{x} - \varepsilon).$$

□

Remarque 30
Un résultat analogue est bien sûr vrai pour les temps négatifs. □

Exemples.

1. Advection à vitesse constante : $f(u) = au$. Alors $f''(u) = 0$ et l'on a une solution pour tout $t \in \mathbb{R}$, ce qui ne fait que confirmer ce que l'on a déjà remarqué, à savoir que $u^0(x - at)$ est solution (pour $u^0 \in \mathcal{C}^1(\mathbb{R})$). Remarquer que la proposition 10 nous apprend l'unicité de la solution régulière.
2. Équation de Burgers : $f(u) = \frac{u^2}{2}$. On a $f''(u) = 1$. La proposition 10 donne l'existence et l'unicité d'une solution régulière si et seulement si $u^{0'}(x) \geq 0 \ \forall x \in \mathbb{R}$. Cela est très réducteur.

Pour sortir de l'impasse soulignée par le second exemple ci-dessus, nous allons généraliser la notion de solution à des fonctions non différentiables (et même non continues). Ce choix est motivé par des expériences physiques montrant que des variables physiques (densité, vitesse, pression pour ne parler que de l'hydrodynamique) peuvent être discontinues. Dans le cas de l'advection à vitesse constante, on peut aussi remarquer qu'il est naturel de considérer $u^0(x - at)$ comme solution même si u^0 n'est pas régulière.

Solutions au sens des distributions

Soit $u \in \mathcal{C}^1(\mathbb{R}_+ \times \mathbb{R})$ solution de (3.7) associé à la donnée initiale u^0 :

$$\begin{cases} \partial_t u(t,x) + \partial_x f(u)(t,x) = 0, \\ u(0,x) = u^0(x). \end{cases} \tag{3.10}$$

Soit $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. On a

$$\varphi(t,x)\partial_t u(t,x) + \varphi(t,x)\partial_x f(u)(t,x) = 0 \quad \forall (t,x) \in \mathbb{R}_+ \times \mathbb{R}.$$

Donc

$$\int_0^\infty \int_{-\infty}^\infty \varphi(t,x)\partial_t u(t,x) + \varphi(t,x)\partial_x f(u)(t,x)\, dx\, dt = 0.$$

En intégrant par parties, on en déduit que

$$\int_{-\infty}^\infty \left([\varphi(\cdot,x)u(\cdot,x)]_0^\infty - \int_0^\infty u(t,x)\partial_t\varphi(t,x)\, dt \right) dx \\ + \int_0^\infty \left([\varphi(t,\cdot)u(t,\cdot)]_{-\infty}^\infty - \int_{-\infty}^\infty f(u(t,x))\partial_x\varphi(t,x)\, dx \right) dt = 0.$$

Puisque φ est à support compact dans $\mathbb{R}_+ \times \mathbb{R}$, cela se simplifie en

$$\int_0^\infty \int_{-\infty}^\infty u(t,x)\partial_t\varphi(t,x) + f(u(t,x))\partial_x\varphi(t,x)\, dx\, dt = -\int_{-\infty}^\infty u^0(x)\varphi(0,x)\, dx. \tag{3.11}$$

Cette équation a un sens sans que u soit régulière. Il suffit que $u \in L^\infty(\mathbb{R}_+ \times \mathbb{R})$.

Définition 6

On appelle *solution faible* de (3.10) avec $u^0 \in L^\infty(\mathbb{R})$ une fonction $u \in L^\infty(\mathbb{R}_+ \times \mathbb{R})$ vérifiant (3.11) pour toute fonction $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. □

Exemples

1 On a montré que toute solution forte (régulière) est une solution faible.

2 La réciproque est fausse. Posons $f(u) = au$, avec $a \in \mathbb{R}$. Soit $u^0 \in L^\infty(\mathbb{R})$. Posons $u(t,x) = u^0(x - at)$. Nous allons vérifier que la notion de solution faible permet de définir cette fonction comme solution du problème d'advection à vitesse constante, ce que l'intuition nous conseillait. Soit $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. Nous voulons nous assurer du fait que $u(t,x) = u^0(x - at)$ vérifie (3.11).

$$\int_0^\infty \int_{-\infty}^\infty u\partial_t\varphi + f(u)\partial_x\varphi\, dx\, dt = \int_0^\infty \int_{-\infty}^\infty u\left(\partial_t\varphi + a\partial_x\varphi\right)\, dx\, dt \\ = \int_0^\infty \int_{-\infty}^\infty u^0(x - at)\left(\partial_t\varphi(t,x) + a\partial_x\varphi(t,x)\right)\, dx\, dt \\ = \int_0^\infty \int_{-\infty}^\infty u^0(y)\left(\partial_t\varphi(t,y+at) + a\partial_x\varphi(t,y+at)\right)\, dy\, dt.$$

Posons $\tilde{\varphi}(t,x) = \varphi(t, x+at)$. On a

$$\begin{aligned}\int_0^\infty \int_{-\infty}^\infty u\partial_t\varphi + f(u)\partial_x\varphi \, dx \, dt &= \int_0^\infty \int_{-\infty}^\infty u^0(x)\partial_t\tilde{\varphi}(t,x) \, dx \, dt \\ &= \int_{-\infty}^\infty u^0(x) \int_0^\infty \partial_t\tilde{\varphi}(t,x) \, dt \, dx = \int_{-\infty}^\infty u^0(x) \left[\tilde{\varphi}(\cdot,x)\right]_0^\infty \, dx \\ &= -\int_{-\infty}^\infty u^0(x)\tilde{\varphi}(0,x) \, dx = -\int_{-\infty}^\infty u^0(x)\varphi(0,x) \, dx.\end{aligned}$$

Donc u est solution faible.

Remarque 31
Une question naturelle qui se pose maintenant est : si u est une solution faible de (3.10) qui de plus est régulière, u est-elle une solution forte de (3.10) ? La réponse est positive, cela vient du fait que si u régulière vérifie (3.11) pour tout $\varphi \in \mathcal{C}_c^\infty$, en intégrant par parties, on en déduit que

$$\int_0^\infty \int_{-\infty}^\infty \left(\partial_t u + \partial_x f(u)\right) \varphi \, dx \, dt = 0$$

pour tout $\varphi \in \mathcal{C}_c^\infty$ et un résultat « classique » permet de conclure que $\partial_t u + \partial_x f(u) = 0$. Ce résultat classique sera utilisé par la suite, il s'agit du lemme 4. □

Remarque 32
Il est intéressant de constater que la formulation faible de (3.10), à savoir, (3.11), est très proche des équations de bilan de la mécanique des milieux continus. Elle consiste dans ce cadre en un retour aux premières étapes de la modélisation physique. Nous ne développons pas cette remarque ici, mais la lecture de [4] apportera à ce sujet de grands éclaircissements. □

Solutions discontinues en translation et relations de Rankine-Hugoniot

La section précédente nous a informés de l'existence de solutions discontinues en translation. Le but de cette section est d'analyser et de caractériser précisément ce type de solutions pour des équations hyperboliques plus générales que l'advection à vitesse constante. On considère l'équation (3.7) avec la condition initiale

$$u^0(x) = \begin{cases} u_G \text{ si } x \leq 0 \\ u_D \text{ si } x > 0 \end{cases}$$

(le problème est à comprendre au sens faible de (3.11)). On cherche une solution (faible) u de ce problème qui soit de la forme

$$u(t,x) = u^0(x - \sigma t)$$

où $\sigma \in \mathbb{R}$ est à déterminer, c'est-à-dire que l'on cherche une solution en translation à la vitesse σ. Lorsque $f(u) = au$, nous avons vu que $\sigma = a$ était une possibilité. Nous voulons montrer que c'est la seule et généraliser ce résultat au cas non linéaire.

Soit u une solution faible : pour tout $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$,

$$\int_0^\infty \int_{-\infty}^\infty u(t,x)\partial_t\varphi(t,x) + f(u(t,x))\partial_x\varphi(t,x)\,dx\,dt = -\int_{-\infty}^\infty u^0(x)\varphi(0,x)\,dx.$$

Le problème est donc de trouver $\sigma \in \mathbb{R}$ tel que

$$\int_0^\infty \int_{-\infty}^\infty u^0(x-\sigma t)\partial_t\varphi(t,x) + f(u^0(x-\sigma t))\partial_x\varphi(t,x)\,dx\,dt = -\int_{-\infty}^\infty u^0(x)\varphi(0,x)\,dx.$$

pour tout $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. Soit $t > 0$; notons, pour $\varepsilon > 0$ quelconque,

$$\begin{aligned} B_\varepsilon &= B\left((t,\sigma t),\varepsilon\right) \subset \mathbb{R}\times\mathbb{R}, \\ D_{\varepsilon,D} &= \{(s,x) \in B_\varepsilon \text{ t. q. } \sigma s < x\} \subset \mathbb{R}\times\mathbb{R}, \\ D_{\varepsilon,G} &= \{(s,x) \in B_\varepsilon \text{ t. q. } \sigma s > x\} \subset \mathbb{R}\times\mathbb{R} \end{aligned}$$

(voir la figure 3.1).

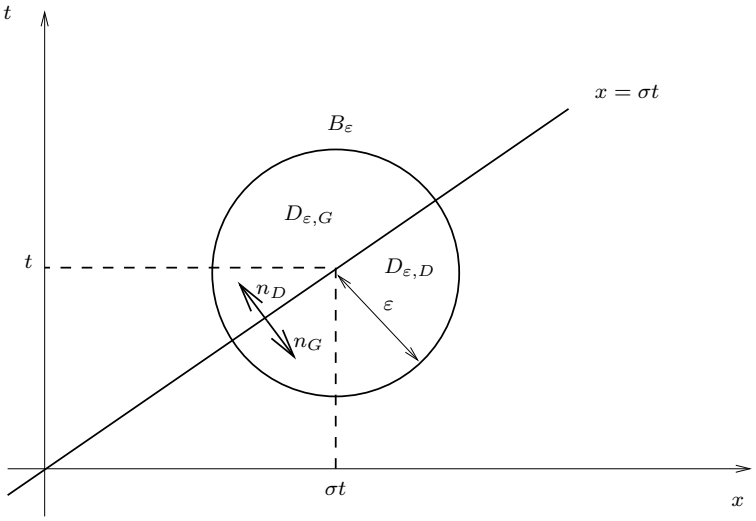

FIGURE 3.1 – Support de la fonction-test dans le plan espace-temps.

Choisissons $\varepsilon < t$ et $\varphi \in \mathcal{C}_c^\infty(B_\varepsilon)$. On a alors $\varphi(0,x) = 0\ \forall x \in \mathbb{R}$. L'équation faible devient donc

$$\int_{B_\varepsilon} u(t,x)\partial_t\varphi(t,x) + f(u(t,x))\partial_x\varphi(t,x)\,dx\,dt = 0.$$

Or $\overline{B_\varepsilon} = \overline{D_{\varepsilon,G}} \cup \overline{D_{\varepsilon,D}}$, donc l'équation est

$$\begin{aligned} \int_{D_{\varepsilon,G}} u(t,x)\partial_t\varphi(t,x) + f(u(t,x))\partial_x\varphi(t,x)\,dx\,dt \\ + \int_{D_{\varepsilon,D}} u(t,x)\partial_t\varphi(t,x) + f(u(t,x))\partial_x\varphi(t,x)\,dx\,dt = 0 \end{aligned}$$

et puisque $u(t,x) = u_G \ \forall (t,x) \in D_{\varepsilon,G}$ et $u(t,x) = u_D \ \forall (t,x) \in D_{\varepsilon,D}$, on doit finalement trouver σ tel que

$$\int_{D_{\varepsilon,G}} u_G \partial_t \varphi(t,x) + f(u_G) \partial_x \varphi(t,x)\, dx\, dt + \int_{D_{\varepsilon,D}} u_D \partial_t \varphi(t,x) + f(u_D) \partial_x \varphi(t,x)\, dx\, dt = 0.$$

On utilise maintenant une formule de Green sur $D_{\varepsilon,G}$ et sur $D_{\varepsilon,D}$ [6] :

$$\int_{\partial D_{\varepsilon,G}} (u_G n_{G,t} + f(u_G) n_{G,x})\, \varphi\, dS + \int_{\partial D_{\varepsilon,D}} (u_D n_{D,t} + f(u_D) n_{D,x})\, \varphi\, dS = 0$$

où $n_{G,t}$, $n_{G,x}$, $n_{D,t}$, $n_{D,x}$ sont les composantes suivant t et x des normales extérieures unitaires de $D_{\varepsilon,G}$ et $D_{\varepsilon,D}$. Or pour $(t,x) \in \partial D_{\varepsilon,G}$ et pour $(t,x) \in \partial D_{\varepsilon,D}$ on a $\varphi(t,x) = 0$ sauf si $x = \sigma t$. De plus, sur ce segment $\{x = \sigma t\}$, on a, à un facteur multiplicatif (de normalisation) près,

$$n_G = \begin{pmatrix} -\sigma \\ 1 \end{pmatrix}, \quad n_D = \begin{pmatrix} \sigma \\ -1 \end{pmatrix},$$

de sorte que σ doit vérifier

$$\int_{\partial D_{\varepsilon,G} \cap \{(t,x) \text{ t. q. } x=\sigma t\}} (-\sigma\,(u_G - u_D) + f(u_G) - f(u_D))\, \varphi\, dS = 0$$

pour tout $\varphi \in \mathcal{C}_c^\infty(B_\varepsilon)$, c'est-à-dire encore

$$(-\sigma\,(u_G - u_D) + f(u_G) - f(u_D)) \int_{\partial D_{\varepsilon,G} \cap \{(t,x) \text{ t. q. } x=\sigma t\}} \varphi\, dS = 0.$$

Une conséquence immédiate est que

$$-\sigma\,(u_G - u_D) + f(u_G) - f(u_D) = 0.$$

Cette équation est une relation de compatibilité entre σ et (u_G, u_D) pour que la fonction discontinue en translation à vitesse σ u soit solution de l'équation. Elle est appelée *relation de Rankine-Hugoniot* [7]. Noter que le fait que l'EDP soit scalaire n'a pas été utilisé : les relations de Rankine-Hugoniot sont vérifiées aussi par les solutions de systèmes d'EDP hyperboliques.

Lorsque $u_G \neq u_D$ ($u_G = u_D$ est un cas trivial de solution de l'EDP), la relation de Rankine-Hugoniot donne l'expression de σ

$$\sigma = \frac{f(u_D) - f(u_G)}{u_D - u_G}.$$

6. L'unique formule à se rappeler, pour un ouvert suffisamment régulier $\Omega \in \mathbb{R}^d$ et $u, v \in \mathcal{C}^1(\overline{\Omega})$, est $\int_\Omega u \partial_i v\, dx = \int_{\partial\Omega} uv n_i\, dS - \int_\Omega v \partial_i u\, dx$ où n_i est la n^{e} coordonnée de la normale unitaire extérieure n à $\partial\Omega$. Dans cette formule, dx représente l'élément de volume dans Ω et dS l'élément de surface sur $\partial\Omega$.

7. On laisse le soin au lecteur de vérifier qu'elle est aussi une condition *suffisante* pour que la fonction soit solution.

Remarque 33

1 Dans le cas $f(u) = au$, la formule précédente donne $\sigma = a$ (nous savions déjà que c'était convenable).

2 La relation de Rankine-Hugoniot implique que les discontinuités d'amplitude très petite se déplacent à une vitesse proche de la vitesse caractéristique :

$$\sigma = \frac{f(u_D) - f(u_G)}{u_D - u_G} = f'(u_G) + \mathcal{O}(u_D - u_G)$$

si f est de classe $\mathcal{C}^1$.

3 Dans le cas $f(u) = u^2/2$, cas de l'équation de Burgers, on a

$$\sigma = \frac{u_G + u_D}{2}.$$

4 La relation de Rankine-Hugoniot peut se montrer de manière très simple si l'on est un peu familier avec les distributions. On a supposé que $u(t,x) = u_G + (u_D - u_G)H(x - \sigma t)$ où $H(x) = \mathbb{1}_{\mathbb{R}_+}(x)$ est la fonction de Heaviside. On sait que la dérivée au sens des distributions de H est δ, masse de Dirac en 0. On a $\partial_t u = -\sigma(u_D - u_G)\delta(x - \sigma t)$. Par ailleurs, $f(u(t,x)) = f(u_G) + (f(u_D) - f(u_G))H(x - \sigma t)$, d'où $\partial_x f(u)(t,x) = (f(u_D) - f(u_G))\delta(x - \sigma t)$. L'équation $\partial_t u + \partial_x f(u) = 0$ est ainsi équivalente à $-\sigma(u_D - u_G) + f(u_D) - f(u_G) = 0$.

5 Dans la démonstration, nous n'avons nulle part utilisé le fait que u était scalaire : le résultat est vrai pour des systèmes d'ordre 1. Cependant, une différence notable est que dans le cas scalaire, pour tout couple (u_G, u_D), $u(t,x) = u_G + (u_D - u_G)H(x - \sigma t)$ est solution faible de l'équation, avec $\sigma = (f(u_D) - f(u_G))/(u_D - u_G)$, et que ceci est faux dans le cas d'un système, puisque dans ce cas toutes les composantes des vecteurs u_G et u_D doivent vérifier une condition de compatibilité impliquant (permettant) qu'elles se déplacent à la *même* vitesse σ. Précisément, pour un système de dimension n, il faut que $(f_j(u_D) - f_j(u_G))(u_{Di} - u_{Gi}) = (f_i(u_D) - f_i(u_G))(u_{Dj} - u_{Gj})$ pour tous $i, j \in \{1, \ldots, n\}$, et la vitesse de translation est alors $\sigma = (f_j(u_D) - f_j(u_G))/(u_{Dj} - u_{Gj})$, indépendante de j (pour j tel que $u_{Dj} \neq u_{Gj}$).

□

3.3.2 Équation de Burgers

On aborde ici l'EDP

$$\partial_t u + \partial_x \frac{u^2}{2} = 0 \tag{3.12}$$

qui, en ce qui concerne ses solutions régulières, est équivalente à

$$\partial_t u + u\partial_x u = 0 \ :$$

u, solution de l'équation, est aussi la pente des courbes caractéristiques sur lesquelles elle est constante. La vitesse σ d'une discontinuité en translation solution de cette équation est donnée par

$$\sigma = \frac{u_G + u_D}{2}$$

(si u_G et u_D sont les états situés à gauche et à droite de la discontinuité).

Non-unicité des solutions

Nous avons montré dans le cas général (scalaire) l'unicité de la solution forte du problème avec donnée initiale. Nous allons voir ici qu'il n'y a pas unicité de la solution dans le cas non régulier. Autrement dit, la notion de solution faible, qui permet d'avoir existence de solutions dans un cadre plus général, est hélas trop faible pour garantir l'unicité (dans le cas du moins de l'équation non linéaire de Burgers).

Afin d'exhiber deux solutions faibles d'un même problème, nous choisissons la condition initiale

$$u^0(x) = \begin{cases} 0 \text{ si } x \leq 0, \\ 1 \text{ si } x > 0. \end{cases} \tag{3.13}$$

Une première solution faible de (3.12) avec la condition initiale (3.13) est alors fournie par la section précédente : il s'agit de la solution discontinue en translation

$$u(t,x) = \begin{cases} 0 \text{ si } x - \dfrac{t}{2} \leq 0, \\ 1 \text{ si } x - \dfrac{t}{2} > 0. \end{cases}$$

Une autre solution faible du problème est donnée par

$$u(t,x) = \begin{cases} 0 \text{ si } x \leq 0, \\ \dfrac{x}{t} \text{ si } 0 < x \leq t, \\ 1 \text{ si } x > t \end{cases}$$

pour tout $t \geq 0$. Cette solution est appelée « détente ». Montrons que c'est une solution faible. Soit $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. D'après le théorème de Fubini,

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi \, dt \, dx = \int_{\mathbb{R}_+} \int_{\mathbb{R}} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi \, dx \, dt,$$

donc

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi \, dt \, dx = \int_{\mathbb{R}_+} \int_{-\infty}^{0} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi \, dx \, dt + \int_{\mathbb{R}_+} \int_{0}^{t} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi \, dx \, dt$$
$$+ \int_{\mathbb{R}_+} \int_{t}^{\infty} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi \, dx \, dt,$$

soit

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi\,dt\,dx = 0 + \int_{\mathbb{R}_+}\int_0^t \frac{x}{t}\partial_t\varphi + \frac{x^2}{2t^2}\partial_x\varphi\,dx\,dt \\ + \int_{\mathbb{R}_+}\int_t^\infty \partial_t\varphi + \frac{1}{2}\partial_x\varphi\,dx\,dt.$$

Ainsi

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi\,dt\,dx = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \frac{x}{t}\partial_t\varphi\mathbb{1}_{[0,t]}(x)\,dx\,dt \\ + \int_{\mathbb{R}_+}\left(\left[\frac{x^2}{2t^2}\varphi\right]_{x=0}^t - \int_0^t \frac{x}{t^2}\varphi\,dx\right)\,dt \\ + \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \partial_t\varphi\mathbb{1}_{[t,+\infty[}(x)\,dx\,dt + \int_{\mathbb{R}_+}\frac{1}{2}\left[\varphi\right]_{x=t}^\infty\,dt$$

et

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi\,dt\,dx \\ = \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \frac{x}{t}\partial_t\varphi\mathbb{1}_{[x,+\infty[}(t)\,dx\,dt + \int_{\mathbb{R}_+}\frac{1}{2}\varphi(t,t)\,dt - \int_{\mathbb{R}_+}\int_0^t \frac{x}{t^2}\varphi\,dx\,dt \\ + \int_{\mathbb{R}_+}\int_{\mathbb{R}_+} \partial_t\varphi\mathbb{1}_{[0,x]}(t)\,dx\,dt - \int_{\mathbb{R}_+}\frac{1}{2}\varphi(t,t)\,dt \\ = \int_{\mathbb{R}_+}\left(\left[\frac{x}{t}\varphi\right]_{t=x}^\infty + \int_x^\infty \frac{x}{t^2}\varphi\,dt\right)\,dx - \int_{\mathbb{R}_+}\int_0^t \frac{x}{t^2}\varphi\,dx\,dt + \int_{\mathbb{R}_+}\left[\varphi\right]_{t=0}^x\,dx$$

et enfin

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\partial_t\varphi + \frac{u^2}{2}\partial_x\varphi\,dt\,dx \\ = -\int_{\mathbb{R}_+}\varphi(x,x)\,dx + \int_{\mathbb{R}_+}\int_x^{+\infty}\frac{x}{t^2}\varphi\,dt\,dx - \int_{\mathbb{R}_+}\int_0^t\frac{x}{t^2}\varphi\,dx\,dt \\ + \int_{\mathbb{R}_+}\varphi(x,x)\,dx - \int_{\mathbb{R}_+}\varphi(0,x)\,dx \\ = -\int_{\mathbb{R}_+}\varphi(0,x)\,dx = -\int_{\mathbb{R}} u^0(x)\varphi(0,x)\,dx.$$

C'est ce qu'il fallait démontrer. On en retiendra que les solutions faibles de l'équation de Burgers ne sont pas uniques. Il existe plusieurs critères permettant de sélectionner la solution « physique » parmi toutes les solutions faibles possibles : critère de dissipation, critère d'entropie, critère d'Oleinik, de Lax, de Liu... Nous utiliserons dans la suite le critère d'Oleinik. Il permet de sélectionner une unique solution dans le cas où le flux de l'équation, f, est convexe. En d'autres

termes, lorsque f est convexe, il existe une unique solution faible du problème (3.10) vérifiant de surcroît le critère d'Oleinik que nous introduirons dans la suite.

Noter cependant que la non-unicité que nous avons montrée est due à la non-linéarité du flux f. En effet, dans le cas linéaire, on a la

Proposition 11
Soit $a \in \mathbb{R}$, soit $u^0 \in L^\infty(\mathbb{R})$. Le problème

$$\begin{aligned} &\partial_t u + a\partial_x u = 0, \\ &u(0,x) = u^0(x) \end{aligned}$$

admet une unique solution faible. Cette solution est donnée par

$$u(t,x) = u^0(x - at).$$

□

Démonstration
Nous avons déjà montré que $u^0(x-at)$ est solution faible, il faut maintenant montrer que c'est l'unique solution. Puisque l'équation est linéaire, il suffit de montrer que l'unique solution du problème avec condition initiale $u^0 = 0$ est $u(t,x) = 0$. Soit v solution de

$$\begin{aligned} &\partial_t v + a\partial_x v = 0, \\ &v(0,x) = 0 \text{ presque partout.} \end{aligned}$$

Pour tout $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$,

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} v \left(\partial_t \varphi + a\partial_x \varphi\right) \, dt \, dx = 0.$$

L'idée de la démonstration est de montrer que v vérifie

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} v\psi \, dt \, dx = 0 \quad \forall \psi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R}).$$

Soit donc $\psi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. Soit T et A tels que

$$\psi(t,x) = 0 \quad \forall (t,x) \notin [0,T] \times [-A,A].$$

Posons

$$\varphi(t,x) = \int_0^t \psi(s, x + a(s-t)) \, ds - \int_0^T \psi(s, x + a(s-t)) \, ds.$$

On a alors :

— φ est de classe $\mathcal{C}^\infty$;
— $\partial_t \varphi + a\partial_x \varphi = \psi$ (à vérifier en exercice) ;
— $\varphi(t,x) = 0 \ \forall (t,x) \notin [0,T] \times [-A, A + aT]$ si $a \geq 0$, ou $\forall (t,x) \notin [0,T] \times [-A - aT, A]$ si $a \leq 0$ (donc φ est à support compact).

(Remarquer que $\int_0^t \psi(s, x + a(s-t))\,ds + f(x-at)$ est de classe $\mathcal{C}^\infty$ et vérifie $\partial_t\varphi + a\partial_x\varphi = \psi$ pour toute fonction f de classe $\mathcal{C}^\infty$, mais n'est pas forcément à support compact). Donc $\int_{\mathbb{R}}\int_{\mathbb{R}_+} v\psi\,dt\,dx = 0\ \forall\psi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. On en déduit que $v = 0$ presque partout par application du lemme 4. □

Remarque 34
Le principe de la démonstration que nous venons d'effectuer est de montrer que l'application

$$\begin{aligned} &\mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R}) \longrightarrow \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R}) \\ &\psi \longmapsto \partial_t\psi + a\partial_x\psi \end{aligned}$$

est une bijection. C'est par un raisonnement tout à fait semblable que nous avons montré l'existence et l'unicité des solutions fortes (en temps petit) puisque nous avons utilisé le fait que les caractéristiques recouvraient tout $\mathbb{R}_+ \times \mathbb{R}$. □

Lemme 4
Soit $\Omega \in \mathbb{R}^n$ un ouvert, soit $v \in L^1_{loc}(\Omega)$ [8]. Supposons que

$$\int_\Omega v\psi\,dx = 0 \quad \forall\psi \in \mathcal{C}_c^\infty(\Omega).$$

Alors $v = 0$ presque partout. □

Il s'agit d'un résultat classique dont la démonstration pourra être trouvée dans [1].

Il est temps maintenant de chercher un critère permettant de lever le problème de la non-unicité des solutions dans le cas non linéaire. Nous avons déjà évoqué le mot de solution « physique ». Bien entendu, ce terme n'a pas de sens clair lorsqu'il s'agit de l'équation de Burgers. Cependant il est naturellement lié à une propriété mathématique bien définie : la continuité de la solution par rapport à la donnée initiale. Nous allons voir que la solution qui propage la discontinuité initiale de (3.13) n'est pas stable par perturbation régulière de la donnée initiale. Pour cela, nous considérons la donnée initiale régulière (de classe $\mathcal{C}^1$)

$$v(x) = \begin{cases} 0 \text{ si } x < -\dfrac{1}{2}, \\ 2(x+1/2)^2 \text{ si } x \in [-\dfrac{1}{2}, 0[, \\ 1 - 2(x-1/2)^2 \text{ si } x \in [0, \dfrac{1}{2}[, \\ 1 \text{ si } x \geq \dfrac{1}{2} \end{cases}$$

et proposons la suite de conditions initiales $\left(u_n^0\right)_{n\in\mathbb{N}^*}$ définie par

$$u_n^0(x) = v(nx) \quad \forall n \in \mathbb{N}^*, \forall x \in \mathbb{R}$$

8. Pour tout compact $K \subset \Omega$, $v\mathbb{1}_K \in L^1(\Omega)$.

(voir la figure 3.2). On a $\lim_{n\to+\infty} u_n^0 = u^0$ presque partout et dans $L^1(\mathbb{R})$, et u_n^0 est de classe $\mathcal{C}^1(\mathbb{R})$ pour tout $n \in \mathbb{N}^*$.

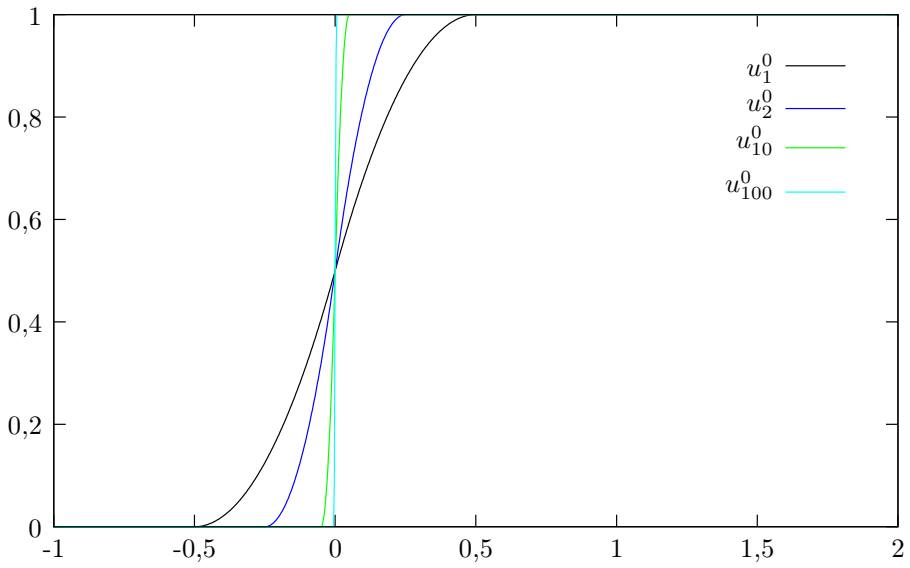

FIGURE 3.2 – Suite de conditions initiales.

Puisque pour tout $n \in \mathbb{N}^*$ u_n^0 est de classe $\mathcal{C}^1$ et croissante, il existe une unique solution globale [9], de classe $\mathcal{C}^1$, au problème initial, cette solution étant donnée par la méthode des caractéristiques. Pour calculer cette solution pour tout n, remarquons que si u_1 est la solution avec donnée initiale u_1^0, $u_n(t,x) = u_1(nt,nx)$ l'est pour le problème avec donnée initiale u_n^0 (c'est une propriété d'auto-similarité [10]). En effet, posons $u_n(t,x) = u_1(nt,nx)$. On a alors naturellement $u_n(0,x) = u_n^0(x)$ et

$$\begin{aligned}\partial_t u_n(t,x) + \partial_x \frac{u_n^2}{2}(t,x) &= \partial_t u_n(t,x) + u_n(t,x)\partial_x u_n(t,x) \\ &= n\partial_t u_1(nt,nx) + nu_1(nt,nx)\partial_x u_1(nt,nx) = 0\end{aligned}$$

donc u_n est solution. Il suffit donc de calculer la solution associée à la donnée initiale $u_1^0 = v$, en remontant les caractéristiques. Soit $t > 0$.

1 Pour $x < -1/2$, $u(t,x) = 0$.

2 Pour $x \in [-1/2, t/2[$, il existe $y \in [-1/2, 0]$ tel que x est atteint au temps t par la caractéristique issue de y :

$$x = y + t\left(2\left(y + \frac{1}{2}\right)^2\right)$$

9. D'après la proposition 10.

10. La solution $u_n(t,x)$ ne dépend que du rapport x/t. Pour un léger approfondissement sur les solutions autosimilaires, on pourra se référer à l'examen du 3 juin 2005. Les solutions autosimilaires jouent un rôle primordial dans l'étude des systèmes hyperboliques.

et la solution au point (t,x) vaut sa valeur au pied de la caractéristique,

$$u(t,x) = u_1^0(y) = 2\left(y+\frac{1}{2}\right)^2.$$

Calculons y. Il suffit pour cela de trouver la racine située dans $[-1/2,0]$ de l'équation polynomiale

$$2ty^2 + (1+2t)y + \frac{t}{2} - x = 0.$$

Cette racine est donnée par

$$y = \frac{-1-2t+\sqrt{(1+2t)^2 - 8t\left(\frac{t}{2}-x\right)}}{4t} = \frac{-1-2t+\sqrt{1+4t+8tx}}{4t}.$$

3 Pour $x \in [t/2, 1/2+t[$, il existe $y \in [0,1/2]$ tel que x est atteint au temps t par la caractéristique issue de y :

$$x = y + t\left(1 - 2\left(y-\frac{1}{2}\right)^2\right)$$

et la solution au point (t,x) vaut sa valeur au pied de la caractéristique,

$$u(t,x) = u_1^0(y) = 1 - 2\left(y-\frac{1}{2}\right)^2.$$

Calculons y. Il suffit pour cela de trouver la racine située dans $[0,1/2]$ de l'équation polynomiale

$$2ty^2 - (1+2t)y + x - \frac{t}{2} = 0.$$

Cette racine est donnée par

$$y = \frac{1+2t-\sqrt{(1+2t)^2 + 8t\left(\frac{t}{2}-x\right)}}{4t} = \frac{1+2t-\sqrt{1+8t^2+4t-8tx}}{4t}.$$

4 Pour $x \geq 1/2+t$, $u(t,x) = 1$.

Récapitulons.

$$u(t,x) = \begin{cases} 0 \text{ si } x < -\dfrac{1}{2}, \\ 2\left(\dfrac{-1-2t+\sqrt{1+4t+8tx}}{4t} + \dfrac{1}{2}\right)^2 \text{ si } x \in [-\dfrac{1}{2}, \dfrac{t}{2}[, \\ 1 - 2\left(\dfrac{1+2t-\sqrt{1+8t^2+4t-8tx}}{4t} - \dfrac{1}{2}\right)^2 \text{ si } x \in [\dfrac{t}{2}, \dfrac{1}{2}+t[, \\ 1 \text{ si } x \geq \dfrac{1}{2}+t. \end{cases}$$

soit

$$u(t,x) = \begin{cases} 0 \text{ si } x < -\dfrac{1}{2}, \\ 2\left(\dfrac{-1+\sqrt{1+4t+8tx}}{4t}\right)^2 \text{ si } x \in [-\dfrac{1}{2}, \dfrac{t}{2}[, \\ 1-2\left(\dfrac{1-\sqrt{1+8t^2+4t-8tx}}{4t}\right)^2 \text{ si } x \in [\dfrac{t}{2}, \dfrac{1}{2}+t[, \\ 1 \text{ si } x \geq \dfrac{1}{2}+t. \end{cases}$$

Ainsi

$$u_n(t,x) = \begin{cases} 0 \text{ si } x < -\dfrac{1}{2n}, \\ 2\left(\dfrac{-1+\sqrt{1+4nt+8n^2tx}}{4nt}\right)^2 \text{ si } x \in [-\dfrac{1}{2n}, \dfrac{t}{2}[, \\ 1-2\left(\dfrac{1-\sqrt{1+8n^2t^2+4nt-8n^2tx}}{4nt}\right)^2 \text{ si } x \in [\dfrac{t}{2}, \dfrac{1}{2n}+t[, \\ 1 \text{ si } x \geq \dfrac{1}{2n}+t. \end{cases}$$

On vérifie très facilement que pour tout x et tout t (strictement positif), $u_n(t,x)$ converge vers $u(t,x)$ définie par

$$u(t,x) = \begin{cases} 0 \text{ si } x < 0, \\ \dfrac{x}{t} \text{ si } x \in [0, \dfrac{t}{2}[, \\ 1-\dfrac{t-x}{t} \text{ si } x \in [\dfrac{t}{2}, t[, \\ 1 \text{ si } x \geq t, \end{cases}$$

c'est-à-dire enfin

$$u(t,x) = \begin{cases} 0 \text{ si } x < 0, \\ \dfrac{x}{t} \text{ si } x \in [0, t[, \\ 1 \text{ si } x \geq t. \end{cases}$$

C'est la solution que nous avons nommée « détente » et non celle qui propage la discontinuité initiale.

Quelques éléments de la suite de solutions au temps $t = 1$ sont représentés sur la figure 3.3. La figure 3.4 présente aussi quelques uns de ces éléments ainsi que la solution détente en fonction de t et x. Sur cette dernière, on a tracé des isolignes (en temps-espace) des solutions. On constate que ce sont des droites : on retrouve bien les caractéristiques.

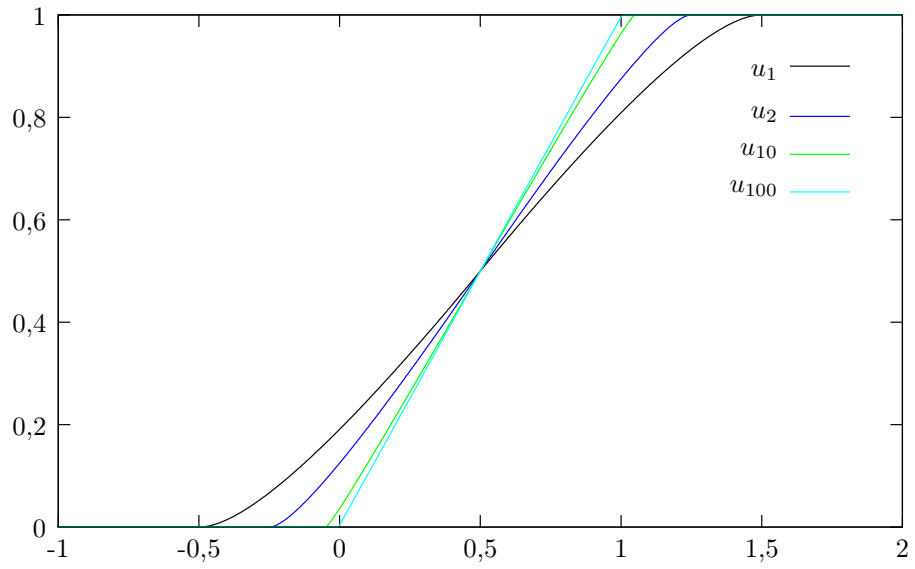

FIGURE 3.3 – Suite de solutions au temps $t = 1$.

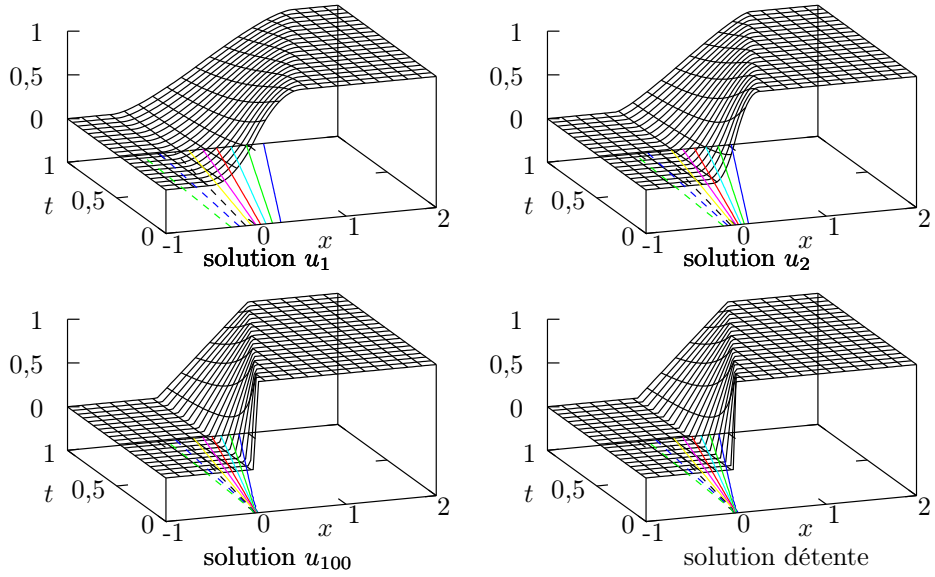

FIGURE 3.4 – Quelques éléments u_n dans le plan espace-temps.

Remarque 35

La conclusion de ces dernières pages pourrait être obtenue avec n'importe quelle régularisation de la donnée initiale. Celle que nous avons utilisée ici présente sur les autres l'avantage de donner lieu à des calculs explicites aisés, notamment en ce qui concerne la position du pied des caractéristiques. □

Voilà qui nous invite à considérer la solution

$$u(t,x) = \begin{cases} 0 \text{ si } x < \dfrac{t}{2} \\ 1 \text{ si } x \geq \dfrac{t}{2} \end{cases}$$

comme indésirable. Il faut remarquer qu'il en va tout autrement de la condition initiale

$$u^0 = \begin{cases} 1 \text{ si } x < 0 \\ 0 \text{ si } x \geq 0. \end{cases}$$

En effet, si l'on considère comme condition initiale une régularisation de classe $\mathcal{C}^1$ de ce u^0, la solution développera un choc en temps fini [11]. La différence provient du fait que dans ce nouveau cas, les caractéristiques se croisent, alors que dans le cas précédent, elles s'écartent. C'est cette constatation qui nous incite à définir la notion suivante d'admissibilité d'une solution dans le cas où il n'y a pas unicité faible [12]

Définition 7
Une solution discontinue (u_G, u_D, σ) est dite *admissible au sens de Lax* si et seulement si elle vérifie

$$f'(u_G) \geq \sigma \geq f'(u_D).$$

□

On pourrait montrer que ce critère suffit à assurer l'unicité des solutions faibles dans le cas où f est convexe. Nous allons pourtant nous pencher plutôt sur un autre critère de sélection. Il est, lui, basé sur le fait que, dans notre exemple du moins, seules les discontinuités décroissantes paraissent admissibles [13]. Une manière plus précise de formaliser ceci est d'écrire que

$$u(t,y) - u(t,x) \leq \frac{1}{t}(y-x) \quad \forall y \geq x.$$

On dit que $u(t,\cdot)$ est *lipschitzienne d'un côté* [14] (en l'occurrence, lipschitzienne à droite).

Définition 8
Soit $u \in L^\infty([0,T[\times\mathbb{R})$. On dit que u satisfait à une inégalité d'Oleinik si et seulement si $\exists C :]0,T[\longrightarrow \mathbb{R}$ décroissante telle que

$$u(t,y) - u(t,x) \leq C(t)(y-x) \quad \forall y \geq x, \forall t > 0.$$

□

11. Nous démontrerons bientôt que dans ce cas, la solution-choc vérifiant la relation de Rankine-Hugoniot est la seule possible.

12. Dans le cas d'une solution forte, régulière, nous savons qu'il y a unicité, donc un critère d'admissibilité n'est nécessaire que dans le cas non régulier.

13. Il faut noter que cela est dû à la convexité de f et que ce serait le contraire avec $f(u) = -u^2/2$.

14. En anglais : one-sided Lipschitz continuous.

Remarquer que l'on n'exclut pas la possibilitÃ© que C ne soit pas borné. Ceci sera essentiel dans l'application à l'équation de Burgers que nous ferons du résultat d'unicité suivant (dû à Oleinik).

Théorème 9 (Oleinik)
Soit $u \in L^\infty([0,T[\times\mathbb{R})$ solution de

$$\begin{cases} \partial_t u(t,x) + \partial_x \left(au\right)(t,x) = 0 \\ u(0,x) = 0 \end{cases}$$

où $a \in L^\infty([0,T[\times\mathbb{R})$ et $T \in \overline{\mathbb{R}_+}$ sont donnés. Si a satisfait à une inégalité d'Oleinik, $u = 0$ presque partout. □

Ce résultat d'unicité pour une équation de transport sera plus tard appliqué aux équations non linéaires qui nous concernent.

Remarque 36
Il s'agit d'une généralisation de la partie « unicité » de la proposition 11. □

Remarque 37
La question de l'*existence* d'une solution à l'équation $\partial_t u + \partial_x(au) = 0$ sous les hypothèses de ce théorème est très délicate et ne sera pas abordée dans ce cours. □

Remarque 38
Voici un contre-exemple dans le cas où la vitesse de transport ne vérifie pas de condition d'Oleinik. On choisit

$$a(t,x) = \begin{cases} -1 \text{ si } x \leq 0, \\ 1 \text{ si } x > 0 \end{cases} \qquad \forall t \in \mathbb{R}.$$

Pour $u^0 = 0$, $u = 0$ est bien sûr une solution, mais il en existe d'autres, par exemple[15] les fonctions du type

$$u(t,x) = \begin{cases} 0 \text{ si } x < -t, \\ -\alpha \text{ si } -t \leq x \leq 0, \\ \alpha \text{ si } 0 < x \leq t, \\ 0 \text{ si } t \leq x, \end{cases}$$

pour tout $\alpha \in \mathbb{R}$. □

Démonstration
Le principe en est le même que pour la proposition 11 mais il faut être ici beaucoup plus fin. Puisque u est solution faible du problème,

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u\left(\partial_t \varphi + a\partial_x \varphi\right) dt\, dx = 0$$

15. Cela se démontre facilement...

pour tout $\varphi \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$. Nous allons montrer que

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\psi\, dt\, dx = 0 \quad \forall \psi \in \mathcal{C}_c^\infty(]0,T[\times\mathbb{R}).$$

Attention, le support de ψ en temps est strictement inclus dans $[0,T]$, en particulier il ne contient pas 0. On suppose que $\psi(t,x) = 0$ pour tout $t \notin [\delta, T-\delta]$. Pour cela, il suffirait de montrer que pour tout $\psi \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$, il existe $\varphi \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$ tel que

$$\psi = \partial_t\varphi + a\partial_x\varphi.$$

Cependant, c'est hors de question puisque a n'est pas régulier. On procède par régularisation de la vitesse a. Posons $B = ||a||_{L^\infty}$ et soit $C :]0,T[\longrightarrow \mathbb{R}$ décroissante telle que

$$a(t,y) - a(t,x) \leq C(t)(y-x) \quad \forall y \geq x, \forall t > 0.$$

Soit $(a_n)_{n\in\mathbb{N}}$ une suite de fonctions de $\mathcal{C}^\infty([0,T[\times\mathbb{R})$ telle que [16]

$$\begin{aligned}
&\forall n \in \mathbb{N}, ||a_n||_{L^\infty} \leq B,\\
&\forall n \in \mathbb{N}, a_n(t,y) - a_n(t,x) \leq C(t)(y-x) \quad \forall y \geq x, \forall t > 0,\\
&\lim_{n\to\infty} a_n = a \text{ presque partout.}
\end{aligned}$$

Pour tout $n \in \mathbb{N}$, soit $X_n(t,t_0,x)$ défini par

$$\begin{cases} \partial_1 X_n(t,t_0,x) = a_n(t, X(t,t_0,x)) \\ X_n(t_0,t_0,x) = x \end{cases}$$

(pour $(t,t_0,x) \in [0,T[^2\times\mathbb{R}$). Une solution de

$$\partial_t\varphi + a_n\partial_x\varphi = \psi$$

est alors donnée [17] par

$$\varphi_n(t,x) = \int_0^t \psi(s, X_n(s,t,x))\, ds - \int_0^T \psi(s, X_n(s,t,x))\, ds,$$

ce qui peut aussi s'écrire

$$\varphi_n(t,x) = \int_T^t \psi(s, X_n(s,t,x))\, ds,$$

ce qui prouve que φ_n est de classe $\mathcal{C}^\infty$ et est à support compact en temps. D'autre part, puisque $||a_n||_{L^\infty} \leq B$,

$$X_n(s,t,x) \in [x - TB, x + TB] \quad \forall (s,t,x) \in [0,T[^2\times\mathbb{R}.$$

16. Montrer qu'une telle suite existe !
17. Une vérification s'impose.

Donc φ_n est à support compact en espace [18]. En définitive, $\varphi_n \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$. Nous avons montré que pour tout $n \in \mathbb{N}$, $\forall \psi \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$, $\exists \varphi_n \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$ tel que

$$\partial_t \varphi_n + a_n \partial_x \varphi_n = \psi.$$

Donc

$$\begin{aligned}\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\psi \, dt\, dx &= \int_{\mathbb{R}}\int_{\delta}^{T-\delta} u\psi \, dt\, dx = \int_{\mathbb{R}}\int_{\delta}^{T-\delta} u\left(\partial_t\varphi_n + a_n\partial_x\varphi_n\right) dt\, dx \\ &= \int_{\mathbb{R}}\int_{\delta}^{T-\delta} u\left(-a\partial_x\varphi_n\right) + u\left(a_n\partial_x\varphi_n\right) dt\, dx = \int_{\mathbb{R}}\int_{\delta}^{T-\delta} u\left(a_n - a\right)\partial_x\varphi_n \, dt\, dx.\end{aligned}$$

Nous voulons montrer que ce dernier terme converge vers 0 lorsque n tend vers ∞. C'est une conséquence du lemme 5, ci-après énoncé et démontré. En appliquant ce lemme à $X_n(s,t,x)$, nous obtenons en effet

$$|\partial_3 X_n(s,t,x)| \leq e^{C(t)(s-t)} \text{ pour } s \geq t > 0$$

car pour $s \geq t$, $C(s) \leq C(t)$. Or φ_n, donnée par $\varphi_n(t,x) = \int_{T-\delta}^{max(t,\delta)} \psi(s, X_n(s,t,x))\, ds$, vérifie

$$\partial_x\varphi_n(t,x) = \int_{T-\delta}^{\max(t,\delta)} \partial_x\psi(s, X_n(s,t,x))\partial_3 X_n(s,t,x)\, ds,$$

donc

$$\partial_x\varphi_n(t,x) \leq ||\partial_x\psi||_{L^\infty} \int_{\max(t,\delta)}^{T-\delta} |\partial_3 X_n(s,t,x)| \, ds,$$

d'où

$$|\partial_x\varphi_n(t,x)| \leq \begin{cases} ||\partial_x\psi||_{L^\infty} \left(\frac{1}{C(\max(t,\delta))}\left|e^{C(\max(t,\delta))(T-t)} - 1\right|\right) \text{ si } C(\max(t,\delta)) > 0 \\ T\,||\partial_x\psi||_{L^\infty} \text{ sinon.}\end{cases}$$

Le second membre ci-dessus est majorÃ© par un rÃ©el indÃ©pendant de t que nous notons C. Nous avons donc

$$\left|\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\psi \, dt\, dx\right| \leq \int_{\mathbb{R}}\int_{\mathbb{R}_+} |u(a_n - a)|\,|\partial_x\varphi_n| \, dt\, dx \leq C\int_K\int_0^T |u(a_n - a)| \, dt\, dx$$

où K est un compact de $\mathbb{R}$ tel que $[0,T]\times K$ englobe le support de φ_n pour tout n (par exemple $[-A-BT, A+BT]$ si $[0,T]\times[-A,A]$ englobe celuis de ψ). On a

- $\lim_{n\to\infty} u(a_n - a) = 0$ presque partout ;
- $u(a_n - a) \in L^1_{loc}$ $\forall n \in \mathbb{N}$ (car ces fonctions sont dans L^∞) ;
- $u(a_n - a) \leq ||u||_{L^\infty}\left(||a_n||_{L^\infty} + ||a||_{L^\infty}\right) \leq 2B\,||u||_{L^\infty}$ $\forall n \in \mathbb{N}$.

18. Le compact en question est d'ailleurs inclus, indÃ©pendamment de n, dans $[0,T]\times[-A-BT, A+BT]$ si A est un rÃ©el tel que le support de ψ est inclus dans $[0,T]\times[-A,A]$.

Par application du théorème de convergence dominée de Lebesgue, nous en déduisons que

$$\lim_{n\to\infty} \int_{\mathbb{R}} \int_{\mathbb{R}_+} u(a_n - a)\, dt\, dx = 0.$$

Ainsi,

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u\psi\, dt\, dx = 0 \quad \forall \psi \in \mathcal{C}_c^\infty([0,T[\times\mathbb{R})$$

et, grâce au lemme 4, $u = 0$ presque partout. □

Lemme 5
Soit $t_0 \in \mathbb{R}$. Soit $X : [0,T[\times[0,T[\times\mathbb{R}$ définie par

$$\begin{cases} \partial_1 X(t,t_0,x) = a(t, X(t,t_0,x)) \\ X(t_0,t_0,x) = x \end{cases}$$

où a est continue par rapport à ses deux variables, globalement lipschitzienne par rapport à sa seconde variable, et vérifie : $\exists C \in \mathbb{R}$ tel que

$$a(t,y) - a(t,x) \leq C(y-x) \quad \forall y \geq x, \forall t \geq t_0.$$

$X(t,t_0,x)$ vérifie

$$|X(t,t_0,y) - X(t,t_0,x)| \leq e^{C(t-t_0)}\, |y-x| \quad \forall t \geq t_0.$$

□

Ce qui est remarquable dans ce lemme est que l'estimation ne dépend pas de la constante de Lipschitz de a mais seulement de sa constante de Lipschitz à droite, C.

Démonstration
Soit t_0, x et y fixés. Posons

$$D(t) = |X(t,t_0,y) - X(t,t_0,x)|\,.$$

On a

$$\begin{aligned} \frac{1}{2}\partial_t\left(D^2\right)(t) &= \left(X(t,t_0,y) - X(t,t_0,x)\right)\partial_1\left(X(t,t_0,y) - X(t,t_0,x)\right) \\ &= \left(X(t,t_0,y) - X(t,t_0,x)\right)\left(a(t,X(t,t_0,y)) - a(X(t,t_0,x))\right) \\ &\leq C\left(X(t,t_0,y) - X(t,t_0,x)\right)^2 = CD^2(t) \end{aligned}$$

donc $\partial_t\left(D^2\right)(t) \leq 2CD^2(t)$. Par application du lemme de Gronwall, si $t \geq t_0$, nous avons donc $D^2(t) \leq D^2(0)e^{2C(t-t_0)}$, soit

$$D(t) \leq D(0)e^{C(t-t_0)}.$$

□

Une conséquence immédiate et très importante du théorème 9 est qu'il existe au plus une solution faible $u \in L^\infty([0,T[\times\mathbb{R})$ satisfaisant à une inégalité d'Oleinik au problème (3.12) avec donnée initiale $u^0 \in L^\infty(\mathbb{R})$. En effet, supposons qu'il existe 2 telles solutions distinctes avec même donnée initiale, u et v. Alors, la fonction $u-v$ vérifie (au sens faible) l'EDP

$$\partial_t(u-v) + \partial_x\left(\frac{u^2-v^2}{2}\right) = 0,$$

que l'on peut écrire aussi

$$\partial_t(u-v) + \partial_x\left(\frac{u+v}{2}(u-v)\right) = 0.$$

Puisque u et v sont dans L^∞ et vérifient chacune une inégalité d'Oleinik, il en est de même de $\frac{u+v}{2}$ qui joue ici le rôle d'une vitesse de transport. Le théorème 9 permet d'affirmer que $u = v$ presque partout.

Ce résultat se généralise au cas d'une équation scalaire à flux convexe.

Théorème 10

On considère le problème

$$\begin{cases} \partial_t u + \partial_x f(u) = 0 \\ u(0,x) = u^0(x) \end{cases}$$

où le flux f est convexe. Il a au plus une solution $u \in L^\infty([0,T[\times\mathbb{R})$ satisfaisant à une inégalité d'Oleinik. □

La démonstration de ce résultat est laissée en exercice. Néanmoins, quelques indications figurent dans le sujet de partiel d'avril 2004 (et son corrigé contient une démonstration détaillée).

Remarque 39

On peut se convaincre facilement que les deux hypothèses faites dans l'énoncé de ce théorème ($u \in L^\infty$ et vérifie une inégalité d'Oleinik) sont nécessaires. Considérons en effet le problème de Burgers avec donnée initiale nulle, dont la solution triviale (et naturelle) est $u(t,x) = 0$ pour tout t et tout x. On peut facilement montrer alors que

$$u(t,x) = \begin{cases} 0 \text{ si } x < -t/2 \\ -1 \text{ si } x \in [-t/2, 0[\\ 1 \text{ si } x \in [0, t/2] \\ 0 \text{ si } x > t/2 \end{cases} \qquad \forall t > 0$$

est une autre solution faible du problème (cette fonction est constante par morceaux et toutes ses discontinuités vérifient la relation de Rankine-Hugoniot). Celle-ci ne vérifie aucune inégalité d'Oleinik.

Par ailleurs, une autre solution faible du problème est donnée par

$$u(t,x) = \begin{cases} 0 \text{ si } x < -\sqrt{t} \\ x/t \text{ si } x \in [-\sqrt{t}, \sqrt{t}] \\ 0 \text{ si } x > \sqrt{t} \end{cases} \qquad \forall t > 0.$$

Celle-ci vérifie une inégalité d'Oleinik (avec $C(t) = 1/t$) mais n'est pas dans L^∞ (cf. partiel du 19 avril 2005). Les deux solutions faibles exhibées ici sont tracées pour différents temps sur la figure 3.5.

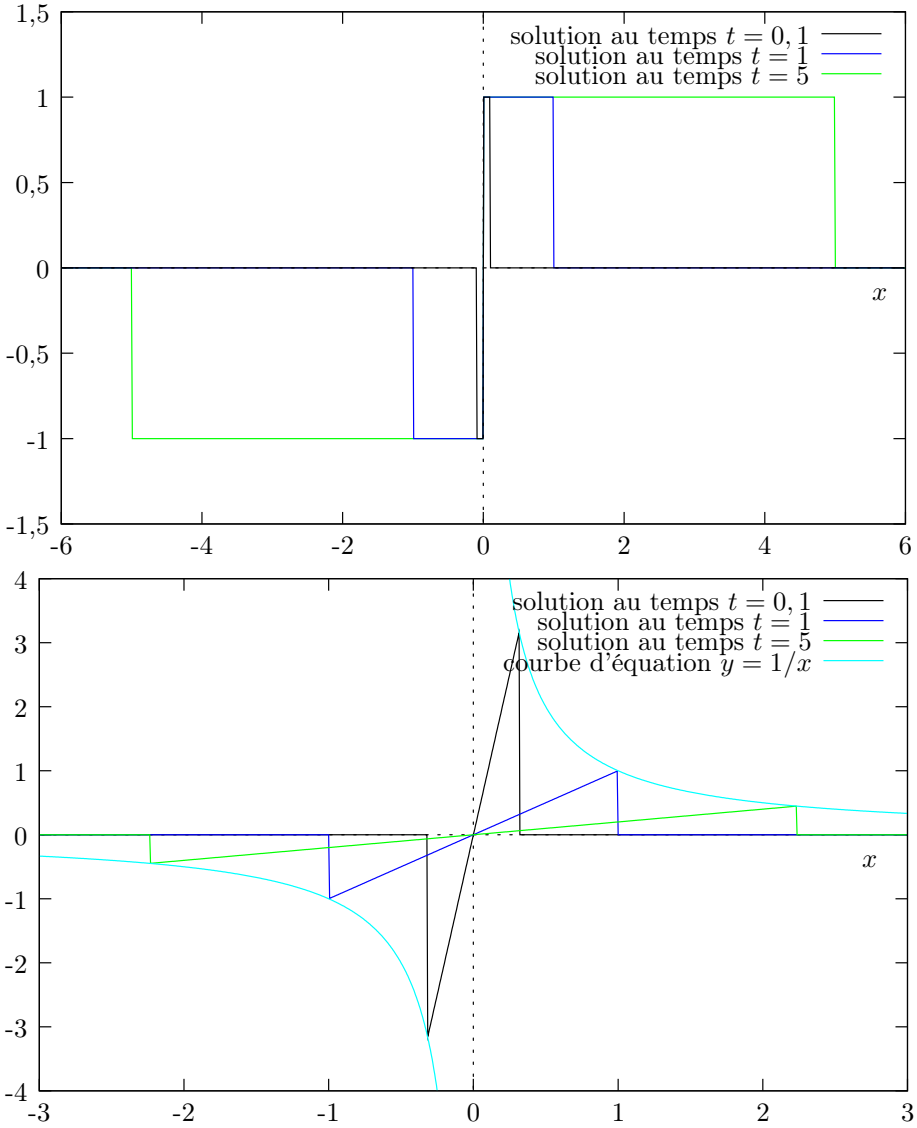

FIGURE 3.5 – Deux solutions non admissibles de l'équation de Burgers avec donnée initiale nulle.

□

Nous pouvons maintenant nous attaquer à l'*existence* d'une solution. Nous proposons ici une démonstration d'existence pour le problème de Burgers. Une autre démonstration de l'existence sera vue ensuite, dans la section concernant l'approximation numérique, pour un problème aux dérivées partielles avec flux convexe quelconque.

Théorème 11
Soit $u^0 \in L^\infty(\mathbb{R})$. Il existe une unique solution $u \in L^\infty(\mathbb{R}_+ \times \mathbb{R})$ de (3.12) avec donnée initiale u^0 satisfaisant à une inégalité d'Oleinik. Cette solution est la limite presque partout, lorsque κ tend vers 0, des fonctions $(u_\kappa)_{\kappa \in \mathbb{R}_+^*}$ solutions de

$$\partial_t u_\kappa + \partial_x \frac{u_\kappa^2}{2} = \kappa \partial_{x,x}^2 u_\kappa.$$

De plus, l'inégalité d'Oleinik à laquelle satisfait la solution est

$$u(t,y) - u(t,x) \leq \frac{1}{t}(y-x) \quad y \geq x, t > 0.$$

□

Démonstration
On s'intéresse donc à l'*équation de Burgers visqueuse*

$$\partial_t u_\kappa + \partial_x \frac{u_\kappa^2}{2} = \kappa \partial_{x,x}^2 u_\kappa \tag{3.14}$$

avec $\kappa > 0$, dont on cherche une solution forte dans $\mathbb{R}_+ \times \mathbb{R}$. Pour exhiber une solution de cette équation, nous allons d'abord transformer celle-ci pour la mettre sous une forme connue : l'équation de la chaleur. Cette transformation est due à Hopf. La solution u_κ vérifie

$$\partial_t u_\kappa = \partial_x \left(\kappa \partial_x u_\kappa - \frac{1}{2} u_\kappa^2 \right).$$

Soit v_κ une primitive en espace de u_κ : par exemple,

$$v_\kappa(t,x) = \int_0^x u_\kappa(t,y)\, dy.$$

Posons

$$\varphi_\kappa = e^{-\frac{1}{2\kappa} v_\kappa}.$$

Nous avons alors

$$\partial_x \varphi_\kappa = -\frac{1}{2\kappa} \varphi_\kappa u_\kappa,$$
$$\partial_{x,x}^2 \varphi_\kappa = \frac{1}{4\kappa^2} \varphi_\kappa u_\kappa^2 - \frac{1}{2\kappa} \varphi_\kappa \partial_x u_\kappa,$$

d'où $u_\kappa = -2\kappa \frac{\partial_x \varphi_\kappa}{\varphi_\kappa}$ et $\kappa \partial_x u_\kappa - \frac{1}{2} u_\kappa^2 = -2\kappa^2 \frac{\partial_{x,x}^2 \varphi_\kappa}{\varphi_\kappa}$. Donc

$$\partial_t \left(\frac{\partial_x \varphi_\kappa}{\varphi_\kappa} \right) = \kappa \partial_x \left(\frac{\partial_{x,x}^2 \varphi_\kappa}{\varphi_\kappa} \right).$$

Or

$$\partial_t \left(\frac{\partial_x \varphi_\kappa}{\varphi_\kappa} \right) = \frac{\varphi_\kappa \partial_{x,t}^2 \varphi_\kappa - \partial_t \varphi_\kappa \partial_x \varphi_\kappa}{\varphi_\kappa^2} = \partial_x \left(\frac{\partial_t \varphi_\kappa}{\varphi_\kappa} \right),$$

donc

$$\partial_x \left(\frac{\partial_t \varphi_\kappa}{\varphi_\kappa} - \kappa \frac{\partial^2_{x,x} \varphi_\kappa}{\varphi_\kappa} \right) = 0.$$

Finalement,

$$\frac{\partial_t \varphi_\kappa}{\varphi_\kappa} - \kappa \frac{\partial^2_{x,x} \varphi_\kappa}{\varphi_\kappa} = \gamma(t)$$

où γ est une constante d'intégration. Posons [19] maintenant

$$\psi_\kappa(t,x) = \varphi_\kappa(t,x) e^{-\int_0^t \gamma(s)\, ds}.$$

Cette nouvelle fonction vérifie

$$\partial_t \psi_\kappa(t,x) = \partial_t \varphi_\kappa(t,x) e^{-\int_0^t \gamma(s)\, ds} - \varphi_\kappa(t,x) e^{-\int_0^t \gamma(s)\, ds} \gamma(t)$$

et

$$\partial^2_{x,x} \psi_\kappa(t,x) = \partial^2_{x,x} \varphi_\kappa(t,x) e^{-\int_0^t \gamma(s)\, ds}.$$

Donc

$$\frac{\partial_t \psi_\kappa}{\psi_\kappa} - \kappa \frac{\partial^2_{x,x} \psi_\kappa}{\psi_\kappa} = \frac{\partial_t \varphi_\kappa}{\varphi_\kappa} - \gamma(t) - \kappa \frac{\partial^2_{x,x} \varphi_\kappa}{\varphi_\kappa} = 0.$$

ψ_κ est donc solution de l'équation de la chaleur

$$\partial_t \psi_\kappa - \kappa \partial^2_{x,x} \psi_\kappa = 0$$

dans $\mathbb{R}_+ \times \mathbb{R}$. La différence avec l'EDP du chapitre 2 est qu'elle est ici posée en domaine non borné (et sans conditions aux limites en espace). Le lemme suivant donne une solution du problème.

Lemme 6

Soit $\psi^0 \in L^\infty(\mathbb{R})$. Posons

$$\psi(t,x) = \int_{\mathbb{R}} \frac{1}{\sqrt{4\kappa\pi t}} e^{-\frac{(x-y)^2}{4\kappa t}} \psi^0(y)\, dy$$

pour tout $(t,x) \in \mathbb{R}_+^* \times \mathbb{R}$. On a

— $\psi(t,x) \in \mathcal{C}^\infty(]0,+\infty[\times\mathbb{R})$;

— $\partial_t \psi(t,x) = \kappa \partial^2_{x,x} \psi(t,x)$ pour tout $(t,x) \in]0,+\infty[\times\mathbb{R}$;

— $\lim_{t\to 0^+} \psi(t,x) = \psi^0(x)$ presque partout.

De plus, ψ est l'unique solution de $\partial_t \psi = \kappa \partial^2_{x,x} \psi$ avec donnée initiale ψ^0. □

Remarque 40

La fonction

$$N(t,x,y) = \frac{1}{(4\kappa\pi t)^{d/2}} e^{-\frac{|x-y|^2}{4\kappa t}}$$

19. Cela revient à changer la constante d'intégration dans la définition de la primitive v_κ.

est appelée *noyau de la chaleur dans* $\mathbb{R}^d$. Ce noyau vérifie

$$\begin{aligned}\partial_t N &= \kappa \partial^2_{x,x} N,\\ N(0,\cdot,y) &= \delta_y(\cdot)\end{aligned}$$

où δ_y est la masse de Dirac de masse 1 au point y. C'est la *solution élémentaire* de l'équation de la chaleur. □

La démonstration de ce lemme est laissée en exercice. Elle consiste à vérifier que

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} N\left(\partial_t\varphi + \kappa\partial^2_{x,x}\varphi\right)\, dt\, dx = \varphi(0,y)$$

pour tout $\varphi \in \mathcal{C}^\infty_c(\mathbb{R}_+ \times \mathbb{R})$. Ce calcul est d'ailleurs fait (dans un cadre légèrement différent) aux pages suivantes.
Revenons à notre démoutonstration. Soit donc

$$\psi_\kappa(t,x) = \frac{1}{\sqrt{4\kappa\pi t}}\int_{\mathbb{R}} \psi(0,y) e^{-\frac{(x-y)^2}{4\kappa t}}\, dy.$$

Une solution de (3.14) est alors donnée par

$$u_\kappa(t,x) = -2\kappa\frac{\partial_x\psi_\kappa(t,x)}{\psi_\kappa(t,x)}.$$

Or

$$\partial_x\psi_\kappa(t,x) = \frac{1}{\sqrt{4\kappa\pi t}}\int_{\mathbb{R}} -\frac{2(x-y)}{4\kappa t}\psi(0,y)e^{-\frac{(x-y)^2}{4\kappa t}}\, dy,$$

donc

$$-2\kappa\frac{\partial_x\psi_\kappa(t,x)}{\psi_\kappa(t,x)} = \frac{1}{t}\frac{\displaystyle\int_{\mathbb{R}}(x-y)\psi(0,y)e^{-\frac{(x-y)^2}{4\kappa t}}\, dy}{\displaystyle\int_{\mathbb{R}}\psi(0,y)e^{-\frac{(x-y)^2}{4\kappa t}}\, dy}.$$

D'autre part, on exprime $\psi(0,y)$ en fonction de u^0 :

$$\psi(0,y) = e^{-\frac{1}{2\kappa}\int_0^y u^0(x)\,dx} e^{-\int_0^0 \gamma(s)\,ds} = e^{-\frac{1}{2\kappa}\int_0^y u^0(x)\,dx}.$$

Ainsi

$$u_\kappa(t,x) = \frac{\displaystyle\int_{\mathbb{R}} \frac{x-y}{t} e^{-\left[\frac{(x-y)^2}{4\kappa t} + \frac{1}{2\kappa}\int_0^y u^0(z)\,dz\right]}\, dy}{\displaystyle\int_{\mathbb{R}} e^{-\left[\frac{(x-y)^2}{4\kappa t} + \frac{1}{2\kappa}\int_0^y u^0(z)\,dz\right]}\, dy},$$

que l'on réécrit

$$u_\kappa(t,x) = \frac{\displaystyle\int_{\mathbb{R}} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy}{\displaystyle\int_{\mathbb{R}} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy}$$

en définissant F par

$$F(t,x,y) = \frac{(x-y)^2}{2t} + \int_0^y u^0(z)\,dz.$$

La fin de cette démonstration consiste à montrer que u_κ a une limite [20] lorsque κ tend vers 0 et que cette limite est solution de l'équation de Burgers non visqueuse (3.12) vérifiant une inégalité d'Oleinik. Ceci est un peu technique.
Pour tout $(t,x) \in \mathbb{R}_+^* \times \mathbb{R}$, notons $y_-(t,x)$ et $y_+(t,x)$ le y minimal et le y maximal tels que $F(t,x,y) = \inf_{y\in\mathbb{R}} F(t,x,y)$:

$$\begin{aligned} y_-(t,x) &= \min_{z\in\mathbb{R}} \{z \text{ t. q. } F(t,x,z) = \min_{y\in\mathbb{R}} F(t,x,y)\} \\ y_+(t,x) &= \max_{z\in\mathbb{R}} \{z \text{ t. q. } F(t,x,z) = \min_{y\in\mathbb{R}} F(t,x,y)\} \end{aligned}$$

Ces réels existent car $F(t,x,y)$ est continue par rapport à y et $\lim_{|y|\to+\infty} F(t,x,y) = +\infty\ \forall (t,x)$. Posons encore, pour tout (t,x), $m(t,x) = \min_{y\in\mathbb{R}} F(t,x,y)$. On a avec ces définitions

$$F(t,x,y) > m(t,x) \quad \forall y \notin [y_-(t,x), y_+(t,x)].$$

Les 3 lemmes suivants vont nous permettre de conclure.

Lemme 7
Avec les notations introduites ci-dessus, on a

$$\frac{x-y_+(t,x)}{t} \le \liminf_{\kappa\to 0^+} u_\kappa(t,x) \le \limsup_{\kappa\to 0^+} u_\kappa(t,x) \le \frac{x-y_-(t,x)}{t}$$

pour tout $(t,x) \in \mathbb{R}_+^* \times \mathbb{R}$. □

Lemme 8
Soit $t > 0$. Avec les notations introduites ci-dessus, si $x_1 < x_2$, $y_+(t,x_1) \le y_-(t,x_2)$. $y_-(t,\cdot)$ est continue à gauche et $y_+(t,\cdot)$ est continue à droite, $\forall t \in \mathbb{R}_+$. □

Lemme 9
Avec les notations introduites ci-dessus, pour tout $t \in \mathbb{R}_+^*$, $y_-(t,x) = y_+(t,x)$ presque partout et $y_-(t,\cdot)$ et $y_+(t,\cdot)$ sont continues presque partout. □

La démonstration de ces trois lemmes est repoussée à la fin de la démonstration du théorème, que nous poursuivons maintenant.

Pour tout $(t,x) \in \mathbb{R}_+^* \times \mathbb{R}$, posons

$$u(t,x) = \frac{x-y_+(t,x)}{t}.$$

20. Noter la similitude avec le résultat du lemme 6 concernant la donnée initiale : on fait ici tendre κ vers 0 au lieu de t dans ce lemme.

D'après le lemme 9, u est continue presque partout et

$$u(t,x) = \frac{x - y_-(t,x)}{t}$$

pour presque tout (t,x). D'après le lemme 7,

$$\lim_{\kappa \to 0^+} u_\kappa(t,x) = u(t,x)$$

presque partout. La fonction u_κ vérifie

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u_\kappa \partial_t \varphi + \frac{u_\kappa^2}{2} \partial_x \varphi + \kappa u_\kappa \partial^2_{x,x} \varphi \, dt \, dx = \int_{\mathbb{R}} u^0 \varphi(0,\cdot) \, dx$$

pour tout $\varphi \in \mathcal{C}_c^\infty([0,+\infty[\times\mathbb{R})$. D'après le théorème de convergence dominée de Lebesgue, u vérifie

$$\int_{\mathbb{R}} \int_{\mathbb{R}_+} u \partial_t \varphi + \frac{u^2}{2} \partial_x \varphi \, dt \, dx = \int_{\mathbb{R}} u^0 \varphi(0,\cdot) \, dx$$

pour tout $\varphi \in \mathcal{C}_c^\infty([0,+\infty[\times\mathbb{R})$. C'est donc une solution faible de (3.12). Il nous reste à montrer que u vérifie une inégalité d'Oleinik.

Soit $t > 0$. Soit x_1, x_2 tels que $x_2 \geq x_1$. On a

$$u(t,x_2) - u(t,x_1) = \frac{x_2 - y_+(t,x_2)}{t} - \frac{x_1 - y_+(t,x_1)}{t} = \frac{x_2 - x_1}{t} - \frac{y_+(t,x_2) - y_+(t,x_1)}{t}.$$

Or $y_+(t,\cdot)$ est croissante car si $x_2 > x_1$, $y_+(t,x_1) \leq y_-(t,x_2)$ d'après le lemme 8 et $y_-(t,x_2) \leq y_+(t,x_2)$. Donc

$$u(t,x_2) - u(t,x_1) \leq \frac{x_2 - x_1}{t} \quad \forall x_2 \geq x_1.$$

Ceci termine la démonstration. □

Démonstration (lemme 7)

On a posé

$$F(t,x,y) = \frac{(x-y)^2}{2t} + \int_0^y u^0(z) \, dz.$$

Puisque $u^0 \in L^\infty(\mathbb{R})$, $\exists Y \in \mathbb{R}$ tel que

$$F(t,x,y) \geq \frac{(x-y)^2}{3t} \quad \forall y \text{ t. q. } |y| \geq Y.$$

Soit $\varepsilon > 0$. Si $y \notin [y_-(t,x) - \varepsilon, y_+(t,x) + \varepsilon]$, $\exists \eta > 0$ tel que $F(t,x,y) > m(t,x) + \eta$. Soit Z tel que $Z \geq Y$, $Z \geq |x|$ et $Z \geq \max(y_-(t,x) - \varepsilon, y_+(t,x) + \varepsilon)$. Observons le numérateur dans l'expression de u_κ.

$$\begin{aligned}
\int_{-\infty}^{\infty} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}} \, dy = &\int_{-\infty}^{-Z} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}} \, dy \\
&+ \int_{-Z}^{y_-(t,x)-\varepsilon} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}} \, dy + \int_{y_-(t,x)-\varepsilon}^{y_+(t,x)+\varepsilon} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}} \, dy \\
&+ \int_{y_+(t,x)+\varepsilon}^{Z} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}} \, dy + \int_{Z}^{\infty} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}} \, dy
\end{aligned}$$

Nous allons montrer que le terme $\int_{y_-(t,x)-\varepsilon}^{y_+(t,x)+\varepsilon} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy$ est dominant lorsque $\kappa \to 0^+$.
Sur l'intervalle $]-\infty, -Z]$, on a $F(t,x,y) \geq \frac{(x-y)^2}{3t}$, d'où

$$\int_{-\infty}^{-Z} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy \leq \int_{-\infty}^{-Z} \frac{|x-y|}{t} e^{-\frac{(x-y)^2}{6\kappa t}}\, dy = \int_{-\infty}^{-Z} \frac{x-y}{t} e^{-\frac{(x-y)^2}{6\kappa t}}\, dy \text{ car } x-y \geq 0.$$

Donc

$$\int_{-\infty}^{-Z} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy \leq \left[3\kappa e^{-\frac{(x-y)^2}{6\kappa t}} \right]_{y=-\infty}^{-Z} = 3\kappa e^{-\frac{(x-y)^2}{6\kappa t}}.$$

Sur l'intervalle $]-Z, y_-(t,x)-\varepsilon]$, $F(t,x,y) > m(t,x)+\eta$ (avec $\eta > 0$). Donc

$$\begin{aligned}\int_{-Z}^{y_-(t,x)-\varepsilon} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy &\leq \int_{-Z}^{y_-(t,x)-\varepsilon} \frac{|x-y|}{t} e^{-\frac{m(t,x)+\eta}{2\kappa}}\, dy \\ &\leq (y_-(t,x)+Z)\, \frac{\max\left(|x+Z|, |x-y_-(t,x)+\varepsilon|\right)}{t} e^{-\frac{m(t,x)}{2\kappa}} e^{-\frac{\eta}{2\kappa}}.\end{aligned}$$

On obtient bien entendu des majorations similaires pour les termes

$$\int_{y_+(t,x)+\varepsilon}^{Z} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy \text{ et } \int_{Z}^{\infty} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy.$$

Observons maintenant le dénominateur dans l'expression de u_κ. On peut découper l'intégrale en 4 intégrales avec les mêmes bornes d'intégration que pour le numérateur et on en déduit, avec les mêmes raisonnements, que

$$\int_{-\infty}^{\infty} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy \approx_{\kappa\to 0^+} \int_{y_-(t,x)-\varepsilon}^{y_+(t,x)+\varepsilon} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy \geq \delta e^{-\frac{m(t,x)}{2\kappa}} e^{-\frac{\eta}{4\kappa}}$$

où δ est la longueur d'un intervalle sur lequel $F(t,x,y) \leq m(t,x)+\frac{\eta}{2}$; $\delta > 0$ car F est continue. Ainsi

$$u_\kappa(t,x) \approx_{\kappa\to 0^+} \frac{\displaystyle\int_{y_-(t,x)-\varepsilon}^{y_+(t,x)+\varepsilon} \frac{x-y}{t} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy}{\displaystyle\int_{y_-(t,x)-\varepsilon}^{y_+(t,x)+\varepsilon} e^{-\frac{F(t,x,y)}{2\kappa}}\, dy}$$

En conclusion, on a bien

$$\frac{x-y_+(t,x)}{t} \leq \liminf_{\kappa\to 0^+} u_\kappa(t,x) \leq \limsup_{\kappa\to 0^+} u_\kappa(t,x) \leq \frac{x-y_-(t,x)}{t}$$

□

Démonstration (lemme 8)
Soit $(t,x_1) \in \mathbb{R}_+ \times \mathbb{R}$. Pour simplifier les notations, posons $y_+ = y_+(t,x_1)$. Soit $y < y_+$ et $x_2 > x_1$.

$$F(t,x_2,y) - F(t,x_2,y_+) \geq F(t,x_2,y) - F(t,x_1,y) + F(t,x_1,y_+) - F(t,x_2,y_+)$$

car $F(t,x_1,y) \geq F(t,x_1,y_+)\ \forall y \in \mathbb{R}$, et donc

$$\begin{aligned} F(t,x_2,y) - F(t,x_2,y_+) &\geq \frac{(x_2-y)^2}{2t} - \frac{(x_1-y)^2}{2t} + \frac{(x_1-y_+)^2}{2t} - \frac{(x_2-y_+)^2}{2t} \\ &= \frac{(y-y_+)(x_1-x_2)}{t} > 0. \end{aligned}$$

Donc $F(t,x_2,y) > F(t,x_2,y_+)\ \forall y < y_+$. La première affirmation du lemme,

$$y_-(t,x_2) \geq y_+(t,x_1) \quad \forall x_2 > x_1,$$

est donc démontrée. Pour $\varepsilon > 0$, on a donc en particulier $y_-(t,x_1-\varepsilon) \leq y_-(t,x_1)$. D'autre part, le minimum de F, $m(t,x)$, est une fonction continue de (t,x), donc $\lim_{\varepsilon\to 0^+} m(t,x_1-\varepsilon) = m(t,x_1)$, d'où $\lim_{\varepsilon\to 0^+} y_-(t,x_1-\varepsilon) = y_-(t,x_1)$. Supposons en effet que ce soit faux : il existe une suite réelle $(\varepsilon_n)_{n\in\mathbb{N}}$ telle que $\varepsilon_n > 0\ \forall n$ et $\lim_{n\to+\infty} \varepsilon_n = 0$ et $\lim_{n\to+\infty} y_-(t,x_1-\varepsilon_n) = y_0 < y_-(t,x_1)$. On a alors

$$F(t,x_1-\varepsilon_n, y_-(t,x_1-\varepsilon_n)) = m(t,x_1-\varepsilon_n).$$

Or $\lim_{n\to+\infty} F(t,x_1-\varepsilon_n,y_-(t,x_1-\varepsilon_n)) = F(t,x_1,y_0) < m(t,x_1)$ tandis que $\lim_{n\to+\infty} m(t,x_1-\varepsilon_n) = m(t,x_1)$. C'est une contradiction. La fonction $y_-(t,\cdot)$ est donc continue à gauche. On montre de la même façon que $y_+(t,\cdot)$ l'est à droite. □

Démonstration (lemme 9)
Soit a et b des réels tels que $b \geq a$. Puisque $y_+(t,x) \geq y_-(t,x)\ \forall (t,x) \in \mathbb{R}_+ \times \mathbb{R}$,

$$\int_a^b y_+(t,x) - y_-(t,x)\,dx \geq 0.$$

D'autre part, pour tout $\varepsilon > 0$,

$$\begin{aligned} \int_a^b y_+(t,x) - y_-(t,x)\,dx &= \int_{a-\varepsilon}^{b-\varepsilon} y_+(t,x+\varepsilon) - y_-(t,x+\varepsilon)\,dx \\ &= \int_{a-\varepsilon}^{b-\varepsilon} y_+(t,x+\varepsilon) - y_+(t,x)\,dx + \int_{a-\varepsilon}^{b-\varepsilon} y_+(t,x) - y_-(t,x+\varepsilon)\,dx \end{aligned}$$

Or d'après le lemme 8, $\int_{a-\varepsilon}^{b-\varepsilon} y_+(t,x) - y_-(t,x+\varepsilon)\,dx \leq 0$. Par ailleurs,

$$\int_{a-\varepsilon}^{b-\varepsilon} y_+(t,x+\varepsilon) - y_+(t,x)\,dx = \int_{\mathbb{R}} \mathbb{1}_{[a-\varepsilon,b-\varepsilon]}(x)\,(y_+(t,x+\varepsilon) - y_+(t,x))\ dx$$

et

— $\lim_{\varepsilon\to 0} \mathbb{1}_{[a-\varepsilon,b-\varepsilon]}(x) = \mathbb{1}_{[a,b]}(x)$ presque partout ;
— $\lim_{\varepsilon\to 0^+} y_+(t,x+\varepsilon) - y_+(t,x) = 0$ d'après le lemme 8 ;

— le produit de ces deux fonctions est borné par une fonction intégrable, car $y_+ \in L^\infty$ et ce produit est à support compact.

Par application du théorème de convergence dominée de Lebesgue,

$$\lim_{\varepsilon \to 0^+} \int_{a-\varepsilon}^{b-\varepsilon} y_+(t, x+\varepsilon) - y_+(t,x)\, dx = 0.$$

Donc

$$\int_a^b y_+(t,x) - y_-(t,x)\, dx \leq 0.$$

Finalement, $y_+(t,x) = y_-(t,x)$ presque partout sur $[a,b]$ $(\forall [a,b])$. □

Le principe du maximum pour l'équation de Burgers est une conséquence directe du théorème 11.

Exercice 3

Soit u la solution évoquée au théorème 11. Montrer qu'elle vérifie

$$||u(t,\cdot)||_{L^\infty(\mathbb{R})} \leq \left|\left|u^0\right|\right|_{L^\infty(\mathbb{R})} \quad \forall t \in \mathbb{R}_+.$$

3.4 Schémas de volumes finis

Nous voulons calculer une solution approchée[21] de

$$\begin{cases} \partial_t u + \partial_x f(u) = 0 \\ u(0,x) = u^0(x) \end{cases} \tag{3.15}$$

où f est régulière. Nous n'avons montré pour l'instant l'existence d'une solution à ce problème que lorsque $f(u) = au$ et $f(u) = u^2/2$. La présente section va combler une lacune en montrant l'existence d'une solution dans le cas où f est convexe (cas dans lequel on sait déjà que la solution faible satisfaisant à une inégalité d'Oleinik est unique).

3.4.1 Généralités

On se donne un maillage de $\mathbb{R}$:

$$\mathbb{R} = \bigcup_{j \in \mathbb{Z}} [x_{j-1/2}, x_{j+1/2}[$$

avec, pour simplifier, $x_{j+1/2} = (j+1/2)\Delta x$ $\forall j \in \mathbb{Z}$, avec $\Delta x \in \mathbb{R}_+^*$ (il s'agit donc d'un maillage régulier). On se donne de même un maillage de $\mathbb{R}_+$:

$$\mathbb{R}_+ = \bigcup_{n \in \mathbb{N}} [t^n, t^{n+1}[$$

21. Solution faible vérifiant une inégalité d'Oleinik

avec $t^n = n\Delta t\ \forall n \in \mathbb{N}$, avec $\Delta t \in \mathbb{R}_+^*$. Soit u une solution faible de (3.15). Elle vérifie

$$\int_{\mathbb{R}}\int_{\mathbb{R}_+} u\partial_t\varphi + f(u)\partial_x\varphi\, dt\, dx = 0$$

$\forall \varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+^* \times \mathbb{R})$. On en « déduit [22] » que

$$\int_{x_{j-1/2}}^{x_{j+1/2}} u(t^{n+1},x)\, dx - \int_{x_{j-1/2}}^{x_{j+1/2}} u(t^n,x)\, dx + \int_{t^n}^{t^{n+1}} f(u(t,x_{j+1/2}))\, dt - \int_{t^n}^{t^{n+1}} f(u(t,x_{j-1/2}))\, dt = 0$$

pour tout n et tout j. Le principe de base d'un schéma de volumes finis est de calculer des valeurs approchées de

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

pour tout n et tout j. Pour cela, on définit la condition initiale discrète par

$$u_j^0 = \frac{1}{\Delta x}\int_{x_{j-1/2}}^{x_{j+1/2}} u^0(x)\, dx \quad \forall j \in \mathbb{Z} \tag{3.16}$$

et on remplace l'EDP par

$$\Delta x\left(u_j^{n+1} - u_j\right) + \Delta t\left(f_{j+1/2}^n - f_{j-1/2}^n\right) = 0 \quad \forall n \in \mathbb{N}, j \in \mathbb{Z}, \tag{3.17}$$

où les *flux numériques* $f_{j+1/2}^n$ sont des valeurs approchées de $1/\Delta t \int_{t^n}^{t^{n+1}} f(u(t,x_{j+1/2}))\, dt$. Il existe de très nombreuses manières efficaces de définir les flux numériques. Nous n'étudierons que celle du

3.4.2 Schéma de Lax-Friedrichs

Le schéma de Lax-Friedrichs s'écrit de manière condensée

$$u_j^{n+1} = \frac{u_{j-1}^n + u_{j+1}^n}{2} - \frac{\Delta t}{2\Delta x}\left(f(u_{j+1}^n) - f(u_{j-1}^n)\right) \tag{3.18}$$

et l'on vérifie que c'est bien un schéma de la forme (3.17) en écrivant

$$u_j^{n+1} - u_j^n = -\frac{\Delta t}{\Delta x}\left[\frac{f(u_{j+1}^n) + f(u_j^n)}{2} + \left(\frac{u_j^n - u_{j+1}^n}{2}\right)\frac{\Delta x}{\Delta t} - \left(\frac{f(u_j^n) + f(u_{j-1}^n)}{2} + \left(\frac{u_{j-1}^n - u_j^n}{2}\right)\frac{\Delta x}{\Delta t}\right)\right],$$

ce qui revient à poser

$$f_{j+1/2}^n = \frac{1}{2}\left(f(u_{j+1}^n) + f(u_j^n) + \left(u_j^n - u_{j+1}^n\right)\frac{\Delta x}{\Delta t}\right) \quad \forall n \in \mathbb{N}, j \in \mathbb{Z} \tag{3.19}$$

dans (3.17).

22. En prenant… $\varphi = \mathbb{1}_{[t^n,t^{n+1}]\times[x_{j-1/2},x_{j+1/2}]}$ (qui n'est pas une fonction-test).

Remarque 41
Le schéma (3.18) est équivalent à

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

Il est donc, au sens des différences finies, consistant à l'ordre 1 avec l'EDP considérée, et consistant à l'ordre 2 en espace avec l'EDP

$$\partial_t u+\partial_x f(u)=\frac{\Delta x^2}{2\Delta t}\partial_{x,x}^2 u$$

qui est une régularisation parabolique de l'EDP avec un « coefficient de conduction » (ou de diffusion, ou de viscosité) tendant vers 0 lorsque l'on raffine le maillage si par exemple Δt et Δx tendent vers 0 à la même vitesse [23]. C'est par ce type de régularisation que l'on a montré l'existence d'une solution à l'équation de Burgers (par la méthode de Hopf). □

Procédons à l'étude du schéma de Lax-Friedrichs, c'est-à-dire à l'étude de sa convergence [24].

Nous supposerons dans toute la suite que $u^0\in L^\infty(\mathbb{R})$. *Alors* $\left(u_j^0\right)_{j\in\mathbb{Z}}\in l^\infty$. *Nous supposerons aussi que* $f\in\mathcal{C}^2(\mathbb{R})$ *et est convexe.*

Proposition 12
On considère le schéma de Lax-Friedrichs (3.18). Sous la condition, dite de Courant-Friedrichs-Lewy [25],

$$\sup_{j\in\mathbb{Z}}\left|f'(u_j^0)\right|\frac{\Delta t}{\Delta x}\le 1,$$

il est tel que

$$\inf_{j\in\mathbb{Z}}u_j^0\le u_j^n\le\sup_{j\in\mathbb{Z}}u_j^0$$

pour tout $j\in\mathbb{Z}$ et tout $n\in\mathbb{N}$. □

C'est une propriété de stabilité l^∞ ou encore un principe du maximum discret.

Démonstration
Encore une fois, l'idée est de montrer que sous la condition de CFL u_j^{n+1} est une combinaison convexe des u_k^n, $k\in\mathbb{Z}$. L'algorithme est

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

23. Une étude précise du comportement de la solution numérique en fonction des vitesses relatives de convergence vers 0 de Δt et Δx est faite dans l'examen du 18 mai 2006.

24. Noter que nous ne sommes pas assurés de l'existence d'une solution, donc que l'étude de la convergence ne peut être basée sur une estimation de l'*erreur*.

25. CFL dans la suite.

Puisque f est de classe $\mathcal{C}^2$ et est convexe, $\exists v \in [u_{j-1}^n, u_{j+1}^n]$ (ou $[u_{j+1}^n, u_{j-1}^n]$) tel que

$$f(u_{j+1}^n) = f(u_{j-1}^n) + \left(u_{j+1}^n - u_{j-1}^n\right) f'(u_{j-1}^n) + \frac{\left(u_{j+1}^n - u_{j-1}^n\right)^2}{2} f''(v),$$

donc, puisque f est convexe, on a

$$f(u_{j+1}^n) \geq f(u_{j-1}^n) + \left(u_{j+1}^n - u_{j-1}^n\right) f'(u_{j-1}^n),$$

d'où

$$u_j^{n+1} \leq u_{j-1}^n \left(\frac{1}{2} + \frac{\Delta t}{2\Delta x} f'(u_{j-1}^n)\right) + u_{j+1}^n \left(\frac{1}{2} - \frac{\Delta t}{2\Delta x} f'(u_{j-1}^n)\right).$$

Sous la condition $\sup_{j\in\mathbb{Z}} \left|f'(u_j^n)\right| \Delta t/\Delta x \leq 1$, chacun des termes multiplicatifs de u_{j-1}^n et u_{j+1}^n est positif, et leur somme vaut 1, donc u_j^{n+1} est majoré par une combinaison convexe des $\left(u_j^n\right)_{j\in\mathbb{Z}}$, donc $u_j^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n$. La minoration $u_j^{n+1} \geq \inf_{j\in\mathbb{Z}} u_j^n$ s'obtient de manière analogue. Il reste à vérifier que la condition $\sup_{j\in\mathbb{Z}} \left|f'(u_j^0)\right| \Delta t/\Delta x \leq 1$ est propagée en temps, c'est-à-dire qu'elle assure $\sup_{j\in\mathbb{Z}} \left|f'(u_j^n)\right| \Delta t/\Delta x \leq 1$ pour tout n. Cela est encore une conséquence de la convexité de f :

$$\sup_{j\in\mathbb{Z}} \left|f'(u_j^{n+1})\right| = \max\left(\left|f'(\sup_{j\in\mathbb{Z}} u_j^{n+1})\right|, \left|f'(\inf_{j\in\mathbb{Z}} u_j^{n+1})\right|\right)$$

car f' est croissante. Or si $\sup_{j\in\mathbb{Z}} \left|f'(u_j^n)\right| \Delta t/\Delta x \leq 1$,

$$\inf_{j\in\mathbb{Z}} u_j^n \leq \inf_{j\in\mathbb{Z}} u_j^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^{n+1} \leq \sup_{j\in\mathbb{Z}} u_j^n$$

(d'après la première partie de la démonstration) et ainsi $\sup_{j\in\mathbb{Z}} \left|f'(u_j^{n+1})\right| \leq \sup_{j\in\mathbb{Z}} \left|f'(u_j^n)\right|$ car f' est croissante. On a donc, par récurrence, le résultat voulu. □

Il nous faut montrer que la solution numérique « converge » vers une solution de (3.15) lorsque Δt et Δx tendent vers 0. Pour cela, on définit une fonction associée à la solution numérique :

$$u_{\Delta t,\Delta x} = \sum_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} u_j^n \mathbb{1}_{[n\Delta t,(n+1)\Delta t[}(t) \mathbb{1}_{[(j-1/2)\Delta x,(j+1/2)\Delta x[}(x) \tag{3.20}$$

pour $(t,x) \in \mathbb{R}_+ \times \mathbb{R}$, où $\left(u_j^n\right)_{(n,j)\in\mathbb{N}\times\mathbb{Z}}$ est donnée par (3.16,3.18). On veut montrer que

$$\lim_{\Delta t,\Delta x\to 0} u_{\Delta t,\Delta x} = u$$

en un certain sens, où u est solution de (3.15). Mais nous ne savons pas si ce problème a une solution. Deux possibilités se présentent pour obtenir un tel résultat :

— soit montrer que $(u_{\Delta t_k,\Delta x_k})$ est une suite de Cauchy pour une suite de pas $(\Delta t_k, \Delta x_k)$ tendant vers 0 ;

— soit montrer que cette suite reste dans un compact.

La première solution semble plus difficile car elle implique une comparaison de solutions numériques calculées avec des maillages différents [26]. Nous allons utiliser une méthode de compacité [27]. Cela nous amène à introduire quelques définitions et résultats que nous admettrons.

Fonctions à variation bornée

Soit Ω un ouvert de $\mathbb{R}^d$. Soit $g \in L^1_{loc}(\Omega)$.

Définition 9
On appelle *variation totale de g sur Ω* et on note $TV_\Omega(g)$ l'élément de $\overline{\mathbb{R}}$ défini par

$$TV_\Omega(g) = \sup_{\varphi \in (\mathcal{C}_c^\infty(\Omega))^d \text{ t. q. } ||\varphi||_{L^\infty(\Omega)} \leq 1} \int_\Omega g \, \mathrm{div}\varphi \, dx.$$

□

Exemples

1 Si $g \in \mathcal{C}^1(\Omega)$,

$$TV_\Omega(g) = \int_\Omega \sum_{i=1}^d |\partial_i g| \, dx.$$

Pour le montrer, définir une suite d'approximations régulières à support compact de $-\left(\mathrm{signe}\left(\partial_i g\right)\right)_{i=1}^d$.

2 Si $g(x) = \sum_{j\in\mathbb{Z}} \alpha_i \mathbb{1}_{I_j}(x)$ où les α_i sont des réels et les intervalles $\overline{I_j}$ forment un recouvrement de $\mathbb{R}$,

$$TV_\mathbb{R}(g) = \sum_{j\in\mathbb{Z}} |\alpha_{j+1} - \alpha_j| \, .$$

Noter que $mes(I_j)$ ne joue aucun rôle. L'idée pour montrer ce résultat est suggérée par la figure 3.6.

26. Voir cependant l'examen du 7 mai 2007, qui propose une telle méthode dans le cas simplifié de l'équation de transport linéaire pour le schéma upwind.

27. On peut noter que la stabilité l^∞, résultat déjà démontré, ne nous apporte pas de compacité : les bornés de L^∞ ne sont pas des compacts de L^∞, ni de L^1, etc. Il va falloir travailler un peu plus.

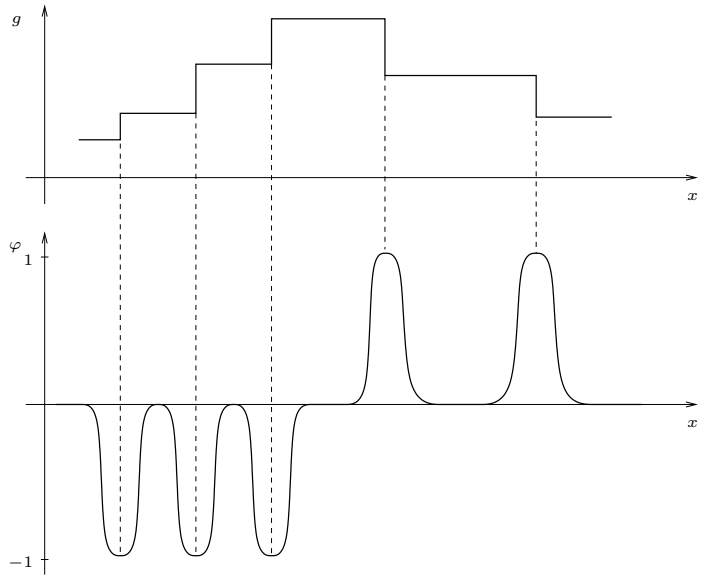

FIGURE 3.6 – Suggestion de définition de la fonction-test φ.

3 Si $u_{\Delta t,\Delta x} \in L^1_{loc}(\mathbb{R}_+ \times \mathbb{R})$ et

$$u_{\Delta t,\Delta x} = \sum_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} u_j^n \mathbb{1}_{[n\Delta t,(n+1)\Delta t[}(t) \mathbb{1}_{[(j-1/2)\Delta x,(j+1/2)\Delta x[}(x),$$

$$TV_{\mathbb{R}_+\times\mathbb{R}}(u_{\Delta t,\Delta x}) = \Delta t \sum_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} \left|u_{j+1}^n - u_j^n\right| + \Delta x \sum_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} \left|u_j^{n+1} - u_j^n\right|.$$

Définition 10
On dit que $g \in L^1_{loc}(\Omega)$ est *à variation bornée sur* Ω si et seulement si $TV_\Omega(g) < +\infty$ et on note $BV(\Omega)$ l'ensemble des fonctions à variation bornée sur Ω :

$$BV(\Omega) = \left\{g \in L^1_{loc}(\Omega) \text{ t. q. } TV_\Omega(g) < +\infty\right\}.$$

□

Voici quelques résultats utiles pour la suite.

Théorème 12
$L^1(\Omega) \cap BV(\Omega)$ est un espace de Banach pour la norme définie par

$$||\cdot||_{L^1\cap BV(\Omega)} = ||\cdot||_{L^1(\Omega)} + TV_\Omega(\cdot).$$

□

Proposition 13
Soit $g \in L^1_{loc}(\mathbb{R})$.

$$TV_{\mathbb{R}}(g) = \sup_{\Delta x>0} \int_{\mathbb{R}} \frac{|g(x+\Delta x) - g(x)|}{\Delta x}\, dx.$$

□

Théorème 13 (Helly)
Supposons que Ω est un ouvert borné de $\mathbb{R}^d$ à frontière lipschitzienne. Alors l'injection de $L^1 \cap BV(\Omega)$ dans $L^1(\Omega)$ est compacte. □

En conséquence, de toute suite bornée de $L^1 \cap BV(\Omega)$ [28] on peut extraire une sous-suite qui converge dans $L^1(\Omega)$ [29]. L'utilisation que nous allons faire de ce résultat est claire : nous allons montrer que $(u_{\Delta t,\Delta x})_{\Delta t,\Delta x\in\mathbb{R}}$ est un borné de $L^1 \cap BV(\Omega)$ pour tout Ω ouvert borné de $\mathbb{R}_+ \times \mathbb{R}$. Noter que si $u_{\Delta t,\Delta x} \in L^1 \cap BV(\mathbb{R}_+ \times \mathbb{R})$, l'expression de la norme de $u_{\Delta t,\Delta x}$ est

$$||u_{\Delta t,\Delta x}||_{L^1\cap BV(\Omega)} = \Delta t\Delta x \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} \left|u_j^n\right| + \Delta t \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} \left|u_{j+1}^n - u_j^n\right| + \Delta x \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} \left|u_j^{n+1} - u_j^n\right|.$$

Soit Ω un ouvert borné de $\mathbb{R}_+ \times \mathbb{R}$. Il existe $T, A \in \mathbb{R}_+$ tels que $\Omega \subset [0,T] \times [-A,A]$. Alors

$$\begin{aligned}||u_{\Delta t,\Delta x}||_{L^1\cap BV(\Omega)} \leq \Delta t\Delta x &\sum_{n=0}^{E(T/\Delta t)+1} \sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1} \left|u_j^n\right| \\ &+ \Delta t \sum_{n=0}^{E(T/\Delta t)+1} \sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1} \left|u_{j+1}^n - u_j^n\right| \\ &+ \Delta x \sum_{n=0}^{E(T/\Delta t)+1} \sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1} \left|u_j^{n+1} - u_j^n\right|.\end{aligned}$$

où $E(x)$ désigne la partie entière de x, pour $x \in \mathbb{R}_+$.

Nous avons déjà montré que si $\sup_{j\in\mathbb{Z}} \left|f'(u_j^0)\right| \Delta t/\Delta x \leq 1$, $\sup_{j\in\mathbb{Z}} \left|u_j^n\right| \leq \sup_{j\in\mathbb{Z}} \left|u_j^0\right|$. Donc

$$\begin{aligned}||u_{\Delta t,\Delta x}||_{L^1(\Omega)} &\leq \sup_{j\in\mathbb{Z}} \left|u_j^0\right| (E(T/\Delta t)+2)\,(2E(A/\Delta x)+3)\,\Delta t\Delta x \\ &\leq \sup_{j\in\mathbb{Z}} \left|u_j^0\right| (T+2\Delta t)\,(2A+3\Delta x)\end{aligned}$$

qui est borné indépendamment de Δt et Δx (lorsqu'ils tendent vers 0) puisque $u^0 \in L^\infty(\mathbb{R})$. L'approximation $u_{\Delta t,\Delta x}$ est donc bornée (indépendamment du maillage) dans $L^1(\Omega)$ pour tout Ω borné. Il ne reste plus qu'à évaluer la variation totale de $u_{\Delta t,\Delta x}$ sur Ω. Nous allons supposer pour simplifier que $u^0 \in BV(\mathbb{R})$. On en déduit alors que la variation totale (en espace) de l'approximation initiale est bornée :

$$\sum_{j\in\mathbb{Z}} \left|u_{j+1}^0 - u_j^0\right| < +\infty.$$

28. Bornée : pour la norme que nous avons définie sur cet espace, norme qui le rend complet.
29. C'est d'ailleurs uniquement la compacité *séquentielle* que nous utiliserons.

En effet,

$$\sum_{j\in\mathbb{Z}} \left|u^0_{j+1} - u^0_j\right| = \sum_{j\in\mathbb{Z}} \left| \frac{1}{\Delta x} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} u^0(x+\Delta x) - u^0(x)\, dx \right|$$
$$\leq \sum_{j\in\mathbb{Z}} \int_{(j-1/2)\Delta x}^{(j+1/2)\Delta x} \left| \frac{u^0(x+\Delta x) - u^0(x)}{\Delta x} \right| dx \leq TV_{\mathbb{R}}(u^0)$$

d'après la proposition 13.

Lemme 10 (Le Roux, Harten)
Un schéma de la forme

$$u^{n+1}_j = u^n_j + (u^n_{j+1} - u^n_j)C^n_{j+1/2} - (u^n_j - u^n_{j-1})D_{j-1/2} \tag{3.21}$$

où

$$\begin{aligned} &C^n_{j+1/2} \geq 0 \quad \forall (n,j) \in \mathbb{N}\times\mathbb{Z}, \\ &D^n_{j+1/2} \geq 0 \quad \forall (n,j) \in \mathbb{N}\times\mathbb{Z}, \\ &C^n_{j+1/2} + D^n_{j+1/2} \leq 1 \quad \forall (n,j) \in \mathbb{N}\times\mathbb{Z} \end{aligned}$$

est *à variation totale décroissante*, c'est-à-dire qu'il vérifie

$$\sum_{j\in\mathbb{Z}} \left|u^{n+1}_{j+1} - u^{n+1}_j\right| \leq \sum_{j\in\mathbb{Z}} \left|u^n_{j+1} - u^n_j\right| \quad \forall n \in \mathbb{N}.$$

□

Remarque 42
La forme (3.21) est dite *forme incrémentale*. Un schéma à variation totale décroissante est souvent dit TVD (pour « Total Variation Diminishing » en anglais). □

Démonstration

$$\begin{aligned} u^{n+1}_{j+1} - u^{n+1}_j = \; & u^n_{j+1} + (u^n_{j+2} - u^n_{j+1})C^n_{j+3/2} \\ & -(u^n_{j+1} - u^n_j)D^n_{j+1/2} \\ & -u^n_j - (u^n_{j+1} - u^n_j)C^n_{j+1/2} \\ & +(u^n_j - u^n_{j-1})D^n_{j-1/2} \end{aligned}$$

soit

$$\begin{aligned} u^{n+1}_{j+1} - u^{n+1}_j = \; & (u^n_{j+1} - u^n_j)\left(1 - C^n_{j+1/2} - D^n_{j+1/2}\right) \\ & +(u^n_{j+2} - u^n_{j+1})C^n_{j+3/2} \\ & +(u^n_j - u^n_{j-1})D^n_{j-1/2}. \end{aligned}$$

D'après les hypothèses sur les coefficients de la forme incrémentale, on a

$$\begin{aligned} \left|u^{n+1}_{j+1} - u^{n+1}_j\right| \leq \; & \left|u^n_{j+1} - u^n_j\right|\left(1 - C^n_{j+1/2} - D^n_{j+1/2}\right) \\ & + \left|u^n_{j+2} - u^n_{j+1}\right| C^n_{j+3/2} \\ & + \left|u^n_j - u^n_{j-1}\right| D^n_{j-1/2}. \end{aligned}$$

Donc

$$\begin{aligned}\sum_{j\in\mathbb{Z}}\left|u_{j+1}^{n+1}-u_j^{n+1}\right| \le \sum_{j\in\mathbb{Z}}\left|u_{j+1}^n-u_j^n\right|\left(1-C_{j+1/2}^n-D_{j+1/2}^n\right)\\ +\sum_{j\in\mathbb{Z}}\left|u_{j+2}^n-u_{j+1}^n\right|C_{j+3/2}^n\\ +\sum_{j\in\mathbb{Z}}\left|u_j^n-u_{j-1}^n\right|D_{j-1/2}^n\\ =\sum_{j\in\mathbb{Z}}\left|u_{j+1}^n-u_j^n\right|\left(1-C_{j+1/2}^n-D_{j+1/2}^n+C_{j+1/2}^n+D_{j+1/2}^n\right).\end{aligned}$$

Cela termine la démonstration. □

Remarque 43

La notion de schéma TVD est très importante en analyse numérique des équations hyperboliques, elle permet souvent, entre autres, de démontrer la convergence de schémas. Elle paraît intuitivement garantir que le schéma ne crée pas d'oscillation. Ceci est cependant faux au sens strict. Considérons par exemple la donnée initiale $u_j^0=\delta_0(j)$ et un schéma tel que $u_j^1=1/2\delta_{-1}(j)+1/2\delta_1(j)$ [30]. On a alors

$$\sum_{j\in\mathbb{Z}}\left|u_{j+1}^1-u_j^1\right|=2=\sum_{j\in\mathbb{Z}}\left|u_{j+1}^0-u_j^0\right|,$$

donc le schéma est TVD. Cependant, des oscillations ont été créées en 1 pas de temps. Noter que le schéma de Lax-Friedrichs a exactement ce comportement pour cette condition initiale et une équation d'advection à vitesse constante ($f(u)=au$) par exemple. □

Proposition 14

Le schéma de Lax-Friedrichs peut se mettre sous la forme incrémentale (3.21) avec des coefficients qui vérifient les hypothèses du lemme 10 sous la condition de CFL $\sup_{j\in\mathbb{Z}}\left|f'(u_j^0)\right|\Delta t/\Delta x\le 1$. Il est TVD sous cette condition. □

Démonstration

Le schéma de Lax-Friedrichs est défini par

$$u_j^{n+1}=\frac{u_{j-1}^n+u_{j+1}^n}{2}-\frac{\Delta t}{2\Delta x}\left(f(u_{j+1}^n)-f(u_{j-1}^n)\right),$$

c'est-à-dire

$$\begin{aligned}u_j^{n+1}=u_j^n \quad &+\tfrac{u_{j+1}^n-u_j^n}{2}-\tfrac{\Delta t}{2\Delta x}\left(f(u_{j+1}^n)-f(u_j^n)\right)\\ &-\tfrac{u_j^n-u_{j-1}^n}{2}-\tfrac{\Delta t}{2\Delta x}\left(f(u_j^n)-f(u_{j-1}^n)\right).\end{aligned}$$

Posons

$$a_{j+1/2}^n=\int_0^1 f'(\theta u_{j+1}^n+(1-\theta)u_j^n)\,d\theta\quad\forall(n,j)\in\mathbb{N}\times\mathbb{Z}.$$

30. Par exemple un schéma écrit sous forme incrémentale avec $C_{j+1/2}^n=1/2=D_{j+1/2}^n\ \forall n\in\mathbb{N},j\in\mathbb{Z}$.

On a

$$a^n_{j+1/2}(u^n_{j+1}-u^n_j)=f(u^n_{j+1})-f(u^n_j)$$

(à vérifier à titre d'exercice, voir le partiel du 6 avril 2004 pour des détails) et

$$\begin{aligned}u^{n+1}_j = u^n_j \quad &+(u^n_{j+1}-u^n_j)\left(\tfrac{1}{2}-\tfrac{\Delta t}{2\Delta x}a^n_{j+1/2}\right)\\ &-(u^n_j-u^n_{j-1})\left(\tfrac{1}{2}+\tfrac{\Delta t}{2\Delta x}a^n_{j-1/2}\right).\end{aligned}$$

C'est la forme incrémentale du schéma, en posant

$$\begin{aligned}C^n_{j+1/2}&=\tfrac{1}{2}\left(1-a^n_{j+1/2}\tfrac{\Delta t}{\Delta x}\right),\\ D^n_{j+1/2}&=\tfrac{1}{2}\left(1+a^n_{j+1/2}\tfrac{\Delta t}{\Delta x}\right)\end{aligned}$$

pour tout $(n,j)\in\mathbb{N}\times\mathbb{Z}$. On a trivialement

$$\begin{aligned}&C^n_{j+1/2}+D^n_{j+1/2}=1 \quad \forall (n,j)\in\mathbb{N}\times\mathbb{Z},\\ &\left|a^n_{j+1/2}\right|\le \sup_{j\in\mathbb{Z}}\left|f'(u^n_j)\right| \quad \forall (n,j)\in\mathbb{N}\times\mathbb{Z} \text{ (par convexité de } f),\end{aligned}$$

d'où $C^n_{j+1/2}\ge 0$ et $D^n_{j+1/2}\ge 0$ sous la condition de CFL. □

Comme précédemment, soit Ω un ouvert borné de $\mathbb{R}_+\times\mathbb{R}$ et soit T et A tels que $\Omega\subset[0,T]\times[-A,A]$.
Sous la condition de CFL, nous avons, pour tout $n\in\mathbb{N}$, $\sum_{j\in\mathbb{Z}}\left|u^n_{j+1}-u^n_j\right|\le\sum_{j\in\mathbb{Z}}\left|u^0_{j+1}-u^0_j\right|$. Donc

$$\begin{aligned}\Delta t\sum_{n=0}^{E(T/\Delta t)+1}\sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1}\left|u^n_{j+1}-u^n_j\right|&\\ \le \Delta t\sum_{n=0}^{E(T/\Delta t)+1}\sum_{j\in\mathbb{Z}}\left|u^n_{j+1}-u^n_j\right| &\le \Delta t\sum_{n=0}^{E(T/\Delta t)+1} TV_{\mathbb{R}}(u^0).\end{aligned}$$

D'autre part, $u^{n+1}_j-u^n_j=C^n_{j+1/2}(u^n_{j+1}-u^n_j)-D^n_{j-1/2}(u^n_j-u^n_{j-1})$, donc

$$\left|u^{n+1}_j-u^n_j\right|\le C^n_{j+1/2}\left|u^n_{j+1}-u^n_j\right|+D^n_{j-1/2}\left|u^n_j-u^n_{j-1}\right|$$

et

$$\begin{aligned}\Delta x\sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1}\sum_{n=0}^{E(T/\Delta t)+1}\left|u^{n+1}_j-u^n_j\right|&\\ \le \Delta x\sum_{n=0}^{E(T/\Delta t)+1}\sum_{j\in\mathbb{Z}}\left(C^n_{j+1/2}\left|u^n_{j+1}-u^n_j\right|+D^n_{j-1/2}\left|u^n_j-u^n_{j-1}\right|\right)&\end{aligned}$$

donc finalement

$$\Delta x \sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1} \sum_{n=0}^{E(T/\Delta t)+1} \left|u_j^{n+1} - u_j^n\right| \leq \Delta x \sum_{n=0}^{E(T/\Delta t)+1} \sum_{j\in\mathbb{Z}} \left|u_{j+1}^n - u_j^n\right| \leq \Delta x \sum_{n=0}^{E(T/\Delta t)+1} TV_{\mathbb{R}}(u^0).$$

On suppose dorénavant que $\Delta t = \lambda \Delta x$ où $\lambda \in \mathbb{R}_+^*$ (Δt et Δx tendent vers 0 à la même vitesse). La condition de CFL s'exprime donc $\lambda \sup_{j\in\mathbb{R}} \left|f'(u_j^0)\right| \leq 1$. Nous avons

$$\Delta x \sum_{j=-E(A/\Delta x)-1}^{E(A/\Delta x)+1} \sum_{n=0}^{E(T/\Delta t)+1} \left|u_j^{n+1} - u_j^n\right| \leq \frac{\Delta t}{\lambda} \sum_{n=0}^{E(T/\Delta t)+1} TV_{\mathbb{R}}(u^0).$$

Nous sommes donc en mesure de majorer la norme de $u_{\Delta t,\Delta x}$ dans $L^1 \cap BV(\Omega)$:

$$TV_\Omega(u_{\Delta t,\Delta x}) \leq \Delta t \sum_{n=0}^{E(T/\Delta t)+1} TV_{\mathbb{R}}(u^0) + \frac{\Delta t}{\lambda} \sum_{n=0}^{E(T/\Delta t)+1} TV_{\mathbb{R}}(u^0) = \left(1 + \frac{1}{\lambda}\right)(T + 2\Delta t)\, TV_{\mathbb{R}}(u^0)$$

de sorte que

$$\left|\left|u_{\Delta t,\Delta t/\lambda}\right|\right|_{L^1\cap BV(\Omega)} \leq (T + 2\Delta t)(2A + 3\Delta x)\left|\left|u^0\right|\right|_{L^\infty(\mathbb{R})} + \left(1 + \frac{1}{\lambda}\right)(T + 2\Delta t)\, TV_{\mathbb{R}}(u^0)$$

qui est borné indépendamment de Δt (lorsqu'il tend vers 0).

D'après le théorème 13, il existe donc une suite de réels strictement positifs $(\Delta t_n)_{n\in\mathbb{N}}$ convergeant vers 0 et $u \in L^1(\Omega)$ tels que

$$\lim_{n\to+\infty} \left|\left|u_{\Delta t_n, 1/\lambda\Delta t_n} - u\right|\right|_{L^1(\Omega)} = 0.$$

Par un procédé d'extraction diagonale, on en déduit l'existence d'une sous-suite et de $u \in L^1_{loc}(\mathbb{R}_+ \times \mathbb{R})$ tels que

$$\lim_{n\to+\infty} \left|\left|u_{\Delta t_n, 1/\lambda\Delta t_n} - u\right|\right|_{L^1_{loc}(\mathbb{R}_+\times\mathbb{R})} = 0.$$

La question est maintenant : u est-il solution de (3.15) [31] ? La réponse, positive, est donnée par le théorème de Lax-Wendroff.

Définition 11

On dit que le schéma (3.17) est *consistant* au sens des volumes finis avec f si et seulement si $\exists K \in \mathbb{N}$ et $F \in \mathcal{C}^0(\mathbb{R}^{2K})$ tels que

$$f_{j+1/2}^n = F(u_{j-K+1}^n, u_{j-K+2}^n, \ldots, u_j^n, \ldots, u_{j+K-1}^n, u_{j+K}^n),$$
$$F(u, u, \ldots, u) = f(u) \quad \forall u \in \mathbb{R}$$

31. Une autre question naturelle est encore : si oui, u vérifie-t-il une inégalité d'Oleinik ? Nous y répondrons par la suite.

(on dit qu'il s'agit d'un flux à $2K$ points, ou d'un schéma à $2K+1$ points). □

Théorème 14 (Lax-Wendroff)
On considère le schéma (3.16,3.17) où l'on suppose que le flux dans (3.17) est consistant avec f. Pour tout Δt et tout Δx on définit une solution approchée par (3.20). Soit $\lambda \in \mathbb{R}$. Supposons qu'il existe une suite $(\Delta t_k)_{k\in\mathbb{N}}$ convergeant vers 0 telle que

$$\exists C \in \mathbb{R} \text{ t. q. } \left|\left|u_{\Delta t_k, \Delta t_k/\lambda}\right|\right|_{L^\infty(\mathbb{R}_+\times\mathbb{R})} \leq C \quad \forall k,$$
$$\exists u \in L^1_{loc}(\mathbb{R}_+^* \times \mathbb{R}) \text{ t. q. } \lim_{k\to+\infty} u_{\Delta t_k, \Delta t_k/\lambda} = u \text{ dans } L^1_{loc}(\mathbb{R}_+^* \times \mathbb{R}).$$

Alors u est solution faible de

$$\begin{cases} \partial_t u + \partial_x f(u) = 0, \\ u(0,x) = u^0(x). \end{cases}$$

□

Démonstration
On pose $\Delta x = \Delta t/\lambda$. Soit $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. On veut montrer que

$$\int_0^\infty \int_{-\infty}^\infty u\partial_t\varphi + f(u)\partial_x\varphi \, dx \, dt = -\int_{-\infty}^\infty u^0(x)\varphi(0,x)\, dx.$$

Posons, pour tout $j \in \mathbb{Z}$ et tout $n \in \mathbb{N}$,

$$\varphi_j^n = \varphi(n\Delta t, j\Delta x)$$

et, pour tout $t \in \mathbb{R}_+$ et $x \in \mathbb{R}$,

$$\varphi_{\Delta t,\Delta x}(t,x) = \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} \varphi_j^n \mathbb{1}_{[n\Delta t,(n+1)\Delta t[}(t)\mathbb{1}_{[(j-1/2)\Delta x,(j+1/2)\Delta x[}(x).$$

Multiplions les termes de (3.17) par $\varphi_j^n \Delta t\Delta x$, nous obtenons

$$\Delta x\left(u_j^{n+1} - u_j^n\right)\varphi_j^n + \Delta t\left(f_{j+1/2}^n - f_{j-1/2}^n\right)\varphi_j^n = 0,$$

et sommons ces expressions sur n et j :

$$\Delta x\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}}\left(u_j^{n+1} - u_j^n\right)\varphi_j^n + \Delta t\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}}\left(f_{j+1/2}^n - f_{j-1/2}^n\right)\varphi_j^n = 0.$$

Ceci est équivalent à

$$\Delta x\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^n - \varphi_j^{n+1}\right) - \Delta x\sum_{j\in\mathbb{Z}} u_j^0\varphi_j^0 + \Delta t\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} f_{j+1/2}^n\left(\varphi_j^n - \varphi_{j+1}^n\right) = 0,$$

soit encore

$$\Delta x\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^{n+1} - \varphi_j^n\right) + \Delta t\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} f_{j+1/2}^n\left(\varphi_{j+1}^n - \varphi_j^n\right) = -\Delta x\sum_{j\in\mathbb{Z}} u_j^0\varphi_j^0.$$

Nous allons étudier la convergence de chaque terme de cette égalité lorsque l'on remplace Δt par Δt_k qui converge vers 0. Bien entendu, les termes u_j^n, φ_j^n et $f_{j+1/2}^n$ dépendent de Δt ; cette dépendance n'a pas été explicitement indiquée dans les notations afin de simplifier celles-ci. On note encore $\Delta x_k = \Delta t_k/\lambda$. Nous allons montrer que

$$\begin{aligned}
&\Delta x_k \sum_{j\in\mathbb{Z}} u_j^0 \varphi_j^0 \longrightarrow_{k\to+\infty} \int_{-\infty}^{\infty} u^0(x)\varphi(0,x)\,dx,\\
&\Delta x_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^{n+1}-\varphi_j^n\right) \longrightarrow_{k\to+\infty} \int_0^{\infty}\int_{-\infty}^{\infty} u\partial_t\varphi\,dx\,dt,\\
&\Delta t_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} f_{j+1/2}^n\left(\varphi_{j+1}^n-\varphi_j^n\right) \longrightarrow_{k\to+\infty} \int_0^{\infty}\int_{-\infty}^{\infty} f(u)\partial_x\varphi\,dx\,dt.
\end{aligned}$$

Puisque Δx_k et Δt_k sont liés, $u_{\Delta t_k,\Delta x_k}$ et $\varphi_{\Delta t_k,\Delta x_k}$ ne dépendent que de k ; nous les notons désormais u_k et φ_k.

Le premier terme

$$\begin{aligned}
\Delta x_k \sum_{j\in\mathbb{Z}} u_j^0\varphi_j^0 &= \sum_{j\in\mathbb{Z}} \varphi_j^0 \int_{(j-1/2)\Delta x_k}^{(j+1/2)\Delta x_k} u^0(x)\,dx\\
&= \sum_{j\in\mathbb{Z}} \int_{(j-1/2)\Delta x_k}^{(j+1/2)\Delta x_k} u^0(x)\varphi_j^0\,dx = \sum_{j\in\mathbb{Z}} \int_{(j-1/2)\Delta x_k}^{(j+1/2)\Delta x_k} u^0(x)\varphi_k(0,x)\,dx.
\end{aligned}$$

Or $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+\times\mathbb{R})$, donc $\varphi(0,\cdot)\in\mathcal{C}_c^\infty(\mathbb{R})$ et

$$\varphi_k(0,\cdot) \longrightarrow_{k\to+\infty} \varphi(0,\cdot) \text{ uniformément.}$$

Donc on a bien

$$\int_{-\infty}^{\infty} u^0(x)\varphi_k(0,x)\,dx \longrightarrow_{k\to+\infty} \int_{-\infty}^{\infty} u^0(x)\varphi(0,\cdot)\,dx.$$

Le deuxième terme

$$\begin{aligned}
&\Delta x_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^{n+1}-\varphi_j^n\right)\\
&\qquad = \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} \int_{(j-1/2)\Delta x_k}^{(j+1/2)\Delta x_k} u_k((n+1)\Delta t_k,x)[\varphi_k((n+1)\Delta t_k,x)-\varphi_k(n\Delta t_k,x)]\,dx.
\end{aligned}$$

Ainsi

$$\begin{aligned}
&\Delta x_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^{n+1}-\varphi_j^n\right)\\
&\qquad = \sum_{n\in\mathbb{N}} \int_{-\infty}^{\infty} u_k((n+1)\Delta t_k,x)[\varphi_k((n+1)\Delta t_k,x)-\varphi_k(n\Delta t_k,x)]\,dx.
\end{aligned}$$

et

$$\Delta x_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^{n+1}-\varphi_j^n\right) = \frac{1}{\Delta t_k}\int_{-\infty}^{\infty}\sum_{n\in\mathbb{N}^*}\int_{n\Delta t_k}^{(n+1)\Delta t_k} u_k(t,x)[\varphi_k(t,x)-\varphi_k(t-\Delta t_k,x)]\,dx.$$

Donc

$$\Delta x_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} u_j^{n+1}\left(\varphi_j^{n+1}-\varphi_j^n\right) = \int_{-\infty}^{\infty}\int_{\Delta t_k}^{\infty} u_k(t,x)\frac{\varphi_k(t,x)-\varphi_k(t-\Delta t_k,x)}{\Delta t_k}\,dx.$$

Remarquons maintenant que cette dernière intégrale peut se décomposer en la somme

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} u_k(t,x)\left(\mathbb{1}_{[\Delta t_k,+\infty[}(t)\frac{\varphi_k(t,x)-\varphi_k(t-\Delta t_k,x)}{\Delta t_k}-\partial_t\varphi(t,x)\right)dx + \int_{-\infty}^{\infty}\int_{0}^{\infty} u_k(t,x)\partial_t\varphi(t,x)\,dt\,dx.$$

La première intégrale de cette somme tend vers 0 lorsque k tend vers $+\infty$ car u_k est borné dans $L^\infty(\mathbb{R}_+\times\mathbb{R})$ et

$$\mathbb{1}_{[\Delta t_k,+\infty[}(\cdot)\frac{\varphi_k(t,\cdot)-\varphi_k(t-\Delta t_k,\cdot)}{\Delta t_k} \longrightarrow_{k\to+\infty} \partial_t\varphi(t,\cdot) \text{ uniformément}$$

et la seconde tend vers

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} u(t,x)\partial_t\varphi(t,x)\,dt\,dx$$

car φ est à support compact et u_k tend vers u dans $L^1_{loc}(\mathbb{R}_+\times\mathbb{R})$.

Le troisième terme Les mêmes manipulations que précédemment donnent

$$\Delta t_k \sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} f_{j+1/2}^n\left(\varphi_{j+1}^n-\varphi_j^n\right) = \int_{-\infty}^{\infty}\int_{0}^{\infty} f_k(t,x)\frac{\varphi_k(t,x+\Delta x_k/2)-\varphi(t,x-\Delta x_k/2)}{\Delta x_k}\,dt\,dx$$

où l'on a posé [32]

$$f_k(t,x)=\sum_{n\in\mathbb{N}}\sum_{j\in\mathbb{Z}} f_{j+1/2}^n\mathbb{1}_{[n\Delta t_k,(n+1)\Delta t_k[}(t)\mathbb{1}_{[j\Delta x_k,(j+1)\Delta x_k[}(x).$$

Comme dans l'étude du deuxième terme, on peut décomposer le troisième terme en la somme

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} f_k(t,x)\left(\frac{\varphi_k(t,x+\Delta x_k/2)-\varphi(t,x-\Delta x_k/2)}{\Delta x_k}-\partial_x\varphi(t,x)\right)dt\,dx + \int_{-\infty}^{\infty}\int_{0}^{\infty} f_k(t,x)\partial_x\varphi(t,x)\,dt\,dx.$$

32. Attention : f_k est décalé d'une demie maille en espace par rapport à φ_k, d'où la translation de φ_k dans l'intégrale précédente.

La première intégrale de cette somme converge vers 0 lorsque k tend vers $+\infty$ car d'une part f_k est borné dans $L^\infty(\mathbb{R}_+ \times \mathbb{R})$ puisque u_k l'est et le flux numérique est consistant (noter donc l'importance de la continuité de la fonction F dans la définition de la consistance au sens des volumes finis), et d'autre part

$$\frac{\varphi_k(t, \cdot + \Delta x_k/2) - \varphi(t, \cdot - \Delta x_k/2)}{\Delta x_k} \longrightarrow_{k\to+\infty} \partial_x\varphi(t, \cdot) \text{ uniformément.}$$

La convergence de la seconde intégrale vers

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} f(u)\partial_x\varphi \, dt \, dx$$

est un peu plus difficile à démontrer car f_k dépend de plusieurs valeurs de u_k : mettons, $2K$, comme dans la définition de la consistance. Notons, pour tout $t \in \mathbb{R}_+$ et tout $x \in \mathbb{R}$, $v_k(t,x) \in \mathbb{R}^{2K}$ le vecteur des $2K$ valeurs de u_k au temps t autour de x :

$$v_k(t,x) = \begin{pmatrix} u_k(t, x - (K+1/2)\Delta x_k) \\ \vdots \\ u_k(t, x - \Delta x_k/2) \\ \vdots \\ u_k(t, x + (K-1/2)\Delta x_k) \end{pmatrix} = \left(v_k^j(t,x)\right)_{j=-K+1}^{K}.$$

On a

$$f_k(t,x) = F((v_k(t,x))$$

en reprenant la notation de la définition de la consistance. Soit $[0,T] \times [-A, A]$ un compact en dehors duquel φ s'annule ($\partial_x\varphi$ s'annule donc aussi en dehors de ce compact). On a

$$\int_{-\infty}^{\infty}\int_{0}^{\infty} f_k(t,x)\partial_x\varphi(t,x) \, dt \, dx = \int_{-A}^{A}\int_{0}^{T} f_k(t,x)\partial_x\varphi(t,x) \, dt \, dx$$

et

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{0}^{\infty} f_k(t,x)\partial_x\varphi(t,x) \, dt \, dx = &\int_{-A}^{A}\int_{0}^{T} f(u(t,x))\partial_x\varphi(t,x) \, dt \, dx \\ &+ \int_{-A}^{A}\int_{0}^{T} \left(f_k(t,x) - f(u(t,x))\right)\partial_x\varphi(t,x) \, dt \, dx.\end{aligned}$$

Rappelons que la seule chose qu'il nous reste à montrer est que la dernière intégrale converge vers 0. Nous nous y attelons.

$$\begin{aligned}&\left|\int_{-A}^{A}\int_{0}^{T} \left(f_k(t,x) - f(u(t,x))\right)\partial_x\varphi(t,x) \, dt \, dx\right| \\ &\qquad \leq ||\partial_x\varphi||_{L^\infty(\mathbb{R}_+\times\mathbb{R}} \int_{-A}^{A}\int_{0}^{T} |f_k(t,x) - f(u(t,x))| \, dt \, dx \\ &\qquad = ||\partial_x\varphi||_{L^\infty(\mathbb{R}_+\times\mathbb{R}} \int_{-A}^{A}\int_{0}^{T} |F(v_k(t,x)) - f(u(t,x))| \, dt \, dx.\end{aligned}$$

Or pour tout $j \in \{-K, \dots, K\}$

$$\int_{-A}^{A} \int_0^T \left| v_k^j(t,x)) - u(t,x) \right| \, dt \, dx = \int_{-A}^{A} \int_0^T |u_k(t, x + (j-1/2)\Delta x_k) - u(t,x)| \, dt \, dx,$$

donc

$$\begin{aligned} \int_{-A}^{A} \int_0^T \left| v_k^j(t,x)) - u(t,x) \right| \, dt \, dx & \\ \leq \int_{-A}^{A} \int_0^T |u_k(t, x + (j-1/2)\Delta x_k) - u(t, x + (j-1/2)\Delta x_k)| \, dt \, dx & \\ + \int_{-A}^{A} \int_0^T |u(t, x + (j-1/2)\Delta x_k) - u(t,x)| \, dt \, dx, & \end{aligned}$$

d'où

$$\begin{aligned} \int_{-A}^{A} \int_0^T \left| v_k^j(t,x)) - u(t,x) \right| \, dt \, dx & \\ \leq \int_{-A+(j-1/2)\Delta x_k}^{A+(j-1/2)\Delta x_k} \int_0^T |u_k(t,x) - u(t,x)| \, dt \, dx & \\ + \int_{-A}^{A} \int_0^T |u(t, x + (j-1/2)\Delta x_k) - u(t,x)| \, dt \, dx. & \end{aligned}$$

Dès que $\Delta x_k \leq 1$,

$$\int_{-A+(j-1/2)\Delta x_k}^{A+(j-1/2)\Delta x_k} \int_0^T |u_k(t,x) - u(t,x)| \, dt \, dx \leq \int_{-A+(j-1/2)}^{A+(j-1/2)} \int_0^T |u_k(t,x) - u(t,x)| \, dt \, dx$$

et cette intégrale tend vers 0 lorsque k tend vers $+\infty$ car u_k tend vers u dans $L^1_{loc}(\mathbb{R}_+ \times \mathbb{R})$ par hypothèse. D'autre part

$$\int_{-A}^{A} \int_0^T |u(t, x + (j-1/2)\Delta x_k) - u(t,x)| \, dt \, dx \longrightarrow_{k \to +\infty} 0$$

(propriété de continuité en moyenne). Donc

$$v_k^j \longrightarrow_{k \to +\infty} u \text{ dans } L^1([0,T] \times [-A, A])$$

pour tout $j \in \{-K, \dots, K\}$. Il existe donc une sous-suite de v_k (notée encore v_k) telle que

$$v_k^j \longrightarrow_{k \to +\infty} u \text{ presque partout dans } [0,T] \times [-A, A]$$

pour tout $j \in \{-K, \dots, K\}$. Puisque F est continue,

$$F \circ v_k \longrightarrow_{k \to +\infty} F(u, \dots, u) = f(u) \text{ presque partout dans } [0,T] \times [-A, A].$$

De plus, u_k est bornée dans L^∞, donc $F \circ v_k$ l'est et u l'est [33] et $f(u)$ l'est aussi. Ainsi

$$\int_{-A}^{A} \int_0^T |F(v_k(t,x)) - f(u(t,x))| \, dt \, dx \longrightarrow_{k\to+\infty} 0$$

d'après le théorème de convergence dominée de Lebesgue. On a donc bien ce que l'on cherchait à démontrer :

$$\Delta t_k \sum_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} f_{j+1/2}^n \left(\varphi_{j+1}^n - \varphi_j^n\right) \longrightarrow_{k\to+\infty} \int_{-\infty}^{\infty} \int_0^{\infty} f(u(t,x)) \partial_x \varphi(t,x) \, dt \, dx.$$

En résumé, la limite u vérifie

$$\int_{-\infty}^{\infty} \int_0^{\infty} u \partial_t \varphi + f(u) \partial_x \varphi \, dt \, dx = - \int_{-\infty}^{\infty} \int_0^{\infty} u^0(x) \varphi(0,x) \, dt \, dx$$

pour tout $\varphi \in \mathcal{C}_c^\infty(\mathbb{R}_+ \times \mathbb{R})$. C'est une solution faible de (3.10). □

On a remarqué que l'hypothèse $\Delta t_k = \lambda \Delta x_k$ n'est pas utile : elle permet seulement de simplifier les notations. Le théorème de Lax-Wendroff est vrai (et démontré) pour Δt_k et Δx_k tendant vers 0 indépendamment.

La compilation des derniers résultats prouvés montre que la solution numérique définie par le schéma de Lax-Friedrichs converge (sous une condition de CFL et sous l'hypothèse de convexité de f) vers une solution faible du problème [34]. Cela montre en particulier l'existence d'une solution faible !

Mais nous savons par expérience que la solution faible n'est pas unique. Nous allons maintenant préciser le résultat de convergence obtenu en montrant que la limite des approximations donnée par le schéma de Lax-Friedrichs vérifie une inégalité d'Oleinik, *en faisant l'hypothèse supplémentaire de l'α-convexité* [35] *du flux f* [36].

Supposons maintenant que f, de classe $\mathcal{C}^2$, est α-convexe : $\exists \alpha \in \mathbb{R}_+^*$ tel que $f''(x) \geq \alpha$ $\forall x \in \mathbb{R}$. Nous allons montrer qu'il existe une solution de (3.10) (avec $u^0 \in L^1 \cap BV(\mathbb{R})$) telle que

$$u(t,y) - u(t,x) \leq \frac{(y-x)}{\alpha t} \text{ pour presque tout } (x,y) \text{ tel que } y \geq x$$

33. En tant que limite dans L^1_{loc} d'une suite bornée dans L^∞.

34. Si le rapport $\Delta t/\Delta x$ est fixé ! Cette hypothèse est ici *nécessaire* car le schéma de Lax-Friedrichs n'est pas *a priori* sous la forme demandée dans le théorème de Lax-Wendroff : en effet, le flux $f_{j+1/2}$ dépend non seulement de u_j et u_{j+1}, mais encore de $\Delta t/\Delta x$ (cf. (3.19)). Fixer le rapport $\Delta t/\Delta x$ est un moyen de lever la dépendance du schéma en ce paramètre. Attention, le schéma de Lax-Friedrichs peut n'être pas convergent (vers la solution du problème hyperbolique qui nous intéresse) si l'on ne fixe pas le rapport $\Delta t/\Delta x$. Par exemple, si l'on choisit $\Delta t = \Delta x^2$, la solution numérique converge (lorsque Δx tend vers 0) vers la solution d'un problème parabolique du type $\partial_t u + \partial_x f(u) = \kappa \partial_{x,x}^2 u$ pour un certain κ non nul (voir l'examen du 18 mai 2006).

35. Uniforme convexité de module de convexité α.

36. Sans cette hypothèse supplémentaire, il n'y a pas en général de solution vérifiant l'inégalité d'Oleinik : considérer par exemple le cas d'un flux linéaire. Mais dans ce cas la solution faible est unique, donc la limite obtenue par le schéma de Lax-Friedrichs est l'unique solution faible du problème. On remarque de surcroît que puisque la limite est unique, toutes les sous-suites extraites de la suite des approximations convergent vers la même limite, donc toute la suite des approximations converge vers la solution.

pour tout $t > 0$ et que cette solution est approchée par le schéma de Lax-Friedrichs. Le plus simple serait d'arriver à montrer que pour tout Δt et tout Δx et pour toute condition initiale discrète,

$$u_{j+1}^n - u_j^n \leq \frac{\Delta x}{\alpha n \Delta t} \quad \forall j \in \mathbb{Z}, \forall n \in \mathbb{N}^*.$$

Ceci est cependant faux. En effet, avec la donnée *numérique* initiale

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

on aurait

$$u_j^1 = \frac{u_{j-1}^0 + u_{j+1}^0}{2} - \frac{\Delta t}{2\Delta x}\left(f(u_{j+1}^0) - f(u_{j-1}^0)\right) = (-1)^{j+1}$$

et (raisonnement par récurrence)

$$u_j^n = (-1)^{j+n} \quad \forall j \in \mathbb{Z}, \forall n \in \mathbb{N}.$$

Bien entendu, une condition initiale discrète définie ainsi pour tout maillage ne converge pas vers une fonction de L^1_{loc}, mais les choses risquent d'être compliquées tout de même. Le plus simple est finalement de ne prendre en considération qu'une maille du maillage (en espace) sur deux.

Proposition 15
Sous les conditions de CFL $\sup_{j\in\mathbb{Z}} \left|f'(u_j^0)\right| \Delta t \leq \Delta x/3$ et $\alpha \Delta t sup_{j\in\mathbb{Z}} \left(u_{j+1}^0 - u_{j-1}^0\right) \leq \Delta x$, la solution de (3.18) vérifie l'inégalité d'Oleinik discrète

$$u_{j+1}^n - u_{j-1}^n \leq \frac{2\Delta x}{\alpha n \Delta t} \quad \forall j \in \mathbb{Z}, \forall n \in \mathbb{N}^*.$$

□

Conséquence Il convient maintenant d'« oublier » une maille sur deux en définissant la fonction approchée par

$$u_{\Delta t, \Delta x}(t,x) = \sum_{n\in\mathbb{N}} \sum_{j\in\mathbb{Z}} u_{2j}^n \mathbb{1}_{[n\Delta t,(n+1)\Delta t[}(t) \mathbb{1}_{[(2j-1)\Delta x,(2j+1)\Delta x[}(x).$$

On peut alors montrer relativement facilement (le faire en exercice) que, si u est une limite de suite extraite de $u_{\Delta t, \Delta t/\lambda}$, u vérifie

$$u(t,y) - u(t,x) \leq \frac{y-x}{\alpha t} \text{ pour presque tout } (x,y) \text{ tel que } y \geq x \tag{3.22}$$

pour tout $t > 0$. Remarquer que dans le cas de l'équation de Burgers, $\alpha = 1$ et on retrouve l'inégalité d'Oleinik connue, qui est optimale : il n'existe pas de fonction $C(t) < 1/t$ telle que

$$u(t,y) - u(t,x) \leq C(t)(y-x) \text{ pour presque tout } (x,y) \text{ tel que } y \geq x$$

pour toute donnée initiale, *cf.* la donnée initiale $H(x)$ pour laquelle la solution développe une détente. En ce sens, l'inégalité d'Oleinik discrète que nous montrons ici est optimale. Cela permet

de conclure à l'existence d'une solution faible de (3.10) vérifiant l'inégalité d'Oleinik (3.22) lorsque f est de classe $\mathcal{C}^2$ et α-convexe et $u^0 \in L^\infty \cap BV(\mathbb{R})$ [37]. On sait par ailleurs qu'une telle solution est unique. Ce résultat (existence et unicité) constitue le point d'orgue du chapitre « hyperbolique » de ce cours.

Démonstration

Elle repose sur un raisonnement par récurrence. Nous supposerons qu'à l'étape n, le résultat est vrai :

$$u_{j+1}^n - u_{j-1}^n \leq \frac{2\Delta x}{\alpha n \Delta t} \quad \forall j \in \mathbb{Z}.$$

À l'étape $n+1$,

$$u_{j+1}^{n+1} = \frac{u_j^n + u_{j+2}^n}{2} - \frac{\Delta t}{2\Delta x}\left(f(u_{j+2}^n) - f(u_j^n)\right),$$
$$u_{j-1}^{n+1} = \frac{u_{j-2}^n + u_j^n}{2} - \frac{\Delta t}{2\Delta x}\left(f(u_j^n) - f(u_{j-2}^n)\right),$$

et ainsi

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

Donc, puisque f est de classe $\mathcal{C}^2$, $\exists y, z \in \mathbb{R}$ tels que

$$u_{j+1}^{n+1} - u_{j-1}^{n+1} = \frac{u_{j+2}^n - u_j^n}{2} - \frac{\Delta t}{2\Delta x}\left((u_{j+2}^n - u_j^n) f'(u_j^n) + \frac{\left(u_{j+2}^n - u_j^n\right)^2}{2} f''(y)\right)$$
$$+ \frac{u_j^n - u_{j-2}^n}{2} + \frac{\Delta t}{2\Delta x}\left((u_j^n - u_{j-2}^n) f'(u_j^n) - \frac{\left(u_j^n - u_{j-2}^n\right)^2}{2} f''(z)\right).$$

L'α-convexité de f implique maintenant que

$$u_{j+1}^{n+1} - u_{j-1}^{n+1} \leq (u_{j+2}^n - u_j^n)\left(\frac{1}{2} - \frac{\Delta t}{2\Delta x} f'(u_j^n)\right) - \frac{1}{2}\frac{\alpha \Delta t}{2\Delta x}(u_{j+2}^n - u_j^n)^2$$
$$+ (u_j^n - u_{j-2}^n)\left(\frac{1}{2} + \frac{\Delta t}{2\Delta x} f'(u_j^n)\right) - \frac{1}{2}\frac{\alpha \Delta t}{2\Delta x}(u_j^n - u_{j-2}^n)^2.$$

Posons, pour tout $j \in \mathbb{Z}$, $\eta_{j+1}^n = \max((u_{j+2}^n - u_j^n), 0)$. On a alors nécessairement $(u_{j+2}^n - u_j^n)^2 \geq {\eta_{j+1}^n}^2$ pour tout $j \in \mathbb{Z}$. En effet,

37. En fait, en dimension 1, $L^\infty \cap BV = BV$: toute fonction de $BV(\mathbb{R})$ est dans $L^\infty(\mathbb{R})$. Donc l'espace dans lequel on doit choisir u^0 est $BV(\mathbb{R})$ (attention : $L^\infty \cap BV(\mathbb{R}^d) \neq BV(\mathbb{R}^d)$ pour $d > 1$). Ce résultat d'existence et d'unicité n'est pas optimal. Le problème de Cauchy pour l'équation scalaire $\partial_t u + \partial_x f(u) = 0$ a une unique solution dès que la donnée initiale est dans $L^\infty(\mathbb{R})$ (on peut encore raffiner ce résultat en autorisant des données initiales mesures, mais ça suffit comme ça !). Ceci est vrai dès que le flux f est localement lipschitzien (sans hypothèse de convexité donc), mais dans ce cas le critère d'unicité à retenir n'est pas la condition d'Oleinik, mais par exemple un critère entropique ou le critère de Lax. Pour des détails, consulter [5].

— si $u^n_{j+2} - u^n_j \geq 0$, $(u^n_{j+2} - u^n_j)^2 = {\eta^n_{j+1}}^2$;
— si $u^n_{j+2} - u^n_j < 0$, $(u^n_{j+2} - u^n_j)^2 > 0$ et $\eta^n_{j+1} = 0$.

La précédente majoration permet donc d'écrire, sous la condition de CFL,

$$u^{n+1}_{j+1} - u^{n+1}_{j-1} \leq \frac{\eta^n_{j+1}}{2}\left(1 - \frac{\Delta t}{\Delta x} f'(u^n_j)\right) - {\eta^n_{j+1}}^2 \frac{\alpha \Delta t}{4\Delta x} + \frac{\eta^n_{j-1}}{2}\left(1 + \frac{\Delta t}{\Delta x} f'(u^n_j)\right) - {\eta^n_{j-1}}^2 \frac{\alpha \Delta t}{4\Delta x}.$$

Sous la condition de CFL classique $\sup_{j\in\mathbb{Z}} \left|f'(u^0_j)\right| \Delta t \leq \Delta x$ (assurée par la condition plus stricte supposée ici), nous avons $1 \pm f'(u^n_j)\frac{\Delta t}{\Delta x} \geq 0$. La fonction définie par

$$g(x) = \frac{1 \pm f'(u^n_j)\frac{\Delta t}{\Delta x}}{2} x - \frac{\alpha \Delta t}{4\Delta x} x^2 \quad \forall x \in \mathbb{R}$$

est croissante sur l'intervalle

$$\left]-\infty, \frac{\left(1 \pm f'(u^n_j)\frac{\Delta t}{\Delta x}\right)\Delta x}{\alpha \Delta t}\right].$$

Donc, en supposant que

$$\frac{2\Delta x}{\alpha n \Delta t} \leq \frac{\left(1 - \sup_{j\in\mathbb{Z}} |f'(u^n_j)|\frac{\Delta t}{\Delta x}\right)\Delta x}{\alpha \Delta t} \quad \forall j \in \mathbb{Z}, \tag{3.23}$$

puisque l'on a $\eta^n_j \leq \frac{2\Delta x}{\alpha n \Delta t}$ par hypothèse de récurrence, on a aussi

$$u^{n+1}_{j+1} - u^{n+1}_{j-1} \leq \frac{2\Delta x}{2\alpha n \Delta t}\left(1 - \frac{\Delta t}{\Delta x} f'(u^n_j)\right) - \left(\frac{2\Delta x}{\alpha n \Delta t}\right)^2 \frac{\alpha \Delta t}{4\Delta x} + \frac{2\Delta x}{2\alpha n \Delta t}\left(1 + \frac{\Delta t}{\Delta x} f'(u^n_j)\right) - \left(\frac{2\Delta x}{\alpha n \Delta t}\right)^2 \frac{\alpha \Delta t}{4\Delta x}.$$

Ainsi

$$u^{n+1}_{j+1} - u^{n+1}_{j-1} \leq \frac{2\Delta x}{\alpha n \Delta t} - \frac{2\Delta x}{\alpha n^2 \Delta t} = \frac{2\Delta x}{\alpha n \Delta t}\left(1 - \frac{1}{n}\right).$$

Or $1 - 1/n = (n-1)/n \leq n/(n+1)$, donc

$$u^{n+1}_{j+1} - u^{n+1}_{j-1} \leq \frac{2\Delta x}{\alpha(n+1)\Delta t},$$

ce qu'il fallait démontrer. Il nous reste à comprendre l'hypothèse faite en (3.23). Nous allons montrer qu'elle est vérifiée automatiquement sous la condition de CFL renforcée $\sup_{j\in\mathbb{Z}} \left|f'(u^0_j)\right| \Delta t \leq \Delta x/3$. L'hypothèse (3.23) est équivalente à

$$\sup_{j\in\mathbb{Z}} |f'(u^n_j)| \frac{\Delta t}{\Delta x} \leq 1 - \frac{2}{n}.$$

Pour $n \geq 3$, la condition de CFL renforcée $\sup_{j\in\mathbb{Z}} \left|f'(u^0_j)\right| \Delta t \leq \Delta x/3$ le garantit. C'est rageant, le seul problème qui reste est celui des deux premiers pas de temps ! Pour le premier pas de

temps : supposons que Δt vérifie aussi $\alpha\Delta t sup_{j\in\mathbb{Z}}\left(u^0_{j+1}-u^0_{j-1}\right)\leq\Delta x$ (c'est bien une condition de type CFL, on peut choisir un tel Δt). Alors, puisque sous la condition de CFL classique on a $\left(u^1_{j+1}-u^1_{j-1}\right)\leq sup_{i\in\mathbb{Z}}\left(u^0_{i+1}-u^0_{i-1}\right)$, on a $\left(u^1_{j+1}-u^1_{j-1}\right)\leq\Delta x/(\alpha\Delta t)\leq 2\Delta x/(\alpha 1\Delta t)$, donc le résultat est vrai pour $n=1$, et ensuite $\left(u^2_{j+1}-u^2_{j-1}\right)\leq sup_{i\in\mathbb{Z}}\left(u^1_{i+1}-u^1_{i-1}\right)\leq sup_{i\in\mathbb{Z}}\left(u^0_{i+1}-u^0_{i-1}\right)$, donc $\left(u^2_{j+1}-u^2_{j-1}\right)\leq\Delta x/(\alpha\Delta t)=2\Delta x/(\alpha 2\Delta t)$, donc le résultat est vrai pour $n=2$. □

Remarque 44
La condition de CFL utilisée n'est pas classique et le résultat obtenu ici n'est sans doute pas optimal. Dans [6], on trouvera une estimation d'Oleinik discrète montrée sous la condition classique. Mais dans cette référence, c'est la condition d'Oleinik elle-même qui n'est pas optimale... D'autre part, il y a fort à parier (vérifier !) que si l'on s'autorise un pas de temps variable $\Delta t^n=t^{n+1}-t^n$, le résultat devient

$$u^n_{j+1}-u^n_{j-1}\leq\frac{2\Delta x}{\alpha t^n}\quad\forall j\in\mathbb{Z},\forall n\in\mathbb{N}^*,$$

sous la condition de stabilité

$$\begin{array}{l}\alpha\Delta t^0 sup_{j\in\mathbb{Z}}\left(u^0_{j+1}-u^0_{j-1}\right)\leq\Delta x\text{ et }\sup_{j\in\mathbb{Z}}|f'(u^0_j)|\frac{\Delta t^0}{\Delta x}\leq 1,\\ \alpha\Delta t^1 sup_{j\in\mathbb{Z}}\left(u^0_{j+1}-u^0_{j-1}\right)\leq\Delta x\text{ et }\sup_{j\in\mathbb{Z}}|f'(u^0_j)|\frac{\Delta t^1}{\Delta x}\leq 1,\\ \sup_{j\in\mathbb{Z}}|f'(u^0_j)|\frac{\Delta t^{n-1}}{\Delta x}\leq 1-\frac{2}{n}\text{ pour }n\geq 2\end{array}$$

et tend donc vers la condition de CFL optimale. □

3.4.3 Quelques résultats numériques

Voici à présent quelques résultats numériques donnés par le schéma de Lax-Friedrichs pour l'équation de Burgers. La condition initiale choisie est périodique de période 1, et est donnée sur $[0,1[$ par

$$u^0(x)=\begin{cases}0\text{ si }x<0,1\\ 1\text{ si }0,1\leq x<0,6\\ 0\text{ si }0,6\leq x.\end{cases}$$

La périodicité de la condition initiale implique la périodicité de l'unique solution vérifiant une inégalité d'Oleinik. D'autre part, la solution approchée donnée par le schéma de Lax-Friedrichs est elle-même périodique, ce qui nous autorise à ne faire le calcul que sur le segment $[0,1[$. On divise ce segment en J intervalles $[j\Delta x,(j+1)\Delta x[$, $j\in\{0,\ldots,J-1\}$ avec $\Delta x=1/J$. Pour $j=0$ et $J=J-1$, le calcul de la solution approchée par la formule

$$u^{n+1}_j=\frac{u^n_{j-1}+u^n_{j+1}}{2}-\frac{\Delta t}{2\Delta x}\left(f(u^n_{j+1})-f(u^n_{j-1})\right)$$

nécessite la connaissance de valeurs u^n_{j-1} et u^n_J qui ne font pas partie des inconnues. Ces valeurs sont cependant données par la périodicité de la solution :

$$\left.\begin{array}{l} u^n_0 = u^n_{J-1} \\ u^n_J = u^n_1 \end{array}\right\} \quad \forall n \in \mathbb{N}.$$

Nous observons la solution au temps final $T = 0,4$. On peut montrer facilement [38] que celle-ci vaut alors

$$u(0,4,x) = \begin{cases} 0 \text{ si } x < 0,1 \\ \frac{x-0,1}{0,4} \text{ si } 0,1 \leq x < 0,5 \\ 1 \text{ si } 0,6 < x \leq 0,8 \\ 0 \text{ si } 0,8 < x. \end{cases}$$

On compare sur la figure 3.7 la solution exacte et les solutions approchées obtenues avec 100 et 1000 mailles. Le nombre de Courant est $\Delta t/\Delta x = 0,4$.

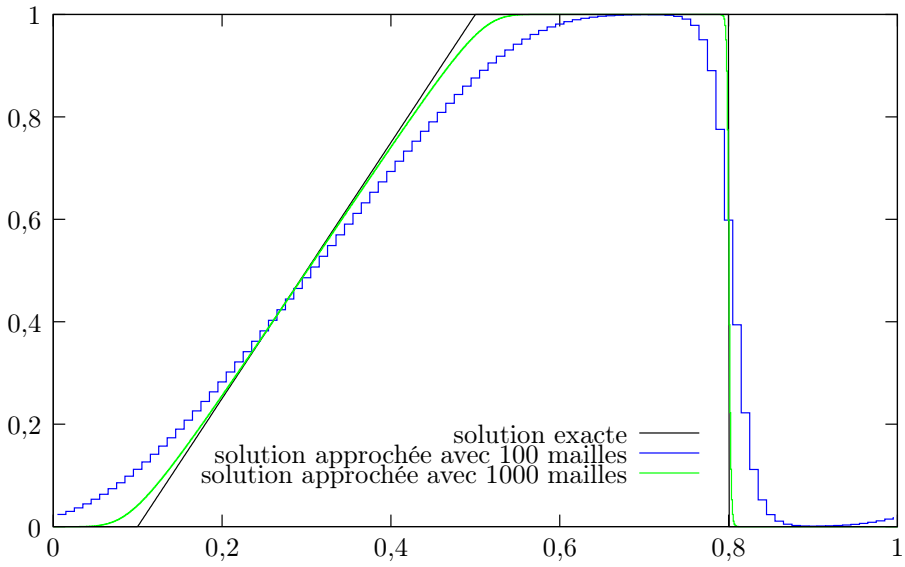

FIGURE 3.7 – Résultats numériques pour l'équation de Burgers.

On vérifie (!) sur cette figure ce que l'on a démontré, à savoir que le schéma est L^∞-stable et qu'il vérifie une inégalité d'Oleinik discrète. On remarque de plus que la discrétisation de Lax-Friedrichs a un effet « régularisant ». Ce phénomène est appelé *diffusion numérique*. Il est rapidement étudié pour le transport dans l'examen du 3 juin 2005 ainsi que dans l'examen du 18 mai 2006 (et leur corrigé).

Voici maintenant le programme (en langage Scilab) avec lequel ont été calculées les solutions représentées ci-dessus.

38. La discontinuité de droite se propage comme un choc, à vitesse 1/2, et celle de gauche forme une détente. Ceci est valable tant que ces deux « ondes » n'interfèrent pas, c'est-à-dire pour des temps inférieurs à 1.

```
        ////////////////////////////////////////////////////////////
        ///// Schéma de Lax-Friedrichs pour l'équation de Burgers. /////
        ////////////////////////////////////////////////////////////

clear();
stacksize(100000000);

function y=f(u)  // Fonction flux.
  y = 0.5*u*u;   // En modifiant cette fonction on
endfunction;     // peut résoudre une autre EDP scalaire.

function y=CI(x) // Conditions initiales.
  n = size(x);
  for i=1:n(1),
    if x(i) < 0.1 then
      y(i) = 0.;
    else
      if x(i) < 0.6 then
        y(i) = 1.;
      else
        y(i) = 0.;
      end;
    end;
    //y(i) = sin(2.*pi*x(i));
  end;
endfunction;

T = 0.4;
J = 1000.;
nbCourant = 0.4; // Ne pas dépasser 0.5.

// Schéma de Lax-Friedrichs : on multiplie le
// nombre de mailles par 2 puisqu'à la fin on n'en
// gardera qu'une sur deux.

J = 2*J;
```

```
dx = 1./J;
x=(1:J+2)'*dx - dx; // Grille en x.
u = CI(x);

t = 0.;

// Boucle en temps.

while t < T
  u(1) = u(J+1);            // Conditions périodiques.
  u(J+2) = u(2);
  cmax = u(2);
  for j=3:J+1
    cmax = max(cmax,u(j)); // Valeur maximale de f'.
  end;
  dt = min(dx/cmax*nbCourant,T - t);
  for j=2:J+1
    utemp(j) = 0.5*(u(j-1) + u(j+1))...
               - 0.5*dt/dx*(f(u(j+1)) - f(u(j-1)));
  end;
  for j=2:J+1,
    u(j) = utemp(j);
  end;
  t = t + dt;

  xset('window',1);
  xbasc();
  plot2d(x,u,leg='solution');
end; // Fin de la boucle en temps.

// Le schéma de Lax-Friedrichs est fait pour que l'on ne garde
// le résultat que dans une maille sur deux.

  ufin = ones(J/2,1);
  xfin = ones(J/2,1);
```

```
for j=1:J/2,
  ufin(j) = u(2*j);
  xfin(j) = x(2*j);
end;

unix('rm -f resultat');                    // Écriture du résultat
plouf = file('open','resultat','unknown'); // dans le fichier
for i=1:J/2,                               // « résultat ».
  fprintf(plouf,'%f   %f\n',xfin(i),ufin(i));
end;
file('close',plouf);
                              // Fin du programme.
```

Chapitre 4

Équations elliptiques

4.1 Introduction, exemples, Généralités

Il s'agit ici d'étudier des EDP du type

$$\sum_{i=1}^{d} b_i(x)\partial_i u(x) + \sum_{i=1}^{d}\sum_{j=1}^{d} c_{i,j}(x)\partial^2_{i,j} u(x) + c(x)u(x) = f(x) \text{ dans } \Omega \text{ ouvert de } \mathbb{R}^d \qquad (4.1)$$

où tous les coefficients sont réels et où la matrice $C(x) = (c_{i,j}(x))_{i,j=1}^{d}$ est symétrique et a ses valeurs propres toutes strictement positives ou toutes strictement négatives.

4.1.1 Exemples

Déformation d'un fil élastique dans $]0,1[$

Les équations mathématiques modélisant ce problème sont (sous certaines hypothèses sur la nature du fil)

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

où u est l'inconnue (la déformation), c une caractéristique du matériau constituant le fil et f le chargement auquel le fil est soumis.

Déformation d'une membrane élastique dans $\Omega \subset \mathbb{R}^2$

Le même modèle, en dimension 2, conduit à

$$\begin{cases} -\Delta u(x) + c(x)u(x) = f(x) \text{ dans } \Omega, \\ u = 0 \text{ sur } \Gamma = \partial\Omega, \end{cases}$$

Ω étant un ouvert de $\mathbb{R}^2$.

Déformation d'une poutre dans $]0,1[$

Lorsque le fil n'est pas souple mais rigide, on parle d'un modèle de poutre. Les équations régissant le déplacement d'un point du segment sont alors par exemple

$$\begin{cases} u^{(4)}(x) + c(x)u(x) = f(x) \text{ dans }]0,1[, \\ u(0) = u(1) = 0, \\ u'(0) = u'(1) = 1. \end{cases}$$

Cette équation n'est pas à proprement parler elliptique mais son étude peut se faire au moyen des mêmes techniques : que nous allons détailler ci-dessous.

4.1.2 Problème modèle et généralités

Le problème modèle que nous allons étudier est

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

où Ω est un ouvert de $\mathbb{R}^d$. Comme nous l'avons fait dans le chapitre sur les équation hyperboliques, on peut en écrire une formulation faible :

$$\int_\Omega \nabla u \nabla \varphi + u\varphi \, dx = \int_\Omega f\varphi \, dx \quad \forall \varphi \in \mathcal{C}^\infty(\Omega).$$

Nous allons dans ce chapitre modifier un peu cette formulation faible. Les outils essentiels de cette partie du cours sont ceux de l'analyse fonctionnelle : les distributions, les espaces de Sobolev et le théorème de Lax-Milgram. Ils sont rappelés dans la section suivante.

Rappels d'analyse fonctionnelle

Théorème 15 (Lax-Milgram)
Soit $(H, \langle\cdot,\cdot\rangle)$ un espace de Hilbert réel et soit a une forme bilinéaire continue coercitive sur $(H, \langle\cdot,\cdot\rangle)$:

$$\begin{cases} \exists C \in \mathbb{R} \text{ t. q. } |a(x,y)| \leq C\,||x||\,||y|| \quad \forall (x,y) \in H^2, \\ \exists \alpha \in \mathbb{R}_+^* \text{ t. q. } a(x,x) \geq \alpha\,||x||^2 \quad \forall x \in H. \end{cases}$$

Soit l une forme linéaire continue sur $(H, \langle\cdot,\cdot\rangle)$:

$$\exists D \in \mathbb{R} \text{ t. q. } |l(x)| \leq D\,||x|| \quad \forall x \in H.$$

Il existe un unique élément $u \in H$ tel que

$$a(u,v) = l(v) \quad \forall v \in H.$$

De plus, si a est symétrique, u est caractérisé par l'égalité

$$\Phi(u) = \min_{v \in H} \Phi(v)$$

où Φ est définie par $\Phi(v) = \frac{1}{2}a(v,v) - l(v)$. □

Nous effectuons ici une démonstration simple de ce résultat dans le cas où a est symétrique. C'est d'ailleurs uniquement dans ce cas-là qu'il sera par la suite utilisé. On verra dans la suite une démonstration générale du théorème de Lax-Milgram qui utilise le théorème de Stampacchia (théorème 22, voir la remarque 47).

Démonstration

Supposons que a est symétrique. Alors, comme par ailleurs $a(u,u) > 0$ pour tout $u \neq 0$, $a(\cdot,\cdot)$ définit un produit scalaire sur H, et ce produit scalaire est équivalent au produit scalaire $(\cdot,\cdot)$ car a est continue et coercitive. Donc, $(H, a(\cdot,\cdot))$ est un espace de Hilbert et l est continue pour ce nouveau produit scalaire. D'après le théorème de représentation de Riesz, l, forme linéaire continue sur $(H, a(\cdot,\cdot))$, peut être représentée (au moyen du produit scalaire) par un unique élément u de H :

$$\exists! u \in H \text{ t. q. } l(v) = a(u,v) \quad \forall v \in H.$$

Cela démontre la première partie du théorème.

Montrons maintenant que cette solution u réalise le minimum de la fonctionnelle Φ définie par $\Phi(x) = \frac{1}{2}a(x,x) - l(x)$ sur H. Soit $h \in H$.

$$\begin{aligned}\Phi(u+h) = \frac{1}{2}a(u,u) - l(u) + \frac{1}{2}a(u,h) + \frac{1}{2}a(h,u) + \frac{1}{2}a(h,h) - l(h) \\ = \Phi(u) + a(u,h) - l(h) + \frac{1}{2}a(h,h) = \Phi(u) + \frac{1}{2}a(h,h).\end{aligned}$$

Donc $\Phi(u+h) \geq \Phi(u) + \alpha/2\,||h||^2$: u réalise le minimum de Φ sur H.

Supposons maintenant que u réalise le minimum de Φ, et montrons que u vérifie $a(u,v) = l(v)$ pour tout $v \in H$. On a $\Phi(u+h) \geq \Phi(u)$ pour tout $h \in H$. D'après le calcul de $\Phi(u+h)$ que nous avons fait ci-dessus, ceci se réécrit

$$\Phi(u) + a(u,h) - l(h) + \frac{1}{2}a(h,h) \geq \Phi(u),$$

soit

$$\frac{1}{2}a(h,h) + a(u,h) - l(h) \geq 0 \quad \forall h \in H.$$

Ainsi, pour tout $\lambda \in \mathbb{R}$,

$$\frac{1}{2}a(\lambda h, \lambda h) + a(u, \lambda h) - l(\lambda h) \geq 0 \quad \forall h \in H,$$

soit, puisque a est bilinéaire et l est linéaire,

$$\frac{\lambda^2}{2}a(h,h) + \lambda\,(a(u,h) - l(h)) \geq 0 \quad \forall h \in H.$$

En choisissant λ assez petit en valeur absolue et du « bon signe », on en déduit que $a(u,h)-l(h)=0$ pour tout $h \in H$. Une autre manière de démontrer cela est de remarquer que Φ est de classe $\mathcal{C}^1(H)$ et que $\nabla\Phi(u).v = a(u,v) - l(v)$. Donc la condition classique de minimalité de Φ en u permet d'affirmer que $a(u,v) = l(v)$ pour tout $v \in H$. □

Afin d'être... Complets ! Rappelons le théorème de représentation de Riesz.

Théorème 16 (de représentation de Riesz)
Soit $(H, \langle\cdot,\cdot\rangle)$ un espace de Hilbert. Soit l une forme linéaire continue sur H. Il existe un unique élément $u \in H$ tel que

$$l(v) = \langle u, v\rangle \quad \forall v \in H.$$

□

Démonstration
Posons $\mathcal{N} = l^{-1}(0)$ (noyau de l). Le noyau de l, image réciproque d'un fermé par une application linéaire continue, est un sous-espace vectoriel fermé de H : on peut donc définir la projection (orthogonale) sur $\mathcal{N}$, notée $P_{\mathcal{N}}$.
Si $\mathcal{N} = H$, il suffit de prendre $u = 0$ pour démontrer le théorème.
Supposons donc maintenant que $\mathcal{N} \neq H$. Soit $x \in H \setminus \mathcal{N}$. Posons $w = (x - P_{\mathcal{N}}(x))/||x - P_{\mathcal{N}}(x)||$. On a

— $w \in H \setminus \mathcal{N}$, donc $l(w) \neq 0$;
— $||w|| = 1$;
— $\langle w, n\rangle = 0$ pour tout $n \in \mathcal{N}$.

Soit $v \in H$. Posons

$$n = v - \frac{l(v)}{l(w)} w$$

(c'est rendu possible par le fait que $l(w) \neq 0$). Ce vecteur vérifie $l(n) = 0$, c'est-à-dire $n \in \mathcal{N}$. Donc $\langle w, n\rangle = 0$, ce qui se réécrit

$$\langle w, v\rangle = \langle w, \frac{l(v)}{l(w)} w\rangle = \frac{l(v)}{l(w)}.$$

On a donc

$$\langle l(w)w, v\rangle = l(v),$$

quel que soit $v \in H$. Le vecteur u de l'énoncé est donc $l(w)w$. L'unicité d'un tel élément est triviale. □

Enchaînons maintenant avec quelques rappels sur les distributions. Dans toute la suite, Ω est un ouvert de $\mathbb{R}^d$.

A) Distributions.

1) Sur $\mathcal{D}(\Omega) = \mathcal{C}_c^\infty(\Omega)$, on définit une pseudo-topologie : une suite $(\varphi_n)_{n\in\mathbb{N}}$ de fonctions de $\mathcal{D}(\Omega)$ *converge* vers $\varphi \in \mathcal{D}(\Omega)$ si et seulement s'il existe un compact $K \subset \Omega$ tel que $\mathrm{supp}(\varphi_n) \subset K$

pour tout n, $\mathrm{supp}(\varphi) \subset K$ et $D^\alpha \varphi_n$ converge uniformément sur K vers $D^\alpha \varphi$ pour tout multi-indice $\alpha \in \mathbb{N}^d$.

2) $\mathcal{D}'(\Omega)$, dual topologique de $\mathcal{D}(\Omega)$, est l'ensemble des formes linéaires continues sur $\mathcal{D}(\Omega)$ pour la pseudo-topologie définie en 1) : soit $T : \mathcal{D}(\Omega) \longrightarrow \mathbb{R}$; $T \in \mathcal{D}'(\Omega)$ si et seulement si

— T est linéaire (donc, en particulier, $T(0) = 0$) ;

— $\forall\, (\varphi_n)_{n\in\mathbb{N}}$ suite de $\mathcal{D}(\Omega)$ convergeant vers 0 au sens donné par 1), $T(\varphi_n)$ converge vers $0 = T(0)$.

$\mathcal{D}'(\Omega)$ est appelé espace des distributions sur Ω.

3) Toute fonction $f \in L^1_{loc}(\Omega)$ définit une distribution, notée habituellement T_f :

$$T_f : \varphi \in \mathcal{D}(\Omega) \longmapsto \int_\Omega f\varphi.$$

On se permet d'écrire que $L^1_{loc}(\Omega) \subset \mathcal{D}'(\Omega)$.

4) Il existe des distributions qui ne peuvent se représenter sous forme d'une fonction de $L^1_{loc}(\Omega)$: $L^1_{loc}(\Omega) \subsetneq \mathcal{D}'(\Omega)$. Un exemple de telle distribution est donné par la masse de Dirac en $x \in \Omega$, δ_x :

$$\delta_x : \varphi \in \mathcal{D}(\Omega) \longmapsto \varphi(x).$$

5) On définit une pseudo-topologie sur $\mathcal{D}'(\Omega)$: $(T_n)_{n\in\mathbb{N}}$, suite de $\mathcal{D}'(\Omega)$, converge vers $T \in \mathcal{D}'(\Omega)$ si et seulement si $T_n(\varphi)$ converge vers $T(\varphi)$ pour tout $\varphi \in \mathcal{D}(\Omega)$ (il s'agit d'une pseudo-topologie faible-$*$).

6) Dérivation des distributions. Lorsque $f \in \mathcal{C}^1(\Omega)$ et $\varphi \in \mathcal{D}(\Omega)$, on a, si Ω est borné et suffisamment régulier, une formule d'intégration par parties (formule de Green) :

$$\int_\Omega \varphi \partial_i f\, dx = \int_{\partial\Omega} f\varphi n_i\, d\sigma - \int_\Omega f\partial_i\varphi\, dx = -\int_\Omega f\partial_i\varphi\, dx$$

(où $d\sigma$ est la mesure d'intégration sur $\partial\Omega$ et n_i est la i^{e} composante de la normale extérieure unitaire à $\partial\Omega$). Par analogie avec cette formule, si $T \in \mathcal{D}'(\Omega)$, on définit la dérivée de T $\partial_i T$, comme la distribution

$$\partial_i T : \varphi \longmapsto -T(\partial_i\varphi).$$

De manière similaire, pour tout multi-indice $\alpha \in \mathbb{N}^d$, on définit $D^\alpha T$ par

$$D^\alpha T : \varphi \longmapsto (-1)^{|\alpha|} T\,(D^\alpha \varphi)$$

où $|\alpha| = \sum_{i=1}^d \alpha_i$. On pourra montrer en exercice que ceci définit effectivement une distribution. On pourra aussi vérifier que la dérivée au sens des distributions de la fonction de Heaviside $H \in L^1_{loc}(\mathbb{R})$ définie par $H(x) = \mathbb{1}_{\mathbb{R}_+}(x)$ est la masse de Dirac en 0 δ_0 et que la dérivée seconde de cette fonction de Heaviside est la distribution qui à $\varphi \in \mathcal{D}(\Omega)$ associe $-\varphi'(0)$.

7) La dérivation est une application linéaire continue de $\mathcal{D}'(\Omega)$ dans $\mathcal{D}'(\Omega)$ pour la pseudo-topologie définie en 5).

B) Espaces de Sobolev.

8) Soit $p \geq 1$. Soit $f \in L^p(\Omega)$. Alors $f \in L^1_{loc}(\Omega)$, f définit donc une distribution $T_f \in \mathcal{D}'(\Omega)$. On dit que $f \in W^{1,p}(\Omega)$ si et seulement si les dérivées partielles premières au sens des distributions de f peuvent se représenter sous la forme de fonctions de $L^p(\Omega)$, c'est-à-dire si et seulement s'il existe des fonctions $g_1, \ldots, g_d \in L^p(\Omega)$ telles que

$$\int_\Omega f \partial_i \varphi = -\int_\Omega g_i \varphi \quad \forall i = 1, \ldots, d, \forall \varphi \in \mathcal{D}(\Omega).$$

On note dans ce cas $\nabla f = (g_i)_{i=1}^d$ et on définit une norme sur $W^{1,p}(\Omega)$ par

$$||f||_{W^{1,p}(\Omega)} = ||f||_{L^p(\Omega)} + ||\nabla f||_{L^p(\Omega)},$$

où $||\nabla f||_p = \left(\sum_{i=1}^d ||\partial_i f||_p^2\right)^{1/2}$. On dit que $f \in W^{m,p}(\Omega)$ si et seulement si toutes ses dérivées partielles d'ordre inférieur ou égal à m peuvent se représenter sous forme de fonctions de $L^p(\Omega)$, et on définit sur $W^{m,p}$ la norme

$$||f||_{W^{m,p}(\Omega)} = \sum_{\alpha \text{ t. q. } 0 \leq |\alpha| \leq m} ||D^\alpha f||_{L^p(\Omega)}.$$

Pour tout $m \in \mathbb{N}$ et tout $p \geq 1$, $W^{m,p}(\Omega)$ est un espace de Banach.

9) Les espaces $W^{m,2}(\Omega)$ sont notés $H^m(\Omega)$. On y définit le produit scalaire

$$\langle f, g \rangle_{H^m(\Omega)} = \sum_{\alpha \text{ t. q. } 0 \leq |\alpha| \leq m} \langle D^\alpha f, D^\alpha g \rangle_{L^2(\Omega)}.$$

La norme qui en dérive,

$$||f||_{H^m(\Omega)} = \left(\sum_{\alpha \text{ t. q. } 0 \leq |\alpha| \leq m} ||D^\alpha f||^2_{L^2(\Omega)} \right)^{1/2},$$

est équivalente à la norme $||\cdot||_{W^{m,2}(\Omega)}$ définie en 8). Pour tout $m \in \mathbb{N}$, $H^m(\Omega)$ est un espace de Hilbert.

10) $\mathcal{D}(\Omega) \subset H^1(\Omega)$ et $\mathcal{D}(\Omega)$ n'est pas dense (en général) dans $H^1(\Omega)$. On pose

$$H^1_0(\Omega) = \overline{\mathcal{D}(\Omega)}^{H^1(\Omega)}.$$

11) Lemme de Poincaré. Si Ω est borné, $\exists C \in \mathbb{R}$ tel que

$$||u||_{L^2(\Omega)} \leq C \, ||\nabla u||_{L^2(\Omega)} \quad \forall u \in H^1_0(\Omega).$$

Une conséquence importante en est que $(\cdot, \cdot)$ défini par

$$(u, v) = \langle \nabla u, \nabla v \rangle_{L^2(\Omega)}$$

est un produit scalaire équivalent à $\langle\cdot,\cdot\rangle_{H^1(\Omega)}$ sur $H^1_0(\Omega)$ si Ω est borné.

12) Deux remarques :

a) $H^1_0(\mathbb{R}^d) = H^1(\mathbb{R}^d)$.

b) $\mathcal{D}(\overline{\Omega}) = \mathcal{C}^\infty_c(\overline{\Omega}) = \mathcal{C}^\infty(\overline{\Omega})$ est dense dans $H^1(\Omega)$ [1] si Ω est borné et de classe $\mathcal{C}^1$.

C) Application trace.

13) Si Ω est borné et de frontière lipschitzienne,

$$\exists C(\Omega) \in \mathbb{R} \text{ t. q. } ||v|_{\partial\Omega}||_{L^2(\partial\Omega)} \leq C(\Omega)\,||v||_{H^1(\Omega)} \quad \forall v \in \mathcal{D}(\overline{\Omega}).$$

Donc l'application qui à $v \in \mathcal{D}(\overline{\Omega})$ associe $v|_{\partial\Omega}$ est linéaire et continue de $\mathcal{D}(\overline{\Omega})$ muni de la norme de $H^1(\Omega)$ dans $L^2(\partial\Omega)$. On prolonge par densité cette application en une application linéaire continue de $H^1(\Omega)$ dans $L^2(\partial\Omega)$. Cette application, l'application trace, notée tr, coïncide avec la trace usuelle pour les fonctions régulières de $\overline{\Omega}$. Noter que les fonctions de $L^2(\Omega)$ n'ont en général pas de trace (au sens où il n'existe pas d'application linéaire continue de $L^2(\Omega)$ dans $L^2(\partial\Omega)$ qui coïncide avec la trace usuelle pour les fonctions régulières).

14) Si Ω est borné et de frontière lipschitzienne,

$$H^1_0(\Omega) = \left\{v \in H^1(\Omega) \text{ t. q. } tr(v) = 0\right\}.$$

On en déduit que $H^1_0(\Omega)$ est dans ces conditions un espace de Hilbert pour le produit scalaire de $H^1(\Omega)$ (en tant qu'image réciproque dans un espace de Hilbert par une application continue (la trace) d'un fermé (le singleton $\{0\}$)). Il est donc un espace de Hilbert pour le produit scalaire $\langle\nabla\cdot,\nabla\cdot\rangle_{L^2(\Omega)}$ en vertu de l'inégalité de Poincaré vue en 11).

15) Formule de Green. Si Ω est borné et de frontière lipschitzienne, on a, comme dans le cas régulier, la formule

$$\int_\Omega v\partial_i u\,dx = \int_{\partial\Omega} tr(u)tr(v)n_i\,d\sigma - \int_\Omega u\partial_i v\,dx \quad \forall u, v \in H^1(\Omega),$$

que l'on réécrit bien vite

$$\int_\Omega v\partial_i u\,dx = \int_{\partial\Omega} uvn_i\,d\sigma - \int_\Omega u\partial_i v\,dx \quad \forall u, v \in H^1(\Omega),$$

où n_i est la i^e composante de la normale extérieure unitaire à $\partial\Omega$. Noter (vérifier, c'est un exercice) que de cette formule découle, entre autres, la formule

$$\int_\Omega v\Delta u\,dx = \int_{\partial\Omega} v\nabla u\cdot n\,d\sigma - \int_\Omega \nabla u\cdot\nabla v\,dx \quad \forall u \in H^2(\Omega), \forall v \in H^1(\Omega).$$

1. Au sens où une fonction de $H^1(\Omega)$ peut Ã[a]tre approchée par une suite de *restrictions* à Ω de fonctions de $\mathcal{D}(\overline{\Omega})$.

4.2 Étude de $-\Delta u + u = f$

4.2.1 Problème de Dirichlet homogène

On considère le problème

$$\begin{cases} -\Delta u + u = f \text{ dans } \Omega, \\ u = 0 \text{ sur } \partial\Omega \end{cases} \tag{4.2}$$

où Ω est un ouvert borné de $\mathbb{R}^d$ de frontière lipschitzienne. Le second membre f est une donnée du problème (fonction dans un espace à préciser pour garantir l'existence d'une solution).

Supposons que u est une solution classique de (4.2) : $u \in \mathcal{C}^2(\overline{\Omega})$ (et $f \in \mathcal{C}^0(\overline{\Omega})$) et soit $\varphi \in \mathcal{D}(\Omega)$. On a

$$\int_\Omega (\nabla u \cdot \nabla \varphi + u\varphi) = \int_\Omega f\varphi.$$

Si maintenant $\varphi \in H_0^1(\Omega)$, il existe une suite $(\varphi_n)_{n\in\mathbb{N}}$ de $\mathcal{D}(\Omega)$ convergeant dans $H_0^1(\Omega)$ vers φ, c'est-à-dire telle que

$$\varphi_n \longrightarrow_{n\to+\infty} \varphi \text{ dans } L^2(\Omega) \text{ et } \nabla\varphi_n \longrightarrow_{n\to+\infty} \nabla\varphi \text{ dans } \left(L^2(\Omega)\right)^d.$$

Or pour tout $n \in \mathbb{N}$

$$\int_\Omega (\nabla u \cdot \nabla \varphi_n + u\varphi_n) = \int_\Omega f\varphi_n.$$

De plus, $\nabla u \in \left(\mathcal{C}^0(\overline{\Omega})\right)^d$ (u est solution classique). Donc $\nabla u \in \left(L^2(\Omega)\right)^d$, $u \in L^2(\Omega)$, et $f \in L^2(\Omega)$, et on peut passer à la limite dans l'équation précédente (la convergence forte implique la convergence faible), d'où

$$\int_\Omega (\nabla u \cdot \nabla \varphi + u\varphi) = \int_\Omega f\varphi.$$

D'autre part, puisque $u \in \mathcal{C}^2(\overline{\Omega})$ et que u vaut 0 sur $\partial\Omega$, $u \in H_0^1(\Omega)$ (Ω étant borné). On appelle *solution faible* de (4.2) un élément $u \in H_0^1(\Omega)$ vérifiant

$$\int_\Omega (\nabla u \cdot \nabla v + uv) = \int_\Omega fv \quad \forall v \in H_0^1(\Omega). \tag{4.3}$$

Théorème 17
Si $f \in L^2(\Omega)$, le problème (4.3) admet une unique solution $u \in H_0^1(\Omega)$. Cette solution est caractérisée par

$$\frac{1}{2}\int_\Omega \left(|\nabla u|^2 + u^2\right) - \int_\Omega fu = \min_{v\in H_0^1(\Omega)} \left(\frac{1}{2}\int_\Omega \left(|\nabla v|^2 + v^2\right) - \int_\Omega fv\right).$$

□

Remarque 45
La caractérisation de la solution comme solution du problème de minimisation porte en mécanique le nom de principe des travaux virtuels (le problème considéré est un problème d'élasticité linéaire stationnaire). □

Démonstration
$H_0^1(\Omega)$, fermé de $H^1(\Omega)$, est un espace de Hilbert pour le produit scalaire de $H^1(\Omega)$. Posons

$$a(u,v) = \int_\Omega (\nabla u \cdot \nabla v + uv) \quad \forall u, v \in H_0^1(\Omega)$$

et

$$l(v) = \int_\Omega fv \quad \forall v \in H_0^1(\Omega)$$

(on vérifie que ces expressions ont bien un sens). Il est évident que l est une forme linéaire et que a est une forme bilinéaire. De plus, l est continue. En effet,

$$\left| \int_\Omega fv \right| \leq ||f||_{L^2(\Omega)} ||v||_{L^2(\Omega)} \leq ||f||_{L^2(\Omega)} ||v||_{H^1(\Omega)} .$$

La forme a est elle aussi continue, et elle est coercitive, car

$$a(u,v) = \langle u, v \rangle_{H^1(\Omega)}$$

(les modules de continuité et de coercitivité de a étant tous deux 1).
Donc, d'après le théorème de Lax-Milgram, il existe un unique $u \in H_0^1(\Omega)$ tel que $a(u,v) = l(v)$ pour tout $v \in H_0^1(\Omega)$. Comme a est symétrique, le théorème de Lax-Milgram indique que cette solution u est caractérisée comme réalisant le minimum de la fonctionnelle $\Phi(v) = \frac{1}{2}a(v,v) - l(v)$.
□

Les questions que nous allons nous poser maintenant sont relatives à
— la dépendance de la solution u de (4.3) vis à vis de la donnée f,
— la régularité de u,
— les propriétés de bornitude $L^\infty(\Omega)$ de u,
— le calcul (approché) de u.

Dépendance de u par rapport au second membre

Théorème 18
1) L'application

$$T_0 : \begin{cases} L^2(\Omega) \longrightarrow H_0^1(\Omega) \\ f \longmapsto u \text{ unique solution de (4.3)} \end{cases}$$

est linéaire et continue.
2) L'application

$$T : \begin{cases} L^2(\Omega) \longrightarrow L^2(\Omega) \\ f \longmapsto u \text{ unique solution de (4.3)} \end{cases}$$

est linéaire, continue et compacte. □

Démonstration
1) T_0 est linéaire. En effet, soit $f_1 \in L^2(\Omega)$, $f_2 \in L^2(\Omega)$, et $u_1 = T_0(f_1)$ et $u_2 = T_0(f_2)$. Soit $\lambda \in \mathbb{R}$. Posons $w = u_1 + \lambda u_2$. On a (au sens faible) $-\Delta w + w = f_1 + \lambda f_2$ et $w = 0$ sur $\partial\Omega$. De plus, la solution (faible) de $-\Delta w + w = f_1 + \lambda f_2$ est unique, et vaut $T_0(f_1 + \lambda f_2)$. Donc on a bien $T_0(f_1) + \lambda T_0(f_2) = u_1 + \lambda u_2 = w = T_0(f_1 + \lambda f_2)$. D'autre part, T_0 est continue, car

$$\int_\Omega \nabla u \cdot \nabla v + uv = \int_\Omega fv \quad \forall v \in H_0^1(\Omega),$$

donc cette égalité est vraie pour $v = u$, ce qui donne

$$\int_\Omega |\nabla u|^2 + |u|^2 = \int_\Omega fu,$$

soit

$$||u||^2_{H^1(\Omega)} = \int_\Omega fu,$$

d'où

$$||u||^2_{H^1(\Omega)} \leq ||f||_{L^2(\Omega)} ||u||_{L^2(\Omega)} \leq ||f||_{L^2(\Omega)} ||u||_{H^1(\Omega)},$$

et finalement

$$||u||_{H^1(\Omega)} \leq ||f||_{L^2(\Omega)}.$$

2) On a $T = Id_{H^1(\Omega)\to L^2(\Omega)} \circ T_0$. T_0 est linéaire continue et le théorème de Rellich (théorème 19) affirme que $Id_{H^1(\Omega)\to L^2(\Omega)}$ est compacte. Ainsi, T est linéaire continue et compacte. □

Théorème 19 (Rellich)
Soit Ω un ouvert borné de $\mathbb{R}^d$ de frontière lipschitzienne. L'inclusion de $H^1(\Omega)$ dans $L^2(\Omega)$ est compacte. □

La démonstration de ce résultat d'analyse fonctionnelle pourra être lue dans [1]. Noter la similitude entre ce théorème et le théorème de Helly (théorème 13).

Régularité de u

Une solution forte de (4.2) est une fonction $u \in \mathcal{C}^2(\overline{\Omega})$ vérifiant (4.2). Une solution faible de (4.2) est une fonction $u \in H_0^1(\Omega)$ vérifiant (4.3). On appelle de plus *solution forte* $L^2(\Omega)$ une fonction $u \in H_0^1(\Omega) \cap H^2(\Omega)$ vérifiant (4.3). L'équation (4.2) a donc un sens dans $L^2(\Omega)$ pour une solution forte $L^2(\Omega)$.

La question que l'on se pose dans ce paragraphe est « la solution u de (4.3) est-elle plus régulière qu'une fonction quelconque de $H_0^1(\Omega)$? ». La réponse, positive sous des hypothèses de régularité de la frontière de Ω, est encore donnée par un résultat d'analyse fonctionnelle que nous ne démontrerons pas ici.

Théorème 20
Supposons que le second membre f de (4.3) est dans $H^m(\Omega)$ et que $\partial\Omega$ est de classe $\mathcal{C}^{m+2}$ pour $m \in \mathbb{N}$. Alors la solution de (4.3) est dans $H^{m+2}(\Omega)$ (par convention, $H^0(\Omega) = L^2(\Omega)$). □

En particulier, si $\partial\Omega$ est de classe $\mathcal{C}^2$ (et $f \in L^2(\Omega)$), la solution du problème (4.3) est dans $H^2(\Omega)$ (et donc dans $H^2(\Omega) \cap H_0^1(\Omega)$). C'est une solution forte $L^2(\Omega)$. Ce résultat est évident en dimension 1. En effet, le problème s'écrit alors (au sens faible)

$$u'' = u - f$$

et puisque u et f sont des fonctions de $L^2(\Omega)$, la distribution u'' peut s'écrire comme une fonction de $L^2(\Omega)$. Résumons : $u \in L^2(\Omega)$, $u' \in L^2(\Omega)$ et $u'' \in L^2(\Omega)$, c'est-à-dire que $u \in H^2(\Omega)$. Ce résultat est pourtant beaucoup moins évident en dimension supérieure, car l'EDP nous dit seulement que

$$\sum_{i=1}^{d} \partial_{i,i}^2 u = u - f,$$

nous n'avons donc des informations que sur $\sum_{i=1}^{d} \partial_{i,i}^2 u$ (c'est une fonction de $L^2(\Omega)$) mais aucune information sur chaque dérivée partielle (croisée ou pas). La démonstration de ce résultat pourra être trouvée dans [1].

Des résultats d'inclusion de Sobolev permettent même dans certains cas d'avoir une solution régulière. On peut démontrer par exemple que si la frontière de Ω est régulière, pour $m > \frac{d}{2} + k$ (avec $k, m \in \mathbb{N}$), on a l'inclusion

$$H^m(\Omega) \subset \mathcal{C}^k(\overline{\Omega}).$$

Principe du maximum

On s'intéresse ici à des estimations de type $L^\infty(\Omega)$ de la solution de (4.3) (Ω est un ouvert borné de frontière lipschitzienne).

Théorème 21
Soit $f \in L^2(\Omega)$ et soit u la solution du problème (4.3) associé. Alors,

$$\min\left(0, \text{infess}_\Omega f\right) \leq u \leq \max\left(0, \text{supess}_\Omega f\right) \text{ presque partout.}$$

□

La démonstration que nous proposons ici, due à Stampacchia, est la même que celle que nous avons donnée pour le principe du maximum concernant la solution (forte) de l'équation de la chaleur (théorème 5).

Démonstration
Nous allons montrer que

$$u \leq \max\left(0, \text{supess}_\Omega f\right) \text{ presque partout.}$$

On se donne[2] une fonction $G \in \mathcal{C}^1(\mathbb{R})$ vérifiant

2. En trouver une !

— $\exists M \in \mathbb{R}$ tel que $|G'(s)| \leq M \ \forall s \in \mathbb{R}$;
— pour tout $s \in]0, +\infty[$, $G'(s) > 0$;
— pour tout $s \in]-\infty, 0]$, $G(s) = 0$.

Posons $K = \max(0, \text{supess}_\Omega f)$. Si $K = +\infty$, le résultat est évident. Supposons donc désormais que $K < +\infty$, et posons $v(x) = G(u(x) - K)$. D'après le lemme 11 énoncé et démontré ci-après, $v \in H^1(\Omega)$ et $\nabla v(x) = G'(u(x) - K)\nabla u(x)$. D'autre part, $v = 0$ sur $\partial\Omega$ car u s'y annule, donc

$$v|_{\partial\Omega} = G(u|_{\partial\Omega} - K) = G(-K),$$

or $K \geq 0$ et G est nulle sur $]-\infty, 0]$. Donc, $v \in H^1_0(\Omega)$. On peut donc prendre cette fonction comme fonction-test dans le problème faible (4.3) : u vérifie

$$\int_\Omega \nabla u \cdot \nabla v + uv = \int_\Omega fv$$

avec $v(x) = G(u(x) - K)$, soit

$$\int_\Omega |\nabla u|^2 G'(u - K) + uG(u - K) = \int_\Omega fG(u - K).$$

Soustrayons $KG(u-K)$ (qui est bien intégrable sur Ω borné) aux deux membres de cette égalité, sous l'intégrale. Il vient

$$\int_\Omega |\nabla u|^2 G'(u - K) + (u - K)G(u - K) = \int_\Omega (f - K)G(u - K).$$

Or $f(x) - K \leq 0$ presque partout sur Ω, et $G(u(x) - K) \geq 0$ pour tout x (propriété de G). Ainsi $\int_\Omega (f - K)G(u - K) \leq 0$ et

$$\int_\Omega (u - K)G(u - K) \leq -\int_\Omega |\nabla u|^2 G'(u - K) \leq 0.$$

D'autre part, c'est encore une propriété de G, on a $sG(s) \geq 0$ pour tout $s \in \mathbb{R}$, donc $(u - K)G(u - K) \geq 0$ sur Ω. On en déduit que

$$(u(x) - K)G(u(x) - K) = 0 \text{ presque partout.}$$

Donc on a presque partout

$$(u(x) - K) = 0 \text{ ou } G(u(x) - K) = 0.$$

Or d'après la définition de G, $G(u(x) - K) = 0$ implique $u(x) - K \leq 0$. Le résultat est démontré. □

Lemme 11

Soit Ω un ouvert borné de frontière lipschitzienne de $\mathbb{R}^d$. Soit $u \in H^1(\Omega)$ et soit $G \in \mathcal{C}^1(\mathbb{R})$ telle que G' est borné. On a $G \circ u \in H^1(\Omega)$ et $\nabla(G \circ u) = (G' \circ u)\nabla u$. □

Démonstration
Il faut vérifier que $G \circ u \in L^2(\Omega)$ et que $\nabla (G \circ u) \in \left(L^2(\Omega)\right)^d$. On sait que $G'(s) \leq M \ \forall s \in \mathbb{R}$, donc $|G(s)| \leq |G(0)| + M\,|s|$. Donc $|G(u(x))| \leq |G(0)| + M\,|u(x)|$, donc $G \circ u \in L^2(\Omega)$. La suite est un peu plus délicate. Puisque $u \in H^1(\Omega)$, il existe une suite $(u_n)_{n\in\mathbb{N}}$ de $\mathcal{D}(\overline{\Omega})$ telle que u_n converge vers u en norme $H^1(\Omega)$, c'est-à-dire telle que

$$\begin{array}{l} u_n \longrightarrow_{n\to+\infty} u \text{ dans } L^2(\Omega) \\ \text{et } \nabla u_n \longrightarrow_{n\to+\infty} \nabla u \text{ dans } \left(L^2(\Omega)\right)^d. \end{array}$$

On peut extraire [3] de cette suite une suite (toujours notée $(u_n)_{n\in\mathbb{N}}$) telle que

$$\begin{array}{l} u_n \longrightarrow_{n\to+\infty} u \text{ dans } L^2(\Omega) \text{ et presque partout} \\ \text{et } \nabla u_n \longrightarrow_{n\to+\infty} \nabla u \text{ dans } \left(L^2(\Omega)\right)^d \text{ et presque partout.} \end{array}$$

Pour tout $n \in \mathbb{N}$, $G \circ u_n \in \mathcal{C}^1(\Omega)$. Soit $\varphi \in \mathcal{D}(\Omega)$ et soit $i \in \{1, \ldots, d\}$.

$$\int_\Omega (G \circ u_n)\partial_i\varphi = -\int_\Omega \varphi\partial_i\left(G \circ u_n\right) = -\int_\Omega \varphi G' \circ u_n \partial_i u_n.$$

Notons $A_n = \int_\Omega (G \circ u_n)\partial_i\varphi$ et $B_n = -\int_\Omega \varphi G' \circ u_n \partial_i u_n$ (pour tout $n \in \mathbb{N}$). On a

$$|G \circ u_n - G \circ u| \leq M\,|u_n - u|\,,$$

donc $G \circ u_n - G \circ u$ tend vers 0 dans $L^2(\Omega)$. Donc A_n converge vers $A = \int_\Omega (G \circ u)\partial_i\varphi$ (car la convergence forte implique la convergence faible). D'autre part,

$$(G' \circ u_n)\partial_i u_n - (G' \circ u)\partial_i u = (G' \circ u_n)\left(\partial_i u_n - \partial_i u\right) + \left(G' \circ u_n - G' \circ u\right)\partial_i u,$$

donc, d'après l'inégalité triangulaire,

$$\begin{aligned} \left|\left|(G' \circ u_n)\partial_i u_n - (G' \circ u)\partial_i u\right|\right|_{L^2(\Omega)} \leq & \left(\int_\Omega \left|(G' \circ u_n)\left(\partial_i u_n - \partial_i u\right)\right|^2\right)^{1/2} \\ & + \left(\int_\Omega \left|\left(G' \circ u_n - G' \circ u\right)\partial_i u\right|^2\right)^{1/2} \end{aligned}$$

La première intégrale ci-dessus, majorée par $M\,||\partial_i u_n - \partial_i u||_{L^2(\Omega)}$, converge vers 0 par hypothèse sur la suite $(u_n)_{n\in\mathbb{N}}$. Quant à la seconde intégrale : on sait que

$$\left|\left(G' \circ u_n - G' \circ u\right)\partial_i u\right|^2$$

tend vers 0 presque partout (car u_n tend vers u presque partout et G' est continue) et que

$$\left|\left(G' \circ u_n - G' \circ u\right)\partial_i u\right|^2 \leq (2M)^2\,|\partial_i u|^2$$

3. Et on le fait.

qui est un terme intégrable. D'après le théorème de convergence dominée de Lebesgue, on a donc

$$\left(G' \circ u_n - G' \circ u\right) \partial_i u \longrightarrow_{n\to+\infty} 0 \text{ dans } L^2(\Omega).$$

Donc $(G' \circ u_n \partial_i u_n - G' \circ u \partial_i u)$ tend vers 0 dans $L^2(\Omega)$. On en déduit que B_n converge vers $B = \int_\Omega \varphi(G' \circ u)\partial_i u$. Ainsi, on a

$$\int_\Omega (G \circ u)\partial_i \varphi = -\int_\Omega \varphi(G' \circ u)\partial_i u,$$

ce qui signifie bien que la distribution $\partial_i (G \circ u)$ peut être représentée par un fonction de $L^2(\Omega)$ et que

$$\partial_i (G \circ u) = (G' \circ u)\partial_i u.$$

□

Remarque 46
L'inégalité de Poincaré, dont une conséquence est que $\langle \nabla \cdot, \nabla \cdot \rangle$ est un produit scalaire équivalent au produit scalaire de $H^1(\Omega)$ sur $H_0^1(\Omega)$, permet de montrer que le problème

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

admet une unique solution sous les mêmes hypothèses. □

4.2.2 Quelques éléments pour le problème de Dirichlet non homogène

On considère ici le problème

$$\begin{cases} -\Delta u + u = f \text{ dans } \Omega, \\ u = g \text{ sur } \partial\Omega \end{cases} \tag{4.4}$$

où Ω est un ouvert borné de $\mathbb{R}^d$ de frontière lipschitzienne. Les seconds membres f et g sont des données du problème.

Nous allons comme dans la section précédente chercher une solution faible de ce problème, solution dans $H^1(\Omega)$. Il est donc naturel d'interpréter la condition de bord comme suit : la solution u a une trace et elle vaut g. Nous supposerons donc que la fonction g est la trace d'une fonction $\tilde{g} \in H^1(\Omega)$ (par exemple : u, lorsque l'on saura qu'il existe une solution au problème !). De la même manière que dans la section précédente, on montre que si u est solution classique de (4.4), elle vérifie pour tout $v \in H_0^1(\Omega)$

$$\int_\Omega \nabla u \cdot \nabla v + uv = \int_\Omega fv.$$

De plus, la trace de u vaut g, donc la trace de $u - \tilde{g}$ vaut 0 : $u - \tilde{g} \in H_0^1(\Omega)$. Posons

$$V = \left\{ v \in H^1(\Omega) \text{ t. q. } v - \tilde{g} \in H_0^1(\Omega) \right\}.$$

On dit que u est solution faible de (4.4) si et seulement si

$$u \in V \text{ et } \int_\Omega \nabla u \cdot \nabla v + uv = \int_\Omega fv \quad \forall v \in H_0^1(\Omega). \tag{4.5}$$

L'application qui à $v \in H_0^1(\Omega)$ associe $tr(v - \tilde{g})$ est continue, ce qui prouve que V est un fermé de $H^1(\Omega)$. Mais V n'est pas un espace vectoriel (c'est un sous-espace affine de $H^1(\Omega)$), on ne peut donc pas appliquer le théorème de Lax-Milgram [4].

On fait appel ici au théorème de Stampacchia.

Théorème 22 (Stampacchia)
Soit $(H, \langle \cdot, \cdot \rangle)$ un espace de Hilbert réel. Soit $V \neq \emptyset$ un sous-ensemble convexe fermé de H. Soit a une forme bilinéaire continue coercitive sur V, et soit l une forme linéaire continue sur V.
Il existe un unique élément $u \in V$ tel que

$$a(u, v - u) \geq l(v - u) \quad \forall v \in V.$$

Si de plus a est symétrique, u est caractérisé par

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

avec $\Phi(v) = \frac{1}{2} a(v, v) - l(v)$. □

Démonstration
H étant un espace de Hilbert, on peut utiliser le théorème de représentation de Riesz : l'application l étant linéaire et continue, il existe un unique élément $L \in H$ tel que

$$l(v) = \langle L, v \rangle \quad \forall v \in H.$$

Soit $u \in H$. L'application qui à $v \in H$ associe $a(u, v)$ est linéaire et continue, donc il existe un unique $A(u) \in H$ tel que

$$a(u, v) = \langle A(u), v \rangle \quad \forall v \in H.$$

Montrons que l'application qui à $u \in H$ associe $A(u) \in H$ est linéaire. On a pour $u_1, u_2 \in H$ et $\lambda \in \mathbb{R}$

$$\begin{cases} a(u_1, v) = \langle A(u_1), v \rangle & \forall v \in H, \\ a(u_2, v) = \langle A(u_2), v \rangle & \forall v \in H, \end{cases}$$

donc

$$\langle A(u_1) + \lambda A(u_2), v \rangle = \langle A(u_1), v \rangle + \lambda \langle A(u_2), v \rangle = a(u_1, v) + \lambda a(u_2, v) = a(u_1 + \lambda u_2, v)$$

d'où finalement $A(u_1 + \lambda u_2) = A(u_1) + \lambda A(u_2)$.
Montrons que l'application qui à $u \in H$ associe $A(u) \in H$ est continue.

$$\langle A(u), v \rangle = a(u, v) \leq C \, ||u|| \, ||v|| \quad \forall u, v \in H$$

4. D'autre part, on remarque que u (la solution) et v (la fonction-test) ne sont pas dans le même espace.

où C est le module de continuité de a. Ainsi

$$||A(u)||^2 = \langle A(u), A(u)\rangle \leq C\,||u||\,||A(u)||\,,$$

d'où $||A(u)|| \leq C\,||u||$.
On note enfin que $\langle A(u), u\rangle \geq \alpha\,||u||^2$ où α est le module de coercitivité de a, pour tout $u \in H$. En effet,

$$\langle A(u), u\rangle = a(u,u) \geq \alpha\,||u||^2\,.$$

On doit montrer qu'il existe u tel que

$$\langle A(u), v-u\rangle \geq \langle L, v-u\rangle \quad \forall v \in V,$$

ce qui revient à

$$\langle L - A(u), v-u\rangle \leq 0 \quad \forall v \in V,$$

encore équivalent à

$$\lambda\langle L - A(u), v-u\rangle \leq 0 \quad \forall v \in V$$

si $\lambda > 0$, soit finalement

$$\langle \lambda L - \lambda A(u) + u - u, v-u\rangle \leq 0 \quad \forall v \in V.$$

Cette dernière inégalité signifie que u est la projection de $\lambda L - \lambda A(u) + u$ sur le convexe fermé V, que l'on note

$$u = P_V(\lambda L - \lambda A(u) + u).$$

Soit S l'application qui à $v \in H$ associe $P_V(\lambda L - \lambda A(v) + v)$. Nous cherchons un point fixe de S. Nous allons montrer que S est une contraction si $\lambda \in \mathbb{R}_+^*$ est correctement choisi. Comme P_V est une projection, on a

$$\begin{aligned} ||S(v_1) - S(v_2)|| &\leq ||\lambda L - \lambda A(v_1) + v_1 - (\lambda L - \lambda A(v_2) + v_2)|| \\ &= ||v_1 - v_2 - \lambda\,(A(v_1) - A(v_2))|| \quad \forall v_1, v_2 \in H. \end{aligned}$$

On a ainsi

$$||S(v_1) - S(v_2)||^2 \leq ||v_1 - v_2||^2 + \lambda^2\,||A(v_1) - A(v_2)||^2 - 2\lambda\langle A(v_1) - A(v_2), v_1 - v_2\rangle$$

et, grâce aux propriétés de l'application qui à $u \in H$ associe $A(u)$ (vues en tête de démonstration),

$$||S(v_1) - S(v_2)||^2 \leq ||v_1 - v_2||^2\left(1 + \lambda^2 C^2 - 2\alpha\lambda\right).$$

Or, si $\lambda \in]0, \frac{2\alpha}{C^2}[$, $1 + \lambda^2 C^2 - 2\alpha\lambda \in]0,1[$. Pour un tel coefficient λ, l'application S est une contraction stricte. Donc elle a un unique point fixe dans H (car H est un espace de Banach). Ce point fixe, noté u, vérifie $S(u) = u$, soit

$$u = P_V(\lambda L - \lambda A(u) + u),$$

c'est la solution du problème.
Supposons maintenant que a est symétrique. Cette forme bilinéaire coercitive définit alors un produit scalaire $(\cdot,\cdot)$ équivalent à $\langle\cdot,\cdot\rangle$ sur H. Il existe donc un unique élément de H, noté w, tel que

$$a(w,v) = l(v) \quad \forall v \in H.$$

La solution u du problème de l'énoncé est donc solution de

$$a(u, v-u) \geq a(w, v-u) \quad \forall v \in V,$$

c'est-à-dire

$$a(w-u, v-u) \leq 0 \quad \forall v \in V,$$

u est donc la projection *selon le produit scalaire* $a(\cdot,\cdot)$, notée Π_V, de w sur V. La solution u réalise donc le minimum, pour $v \in V$, de $(a(w-v, w-v))^{1/2}$, c'est-à-dire de $a(v,v) - 2a(w,v) + a(w,w)$ et donc de

$$a(v,v) - 2a(w,v)$$

puisque $a(w,w)$ ne dépend pas de v. La solution u réalise donc le minimum dans V de

$$a(v,v) - 2l(v).$$

C'est ce qu'il fallait démontrer. □

Remarque 47
Le théorème de Lax-Milgram peut se démontrer grâce à ce théorème. En effet, pour ce dernier, on cherche un élément $u \in H$ tel que $a(u,v) = l(v)$ pour tout $v \in H$. H étant un sous-ensemble convexe fermé de H, le théorème de Stampacchia informe du fait qu'il existe un unique élément $u \in H$ tel que

$$a(u, v-u) \geq l(v-u) \quad \forall v \in H.$$

Avec les notations introduites dans la précédente démonstration, u est l'unique solution de

$$\langle L - A(u), \nu v - u\rangle \leq 0 \quad \forall v \in H, \forall \nu \in \mathbb{R}.$$

Donc $\langle L - A(u), v\rangle = 0$ $\forall v \in H$, c'est-à-dire $a(u,v) = l(v)$ pour tout $v \in H$ (il faudrait encore montrer l'unicité d'un tel u). □

Théorème 23
Le problème faible de Dirichlet non homogène (4.5) admet une unique solution. □

On a bien sûr supposé que $f \in L^2(\Omega)$ et que g est la trace d'une fonction $\tilde{g} \in H^1(\Omega)$.

Démonstration
u est solution faible de (4.4) si et seulement si

$$\int_\Omega (\nabla u \cdot (\nabla v - \nabla u) + u(v-u)) \geq \int_\Omega f(v-u) \quad \forall v \in V.$$

En effet :

— si u est solution faible, $v - u \in H_0^1(\Omega)$ pour tout $v \in V$ donc

$$\int_\Omega (\nabla u \cdot (\nabla v - \nabla u) + u(v-u)) = \int_\Omega f(v-u) \quad \forall v \in V\,;$$

— si u vérifie l'inégalité variationnelle, soit $w \in H_0^1(\Omega)$ et posons $v = u + w$: on a $v \in V$ donc

$$\int_\Omega (\nabla u \cdot (\nabla v - \nabla u) + u(v-u)) = \int_\Omega (\nabla u \cdot \nabla w + uw) \geq \int_\Omega f(v-u) = \int_\Omega fw,$$

puis posons $v = u - w$: on a $v \in V$ donc

$$\int_\Omega (\nabla u \cdot (\nabla v - \nabla u) + u(v-u)) = -\int_\Omega (\nabla u \cdot \nabla w + uw) \geq \int_\Omega f(v-u) = -\int_\Omega fw,$$

d'où

$$\int_\Omega (\nabla u \cdot \nabla w + uw) = \int_\Omega fw \quad \forall w \in H_0^1(\Omega).$$

De plus, V est fermé, non vide et convexe (c'est un sous-espace affine de $H^1(\Omega)$), donc le théorème de Stampacchia s'applique, ceci termine la démonstration. □

Remarque 48
On peut démontrer (exercice) que la solution u de (4.5) ne dépend que de g, pas du *relèvement* $\tilde{g}$ de g. □

Le théorème de régularité déjà utilisé permet de montrer qu'en fait, la solution u est une solution forte $L^2(\Omega)$: $u \in V \cap H^2(\Omega)$.

La technique des troncatures de Stampacchia, employée dans la section 4.2.1, permet de démontrer que si $\partial\Omega$ est de classe $\mathcal{C}^2$, u vérifie le principe du maximum

$$\min\left(\text{infess}_{\partial\Omega} g, \text{infess}_\Omega f\right) \leq u \leq \max\left(\text{supess}_{\partial\Omega} g, \text{supess}_\Omega f\right)$$

presque partout dans Ω.

4.2.3 Quelques éléments pour le problème de Neumann homogène

On considère cette fois le problème

$$\begin{cases} -\Delta u + u = f \text{ dans } \Omega, \\ \nabla u \cdot n = 0 \text{ sur } \partial\Omega \end{cases} \tag{4.6}$$

où Ω est un ouvert borné de $\mathbb{R}^d$ de frontière lipschitzienne et de normale unitaire extérieure n. Le second membre f est une donnée du problème.

Si u est une solution forte de (4.6), pour tout $\varphi \in \mathcal{D}(\overline{\Omega})$ on a

$$\int_\Omega (\nabla u \cdot \nabla \varphi + u\varphi) = \int_\Omega f\varphi$$

car $\int_{\partial\Omega} \varphi \nabla u \cdot n = 0$ à cause de la condition de bord. De plus, $\mathcal{D}(\overline{\Omega})$ est dense dans $H^1(\Omega)$, donc on a

$$\int_\Omega (\nabla u \cdot \nabla v + uv) = \int_\Omega fv \quad \forall v \in H^1(\Omega).$$

On dit que u est solution faible de (4.6) si et seulement si

$$u \in H^1(\Omega) \text{ et } \int_\Omega (\nabla u \cdot \nabla v + uv) = \int_\Omega fv \quad \forall v \in H^1(\Omega). \tag{4.7}$$

Théorème 24
Le problème (4.7) a une unique solution. □

Ce résultat est une conséquence immédiate du théorème de Lax-Milgram.

On peut cependant se demander si la forme faible (4.7) est effectivement reliée à la forme forte (4.6). En effet, à la différence du cas des conditions de Dirichlet homogènes, les conditions aux limites n'apparaissent pas explicitement dans l'espace fonctionnel dans lequel on cherche la solution. La solution faible vérifie-t-elle ces conditions ?

Supposons que $\partial\Omega$ est de classe $\mathcal{C}^2$. Alors la solution faible $u \in H^1(\Omega)$ est une solution forte $L^2(\Omega)$, donc $\Delta u \in L^2(\Omega)$ et

$$-\Delta u + u = f \text{ dans } L^2(\Omega).$$

En multipliant cette équation par $v \in H^1(\Omega)$ et en intégrant sur Ω, on obtient

$$\int_\Omega (-v\Delta u + uv) = \int_\Omega fv$$

et, grâce à une formule de Green,

$$-\int_{\partial\Omega} v\nabla u \cdot n + \int_\Omega (\nabla u \cdot \nabla v + uv) = \int_\Omega fv.$$

Or u est solution faible, ce qui signifie que

$$\int_\Omega (\nabla u \cdot \nabla v + uv) = \int_\Omega fv.$$

Donc on a

$$\int_{\partial\Omega} v\nabla u \cdot n = 0 \quad \forall v \in H^1(\Omega).$$

Or l'image de $H^1(\Omega)$ par l'application trace est dense dans $L^2(\partial\Omega)$ (théorème admis ici). On en déduit que

$$\int_{\partial\Omega} \varphi \nabla u \cdot n = 0 \quad \forall \varphi \in L^2(\partial\Omega).$$

Cela entraîne que $\nabla u \cdot n = 0$ presque partout sur $\partial\Omega$.

4.2.4 Méthode des éléments finis

Nous désirons résoudre de manière approchée le problème

$$\text{trouver } u \in V \text{ tel que } a(u,v) = l(v) \quad \forall v \in V$$

où V est un espace de Hilbert, a une forme bilinéaire continue coercitive sur V et l une forme linéaire continue sur V. Le théorème de Lax-Milgram assure l'existence et l'unicité d'une solution à ce problème.

Le principe de la méthode des éléments finis est de résoudre de manière *exacte* le problème

$$\text{trouver } u_h \in V_h \text{ tel que } a(u_h,v_h) = l(v_h) \quad \forall v_h \in V_h$$

où V_h est un sous-espace vectoriel fermé de V. Là encore, le théorème de Lax-Milgram assure l'existence d'une unique solution u_h car V_h est un espace de Hilbert, a y est une forme bilinéaire continue et coercitive, et l y est une forme linéaire continue. Pour que la résolution exacte du problème soit possible dans V_h, on le choisit de dimension finie : ainsi, le problème linéaire à résoudre n'est que celui de l'inversion d'une matrice. Pour que la solution $u_h \in V_h$ soit proche de la solution u (inconnue), il faut garantir que V_h est « proche » de V en un certain sens (à définir). On est amené à définir une suite d'approximations $(V_h)_{h>0}$ de V et on cherche à définir de manière convenable une « convergence » de V_h vers V lorsque h tend vers 0.

Définition 12
On dit que $(V_h)_{h>0}$ est une suite d'approximations *conforme* de V si et seulement si
- pour tout $h > 0$ (assez petit), V_h est un sous-espace de dimension finie de V ;
- pour tout $v \in V$, $\forall \varepsilon > 0$, $\exists \eta \in \mathbb{R}$ tel que pour tout $h \leq \eta$, $\exists v_h \in V_h$ tel que

$$||v - v_h||_V \leq \varepsilon.$$

□

Lemme 12 (lemme de Céa, ou encore second lemme de Strang)
Soit V un espace de Hilbert et soit V_h un sous-espace vectoriel de dimension finie de V. Soit a une forme bilinéaire continue coercitive sur V de module de continuité C et de module de coercitivité α, soit l une forme linéaire continue sur V de module de continuité D. Soit $u \in V$ la solution de

$$a(u,v) = l(v) \quad \forall v \in V.$$

Soit $u_h \in V_h$ la solution de

$$a(u_h,v_h) = l(v_h) \quad \forall v_h \in V_h.$$

Alors

$$||u - u_h||_V \leq \frac{C}{\alpha} \inf_{v_h \in V_h} ||u - v_h||_V .$$

□

Démonstration
On a $a(u, v_h) = l(v_h)$ pour tout $v_h \in V_h$ et $a(u_h, v_h) = l(v_h)$ pour tout $v_h \in V_h$, donc

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

Puisque $u_h - v_h \in V_h$, on en déduit que $a(u - u_h, u_h - v_h) = 0$. Ainsi

$$a(u - u_h, u - v_h) = a(u - u_h, u - u_h) + a(u - u_h, u_h - v_h) = a(u - u_h, u - u_h).$$

Or

$$a(u - u_h, u - v_h) \leq C \, ||u - u_h||_V \, ||u - v_h||_V$$

par continuité de a et

$$a(u - u_h, u - u_h) \geq \alpha \, ||u - u_h||_V^2$$

par coercitivité de a. Donc

$$||u - u_h||_V \leq \frac{C}{\alpha} \, ||u - v_h||_V \quad \forall v_h \in V_h,$$

ce qu'il fallait démontrer. □

C'est un bon départ pour montrer la convergence de u_h vers u lorsque h tend vers 0, mais il reste à trouver une estimation de

$$\inf_{v_h \in V_h} ||u - v_h||_V ,$$

ce qui est en général assez difficile. Nous verrons dans la suite sur un exemple concret (pour le problème $-u'' + u = f$ dans $]0, 1[$) comment le faire. Pour l'instant, gardons le problème abstrait $a(u, v) = l(v)$ et intéressons-nous de plus près au problème discrétisé par les éléments finis associé.

Soit n_h la dimension de V_h et soit $\left(w_h^1, \ldots, w_h^{n_h}\right)$ une base de V_h. Soit $u_h \in V_h$ la solution de $a(u_h, v_h) = l(v_h)$ pour tout $v_h \in V_h$. Notons $\left\{u_{h1}, \ldots, u_{hn_h}\right\}$ les composantes de u_h :

$$u_h = \sum_{i=1}^{n_h} u_{hi} w_h^i,$$

et notons $U_h = (u_{hi})_{i=1}^{n_h}$ le vecteur de $\mathbb{R}^{n_h}$ formé des composantes de u_h dans la base

$$\left(w_h^1, \ldots, w_h^{n_h}\right).$$

Proposition 16
$u_h \in V_h$ est solution de

$$a(u_h, v_h) = l(v_h) \quad \forall v_h \in V_h$$

si et seulement si

$$A_h U_h = B_h$$

où

$$A_{hi,j} = a(w_h^j, w_h^i) \quad \forall i,j \in \{1,\dots,n_h\},$$
$$B_{hi} = l(w_h^i) \quad \forall i \in \{1,\dots,n_h\}.$$

La matrice A_h est définie positive et le système linéaire admet une unique solution. □

Remarque 49

Noter que l'on a

$$A_{hi,j} = a(w_h^j, w_h^i) \quad \forall i,j \in \{1,\dots,n_h\}$$

et non

$$A_{hi,j} = a(w_h^i, w_h^j) \quad \forall i,j \in \{1,\dots,n_h\},$$

ce qui n'est pas la même formule si a n'est pas symétrique [5] ! □

Démonstration

L'équation $a(u_h, v_h) = l(v_h)$ est linéaire, elle est donc vérifiée pour tout $v_h \in V_h$ si et seulement si elle l'est pour chaque élément de la base $\left(w_h^1, \dots, w_h^{n_h}\right)$ de V_h. Donc u_h est solution du problème si et seulement si

$$a\left(\sum_{j=1}^{n_h} u_{hj} w_h^j, w_h^i\right) = l(w_h^i) \quad \forall i \in \{1,\dots,n_h\},$$

ce qui est équivalent, par bilinéarité de a, à

$$\sum_{j=1}^{n_h} u_{hj} a\left(w_h^j, w_h^i\right) = l(w_h^i) \quad \forall i \in \{1,\dots,n_h\}.$$

Ce système s'écrit bien

$$A_h U_h = B_h$$

avec les expressions de A_h et B_h données dans l'énoncé. Noter que l'on sait déjà que ce système est inversible, puisque le problème de trouver $u_h \in V_h$ tel que $a(u_h, v_h) = l(v_h)$ a une unique solution d'après le théorème de Lax-Milgram.

Montrons que A_h est définie positive. Soit $U = (u_i)_{i=1}^{n_h}$ un vecteur de $\mathbb{R}^{n_h}$. On a

$$U^T A_h U = \sum_{i=1}^{n_h}\sum_{j=1}^{n_h} A_{hi,j} u_j u_i = a\left(\sum_{j=1}^{n_h} u_j w_h^j, \sum_{i=1}^{n_h} u_i w_h^i\right) \geq \alpha \left\|\sum_{j=1}^{n_h} u_j w_h^j\right\|_V^2$$

par coercitivité de a. Donc A_h est définie positive. □

5. Cette remarque n'est pas totalement débile, si on ne l'a pas à l'esprit le jour où l'on programme effectivement la méthode on fait la faute.

Un exemple concret d'éléments finis

Pour simplifier, on se place ici en dimension 1 et on étudie le problème

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

(avec $b > 0$ et $c \geq 0$) dont la formulation variationnelle est

$$\begin{array}{l} \text{trouver } u \in H_0^1(]0,1[) \text{ tel que} \\ a(u,v) = l(v) \quad \forall v \in H_0^1(]0,1[) \end{array}$$

avec

$$a(u,v) = \int_0^1 bu'v' + cuv \quad \forall u, v \in H_0^1(]0,1[)$$

et

$$l(v) = \int_0^1 fv \quad \forall v \in H_0^1(]0,1[).$$

Il nous faut construire la suite d'approximations conforme $(V_h)_{h>0}$, c'est-à-dire trouver des fonctions de base de V_h (pour tout $h > 0$). Ces fonctions doivent être dans $H_0^1(]0,1[)$. En dimension 1, cela impose qu'elles soient continues sur $[0,1]$, mais cela laisse encore un choix très large... On pourrait par exemple penser à définir V_h comme l'espace des polynômes de degré inférieur ou égal à $n_h - 1$. Cependant, la matrice A_h serait alors constituée d'intégrales de polynômes et (sauf cas très particulier) aucun des coefficients de A_h ne serait nul, ce qui rendrait son inversion difficile. On écarte pour cette raison ce choix (ici) au profit de fonctions de base qui rendront la matrice A_h creuse. Puisque

$$A_{hi,j} = \int_0^1 bw_h^{i\,\prime} w_h^{j\,\prime} + cw_h^i w_h^j,$$

cela revient à trouver des fonctions w_h^i dont les supports soient d'intersections « petites ». Un tel exemple de fonctions de base est fourni par les *éléments* $P1$ définis comme suit. On pose, pour tout $i \in \{1, \ldots, n_h\}$,

$$w_h^i(x) = (1 - |(n_h + 1)x - i|)_+ \quad \forall x \in [0,1]$$

où $(y)_+$ est la partie positive de $y \in \mathbb{R} : (y)_+ = \max(0, y)$. Pour $n_h = 9$, on visualise quelques-unes de ces fonctions de base sur la figure 4.1.

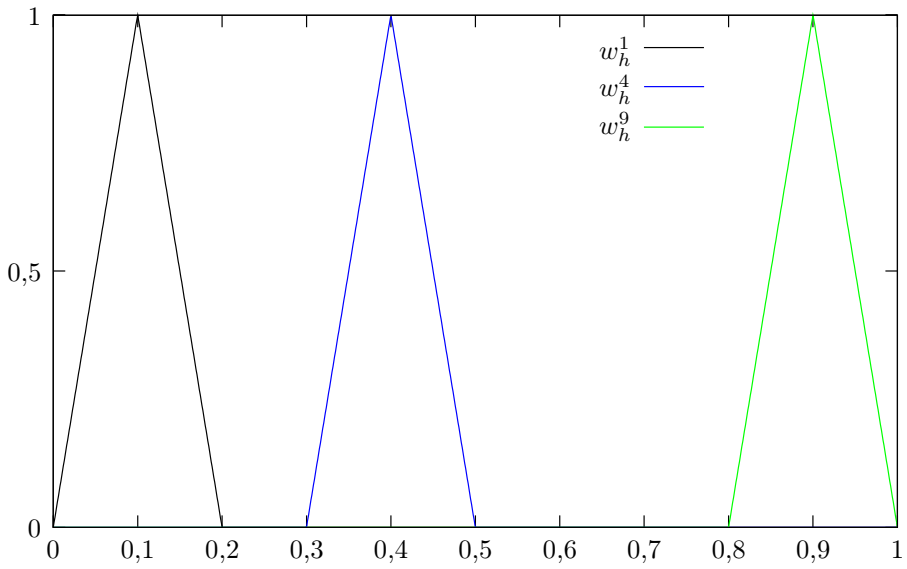

FIGURE 4.1 – Quelques fonctions de base des éléments finis $P1$.

Pour tout $n_h \in \mathbb{N}$, on a ainsi n_h fonctions, et on pose $V_h = vect\{w_h^1, \ldots, w_h^{n_h}\}$.

Lemme 13
V_h est l'ensemble des fonctions continues sur $[0,1]$, nulles en 0 et en 1 et affines sur chaque intervalle du type $[\frac{i}{n_h+1}, \frac{i+1}{n_h+1}]$ pour $i \in \{0, \ldots, n_h\}$. C'est un espace vectoriel de dimension n_h. □

Lemme 14
Les fonctions de V_h sont entièrement déterminées par les valeurs qu'elles prennent aux points du maillage[6] $\frac{i}{n_h+1}$ pour $i \in \{1, \ldots, n_h\}$. □

La démonstration de ces lemmes est (l)ai(s)sée (en exercice).

Lemme 15
Pour tout $n_h \in \mathbb{N}^*$, $V_h \subset V = H_0^1(]0,1[)$. □

Démonstration
Il faut montrer que $V_h \subset L^2(]0,1[)$ et que la dérivée au sens des distributions d'une fonction de V_h peut être représentée par une fonction de $L^2(]0,1[)$.
Le fait que $V_h \subset L^2(]0,1[)$ est une conséquence immédiate de la continuité des fonctions de V_h.
Soit maintenant $v \in V_h$. Cette fonction est dérivable sur chaque intervalle du type $]\frac{i}{n_h+1}, \frac{i+1}{n_h+1}[$:

6. Le maillage en question est un maillage régulier de $[0,1]$ de pas $\frac{1}{n_h+1}$... Que l'on assimilera à h puisque h est un paramètre devant tendre vers 0.

notons v_i cette dérivée (constante) sur cet intervalle. Soit $\varphi \in \mathcal{D}(]0,1[)$.

$$\int_0^1 v\varphi' = \sum_{i=0}^{n_h} \int_{\frac{i}{n_h+1}}^{\frac{i+1}{n_h+1}} v\varphi' = \sum_{i=0}^{n_h} \left([v\varphi]_{\frac{i}{n_h+1}}^{\frac{i+1}{n_h+1}} - \int_{\frac{i}{n_h+1}}^{\frac{i+1}{n_h+1}} \varphi v' \right)$$
$$= \sum_{i=0}^{n_h} [v\varphi]_{\frac{i}{n_h+1}}^{\frac{i+1}{n_h+1}} - \sum_{i=0}^{n_h} \int_{\frac{i}{n_h+1}}^{\frac{i+1}{n_h+1}} \varphi v_i = v(1)\varphi(1) - v(0)\varphi(0) - \int_0^1 \varphi v' = -\int_0^1 \varphi v'$$

où l'on a posé

$$v'(x) = \sum_{i=0}^{n_h} v_i \mathbb{1}_{]\frac{i}{n_h+1}, \frac{i+1}{n_h+1}]}(x).$$

Cette fonction est un élément de $L^2(]0,1[)$, donc v est bien un élément de $H^1(]0,1[)$. D'autre part, $v(0) = v(1) = 0$, donc $v \in H_0^1(]0,1[)$. □

Le sous-espace V_h étant choisi, il nous reste à calculer la matrice A_h.

Lemme 16
Avec les éléments $P1$ définis ci-dessus, la matrice du système linéaire à résoudre est

$$A_h = \begin{pmatrix} a_h & b_h & 0 & \cdots & \cdots & \cdots & 0 \\ b_h & a_h & b_h & 0 & \cdots & \cdots & 0 \\ 0 & b_h & a_h & b_h & 0 & \cdots & 0 \\ \vdots & 0 & b_h & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & b_h \\ 0 & 0 & 0 & \cdots & 0 & b_h & a_h \end{pmatrix}$$

où l'on a posé $h = \frac{1}{n_h+1}$ et $a_h = \frac{2b}{h} + \frac{2ch}{3}$ et $b_h = -\frac{b}{h} + \frac{ch}{6}$. □

La démonstration de ce résultat est encore laissée en exercice (elle consiste en l'intégration de polynômes de degré au plus 2).

Nous allons maintenant montrer que la méthode des éléments finis converge (dans notre cas particulier).

Théorème 25
Soit $f \in L^2(]0,1[)$. Soit $u \in H_0^1(]0,1[)$ la solution faible (forte $L^2(]0,1[)$) de

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

et soit u_h la solution (éléments finis $P1$) de

$$A_h U_h = B_h$$

décrite ci-dessus, avec $n_h = 1/h - 1 \in \mathbb{N}^*$.
Alors, il existe $D \in \mathbb{R}$ ne dépendant pas h tel que

$$||u - u_h||_{H^1(]0,1[)} \leq Dh.$$

□

Démonstration
On sait déjà, d'après le lemme 12, que

$$||u - u_h||_{H^1(]0,1[)} \leq \frac{C}{\alpha} \inf_{v_h \in V_h} ||u - v_h||_{H^1(]0,1[)}$$

(avec la définition de V_h utilisée ci-dessus). Nous allons maintenant trouver une majoration utile de

$$\inf_{v_h \in V_h} ||u - v_h||_{H^1(]0,1[)}.$$

Supposons d'abord que la solution u est de classe $\mathcal{C}^2([0,1])$. Considérons la fonction $I_h u$ définie sur $[0,1]$ comme la fonction continue sur $[0,1]$, affine sur tout intervalle du type $[ih,(i+1)h]$ pour $i \in \{0,\dots,n_h\}$ et telle que $I_h u(ih) = u(ih)$ pour tout $i \in \{0,\dots,n_h+1\}$. C'est la fonction interpolée de u aux points du maillage. On a $I_h u \in V_h$ et

$$I_h u = \sum_{i=1}^{n_h} u(ih) w_h^i.$$

Sur chaque intervalle du type $]ih,(i+1)h[$ $(i \in \{1,\dots,n_h\})$, on a

$$(I_h u)'(x) = \frac{u((i+1)h) - u(ih)}{h},$$

donc

$$(u - I_h u)'(x) = u'(x) - \frac{u((i+1)h) - u(ih)}{h}.$$

Or, u étant supposée de classe $\mathcal{C}^2$, on a

$$u((i+1)h) - u(x) = ((i+1)h - x)\, u'(x) + \int_x^{(i+1)h} u''(y)((i+1)h - y)\, dy$$

et

$$u(ih) - u(x) = (ih - x)\, u'(x) + \int_x^{ih} u''(y)(ih - y)\, dy,$$

d'où

$$u((i+1)h) - u(ih) = hu'(x) + \int_x^{(i+1)h} u''(y)((i+1)h - y)\, dy - \int_x^{ih} u''(y)(ih - y)\, dy.$$

D'après l'inégalité de Cauchy-Schwarz,

$$\left| \int_x^{(i+1)h} u''(y)((i+1)h - y)\, dy \right|^2 \leq \frac{|(i+1)h - x|^3}{3} \int_x^{(i+1)h} u''(y)^2\, dy$$

et

$$\left| \int_x^{ih} u''(y)(ih - y)\, dy \right|^2 \leq \frac{|x - ih|^3}{3} \int_{ih}^x u''(y)^2\, dy.$$

On en déduit que pour tout $x \in]ih, (i+1)h[$,

$$\left| u'(x) - \frac{u((i+1)h) - u(ih)}{h} \right|^2$$
$$\leq 2 \left(\frac{|(i+1)h - x|^3}{3h^2} \int_x^{(i+1)h} u''(y)^2\, dy + \frac{|x - ih|^3}{3h^2} \int_{ih}^x u''(y)^2\, dy \right) \leq \frac{2h}{3} \int_{ih}^{(i+1)h} u''(y)^2\, dy.$$

Ainsi,

$$\int_{ih}^{(i+1)h} \left((u - I_h u)'(y) \right)^2\, dy \leq \frac{2h^2}{3} \int_{ih}^{(i+1)h} u''(y)^2\, dy$$

et en sommant sur $i \in \{1, \ldots, n_h\}$ on obtient

$$\int_0^1 \left((u - I_h u)'(y) \right)^2\, dy \leq \frac{2h^2}{3} \int_0^1 u''(y)^2\, dy.$$

Ceci est vrai si $u \in \mathcal{C}^2([0,1])$. Par densité de $\mathcal{C}^2([0,1])$ dans $H^2(]0,1[)$, cela reste vrai pour la solution générale $u \in H_0^1(]0,1[) \cap H^2(]0,1[)$ du problème de Dirichlet homogène avec $f \in L^2(]0,1[)$.

On peut soit faire un raisonnement similaire soit utiliser l'inégalité de Poincaré pour en déduire une majoration du même type pour $\int_0^1 \left((u - I_h u)(y) \right)^2\, dy$. Cela clôt la démonstration. □

4.2.5 Quelques résultats numériques

Présentons maintenant l'application de la méthode, décrite ci-dessus, des éléments finis $P1$ sur le segment $[0,1]$ pour le problème $-\partial_{x,x}^2 u + u = 1$ avec des conditions de Dirichlet homogènes. La solution exacte est alors donnée par

$$u(x) = 1 - \frac{e^x + e^{1-x}}{1 + e} \quad \forall x \in [0,1].$$

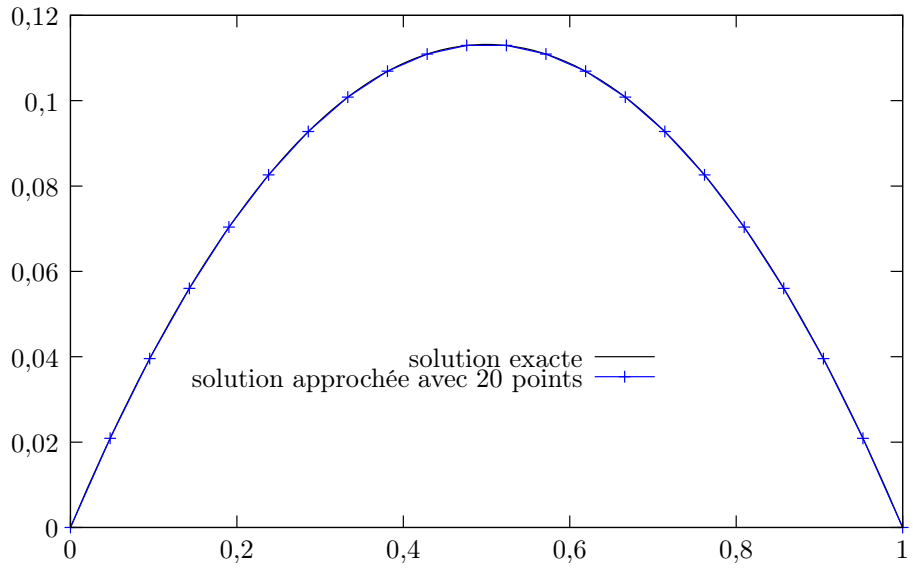

FIGURE 4.2 – Résultat obtenu avec les éléments finis $P1$.

Dans ce cas, où la solution est très régulière et symétrique, l'approximation donnée par les éléments finis avec seulement 20 mailles est déjà excellente. Voici d'autres résultats, avec le second membre

$$f(x) = \begin{cases} 1 \text{ si } x \in [0, 1/3[, \\ 0 \text{ si } x \in [1/3, 2/3[, \\ 2 \text{ si } x \in [2/3, 1]. \end{cases}$$

Un calcul un peu fastidieux (et donc pas inutile) montre que la solution est alors

$$u(x) = \begin{cases} 1 + \alpha e^x + \beta e^{-x} \text{ si } x \in [0, 1/3[, \\ \gamma e^x + \delta e^{-x} \text{ si } x \in [1/3, 2/3], \\ 2 + \epsilon e^x + \zeta e^{-x} \text{ si } x \in]2/3, 1] \end{cases}$$

avec

$$\begin{cases} \zeta = \dfrac{-2 - \frac{g}{d}e}{\frac{1}{e} - \frac{f}{d}e} \\ \epsilon = \dfrac{g - f\zeta}{d} \\ \delta = \dfrac{2 - \frac{c}{a}e^{2/3} + \zeta e^{-2/3} + \epsilon e^{2/3}}{e^{-2/3} - \frac{b}{a}e^{2/3}} \\ \gamma = \dfrac{c - b\delta}{a} \\ \beta = \dfrac{e^{1/3} - 1 + \delta e^{-1/3} + \gamma e^{1/3}}{e^{-1/3} - e^{1/3}} \\ \alpha = -1 - \beta, \end{cases}$$

où les coefficients a, b, c, d, f et g sont donnés par

$$\begin{cases} a = -e^{1/3} - e^{1/3}\dfrac{e^{-1/3}+e^{1/3}}{e^{-1/3}-e^{1/3}} \\ b = e^{-1/3} - e^{-1/3}\dfrac{e^{-1/3}+e^{1/3}}{e^{-1/3}-e^{1/3}} \\ c = e^{1/3} + (e^{1/3}-1)\dfrac{e^{-1/3}+e^{1/3}}{e^{-1/3}-e^{1/3}} \\ d = \dfrac{2}{e^{-2/3} - \frac{b}{a}e^{2/3}} \\ f = -2e^{-2/3} + e^{-2/3}\dfrac{2e^{-2/3}}{e^{-2/3} - \frac{b}{a}e^{2/3}} \\ g = 2 - 2e^{-2/3}\dfrac{2 - \frac{c}{a}e^{2/3}}{e^{-2/3} - \frac{b}{a}e^{2/3}} \end{cases}$$

(je n'en mettrais pas ma main à couper).

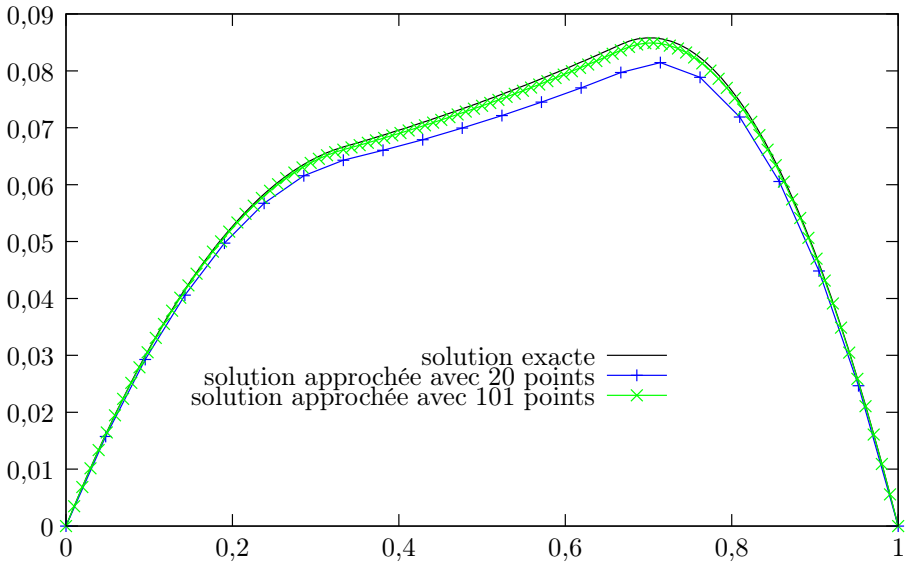

FIGURE 4.3 – Résultat obtenu avec les éléments finis $P1$.

Il est clair sur ce cas-test que l'approximation est d'autant meilleure que le nombre de points est grand. Ces résultats ont été calculés à l'aide du programme (en langage Scilab) suivant.

```
//////////////////////////////////////////////////////////////
////////////// Éléments finis P1 pour l'équation /////////////
///////////////          −Δ u + u = f.          ////////////
//////////////////////////////////////////////////////////////

clear();
```

```
stacksize(100000000);

/////////////// Paramètres. ///////////////

pi = 3.14159;
Xmax = 1;               // Longueur de l'intervalle.
M = 20;                 // Nombre de points du maillage
                        // (à l'intérieur du segment).
dx = Xmax/(M+1);        // Pas en espace.
x=(1:M)'*dx;             // Points de la grille en x.

///////// Remplissage de la matrice creuse A. /////////

a = 2./dx + 2.*dx/3.;
b = dx/6. - 1./dx;
A = zeros(M,M);

for i=2:M-1,
  A(i,i) = a;
  A(i,i-1) = b;
  A(i,i+1) = b;
end;
A(1,1) = a;
A(1,2) = b;
A(M,M) = a;
A(M,M-1) = b;

A = sparse(A);

/////// Création du vecteur B (second membre). //////

B = zeros(M,1);

for i=1:ceil((M+1)/3)-1,
  B(i,1) = dx;
end;
B(ceil((M+1)/3),1) = dx/2.;
for i=ceil((M+1)/3)+1:ceil(2*(M+1)/3)-1,
```

```
  B(i,1) = 0.;
end;
B(ceil(2*(M+1)/3),1) = dx/2.;
for i=ceil(2*(M+1)/3)+1:M,
  B(i,1) = 2.*dx;
end;
//for i=1:M,
//  B(i,1) = dx;
//end;

// On factorise la matrice A pour pouvoir résoudre
// les systèmes linéaires plus rapidement.
[factA,rk] = lufact(A);

u = zeros(M,1);      // vecteur de la solution numérique.

u = lusolve(factA,B); // Résolution du système linéaire.

xset("wwpc"); // Options d'affichage de la solution.
plot(x,u);
xset("wshow");

unix('rm -f resultat');                    // Écriture du résultat
plouf = file('open','resultat','unknown'); // dans le fichier
for i=1:M,                                 // << resultat >> : un peu
                                           // d'originalité ne
                                           // fait pas de mal.
  fprintf(plouf,'%f   %f\n',xfin(i),ufin(i));
end;
file('close',plouf);
                            // Fin du programme...
                     // Qui semble coïncider avec la fin du cours !
```

Bibliographie

[1] H. Brézis : Analyse Fonctionnelle, Masson, 1993.

[2] P.-G. Ciarlet : Introduction à l'analyse numérique matricielle et à l'optimisation, Masson, 1994.

[3] M. Crouzeix et A. L. Mignot, Analyse numérique des équation différentielles, Masson, 1992.

[4] G. Duvaut : Mécanique des Milieux Continus, Masson, 1990.

[5] E. Godlewski et P.-A. Raviart : Hyperbolic systems of conservation laws, Ellipses, 1991.

[6] L. Hörmander : Lectures on Nonlinear Hyperbolic Differential Equations, Springer, 1997.

[7] B. Lucquin : Équations aux dérivées partielles et leurs approximations, Ellipses, 2004.

[8] P.-A. Raviart et J.-M. Thomas : Introduction à l'analyse numérique des équations aux dérivées partielles, Masson, 1983.

[9] H. Reinhard : Équations différentielles, Dunod, 1989.

[10] W. Rudin : Analyse Fonctionnelle, Édiscience, 1995.

[11] L. Schwartz : Méthodes mathématiques pour les sciences physiques, Hermann, 1965.

[12] L. Schwartz : Théorie des distributions, Hermann, 1966.

[13] D. Serre : Matrices, Theory and Applications, Springer, 2002.

[14] D. Serre : Systèmes de lois de conservation (I), Diderot, 1996.

[15] John C. Strikwerda : Finite Difference Schemes and Partial Differential Equations, SIAM, 2004.

Index