Préparation à l'agrégation externe de mathématiques

Épreuve de modélisation – option B : calcul scientifique

Introduction aux équations aux dérivées partielles

Guillaume Delay

Laboratoire Jacques-Louis Lions

Sorbonne Université

26 avril 2022

Avant-propos

Le but de ce cours est de proposer une introduction à la théorie des équations aux dérivées partielles (EDP dans la suite). Nous étudierons plusieurs équations ainsi que leur discrétisation par la méthode des différences finies.

Dans le cadre du programme officiel de l'agrégation de mathématiques (épreuve de modélisation, option B : calcul scientifique), nous aborderons notamment :

— des notions élémentaires portant sur les EDP classiques en dimension 1.

— l'équation de transport linéaire avec la méthode des caractéristiques.

— l'équation des ondes et l'équation de la chaleur. Une résolution par série de Fourier et transformée de Fourier sera proposée ainsi qu'une méthode de séparation des variables. Les aspects qualitatifs seront abordés.

— les équations elliptiques avec l'utilisation du théorème de Lax–Milgram.

— des exemples de discrétisation des EDP en dimension 1 avec la méthode des différences finies. L'étude des propriétés de ces discrétisations sera proposée : notions de consistance, stabilité, convergence et d'ordre.

Vous êtes par ailleurs invités à lire le rapport du jury (disponible sur internet). Vous vous rendrez compte que le jury insiste notamment sur le fait que :

— l'épreuve de modélisation, comme les autres, requiert une démarche rigoureuse de la part des candidats.

— il faut équilibrer sa présentation entre une présentation du modèle étudié, des preuves mathématiques rigoureuses et des illustrations informatiques.

— il attend une prise de recul de la part des candidats. Il faudra donc notamment être capable de critiquer les limites du modèle présenté dans le texte, d'expliquer le comportement qualitatif de celui-ci (par exemple expliquer ce qu'il se passe quand la valeur d'un paramètre change) et être capable de conclure sur la problématique de départ.

Ce cours sera composé de :

— quatre séances de cours de trois heures chacune.

— une séance de programmation de trois heures.

Dans une première partie, nous présenterons les équations étudiées dans ce cours ainsi que les problèmes physiques associés. Chacune des parties suivantes sera consacrée à l'étude plus approfondie d'une EDP. Nous présenterons notamment les principales caractéristiques de cette EDP, les outils d'analyse utilisés ainsi qu'une discrétisation par différences finies. Les EDP étudiées dans la suite seront les équations elliptiques, l'équation de transport, l'équation de la chaleur et enfin l'équation des ondes. Une dernière section sera consacrée à des éléments de cours hors-programme destinés aux candidats souhaitant approfondir leurs connaissances sur ce sujet.

1 Présentation des EDP du cours

Nous présentons dans cette section les EDP étudiées dans la suite du cours. Nous essayons de donner une signification physique aux différents termes. Nous nous intéressons à des EDP de la forme

$$a\frac{\partial^2 u}{\partial x^2} + b\frac{\partial^2 u}{\partial x \partial y} + c\frac{\partial^2 u}{\partial y^2} + d\frac{\partial u}{\partial x} + e\frac{\partial u}{\partial y} + fu = F, \tag{1}$$

où a, b, c, d, e et f sont des réels et où F et u sont des fonctions de x et y.

Une partie du comportement qualitatif de l'EDP peut être déterminée à partir de la valeur de ces coefficients. Considérons l'équation

$$ax^2 + bxy + cy^2 + dx + ey + f = A, \tag{2}$$

avec A un réel tel que l'ensemble des solutions soit non vide. S'il s'agit de l'équation :

— d'une ellipse, on dira que l'équation est elliptique.

— d'une parabole, on dira que l'équation est parabolique.

— d'une hyperbole, on dira que l'équation est hyperbolique.

Cette dénomination n'est pas juste esthétique. En effet, comme nous le verrons plus loin dans ce cours, chacun de ces types d'équations dispose de propriétés spécifiques.

Dans l'équation (1) nous avons considéré un problème qui dépend de deux variables x et y. Les notions d'EDP elliptiques, paraboliques et hyperboliques peuvent aussi être généralisées à un plus grand nombre de variables. Dans le cadre de ce cours, nous nous concentrerons sur l'étude d'équations avec une seule dimension d'espace. On considérera donc une seule variable x dans le cas d'un problème stationnaire et deux variables t (le temps) et x (l'espace) dans le cas d'un problème instationnaire.

Remarque 1. *Les démonstrations de cette section ne sont pas exigibles. On demande simplement aux candidats de se représenter à quoi correspondent les équations et les paramètres introduits. Aucune connaissance en physique n'est requise pour cette section qui peut être lue indépendamment des autres. Les candidats doivent cependant avoir à l'esprit que tous les textes comportent une part de modélisation et que les modèles présentés dans cette section sont classiques (donc susceptibles d'être rencontrés dans un texte).*

1.1 Équations elliptiques

Nous nous intéressons dans cette section à des équations de la forme

$$-\frac{\mathrm{d}}{\mathrm{d}x}\left(k\frac{\mathrm{d}u}{\mathrm{d}x}\right) = f, \tag{3}$$

pour une dimension d'espace x. Si l'on considère plusieurs dimensions d'espace, cette équation devient

$$-\nabla \cdot (k\nabla u) = f, \tag{4}$$

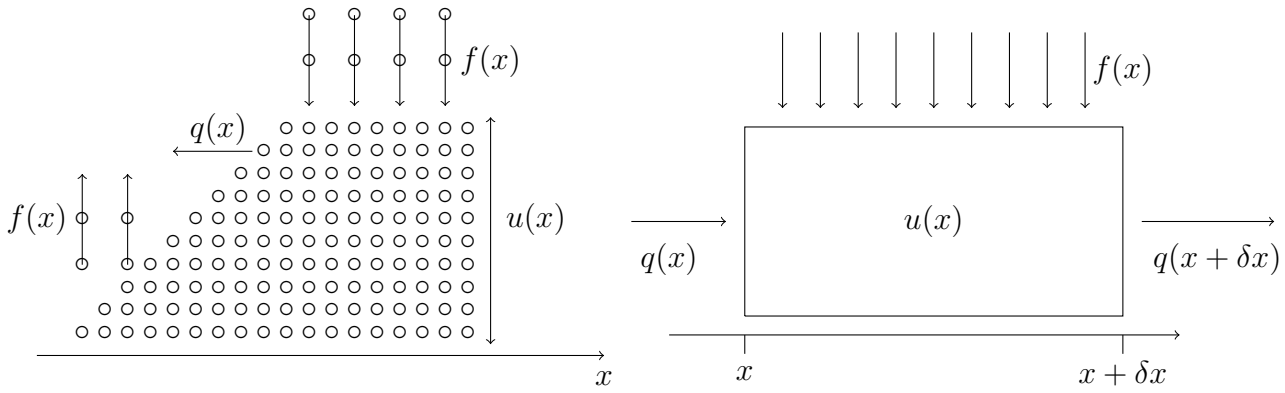

FIGURE 1 – Répartition de particules dans un domaine. Gauche : représentation du problème. Droite : équilibre des flux sur une portion infinitésimale du domaine.

où $\nabla\cdot$ et ∇ sont respectivement les opérateurs divergence et gradient. Dans ces deux équations, k est un coefficient qui peut éventuellement dépendre des variables d'espace et f est une fonction de l'espace.

Nous présentons deux problèmes physiques qui font intervenir cette équation. Le premier est le cas où des particules circulent dans un domaine. Le second est un problème d'équilibre mécanique.

Nous représentons dans la partie gauche de la figure 1 le premier problème. Nous nous intéressons à des particules qui circulent dans un milieu unidimensionnel. La position est repérée par la coordonnée d'espace x. On note $u(x)$ la densité de particules en x. Certaines particules entrent ou sortent du domaine en x, on note $f(x)$ le terme source représentant cette variation ($f(x)$ sera positif si des particules apparaissent en x et négatif si elles disparaissent). De plus, les particules se déplacent à travers le domaine, on note $q(x)$ le flux de particules en x (le nombre de particules qui traversent l'axe vertical d'abscisse x par unité de temps). Ce flux est positif si les particules vont vers la droite et négatif si elles vont vers la gauche.

On s'intéresse au cas où les flux sont à l'équilibre, il n'y a donc pas d'accumulation de particules en aucun point de l'espace. Le problème ne dépend donc pas du temps.

Pour établir les équations de ce problème, considérons une portion infinitésimale du domaine (de taille δx) comme représenté sur la droite de la figure 1. Le nombre de particules à l'intérieur de cette section doit rester constant. On obtient donc la relation de conservation $q(x) - q(x+\delta x) + f(x)\delta x = 0$, ce qui donne la relation

$$\frac{\mathrm{d}q}{\mathrm{d}x} = f. \tag{5}$$

De plus, on considère que les particules fuient les zones de forte densité : le flux $q(x)$ suit la direction opposée au gradient de u. On note donc

$$q(x) = -k(x)\frac{\mathrm{d}u}{\mathrm{d}x}(x). \tag{6}$$

Ici k est un coefficient positif qui peut dépendre de l'espace. Il traduit le rapport de proportionnalité entre le gradient de la densité et le flux qui en résulte. Ainsi, pour une densité fixée, si k est grand alors les particules circuleront facilement et le flux sera important ; à

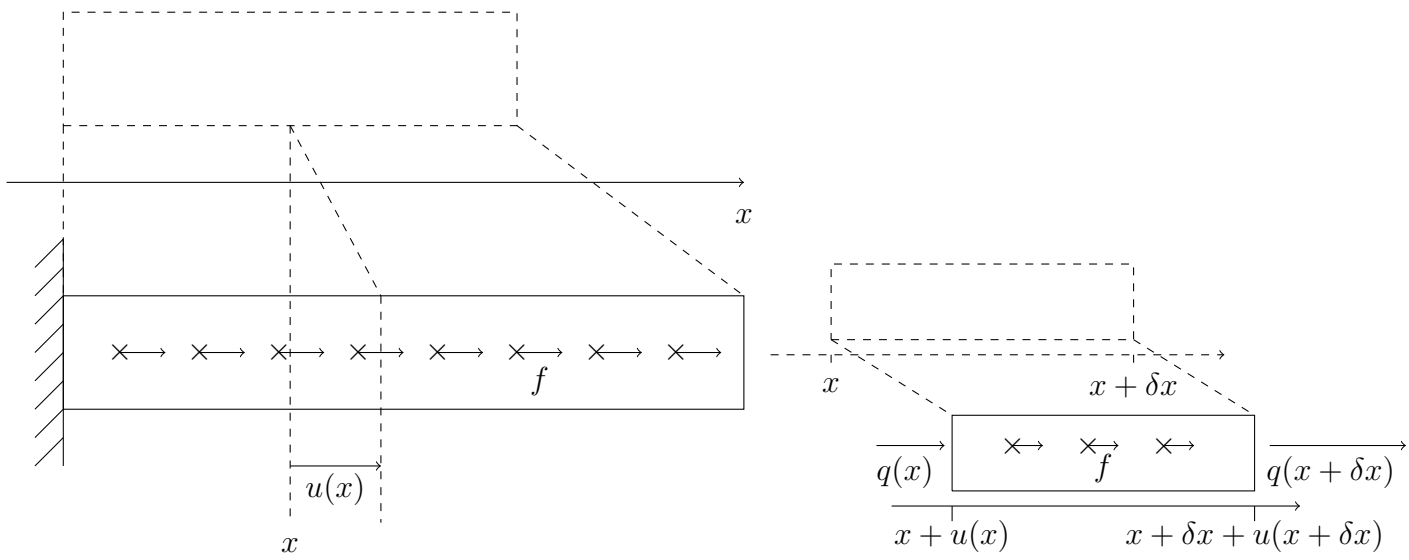

FIGURE 2 – Une barre élastique en équilibre. Gauche : représentation du problème. Droite : déformation d'un élément infinitésimal de matière. Le trait continu représente la configuration soumise à une charge f. Le trait discontinu représente la configuration au repos (sans f).

l'inverse, un k petit traduit le fait que les particules ont du mal à circuler dans le milieu. L'équation finale sur u est donc (3).

En pratique les particules que nous avons considérées peuvent représenter par exemple des molécules. Dans ce cas u sera une concentration chimique, q un flux de molécules, f un terme représentant leur apparition ou leur disparition (par des réactions chimiques) et k sera un coefficient déterminant la facilité avec laquelle les molécules se déplacent.

On peut aussi dire que les particules représentent de l'énergie thermique qui se propage à travers un matériau. Dans ce cas, u sera la température , q un flux thermique, f une source ou un puit de chaleur et k sera la conductivité thermique du matériau considéré.

Nous citons une dernière possibilité selon laquelle les particules sont des individus (humains ou animaux). Dans ce cas, u correspond à une densité de population, q à un flux de population, f représente les naissances et morts et k est un coefficient représentant la facilité avec laquelle la population peut se déplacer.

Nous présentons maintenant un autre problème physique faisant intervenir des équations elliptiques. Considérons un matériau soumis à des contraintes mécaniques. Par exemple, on représente sur la figure 2 une barre élastique en équilibre.

La barre dans son état initial est représentée en pointillés. Sous l'effet d'une force linéique f, cette barre s'allonge et atteint l'état d'équilibre représenté en trait continu. On note $u(x)$ le déplacement de la matière qui a eu lieu en x entre la configuration au repos et la configuration soumise à la charge f.

On considère que la barre est à l'équilibre mécanique. On note $q(x)$ la force qu'exerce la section de gauche sur la section de droite en x. En faisant un bilan de force comme représenté sur la partie droite de la figure 1, les forces s'exerçant sur une portion infinitésimale de barre sont la force $q(x)$ à gauche, la force $-q(x+\delta x)$ à droite et la force linéique $f(x)\delta x$. La barre étant à l'équilibre la somme de ces forces est nulle. On retrouve donc (5).

De plus, la force s'exerçant en x à travers la section de la barre est proportionnelle à l'élongation de la barre et s'oppose au mouvement imposé. Ceci est intuitif, pensez à un élastique : si vous l'allongez, alors il s'exerce une force qui tend à le faire revenir vers sa position initiale. De plus, plus l'élongation est importante, plus l'intensité de la force est grande. On obtient donc la loi d'élasticité (6) où k est un coefficient de raideur : plus k est grand, plus la barre est raide (plus il faut forcer pour la déformer). En pratique, le coefficient de raideur dépend du matériau choisi et de la géométrie de la section de la barre.

Notons que dans (6), la dérivée en espace correspond bien à l'élongation de la barre en x. Pour s'en convaincre, on regardera la partie droite de la figure 2. Un élément de matière de longueur δx dans sa position de référence a pour longueur $x + \delta x + u(x + \delta x) - u(x) - x$ sous charge f. La nouvelle longueur est donc de $\delta x + u(x + \delta x) - u(x) \simeq \left(1 + \dfrac{\partial u}{\partial x}\right) \delta x$ et la dérivée partielle de u en x est donc bien une élongation par unité de longueur.

D'autres problèmes physiques peuvent être modélisés par des équations elliptiques. Nous citerons simplement l'électromagnétisme sans donner plus de détails. Nous verrons par la suite que, dans le cas instationnaire, les deux problèmes exposés ici correspondent à des équations de nature différente. Le premier est représenté par une équation parabolique en instationnaire, tandis que la deuxième est représenté par une équation hyperbolique.

Évoquons maintenant les conditions aux limites les plus classiques que l'on peut associer à ce problème. Tout d'abord, nous pouvons considérer la condition de Dirichlet $u = g$ où g est une donnée du problème. Ceci revient à imposer la valeur de la solution u sur le bord du domaine. Par exemple, dans le cas de l'équation de la chaleur stationnaire, la condition de Dirichlet revient à considérer que le bord du domaine correspond à un élément à forte capacité thermique dont la température restera constante quoi qu'il arrive.

Une autre condition aux limites classique est la condition de Neumann $q = g$ où g est une donnée. Ceci revient à imposer la valeur du flux q sur le bord du domaine. Dans la majorité des cas, la donnée de Neumann g est nulle. Par exemple, dans le cas du problème de la chaleur, ceci revient à considérer que le bord du domaine est adiabatique : quoi qu'il arrive aucun flux de chaleur ne traversera le bord du domaine (pensez par exemple à une bouteille isotherme).

Ces deux conditions aux limites peuvent éventuellement être utilisées simultanément en des points différents du bord (Dirichlet sur une partie de la frontière et Neumann sur le reste). On parle alors de conditions aux limites mixtes.

	Chaleur stationnaire	Barre élastique	Molécules
u	température	déplacement	concentration
$q = -k\nabla u$	flux de chaleur	force interne	flux de mol.
k	conductivité thermique	coeff. de raideur	coeff. de diffusion
f	source/puit de chaleur	force externe	réactions chimiques
$u = g$	temp. constante	déplacement imposé	concentration imposée
$q = 0$	paroi adiabatique	frontière libre	frontière imperméable

TABLE 1 – Sens physique de la représentation mathématique en fonction du problème traité.

Exercice 1. *On se place dans le cas d'un domaine bidimensionnel $\Omega \subset \mathbb{R}^2$. On note x et y les deux variables d'espace. On suppose que k est constant : pour tout $(x, y) \in \Omega$, $k(x, y) = k_0 > 0$. Montrer que l'équation (4) est elliptique.*

Remarque 2. *Dans la plupart des applications, k est une constante. Prenons par exemple $k = 1$. L'équation (4) devient $-\Delta u = f$. On appelle cette équation équation de Laplace ou équation de Poisson. Pour simplifier la présentation, c'est cette équation que nous étudierons par la suite.*

Pour finir cette section, nous illustrons le comportement de la solution de (3) en fonction de k (voir figure 3). Comme nous l'avons évoqué précédemment, une solution avec un k plus grand est plus "plate" (par exemple dans le cas des molécules, une plus grande diffusion entraîne une plus grande homogénéité de la concentration).

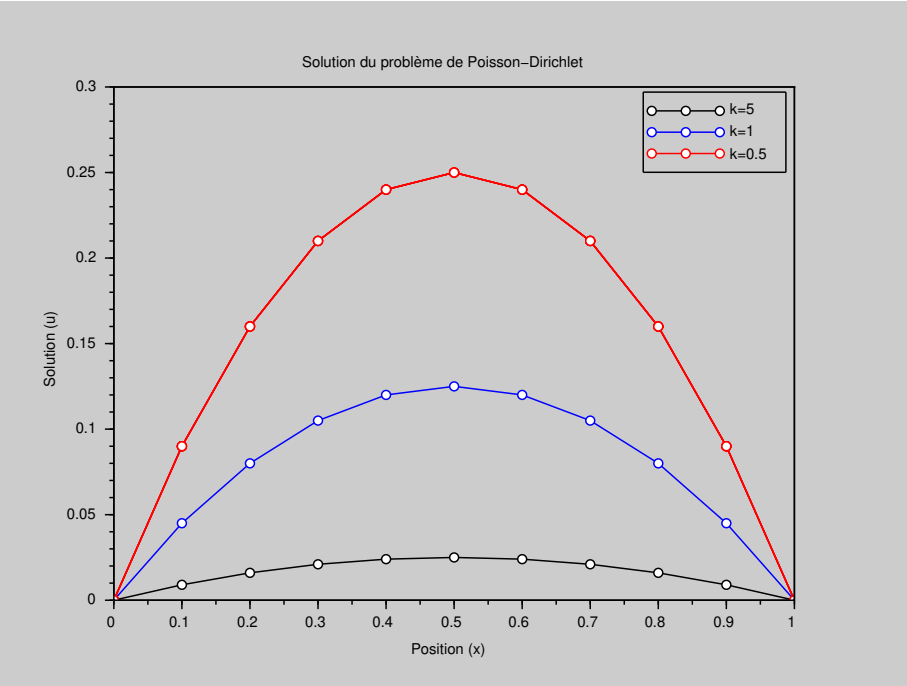

FIGURE 3 – Solution de l'équation elliptique 1D (3) pour différents k (pour $f(x) = 1$) avec les conditions de Dirichlet homogènes $u(0) = u(1) = 0$.

1.2 Équation de transport

La deuxième équation que nous étudions est l'équation de transport (unidimensionelle). Elle correspond à une quantité qui est transportée à vitesse constante dans une direction. Cela peut être par exemple un polluant transporté par une rivière. Dans cette section, nous nous intéressons au cas de tas de sable sur un tapis roulant se déplaçant à vitesse constante a.

Sur la figure 4, nous représentons des tas de sables qui se déplacent à vitesse constante a vers la droite. On note $u(t, x)$ la hauteur du sable en x à l'instant t.

Si l'on considère un temps infinitésimal δt, le sable aura avancé d'une distance $\delta x = a\delta t$. La hauteur en $(t + \delta t, x + \delta x)$ sera la même qu'elle était en (t, x), on traduit cela par $u(t + \delta t, x + a\delta t) = u(t, x)$. On obtient ainsi l'équation de transport

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0. \tag{7}$$

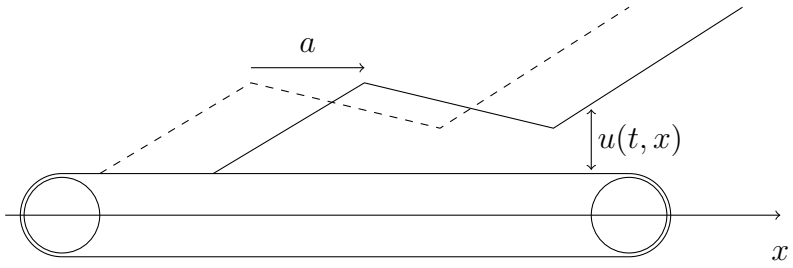

FIGURE 4 – Des tas de sable transportés sur un tapis roulant. La ligne discontinue représente les tas en $t = 0s$ et la ligne continue en $t = 1s$.

Ici, a correspond à la vitesse du transport. Si a est positif, le sable bouge vers la droite ; si a est négatif, le sable bouge vers la gauche ; plus a est grand en valeur absolue, plus le mouvement est rapide.

Remarque 3. *En se référant à la définition, cette équation n'est ni elliptique, ni parabolique, ni hyperbolique. Cependant, le comportement des solutions de cette équation est proche d'un comportement hyperbolique. Notons par ailleurs que l'équation* (2) *associée à l'équation de transport est une droite et les hyperboles ont des droites comme asymptotes (d'où le comportement hyperbolique de l'équation de transport).*

Pour finir la présentation de l'équation de transport, nous donnons des résultats numériques pour différentes valeurs de la vitesse de transport a (voir figure 5). Ces résultats illustrent bien le fait a est la vitesse à laquelle la solution se déplace.

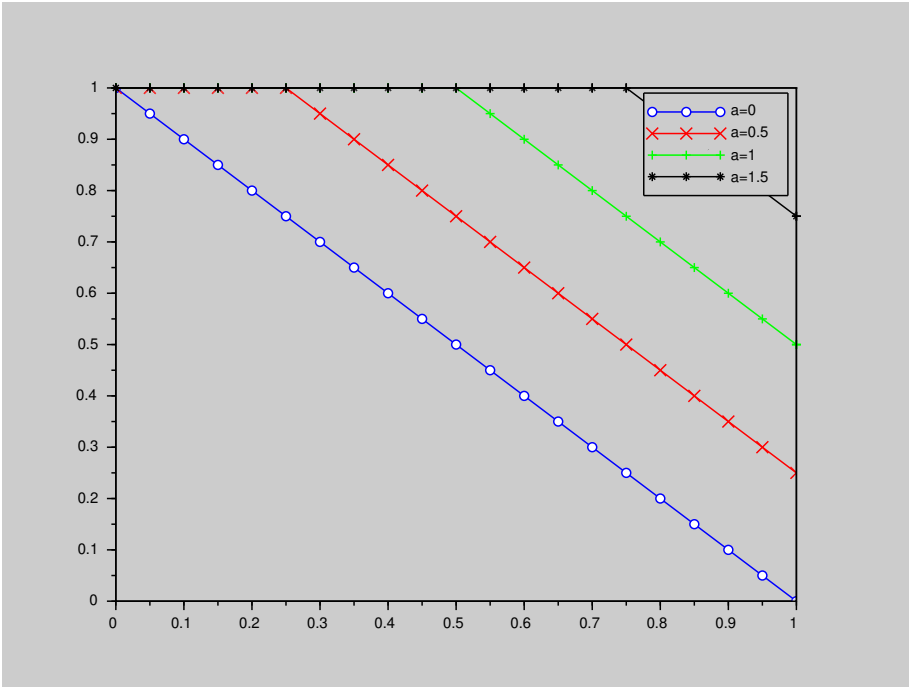

FIGURE 5 – Solution de l'équation de transport à $t = 0.5$ pour différents a. La donnée initiale ($t = 0$) correspond à la solution pour $a = 0$. On a imposé la condition de Dirichlet à gauche $u(t, 0) = 1$.

1.3 Équation de la chaleur

On se place dans le cadre du premier problème que nous avons évoqué dans la section 1.1 (voir figure 1). Pour plus de simplicité, nous considérons que nous sommes dans le cas de la conduction thermique (bien que comme nous l'avons vu précédemment d'autres problèmes comme le mouvement d'une population ou de molécules peuvent être considérés). La différence avec ce qui a été fait dans la section 1.1 est qu'ici les flux de chaleur ne sont pas nécessairement à l'équilibre. On permet donc à la température de changer au cours du temps.

Nous allons maintenant refaire le raisonnement de la section 1.1 avec cette fois-ci une température $u(t,x)$ qui dépend du temps. Si l'on considère la partie droite de la figure 1, l'accumulation d'énergie en (t,x) est due au fait que les flux ne sont pas équilibrés ("ce qui entre n'est pas égal à ce qui sort"). Ainsi, s'il y a plus d'énergie qui entre dans le domaine infinitésimal qu'il n'y en a qui en sort, alors la température augmente. Nous traduisons cela par l'équation $c(x)\delta x\dfrac{\partial u}{\partial t}(t,x) = f(t,x)\delta x + q(t,x) - q(t,x+\delta x)$, ce qui donne

$$c(x)\frac{\partial u}{\partial t}(t,x) = f(t,x) - \frac{\partial q}{\partial x}(t,x). \tag{8}$$

où $c(x)$ est la capacité thermique linéique du matériau, $q(t,x)$ et $f(t,x)$ sont respectivement le flux de chaleur et la source de chaleur en (t,x). Comme précédemment, le flux de chaleur est proportionnel au gradient de température (voir (6)). Pour simplifier la présentation, nous considérons que la capacité thermique c et la conductivité thermique k sont des constantes (qui ne dépendent ni du temps, ni de l'espace). En combinant les équations (6) et (8), on obtient l'équation de la chaleur

$$\frac{\partial u}{\partial t} - \nu\frac{\partial^2 u}{\partial x^2} = \widetilde{f}, \tag{9}$$

où $\nu = k/c > 0$ est la diffusivité thermique (qui dépend du matériau considéré et de sa géométrie).

Les conditions aux limites classiques sont les mêmes que celles évoquées pour le problème stationnaire (voir tableau 1).

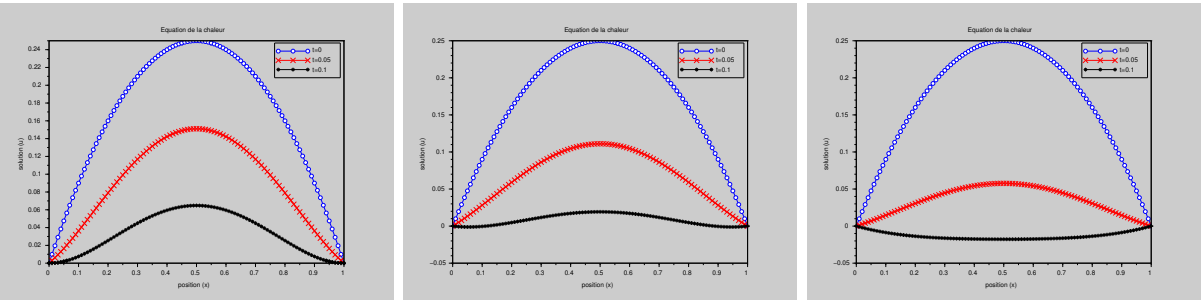

FIGURE 6 – Solution de l'équation de la chaleur pour $f(x) = -1$. Gauche : $\nu = 0.5$. Milieu : $\nu = 1$. Droite : $\nu = 2$. Les courbes représentent la solution à $t = 0$ (bleu), $t = 0.05$ (rouge) et $t = 0.1$ (noir). On a imposé les conditions de Dirichlet $u(t,0) = u(t,1) = 0$. La condition initiale est $u_0(x) = x(1-x)$.

Nous illustrons sur la figure 6 l'influence du paramètre ν. Plus ν est grand, plus la chaleur va se diffuser rapidement et plus la courbe va s'applatir rapidement.

Exercice 2. *Prouver que l'équation* (9) *est une équation parabolique.*

1.4 Équation des ondes

L'équation des ondes peut être obtenue en considérant un modèle mécanique comme celui de la figure 2 où cette fois-ci les forces ne sont pas nécessairement à l'équilibre et où les déplacements $u(t,x)$ peuvent varier au cours du temps.

Le principe fondamental de la dynamique appliqué à une tranche de matière de longueur δx nous dit que l'accélération de cette tranche (la dérivée seconde de sa position $x+u(t,x)$) multipliée par sa masse est égale à la somme des forces qui agissent sur elle. Ainsi, $\rho(x)\delta x\dfrac{\partial^2 u}{\partial t^2} = f(t,x)\delta x + q(t,x) - q(t,x+\delta x)$, ce qui nous donne l'équation

$$\rho(x)\frac{\partial^2 u}{\partial t^2}(t,x) + \frac{\partial q}{\partial x}(t,x) = f(t,x), \tag{10}$$

où $\rho(x)$ correspond à la masse de la barre par unité de longueur. Plus ρ est grand, plus la barre a d'inertie et plus elle accélère lentement. Notons à ce stade qu'il y a, dans cette équation, une dérivée seconde en temps à la place de la dérivée première qu'il y avait dans (8).

La force de compression à travers la barre est donnée par la loi d'élasticité (6). Pour simplifier la présentation, on considère que la raideur k et la masse linéique ρ sont constantes (ne dépendent ni de t ni de x). En combinant (6) et (10) on obtient l'équation des ondes

$$\frac{\partial^2 u}{\partial t^2} - c^2\frac{\partial^2 u}{\partial x^2} = \widetilde{f}, \tag{11}$$

où $c>0$ défini par $c^2 = k/\rho$ correspond à la vitesse de propagation des ondes dans ce milieu.

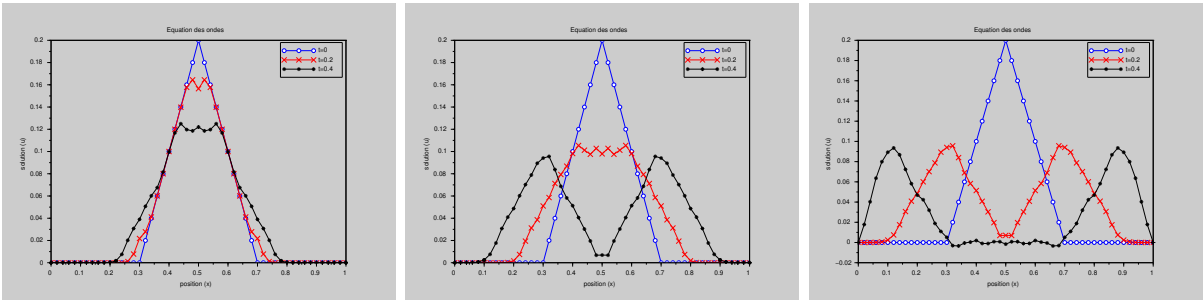

FIGURE 7 – Solution de l'équation des ondes pour $f=0$ et $v_0=0$. Gauche : $c=0.2$. Milieu : $c=0.5$. Droite : $c=1$. Les courbes représentent la solution à $t=0$ (bleu), $t=0.2$ (rouge) et $t=0.4$ (noir). On impose les conditions de Dirichlet $u(t,0)=u(t,1)=0$. La condition initiale est représentée en bleu.

On illustre l'influence de c sur la figure 7. Nous voyons que nous avons une propagation de l'information dans les deux directions (les x positifs et les x négatifs). Plus c est grand, plus l'information se propage rapidement. Notons également que la solution pour $c=0.5$ et $t=0.4$ correspond à la solution pour $c=1.0$ et $t=0.2$. On voit donc bien que c correspond à une vitesse de propagation de l'onde.

Nous avons évoqué le fait que l'équation des ondes pouvait représenter le déplacement d'une barre en compression. On peut aussi modéliser des problèmes comme la propagation du déplacement d'une corde ou la propagation d'une onde électromagnétique. Une bonne façon d'interpréter la figure 7 est d'imaginer une corde tendue sur un intervalle $(0, 1)$. La solution $u(t, x)$ correspond alors à la position verticale de la corde en x à l'instant t.

Rappelons enfin que les conditions aux limites classiques sont les mêmes que celles du cas stationnaire (voir tableau 1).

Exercice 3. *Prouver que l'équation* (11) *est hyperbolique.*

2 Équation de Poisson

On considère Ω un ouvert borné de $\mathbb{R}^d$ avec $d \in \{1, 2, 3\}$. Nous simplifions le cadre évoqué précédemment en considérant $k = 1$. On s'intéresse donc au problème

$$\text{Trouver } u \in C^2(\Omega) \text{ telle que } \begin{cases} -\Delta u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \tag{12}$$

où on note $\partial\Omega$ le bord du domaine Ω.

2.1 Propriétés générales de l'équation

Théorème 1 (Existence et unicité de la solution)**.** *Si $f \in C^0(\Omega)$, alors il existe une unique solution $u \in C^2(\Omega)$ au problème* (12)*. De plus, pour $\ell \in \mathbb{N}$, si $f \in C^\ell(\Omega)$, alors $u \in C^{\ell+2}(\Omega)$.*

Dans ce théorème nous voyons l'effet régularisant de l'équation de Poisson : si le terme source est de classe C^ℓ alors la solution du problème est de classe $C^{\ell+2}$. Les outils utilisés pour établir cette preuve sont présentés dans la section 2.2.

Proposition 1 (Principe du maximum)**.** *On se place dans le cadre du théorème 1. Si $f \leq 0$, alors $u \leq 0$. En particulier, $\max_{x\in\Omega} u(x) = \max_{x\in\partial\Omega} u(x) = 0$. Si, de plus, il existe $m \in \Omega$ tel que $u(m) = 0$, alors $\forall x \in \Omega$, $u(x) = 0$ (principe du maximum fort).*

Définition 1 (Fonctions harmoniques)**.** *Si $u \in C^2(\Omega)$ et $-\Delta u = 0$ dans Ω, on dit que u est une fonction harmonique.*

Il existe des liens entre fonctions harmoniques et fonctions holomorphes. En effet, si φ est une fonction holomorphe, alors $\Re(\varphi)$ et $\Im(\varphi)$ (parties réelle et imaginaire) sont des fonctions harmoniques.

D'autres conditions aux limites peuvent être considérées. Nous pouvons par exemple considérer des conditions de Dirichlet non homogènes.

$$\text{Trouver } u \in C^2(\Omega) \text{ telle que } \begin{cases} -\Delta u = f & \text{dans } \Omega, \\ u = g & \text{sur } \partial\Omega, \end{cases}$$

où $g \in C^2(\partial\Omega)$ est une donnée du problème. Pour résoudre ce problème, on considère un relèvement de g, c'est-à-dire une fonction $\tilde{g} \in C^2(\Omega)$ telle que $\tilde{g} = g$ sur $\partial\Omega$, et on cherche la solution u sous la forme $u = \tilde{u} + \tilde{g}$. La fonction $\tilde{u}$ est alors solution du problème avec conditions de Dirichlet homogènes

$$\text{Trouver } \tilde{u} \in C^2(\Omega) \text{ telle que } \begin{cases} -\Delta \tilde{u} = f + \Delta \tilde{g} & \text{dans } \Omega, \\ \tilde{u} = 0 & \text{sur } \partial\Omega. \end{cases}$$

Moyennant l'existence du relèvement $\tilde{g}$, résoudre le problème non homogène revient à résoudre un problème homogène équivalent.

Nous pouvons également considérer des conditions de Neumann. Le problème de Poisson devient alors

$$\text{Trouver } u \in C^2(\Omega) \text{ telle que } \begin{cases} -\Delta u = f & \text{dans } \Omega, \\ \nabla u \cdot n = 0 & \text{sur } \partial\Omega, \\ \int_\Omega u \, \mathrm{d}x = 0, \end{cases} \tag{13}$$

où n désigne la normale sortante au domaine.

Dans le problème (13), la condition $\int_\Omega u \, \mathrm{d}x = 0$ a été ajoutée pour garantir l'unicité de la solution. En effet, si on enlevait cette condition, pour u solution de (13), $u + c$ serait aussi solution pour tout $c \in \mathbb{R}$. Des alternatives existent à la condition $\int_\Omega u \, \mathrm{d}x = 0$, l'essentiel est d'éliminer l'infinité de solutions générées en ajoutant $c \in \mathbb{R}$.

Terminons cette section en relevant le fait que l'existence d'une solution au problème (13) requiert la condition $\int_\Omega f \, \mathrm{d}x = 0$.

Exercice 4. *Prouver que s'il existe une solution u au problème* (13)*, alors f vérifie $\int_\Omega f \, \mathrm{d}x = 0$.*

2.2 Formulation variationnelle, théorème de Lax–Milgram

Le but de cette section est de présenter les outils utilisés pour prouver le théorème 1. Nous utilisons notamment le théorème de Lax–Milgram.

Théorème 2 (Lax–Milgram)**.** *On fait les hypothèses suivantes.*

- *Soit V un espace de Hilbert.*
- *Soit ℓ une forme linéaire continue sur V (il existe $C_\ell > 0$ tel que $\forall v \in V, |\ell(v)| \leq C_\ell \|v\|_V$).*
- *Soit a une forme bilinéaire continue sur V (il existe $C_a > 0$ tel que $\forall v, w \in V, |a(v,w)| \leq C_a \|v\|_V \|w\|_V$).*
- *On suppose de plus que a est coercive : il existe $\alpha > 0$ tel que $\forall v \in V, a(v,v) \geq \alpha \|v\|_V^2$.*

Sous ces hypothèses le problème suivant est bien posé :

$$\text{Trouver } u \in V, \text{ tel que } \forall v \in V,\ a(u,v) = \ell(v). \tag{14}$$

Ceci signifie que le problème (14) *admet une unique solution $u \in V$ et que celle-ci est bornée, ici par $\|u\|_V \leq \dfrac{C_\ell}{\alpha}$.*

Pour utiliser ce théorème, écrivons le problème (12) sous sa forme variationnelle. On introduit l'espace de Sobolev

$$H_0^1(\Omega) := \{v \in L^2(\Omega) \mid \nabla v \in [L^2(\Omega)]^d \text{ et } v_{|\partial\Omega} = 0 \text{ sur } \partial\Omega\}, \tag{15}$$

où ∇v, le gradient de v, est défini au sens des distributions et $v_{|\partial\Omega}$ est la trace de v sur le bord du domaine.

Si u est une solution de (12), alors pour tout $v \in H_0^1(\Omega)$ on peut écrire

$$\int_\Omega -\Delta u v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x.$$

En intégrant par parties on obtient

$$\int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x - \int_{\partial\Omega} (\nabla u \cdot n) v \, \mathrm{d}s = \int_\Omega f v \, \mathrm{d}x.$$

Puisque $v_{|\partial\Omega} = 0$, le deuxième terme de cette expression est nul. Nous avons donc établi la proposition suivante.

Proposition 2. *Toute solution du problème* (12) *est solution du problème*

$$\text{Trouver } u \in H_0^1(\Omega), \text{ tel que } \forall v \in H_0^1(\Omega), \ \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x = \int_\Omega f v \, \mathrm{d}x. \tag{16}$$

Proposition 3. *Le problème* (16) *est bien posé.*

Démonstration. Le problème (16) correspond au problème (14) avec $V = H_0^1(\Omega)$, $\forall v, w \in H_0^1(\Omega), a(v,w) = \int_\Omega \nabla v \cdot \nabla v \ \mathrm{d}x$ et $\ell(v) = \int_\Omega f v \ \mathrm{d}x$. Nous allons montrer que toutes les hypothèses du théorème de Lax–Milgram sont vérifiées. On admet le fait que $H_0^1(\Omega)$ équipé de la norme $\|v\|_{H^1(\Omega)} := (\int_\Omega v^2 + \nabla v \cdot \nabla v \ \mathrm{d}x)^{1/2}$ est un espace de Hilbert (se reporter à un cours sur les distributions). D'après l'inégalité de Cauchy–Schwarz, on a $|\ell(v)| \leq (\int_\Omega f^2 \ \mathrm{d}x)^{1/2} (\int_\Omega v^2 \ \mathrm{d}x)^{1/2} \leq (\int_\Omega f^2 \ \mathrm{d}x)^{1/2} \|v\|_{H^1(\Omega)}$. La forme linéaire ℓ est donc continue.

De la même façon, l'inégalité de Cauchy–Schwarz permet de prouver que la forme bilinéaire a est continue. Pour finir, nous utilisons l'inégalité de Poincaré : il existe $c > 0$ tel que $\forall v \in H_0^1(\Omega), \int_\Omega v^2 \ \mathrm{d}x \leq c \int_\Omega \nabla v \cdot \nabla v \ \mathrm{d}x$. Avec cette inégalité, on peut prouver que a est coercive. Toutes les hypothèses du théorème de Lax–Milgram sont réunies, le problème (16) est donc bien posé. □

Remarque 4. *Les résultats de cette section prouvent l'unicité de la solution de* (12) *dans* $H_0^1(\Omega)$*. On a également prouvé l'existence d'une solution à* (16)*. Pour prouver le théorème 1, il faut montrer que si* $f \in C^\ell(\Omega)$ *alors la solution de* (16) *est dans* $C^{\ell+2}(\Omega)$ *et est solution de* (12)*. Cependant cette preuve est très délicate et nous ne la développerons pas ici.*

Remarque 5. *On dit qu'une solution de* (12) *est une solution forte du problème de Poisson au sens où les dérivées sont des dérivées usuelles. On dit qu'une solution de* (16) *est une solution faible du problème de Poisson au sens où les dérivées sont des dérivées faibles (au sens des distributions). Nous avons montré qu'une solution forte est une solution faible. La réciproque est plus délicate et nécessite des hypothèses sur la régularité de* f*.*

Exercice 5. *On note* $H^1(\Omega) = \{v \in L^2(\Omega) \mid \nabla v \in [L^2(\Omega)]^d\}$ *et* $H^1_\bullet(\Omega) = \{v \in H^1(\Omega) \mid \int_\Omega v \ \mathrm{d}x = 0\}$*. Prouver que toute solution du problème* (13) *est solution de la formulation variationnelle*

$$\text{Trouver } u \in H^1_\bullet(\Omega), \text{ tel que } \forall v \in H^1_\bullet(\Omega), \ \int_\Omega \nabla v \cdot \nabla v \ \mathrm{d}x = \int_\Omega f v \ \mathrm{d}x. \tag{17}$$

Prouver de plus que le problème (17) *est bien posé.*

Pour cela, on supposera que $H^1_\bullet$ *équipé de la norme* $\|\cdot\|_{H^1(\Omega)}$ *introduite précédemment est un espace de Hilbert. On pourra également utiliser l'inégalité de Poincaré–Wirtinger : il existe une constante* $C > 0$ *telle que*

$$\forall v \in H^1(\Omega), \qquad \int_\Omega \left(v - \frac{1}{|\Omega|} \int_\Omega v(x') \ \mathrm{d}x' \right)^2 \mathrm{d}x \leq C \int_\Omega \nabla v \cdot \nabla v \ \mathrm{d}x.$$

2.3 Discrétisation par la méthode des différences finies

Le but de la méthode des différences finies est d'approcher la solution d'une EDO (équation différentielle ordinaire) ou EDP (équation aux dérivées partielles) par des valeurs censées représenter cette fonction en certains points. Dans toutes les discrétisations abordées dans

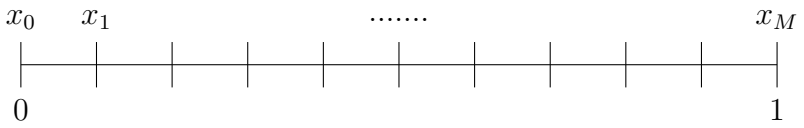

FIGURE 8 – Représentation de la discrétisation de l'intervalle $(0,1)$ en M sous-intervalles.

ce cours, nous ne considérerons que le cas de problèmes à une dimension en espace (les problèmes instationnaires auront une deuxième dimension correspondant au temps).

Par exemple, intéressons nous au problème (12) avec $\Omega = (0,1)$. Notre problème est donc

$$\begin{cases} -\dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2} = f \text{ dans } (0,1), \\ u(0) = u(1) = 0. \end{cases} \tag{18}$$

Nous discrétisons l'intervalle $(0,1)$ en $M > 0$ sous-intervalles. On définit donc les points $x_j = jh$ $(0 \leq j \leq M)$ où $h = 1/M$. On a bien $x_0 = 0$ et $x_M = 1$ (voir figure 8).

Étant donné que nous ne disposons que de valeurs en certains points, on ne peut pas définir de dérivées au sens usuel. On utilise donc des taux d'accroissement. Rappelons qu'une dérivée est la limite d'un taux d'accroissement. Par exemple, on peut montrer que $\dfrac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x) = \lim\limits_{h\to 0} \dfrac{u(x+h) - 2u(x) + u(x-h)}{h^2}$. Avec les notations introduites précédemment, cela signifie que, pour h suffisamment petit,

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x_j) \simeq \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1})}{h^2}. \tag{19}$$

La méthode des différences finies consiste donc à définir une suite $(u_j)_{0\leq j\leq M}$ qui reprenne les éléments du problème (18). En particulier, il faut remplacer les dérivées continues par des taux d'accroissment. Il existe plusieurs façons de définir la suite $(u_j)_{0\leq j\leq M}$ et toutes n'ont pas les mêmes propriétés. La propriété la plus importante est la convergence : on veut que $\lim\limits_{M\to+\infty} \max_{0\leq j\leq M} |u(x_j) - u_j| = 0$. Nous aborderons plus en détails ces aspects dans la section 3.3.

Nous proposons maintenant de discrétiser le probème (18) en utilisant (19). Nous construisons donc une suite $(u_j)_{0\leq j\leq M}$ vérifiant

$$\begin{cases} \forall 1 \leq j \leq M-1, -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \\ u_0 = u_M = 0. \end{cases} \tag{20}$$

Proposition 4. *La relation* (20) *définit une unique suite* $(u_j)_{0\leq j\leq M}$ *dont les valeurs peuvent être déterminées en résolvant le système linéaire* $A_\Delta U = F$*, avec*

$$A_\Delta = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & 0 & \cdots & & \\ -1 & 2 & -1 & 0 & \cdots & \\ 0 & -1 & 2 & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ & & & \ddots & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}, \qquad U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{M-1} \end{pmatrix}, \qquad F = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{M-1}) \end{pmatrix}. \tag{21}$$

Notons que la valeur de u_0 et u_M est déjà connue (elle est nulle).

Démonstration. La relation (20) est équivalente au système linéaire $A_\Delta U = F$. Pour prouver cela, on écrit les équations discrètes

$$\begin{aligned}
\frac{-u_0 + 2u_1 - u_2}{h^2} &= f(x_1),\\
\frac{-u_1 + 2u_2 - u_3}{h^2} &= f(x_2),\\
\frac{-u_2 + 2u_3 - u_4}{h^2} &= f(x_3),\\
&\vdots\\
\frac{-u_{M-2} + 2u_{M-1} - u_M}{h^2} &= f(x_{M-1}).
\end{aligned}$$

En prenant en compte $u_0 = u_M = 0$, on obtient bien $A_\Delta U = F$. La réciproque se prouve de manière similaire.

Maintenant, le fait que ce système admet une unique solution vient de l'inversibilité de la matrice A_Δ qui est une conséquence de la proposition 5. □

Remarque 6. *La matrice A_Δ doit absolument être connue par cœur. En effet, ceci est exigible et le jury du concours se plaint dans ses rapports que beaucoup trop de candidats ne connaissent pas cette matrice.*

Proposition 5. *La matrice A_Δ est symétrique définie positive.*

Exercice 6. *Prouver la proposition 5.*

Proposition 6 (Convergence)**.** *Supposons que la solution u du problème* (18) *soit dans $C^4(0,1)$. Alors il existe $C > 0$ tel que pour tout $M \geq 2$, on a*

$$\max_{0 \leq j \leq M} |u(x_j) - u_j| \leq Ch^2,$$

avec $h = 1/M$.

Démonstration. Étant donné que $u \in C^4(0,1)$, on peut utiliser la formule de Taylor–Young

$$\begin{aligned}
u(x_{j+1}) &= u(x_j) + hu'(x_j) + \frac{h^2}{2}u^{(2)}(x_j) + \frac{h^3}{6}u^{(3)}(x_j) + \frac{h^4}{24}u^{(4)}(x_j) + o(h^4),\\
u(x_{j-1}) &= u(x_j) - hu'(x_j) + \frac{h^2}{2}u^{(2)}(x_j) - \frac{h^3}{6}u^{(3)}(x_j) + \frac{h^4}{24}u^{(4)}(x_j) + o(h^4).
\end{aligned}$$

On obtient ainsi

$$\frac{-u(x_{j+1}) + 2u(x_j) - u(x_{j-1})}{h^2} = -u^{(2)}(x_j) - \frac{h^2}{12}u^{(4)}(x_j) + o(h^2).$$

On a de plus

$$\frac{-u_{j+1} + 2u_j - u_{j-1}}{h^2} = f(x_j) = -u^{(2)}(x_j),$$

et donc

$$\frac{-\delta_{j+1} + 2\delta_j - \delta_{j-1}}{h^2} = \frac{h^2}{12} u^{(4)}(x_j) + o(h^2),$$

où $\delta_j = u_j - u(x_j)$. En notant $\underline{\delta}$ le vecteur des δ_j, on a $\underline{\delta} = A_\Delta^{-1}\underline{\epsilon}$ avec $(\underline{\epsilon})_j = O(h^2)$. En supposant que $\sup_{X \in \mathbb{R}^{M-1}\setminus\{0\}} \dfrac{\max_j |(A_\Delta^{-1}X)_j|}{\max_j |X_j|}$ est borné quand h tend vers 0 (on ne le montre pas ici), on obtient le résultat attendu. □

Le résultat précédent signifie que pour h suffisamment petit on s'attend à ce que l'erreur commise par le schéma numérique décroisse comme le carré de h. On dit que le schéma est d'ordre 2, nous étudierons plus précisément cette notion plus loin dans le cours.

Cette notion d'ordre de convergence peut être illustrée en représentant l'erreur commise par le schéma $\varepsilon = \max_{0 \le j \le M} |u(x_j) - u_j|$ en fonction de h en utilisant une échelle logarithmique pour les axes des abscisses et des ordonnées. Par exemple, considérons $f(x) = x^2$, la solution exacte associée est $u(x) = \frac{x}{12}(1 - x^3)$. L'erreur du schéma commise sur ce cas test est représentée sur la figure 9. On voit que l'on obtient une droite de pente 2. Ceci est en accord avec la proposition 6 : si l'erreur est $\varepsilon = Ch^2$ alors $\log(\varepsilon) = 2\log(h) + \log(C)$.

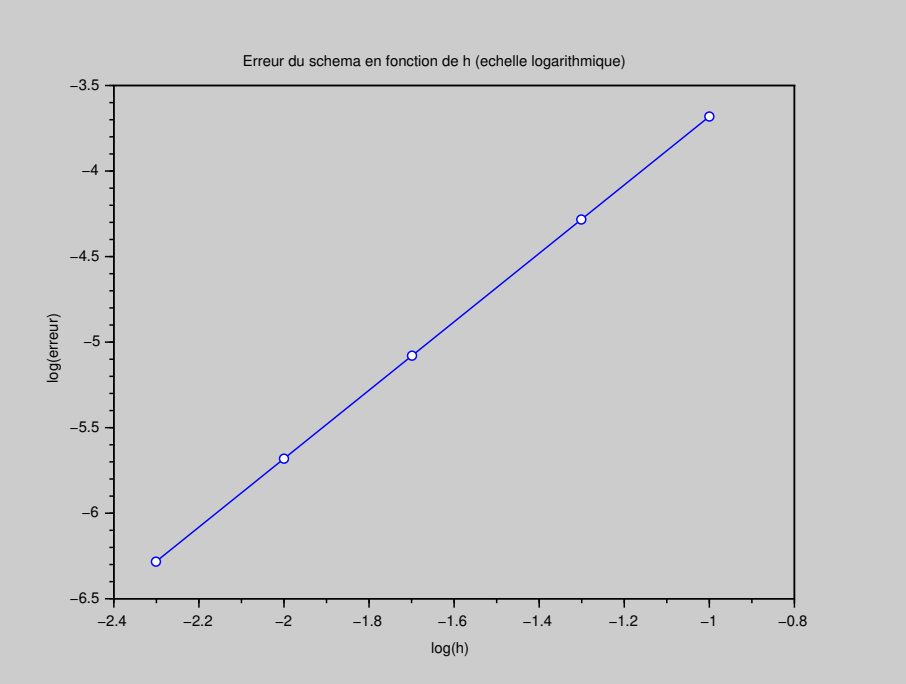

FIGURE 9 – Erreur en fonction de h (échelle logarithmique) pour $f(x) = x^2$.

Exercice 7. *Coder le schéma décrit précédemment et retrouver les résultats de la figure 9 (les différents points ont été obtenus pour $M = 10$, 20, 50, 100 et 200).*

Exercice 8 (Coefficient de raideur)**.** *On peut décider d'ajouter un coefficient de raideur $k > 0$ constant. La nouvelle matrice devient alors kA_Δ où A_Δ est la matrice précédente. La figure 3 a été obtenue pour $f(x) = 1$ et $M = 10$ ($h = 1/10$). Coder le schéma proposé et retrouver les résultats de la figure 3. Calculer la solution exacte. En déduire l'erreur que commet le schéma. Que remarquez-vous ?*

Exercice 9 (Conditions aux limites de Dirichlet non homogènes)**.** *Adapter le schéma numérique précédent pour approcher le problème*

$$\begin{cases} -\dfrac{\mathrm{d}^2u}{\mathrm{d}x^2} = f \text{ dans } (0,1), \\ u(0) = \alpha, \\ u(1) = \beta, \end{cases}$$

où $\alpha, \beta \in \mathbb{R}$. Donner le système linéaire à résoudre.

Exercice 10 (Condition aux limites de Neumann)**.** *On considère le problème de Poisson avec des conditions aux limites de Neumann*

$$\begin{cases} -\dfrac{\mathrm{d}^2u}{\mathrm{d}x^2} = f \text{ dans } (0,1), \\ u'(0) = u'(1) = 0, \\ \int_0^1 u(x)\,\mathrm{d}x = 0. \end{cases}$$

1. *On décide d'adapter la démarche précédente en considérant le schéma*

$$\begin{cases} \forall 1 \leq j \leq M-1, -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \\ \dfrac{u_1 - u_0}{h} = \dfrac{u_M - u_{M-1}}{h} = 0. \end{cases}$$

   Calculer la matrice associée. Montrer qu'elle est positive mais qu'elle n'est pas inversible. Commenter.

2. *On ajoute une condition de moyenne nulle, le schéma devient*

$$\begin{cases} \forall 1 \leq j \leq M-1, -\dfrac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \\ \dfrac{u_1 - u_0}{h} = \dfrac{u_M - u_{M-1}}{h} = 0, \\ \displaystyle\sum_{j=1}^{M-1} u_j = 0. \end{cases}$$

   On impose la condition de moyenne nulle par un multiplicateur de Lagrange $\alpha \in \mathbb{R}$. Le problème devient $\widetilde{A}\widetilde{U} = \widetilde{F}$ avec

$$\widetilde{A} = \frac{1}{h^2}\begin{pmatrix} 1 & -1 & & & & h^2 \\ -1 & 2 & -1 & & & \vdots \\ & \ddots & \ddots & \ddots & & \vdots \\ & & -1 & 2 & -1 & \vdots \\ & & & -1 & 1 & h^2 \\ h^2 & \cdots & \cdots & \cdots & h^2 & 0 \end{pmatrix}, \qquad \widetilde{U} = \begin{pmatrix} u_1 \\ \vdots \\ u_{M-1} \\ \alpha \end{pmatrix}, \qquad \widetilde{F} = \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_{M-1}) \\ 0 \end{pmatrix}.$$

   Montrer que la matrice $\widetilde{A}$ est inversible.

3. *Coder ce schéma et le tester avec $f(x) = 1$. Que se passe-t-il ? Est-ce un comportement normal ? D'après vous, que représente α ?*

4. *Essayons maintenant*

$$f(x) = \begin{cases} 1 & si \quad x \in (0, 0.5), \\ 0 & si \quad x = 0.5, \\ -1 & si \quad x \in (0.5, 1). \end{cases}$$

Calculer la solution exacte à ce problème. Déterminer numériquement l'ordre de convergence du schéma. Est-ce en contradition avec la proposition 6 ?

5. *On pourra également reprendre la question précédente avec* $f(x) = x - \frac{1}{2}$.

3 Équation de transport

Nous étudions ici l'équation de transport dans un espace à une dimension. Le problème dépend donc d'une variable d'espace $x \in [0,1]$ et d'une variable de temps $t \in [0,T]$ avec $T > 0$. Nous considérons le problème

$$\text{Trouver } u \in C^1([0,T]\times[0,1]) \text{ telle que } \begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 \text{ dans } (0,T)\times(0,1), \\ \forall t \in (0,T), u(t,0) = \alpha(t), \\ \forall x \in (0,1), u(0,x) = u_0(x), \end{cases} \tag{22}$$

où $a > 0$ correspond à la vitesse de transport de l'équation, $\alpha(t)$ est la donnée de Dirichlet à gauche et u_0 est la donnée initiale. Dans l'énoncé de notre problème, nous avons noté $C^1([0,T]\times[0,1])$ l'ensemble des fonctions C^1 de $[0,T]\times[0,1]$ à valeur dans $\mathbb{R}$.

3.1 Propriétés générales

Le problème (22) admet une unique solution u. Cette solution est obtenue en "déplaçant" la donnée initiale vers la droite et en "faisant entrer" la donnée de Dirichlet dans le domaine. De manière plus rigoureuse, nous avons le théorème suivant.

Théorème 3. *Supposons que $a > 0$, $u_0 \in C^1([0,1])$ et $\alpha \in C^1([0,T])$. Supposons, de plus, les conditions de compatibilité $u_0(0) = \alpha(0)$ et $\alpha'(0) + au_0'(0) = 0$. Le problème (22) admet alors une unique solution donnée par*

$$u(t,x) = \begin{cases} u_0(x-at) & \text{si } x \geq at, \\ \alpha(t - x/a) & \text{sinon.} \end{cases} \tag{23}$$

Remarque 7. *On n'a besoin d'une donnée de Dirichlet que d'un seul côté du domaine. Puisque l'on a choisi $a > 0$, il faut fixer la donnée de Dirichlet à gauche. On aurait aussi pu prendre $a < 0$ et utiliser la condition de Dirichlet à droite $u(t,1) = \alpha(t)$.*

Nous allons maintenant évoquer une autre propriété intéressante de l'équation de transport : la réversibilité de l'équation. Cela signifie qu'en utilisant la solution au temps final et éventuellement d'autres données (ici la solution qui sort du domaine à droite), on peut reconstituer la solution en tout temps $t \in [0,T]$. Il s'agit en fait d'étudier un problème où le temps s'écoule "en sens inverse".

Proposition 7 (Réversibilité de l'équation de transport)**.** *Soit u une solution du problème (22). Le problème*

$$\textit{Trouver } \widetilde{u} \in C^1([0,T]\times[0,1]) \textit{ telle que } \begin{cases} \dfrac{\partial \widetilde{u}}{\partial t} - a\dfrac{\partial \widetilde{u}}{\partial x} = 0 & \textit{dans } (0,T)\times(0,1), \\ \forall t \in (0,T), & \widetilde{u}(t,1) = u(T-t,1), \\ \forall x \in (0,1), & \widetilde{u}(0,x) = u(T,x), \end{cases} \tag{24}$$

admet une unique solution définie par $\forall t \in [0,T], \forall x \in [0,1]$, $\widetilde{u}(t,x) = u(T-t,x)$.

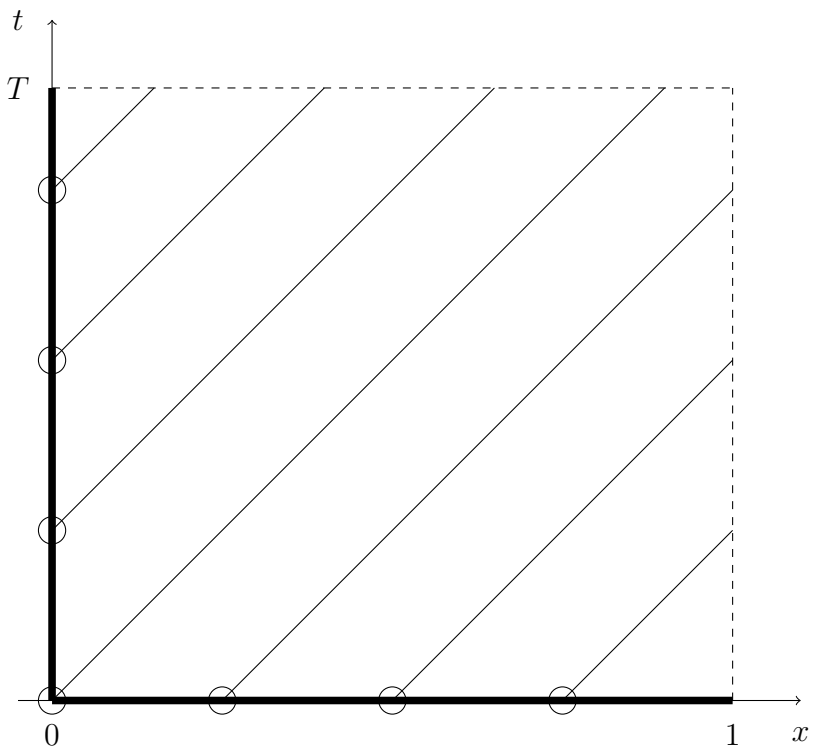

FIGURE 10 – Les droites caractéristiques de l'équation de transport ($a = 1$ et $T = 0.9$). Ces droites sont représentées dans le domaine $(0, T) \times (0, 1)$ (ici le temps est en ordonnée et l'espace en abscisse). La partie de la frontière où on utilise les données de u est représentée en gras. Les cercles indiquent qu'en ces points, on attribue la valeur de u à la droite caractéristique (et cette valeur est ensuite conservée sur toute la droite). Comparer cette figure avec le théorème 3.

Cette propriété signifie que l'information est conservée au cours du temps : le comportement de l'équation ne dégrade pas cette information. On peut donc, avec la connaissance de u à l'instant final dans tout le domaine et en tout temps sur la droite du domaine, retrouver la valeur de u en tout temps dans tout le domaine. Pour cela, on inverse le temps ($\tilde{t} = T - t$) et le sens du transport (il y a un $-a$ au lieu de $+a$ dans (24)).

Notons que toutes les équations instationnaires ne sont pas réversibles, nous verrons notamment la non-réversibilité de l'équation de la chaleur en section 4.2.

3.2 Méthode des caractéristiques

La méthode des caractéristiques consiste à montrer que la solution se conserve sur certaines courbes que l'on appelle trajectoires caractéristiques. Dans le cas monodimensionnel avec a constant, ces courbes sont des droites. On parlera donc de droites caractéristiques. Cependant, dans le cas où plusieurs dimensions d'espace sont considérées, ces courbes peuvent être plus complexes (voir la section 6.1).

Nous pouvons montrer que dans le cas considéré dans cette section, la solution se conserve sur les droites d'équation $x = at + b$ avec a la vitesse de transport dans (22) et $b \in \mathbb{R}$. Ces droites sont donc les droites caractéristiques de notre problème, nous les représentons sur

la figure 10. Le résultat de conservation est énoncé dans la proposition suivante.

Proposition 8. *Soit $u \in C^1([0,T]\times[0,1])$ vérifiant l'équation de transport*

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0 \text{ dans } (0,T)\times(0,1),$$

avec $a \in \mathbb{R}$. Pour tous $t, s \in [0,T]$ et $x \in [0,1]$ tels que $t-s \in [0,T]$ et $x - as \in [0,1]$, on a

$$u(t,x) = u(t-s, x-as).$$

Exercice 11. *Prouver la proposition 8.*

Notons que la proposition 8 est valable pour tout $a \in \mathbb{R}$ et pour n'importe quelles conditions aux limites. Elle montre que si l'on se déplace sur une droite caractéristique tout en restant dans le domaine $(0,T)\times(0,1)$, alors la valeur de u se conserve.

Pour calculer la valeur de $u(t,x)$, il suffit de déterminer la droite caractéristique passant en (t,x) et chercher son intersection avec le bord du domaine sur lequel on connaît la valeur de u (il s'agit de $\big((0,T)\times\{0\}\big)\cup\big(\{0\}\times(0,1)\big)$, en gras sur la figure 10). Nous allons utiliser cette méthode pour prouver le théorème 3.

Preuve du théorème 3. Commençons par montrer que la fonction u donnée dans (23) est solution du problème (22). On note $u^+(t,x) = u_0(x-at)$ et $u^-(t,x) = \alpha(t - x/a)$.

Par construction, les fonctions u^+ et u^- sont C^1 sur leur domaine de définition. Nous cherchons à les raccorder le long de la droite $x = at$. Soit $(t_0, x_0) \in [0,T]\times[0,1]$ qui vérifient $x_0 = at_0$. On a $\lim_{(t,x)\to(t_0,x_0)} u^+(t,x) = \lim_{s\to 0^+} u_0(s) = u_0(0) = \alpha(0) = \lim_{s\to 0^+}\alpha(s) = \lim_{(t,x)\to(t_0,x_0)} u^-(t,x)$. La fonction u est donc continue. De plus, par un raisonnement similaire, la condition $\alpha'(0) = -au_0'(0)$ implique que u est dans $C^1([0,T]\times[0,1])$. De plus, u vérifie la condition initiale $u(0,x) = u_0(x)$ ainsi que la condition de Dirichlet $u(t,0) = \alpha(t)$ et elle vérifie l'équation de transport. Il s'agit donc bien d'une solution de (22).

Supposons maintenant que $v \in C^1([0,T]\times[0,1])$ est solution de (22). Si $x \le at$, on applique la proposition 8 avec $s = x/a$, avec la condition de Dirichlet on obtient $v(t,x) = \alpha(t - x/a)$. Si $x \ge at$, on applique la proposition 8 avec $s = t$, avec la condition initiale on obtient $v(t,x) = u_0(x-at)$. Nous avons donc prouvé l'unicité de la solution. □

Nous voyons donc que la valeur de la solution en (t,x) est à aller chercher sur le bord du domaine espace-temps. C'est-à-dire, en fonction de la valeur de t et x, soit sur la condition initiale, soit sur la donnée de Dirichlet (l'endroit où récupérer l'information est représenté sur la figure 10 par un cercle).

Remarque 8. *Avec un raisonnement similaire, on peut montrer que si $a = 0$, il n'y a pas besoin de condition de Dirichlet et que si $a < 0$, il faut mettre la condition de Dirichlet sur la droite du domaine (en $x = 1$). Le même résultat d'existence et d'unicité s'applique alors.*

On peut utiliser la méthode des caractéristiques pour prouver d'autres résultats comme par exemple ceux énoncés dans les exercices suivants.

Exercice 12. *Prouver la proposition 7.*

Exercice 13 (Conditions aux limites périodiques)**.** *Soit $u_0 \in C^1([0,1])$ qui vérifie $u_0(0) = u_0(1)$ et $u_0'(0) = u_0'(1)$. Prouver que l'équation de transport avec des conditions aux limites périodiques*

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = 0 \text{ dans } (0,T)\times(0,1), \\ \forall t \in (0,T), \quad u(t,0) = u(t,1), \\ \forall x \in (0,1), \quad u(0,x) = u_0(x), \end{cases} \tag{25}$$

avec $a \in \mathbb{R}$ admet pour unique solution $u(t,x) = u_0(f(x-at))$, où $f(x)$ est la partie fractionnaire de x, c'est-à-dire $x = E(x) + f(x)$ avec $E(x) \in \mathbb{Z}$ et $0 \leq f(x) < 1$.

Exercice 14 (Prise en compte d'un terme source)**.** *Montrer que l'unique solution du problème*

$$\begin{cases} \dfrac{\partial u}{\partial t} + a\dfrac{\partial u}{\partial x} = f(t,x) \text{ dans } (0,T)\times\mathbb{R}, \\ \forall x \in \mathbb{R}, \quad u(0,x) = u_0(x), \end{cases}$$

avec $a \in \mathbb{R}$, $f \in C^0([0,T]\times\mathbb{R})$ et $u_0 \in C^1(\mathbb{R})$, est donnée par $u(t,x) = u_0(x-at) + \int_0^t f(s, x+a(s-t))\,\mathrm{d}s$.
NB : Notez que l'on a posé ce problème sur $\mathbb{R}$ entier pour ne pas avoir à se soucier des conditions aux limites.

3.3 Discrétisation par les différences finies et analyse numérique

Nous cherchons maintenant à discrétiser le problème (22) par la méthode des différences finies. Dans la section 2.3, pour approcher l'équation de Poisson, nous avons discrétisé l'espace $(0,1)$ et nous avons approché la solution $u(x)$ par une suite de terme général u_j.

Dans le cas présent, la solution dépend de deux variables t et x, il faut donc discrétiser ces deux dimensions. On découpe l'intervalle de temps $[0,T]$ en N sous-intervalles $[t_n, t_{n+1}]$ avec pour tout $0 \leq n \leq N$, $t_n = nh_t$ et $h_t = T/N$. De même, on découpe l'intervalle d'espace $[0,1]$ en M sous-intervalles $[x_j, x_{j+1}]$ avec pour tout $0 \leq j \leq M$, $x_j = jh_x$ et $h_x = 1/M$. De plus, nous allons approcher la fonction $u(t,x)$ par une suite $(u_j^n)_{\substack{0\leq n\leq N \\ 0\leq j\leq M}}$. Ici, l'indice j donne la position en espace et l'exposant n donne le temps considéré. Ainsi, on cherche à calculer u_j^n de manière à ce que ce soit une approximation de $u(t_n, x_j)$.

3.3.1 Motivation de l'analyse numérique

Comme nous l'avons vu précédemment, lorsque l'on considère l'équation de transport avec une condition de Dirichlet, l'information de la donnée initiale sort progressivement du domaine en étant remplacée par la donnée de Dirichlet. On pourrait vouloir s'intéresser à l'évolution sur le temps long de l'information issue de la donnée initiale. Pour cela, on peut essayer de suivre cette donnée initiale en considérant des conditions périodiques. C'est ce que nous faisons dans cette section. On s'intéresse donc au problème (25).

Avec ces conditions frontières, l'information qui sort en $x = 1$ réentre immédiatement en $x = 0$. On peut donc la suivre au cours de sa propagation dans le domaine et observer la qualité de l'approximation que nous avons faite. Nous allons donc comparer nos approximations numériques avec la solution exacte donnée par $u(t,x) = u_0(f(x-at))$ (voir exercice 13).

Nous proposons maintenant deux discrétisations du problème (25). La méthode consiste à approcher les dérivées partielles (en temps et en espace) par des taux d'accroissement. Nous comparons le comportement de deux discrétisations différentes.

Nous proposons tout d'abord d'approcher la dérivée en temps par une approximation décentrée aval et la dérivée en espace par une approximation centrée comme suit

$$\frac{\partial u}{\partial t}(t_n, x_j) \simeq \frac{u(t_{n+1}, x_j) - u(t_n, x_j)}{h_t}, \qquad \frac{\partial u}{\partial x}(t_n, x_j) \simeq \frac{u(t_n, x_{j+1}) - u(t_n, x_{j-1})}{2h_x}.$$

On obtient le schéma numérique

$$\begin{cases} \forall 0 \leq j \leq M, u_j^0 = u_0(x_j), \\ \forall 1 \leq n \leq N, u_0^n = u_M^n, \\ \forall 1 \leq n \leq N, \forall 0 \leq j \leq M, u_j^n = u_j^{n-1} - \dfrac{ah_t}{2h_x}(u_{j+1}^{n-1} - u_{j-1}^{n-1}), \end{cases} \tag{26}$$

où, pour simplifier l'écriture, nous avons introduit $u_{M+1}^n = u_1^n$ et $u_{-1}^n = u_{M-1}^n$ (rappelons que $u_M^n = u_0^n$).

Nous nous proposons également d'approcher les dérivées en temps et les dérivées en espace respectivement par des approximations décentrées aval et amont comme suit

$$\frac{\partial u}{\partial t}(t_n, x_j) \simeq \frac{u(t_{n+1}, x_j) - u(t_n, x_j)}{h_t}, \qquad \frac{\partial u}{\partial x}(t_n, x_j) \simeq \frac{u(t_n, x_j) - u(t_n, x_{j-1})}{h_x}. \tag{27}$$

On obtient le schéma numérique

$$\begin{cases} \forall 0 \leq j \leq M, u_j^0 = u_0(x_j), \\ \forall 1 \leq n \leq N, u_0^n = u_M^n, \\ \forall 1 \leq n \leq N, \forall 0 \leq j \leq M, u_j^n = u_j^{n-1} - \dfrac{ah_t}{h_x}(u_j^{n-1} - u_{j-1}^{n-1}), \end{cases} \tag{28}$$

où, pour simplifier l'écriture, nous avons introduit $u_{-1}^n = u_{M-1}^n$.

Nous reportons sur la figure 11 les résultats numériques obtenus à partir de ces deux schémas. Lorsque l'on raffine le maillage (quand on augmente N et M), la solution obtenue par le schéma décentré (28) tend vers la solution exacte tandis que la solution obtenue par le schéma centré (26) se dégrade de plus en plus. La solution obtenue par le schéma centré semble diverger quand $N, M \to +\infty$.

Il va de soi que le schéma décentré a un comportement tout à fait acceptable tandis que le schéma centré est totalement inutilisable. Il ne suffit donc pas de remplacer les dérivées partielles de l'EDP par des taux d'accroissement pour obtenir un bon schéma numérique. Pour être sûr que le schéma que l'on conçoit a un bon comportement, il faut vérifier certaines propriétés que nous allons aborder dans la section suivante.

Exercice 15. *La figure 11 a été obtenue pour $a = 1$, $T = 0.5$, $u_0(x) = x^2(1 - x^2)$ et $(N, M) = (15, 20)$, $(30, 40)$, $(60, 80)$ et $(120, 160)$. Dans ces conditions la solution exacte est $u(t, x) = u_0(f(x - at))$ (voir exercice 13). Coder les schémas* (26) *et* (28) *et retrouver la figure 11. On pourra aussi explorer d'autres valeurs de (N, M).*

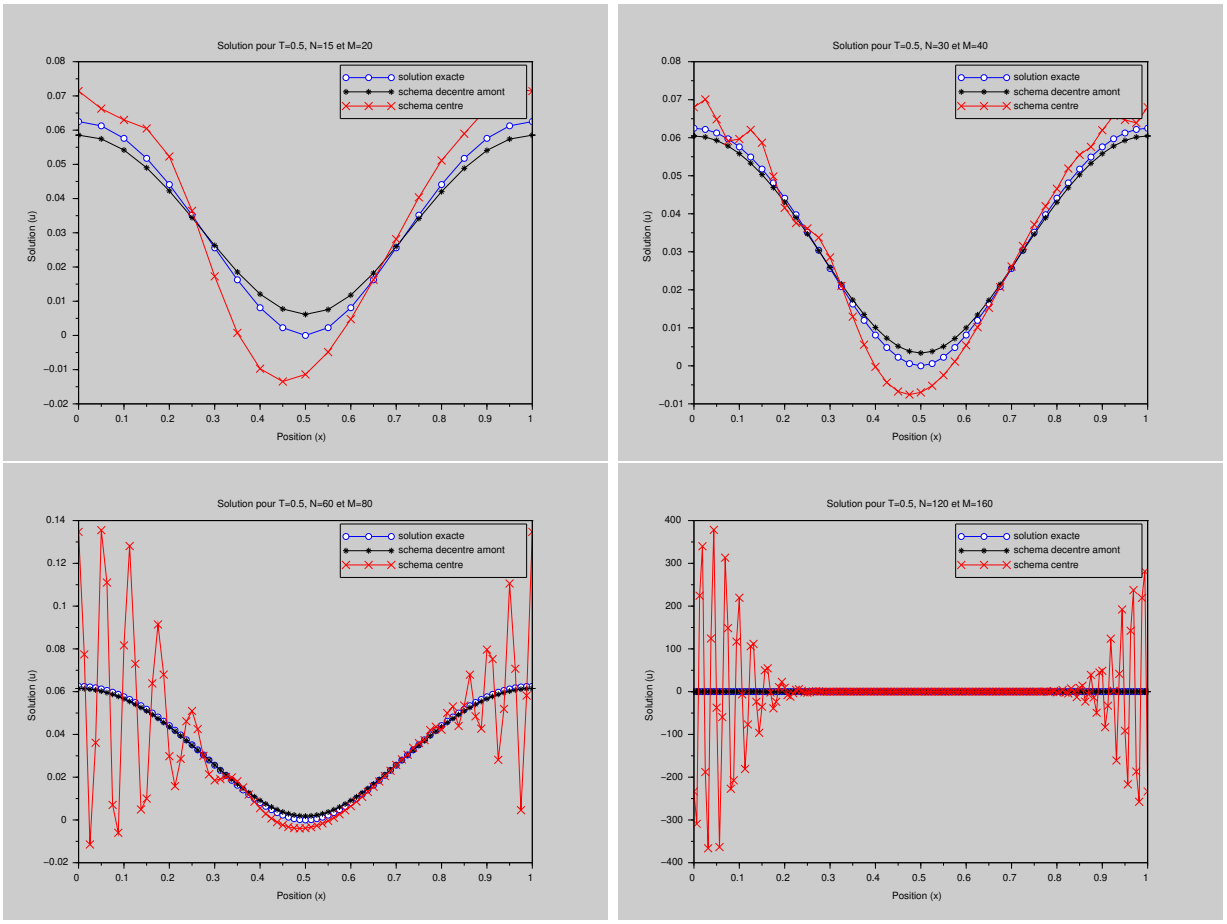

FIGURE 11 – Équation de transport avec conditions périodiques : solution exacte (à l'instant final) et solutions (à l'instant final) obtenues par les schémas décentré et centré pour $T = 0.5$, $a = 1$ et $(N, M) = (15, 20)$, $(30, 40)$, $(60, 80)$ et $(120, 160)$. La condition initiale est $u_0(x) = x^2(1-x)^2$.

3.3.2 Analyse numérique : définitions et théorèmes

Nous définissons maintenant des notions que nous utilisons plus loin pour analyser les deux schémas introduits précédemment. Pour cela, à n donné, nous introduisons $U^n = (u_j^n)_{1 \leq j \leq M}$: u_j^n est égal à la j-ème coordonnée du vecteur U^n. On peut montrer que U^n vérifie la relation de récurrence

$$U^0 = U_0 \qquad \text{et} \qquad \forall n \geq 0, \quad U^{n+1} = AU^n + h_t F^n, \tag{29}$$

où U_0 est donné par $(U_0)_j = u(0, x_j)$, A est une matrice et F^n est un vecteur. Afin de simplifier notre propos, nous considérons que le système à inverser a M inconnues que l'on numérote de 1 à M. Ainsi, pour tout $0 \leq n \leq N$, $U^n \in \mathbb{R}^M$ avec $\forall 1 \leq j \leq M$, $(U^n)_j = u_j^n$ et de même $A \in \mathbb{R}^{M \times M}$ et $F^n \in \mathbb{R}^M$. La valeur de u_0^n n'est pas considérée comme inconnue du problème car elle est donnée par une condition limite (Dirichlet ou périodique). Si le système dispose d'un nombre différent d'inconnues, on pourra aisément adapter notre propos.

Remarque 9. *Dans le cas présent, nous considérons une suite dont la formule de récurrence ne nécessite que le pas de temps précédent (schéma à un pas de temps). En pratique, certains*

schémas nécessitent plusieurs pas de temps précédents. Le schéma (29) *ainsi que les notions que nous abordons dans cette section peuvent être adaptés à ce cas de figure. Nous verrons ceci plus loin dans le cours.*

Définissons maintenant un certain nombre de notions utiles pour étudier le comportement du schéma (29). On introduit notamment $\|\cdot\|$ une norme sur $\mathbb{R}^M$.

Définition 2 (Stabilité)**.** *Soit $\mathcal{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$ tel que $(0,0) \in \overline{\mathcal{S}}$. On dit que le schéma numérique* (29) *est stable pour la norme $\|\cdot\|$ sous la condition $\mathcal{S}$ s'il existe $C_1, C_2 > 0$ qui ne dépendent que de T tels que $\forall (h_x, h_t) \in \mathcal{S}$, $\forall U_0 \in \mathbb{R}^M$, $\forall (F^n)_{0\leq n\leq N} \in \mathbb{R}^{(N+1)\times M}$, on a*

$$\max_{0\leq n\leq N} \|U^n\| \leq C_1 \|U_0\| + C_2 \max_{0\leq n\leq N} \|F^n\|, \tag{30}$$

où (U^n) est l'unique suite vérifiant (29)*.*

On dit également que le schéma (29) *est inconditionnellement stable pour la norme $\|\cdot\|$ si la définition précédente s'applique avec $\mathcal{S} = \mathbb{R}_+^* \times \mathbb{R}_+^*$.*

Définition 3 (Erreur de troncature)**.** *L'erreur de troncature (ou erreur de consistance) du schéma* (29) *au temps t_n est un vecteur $\varepsilon^n \in \mathbb{R}^M$ $(1 \leq n \leq N)$ défini par*

$$\varepsilon^{n+1} := \widetilde{u}^{n+1} - A\widetilde{u}^n - h_t F^n, \tag{31}$$

où $\widetilde{u}^n \in \mathbb{R}^M$ est défini par $(\widetilde{u}^n)_j = u(t_n, x_j)$ avec u la solution exacte du problème associé à (29)*.*

Définition 4 (Consistance)**.** *On dit que le schéma numérique* (29) *est consistant pour la norme $\|\cdot\|$ si pour toute solution exacte régulière u du problème associé à* (29)*, on a*

$$\lim_{\substack{N\to+\infty \\ M\to+\infty}} \max_{0\leq n\leq N} \frac{\|\varepsilon^n\|}{h_t} = 0, \tag{32}$$

où ε^n est l'erreur de troncature (et $h_t = T/N$).

On dit, de plus, que le schéma (29) *est consistant d'ordre $p \in \mathbb{N}^*$ en temps et $q \in \mathbb{N}^*$ en espace pour la norme $\|\cdot\|$ si pour toute solution régulière u (du problème avant discrétisation) il existe une constante $C > 0$ indépendante de h_t et h_x telle que*

$$\forall N, M \geq 2, \quad \max_{0\leq n\leq N} \frac{\|\varepsilon^n\|}{h_t} \leq C(h_t^p + h_x^q). \tag{33}$$

Remarque 10. *Notons que la constante C dans la définition 4 dépend de la solution u du problème associé à* (29)*. La condition importante sur cette constante est qu'elle ne doit pas dépendre de la discrétisation (h_t, h_x, N, M).*

Définition 5 (Convergence)**.** *On dit que le schéma numérique* (29) *converge (ou est convergent) pour la norme $\|\cdot\|$ sous la condition $\mathcal{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$ si*

$$\lim_{\substack{(h_t,h_x)\in\mathcal{S} \\ (h_t,h_x)\to(0,0)}} \max_{0\leq n\leq N} \|\tilde{u}^n - U^n\| = 0, \tag{34}$$

avec $(\tilde{u}^n)_j = u(t_n, x_j)$, $Nh_t = T$ et $Mh_x = 1$ où u est la solution du problème associé à (29)*.*

De manière similaire, on dit que le schéma numérique (29) *converge (ou est convergent) à l'ordre* $p \in \mathbb{N}^*$ *en temps et* $q \in \mathbb{N}^*$ *en espace pour la norme* $\|\cdot\|$ *sous la condition* $\mathcal{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$ *s'il existe une constante* $C > 0$ *indépendante de* h_t *et* h_x *telle que*

$$\forall (h_t, h_x) \in \mathcal{S}, \quad \max_{0 \leq n \leq N} \|\tilde{u}^n - U^n\| \leq C(h_t^p + h_x^q). \tag{35}$$

On dit enfin que le schéma est convergent (resp. convergent à l'ordre p *en temps et* q *en espace) pour la norme* $\|\cdot\|$ *si la définition ci-dessus est vérifiée avec* $\mathcal{S} = \mathbb{R}_+^* \times \mathbb{R}_+^*$.

Ces définitions sont reliées par le théorème de Lax.

Théorème 4 (Théorème de Lax)**.** *Si le schéma* (29) *est stable pour la norme* $\|\cdot\|$ *sous la condition* $\mathcal{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$ *et consistant (respectivement consistant d'ordre* p *en temps et* q *en espace) pour la norme* $\|\cdot\|$*, alors le schéma* (29) *est convergent (respectivement convergent d'ordre* p *en temps et* q *en espace) pour la norme* $\|\cdot\|$ *sous la condition* $\mathcal{S}$*.*

De plus, si le problème exact (le problème vérifié par u*) est bien posé, alors la consistance et la stabilité du schéma sont nécessaires à sa convergence (en norme* $\|\cdot\|$ *pour les trois notions).*

Le théorème 4 s'applique aussi avec $\mathcal{S} = \mathbb{R}_+^* \times \mathbb{R}_+^*$. Avant d'exposer une preuve de ce théorème faisons quelques remarques sur les définitions précédentes.

Concernant la stabilité (cf définition 2), on peut dire que la stabilité du schéma correspond à la continuité de l'application linéaire qui à U_0 et (F^n) associe (u_j^n) avec une constante de continuité que ne dépend pas de h_t et h_x. Cette continuité permet d'obtenir le fait qu'une "petite" modification de la donnée initiale ou des termes F^n ne crée qu'une "petite" modification du résultat obtenu et que cette propriété se conserve lorsque $(h_t, h_x) \to (0, 0)$. Remarquons sur la figure 11 que l'erreur commise par le schéma centré augmente fortement lorsque $(h_t, h_x) \to (0, 0)$, nous verrons plus loin que ce comportement est dû à un défaut de stabilité.

La condition $\mathcal{S}$ permet de traiter des cas où le schéma n'est stable que si on restreint les h_t et h_x utilisés. Nous verrons dans la suite de cette section que le schéma décentré amont (28) n'est stable que sous une certaine condition que nous expliciterons.

Concernant la norme $\|\cdot\|$, nous verrons dans cette section que la norme $\|\cdot\|_\infty$ définie par $\|U\|_\infty := \max_{1 \leq j \leq M} |U_j|$ est adaptée à l'équation de transport. On pourrait considérer d'autres normes pour exprimer cette stabilité. Par exemple, on pourrait considérer la norme $\|U\|_{L^2} := \sqrt{h_x}(\sum_{0 \leq j \leq M} (U_j)^2)^{1/2}$. La définition proposée ne serait alors pas nécessairement équivalente à celle de la définition 2. En effet, si toutes les normes sont équivalentes en dimension finie, rien ne prouve que les constantes de la relation d'équivalence ne dépendent pas de h_t et h_x. La norme $\|\cdot\|_{L^2}$ sera considérée pour mener l'analyse numérique de l'équation de la chaleur.

L'erreur de troncature représente l'erreur que fait le schéma au cours d'un pas de temps en partant de la solution exacte. En effet, on compare (cf définition 3) la solution exacte $\widetilde{u}^{n+1}$ à la solution obtenue en faisant un pas de temps du schéma à partir de la solution exacte au pas de temps précédent $\widetilde{u}^n$.

La consistance d'un schéma (cf définition 4) signifie que l'erreur de troncature commise par unité de temps (on divise par h_t) tend vers 0 lorsque $(h_t, h_x) \to (0, 0)$. De plus, on peut quantifier la notion de consistance grâce à l'ordre de consistance (plus les ordres de

consistance en temps et en espace sont élevés, plus l'erreur de troncature décroît vite vers 0 quand $(h_t, h_x) \to (0,0)$). En quelque sorte, la consistance signifie que le schéma est cohérent avec le problème exact (les termes présents dans le schéma numérique correspondent à des termes du problème exact et vice-versa).

La convergence d'un schéma signifie que la solution approchée obtenue correspond bien à une approximation de la solution exacte au sens où si l'on fait tendre les pas de temps et d'espace vers 0, l'erreur commise par le schéma tend aussi vers 0. Là encore on peut quantifier cette décroissance de l'erreur commise avec la notion d'ordre de convergence. Étant donné que les ordres de consistance et de convergence se correspondent (voir théorème 4), on parle parfois simplement d'ordre d'un schéma (sans préciser 'convergence' ou 'consistance').

En pratique, la convergence est la propriété que l'on veut avoir car elle garantit que le schéma considéré fournit une approximation raisonnable du problème exact (sous réserve que les pas de temps et d'espace sont suffisamment petits). Cependant, cette propriété n'est pas aisée à démontrer directement. C'est pourquoi nous utilisons le théorème 4 qui assure la convergence à partir de la stabilité et de la consistance. Notons également que, si le problème exact est bien posé, alors avoir la convergence est équivalent à avoir la stabilité et la consistance. Ceci signifie qu'en pratique, on ne peut pas espérer avoir un schéma convergent s'il manque la consistance ou la stabilité.

Remarque 11. *Les notions de consistance et de convergence sont différentes. Les jurys se plaignent des candidats qui confondent ces deux notions. Ne faîtes pas cette erreur.*

Remarque 12. *Toutes ces définitions caractérisent le comportement du schéma lorsque* $(h_t, h_x) \to (0,0)$. *La plupart des résultats d'analyse numérique se concentrent sur cette limite.*

Nous allons maintenant prouver la première partie du théorème 4 qui est celle que l'on utilise en pratique.

Preuve partielle du théorème 4. Dans cette preuve, nous utilisons toutes les définitions précédentes. Tout d'abord, par définition de l'erreur de troncature, $\widetilde{u}^{n+1} = A\widetilde{u}^n + h_t F^n + \varepsilon^{n+1}$. On montre ainsi que $w^n = \widetilde{u}^n - U^n$ vérifie $w^0 = 0$ (puisque $(U^0)_j = u(t_0, x_j)$) et $w^{n+1} = Aw^n + \varepsilon^{n+1}$.

Par stabilité du schéma, on obtient $\max_{0\leq n\leq N} \|w^n\| \leq C_2 \max_{0\leq n\leq N} \dfrac{\|\varepsilon^n\|}{h_t}$. On conclut en utilisant la définition de la consistance. □

Comme nous l'avons dit, en pratique, pour prouver la convergence du schéma on prouve sa consistance et sa stabilité. Il se trouve que les preuves de consistances sont généralement plutôt aisées, la difficulté se trouve donc plutôt dans la preuve de la stabilité. Afin de faciliter ces preuves, nous en donnons maintenant des conditions nécessaires et des conditions suffisantes. Pour $A \in \mathbb{R}^{M\times M}$, on note $|\!|\!| A |\!|\!| = \sup_{V\in\mathbb{R}^M\setminus\{0\}} \dfrac{\|AV\|}{\|V\|}$, où $\|\cdot\|$ est la norme considérée sur les vecteurs. Rappelons que $|\!|\!|\cdot|\!|\!|$ est une norme sur $\mathbb{R}^{M\times M}$. Commençons d'abord par le résultat suivant.

Proposition 9. *Le schéma* (29) *est stable pour la norme* $\|\cdot\|$ *si et seulement si il existe* $C > 0$ *dépendant uniquement de* T *tel que*

$$\forall n \in \mathbb{N}, \quad |\!|\!| A^n |\!|\!| \leq C. \tag{36}$$

Corrolaire 1. *Si $|||A||| \leq 1$, alors le schéma est stable pour la norme $\|\cdot\|$.*

Preuve de la proposition 9. Supposons tout d'abord que le schéma est stable. Dans ce cas, appliquons le schéma à $U_0 = V \in \mathbb{R}^M$ quelconque et $F^n = 0$. La solution obtenue est $U^n = A^n V$. On a donc prouvé qu'il existe $C > 0$ dépendant uniquement de T tel que $\forall V \in \mathbb{R}^M, \quad \|A^n V\| \leq C\|V\|$ et donc (36).

Maintenant supposons que (36) est établie. Appliquons le schéma à $U_0 \in \mathbb{R}^M$ et $(F^n) \in \mathbb{R}^{(N+1)\times M}$. Nous pouvons prouver (voir exercice 16) que la solution obtenue est donnée par

$$\forall 0 \leq n \leq N, \quad U^n = A^n U_0 + h_t \sum_{\ell=0}^{n-1} A^{n-\ell-1} F^\ell.$$

Ainsi, $\|U^n\| \leq |||A^n|||\|U_0\| + h_t \sum_{\ell=0}^{n-1} |||A^{n-\ell-1}|||\|F^n\|$. En utilisant (36), on obtient $\|U^n\| \leq C\|U_0\| + Ch_t n \max_{0\leq\ell\leq N} \|F^\ell\|$, avec C la constante de (36) qui dépend uniquement de T. De plus, $nh_t = t_n \leq T$. On a donc établi la stabilité du schéma. □

Exercice 16. *Soient $U_0 \in \mathbb{R}^M$ et $(F^n) \in \mathbb{R}^{(N+1)\times M}$. Montrer que l'unique solution de* (29) *est donnée par*

$$\forall 0 \leq n \leq N, \quad U^n = A^n U_0 + h_t \sum_{\ell=0}^{n-1} A^{n-\ell-1} F^\ell. \tag{37}$$

Remarque 13. *Une conséquence de la proposition 9 est que la stabilité du schéma ne dépend que de A et non de F.*

Nous introduisons une dernière définition : la stabilité au sens de Von Neumann. Lorsque l'on veut prouver qu'un schéma de la forme (29) est instable pour les normes $\|\cdot\|_\infty$ et $\|\cdot\|_{L^2}$, la façon standard de procéder est de montrer qu'il n'est pas stable au sens de Von Neumann. On considère alors $F^n = 0$ et une condition initiale sous la forme d'une onde spatiale : $u_j^0 = e^{2i\pi k x_j}$ avec i la racine carré de l'unité et $k \in \mathbb{Z}$. On étudie ensuite l'effet qu'a l'application du schéma sur cette condition initiale complexe.

L'intérêt de choisir un vecteur de la forme $u_j^0 = e^{2i\pi k x_j}$ est que l'on a $u_j^1 = \sum_{\ell=1}^M a_{j\ell} u_\ell^0$. En prenant $\mathcal{A}_j(k) = \sum_{\ell=1}^M a_{j\ell} e^{2i\pi k(x_\ell - x_j)}$, on montre que $\forall 1 \leq j \leq M,\ u_j^1 = \mathcal{A}_j(k) u_j^0$ avec $\mathcal{A}_j(k) \in \mathbb{C}$.

Définition 6 (Stabilité au sens de Von Neumann)**.** *On dit qu'un schéma* (29) *est stable au sens de Von Neumann si pour tout $1 \leq j \leq M$ et pour tout $k \in \mathbb{Z}$, $|\mathcal{A}_j(k)| \leq 1$.*

Proposition 10. *La stabilité au sens de Von Neumann est nécessaire à la stabilité du schéma pour la norme $\|\cdot\|_\infty$. Dit autrement, la stabilité du schéma pour la norme $\|\cdot\|_\infty$ implique la stabilité au sens de Von Neumann.*

Démonstration. On peut montrer que si on considère $u_j^0 = \Re(e^{2i\pi k x_j})$ et $F^n = 0$ alors $\forall 0 \leq n \leq N, u_j^n = \Re((\mathcal{A}_j(k))^n e^{2i\pi k x_j})$. Idem, si $u_j^0 = \Im(e^{2i\pi k x_j})$, alors $\forall 0 \leq n \leq N, u_j^n = \Im((\mathcal{A}_j(k))^n e^{2i\pi k x_j})$.

Ainsi, si le schéma (29) est stable pour la norme $\|\cdot\|_\infty$, alors $\max_{\substack{0\leq n\leq N\\ 1\leq j\leq M}} |\Re((\mathcal{A}_j(k))^n e^{2i\pi k x_j})| \leq C$ où $C > 0$ ne dépend que de T. En faisant le même raisonnement sur la partie imaginaire, on obtient $\max_{\substack{0\leq n\leq N\\ 1\leq j\leq M}} |(\mathcal{A}_j(k))^n e^{2i\pi k x_j}| \leq \sqrt{2}C$.

Or ceci n'est possible que si $|\mathcal{A}_j(k)| \leq 1$ sinon on dépasse la constante $\sqrt{2}C$ en faisant tendre N vers $+\infty$. Ceci prouve le résultat. □

Par la suite nous utilisons la stabilité de Von Neumann lorsque nous voulons prouver qu'un schéma n'est pas stable pour la norme $\|\cdot\|_\infty$.

3.3.3 Analyse des deux schémas donnés en introduction

Nous allons montrer dans cette section que le schéma numérique proposé en (26) n'est pas stable pour la norme $\|\cdot\|_\infty$ contrairement au schéma (28). C'est ce défaut de stabilité qui explique son mauvais comportement.

On considère comme degrés de liberté $U^n = \begin{pmatrix} u_1^n \\ \vdots \\ u_M^n \end{pmatrix} \in \mathbb{R}^M$. Les deux schémas précédents peuvent être écrits sous la forme matricielle (29) avec $F^n = 0$ et

$$A_C = \frac{1}{2}\begin{pmatrix} 2 & -\mathrm{c} & 0 & \dots & 0 & \mathrm{c} \\ \mathrm{c} & 2 & -\mathrm{c} & & & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & & & \ddots & \ddots & -\mathrm{c} \\ -\mathrm{c} & 0 & & & \mathrm{c} & 2 \end{pmatrix}, \tag{38}$$

pour le schéma centré (26) et

$$A_{AM} = \begin{pmatrix} 1-\mathrm{c} & 0 & \dots & 0 & \mathrm{c} \\ \mathrm{c} & \ddots & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \mathrm{c} & 1-\mathrm{c} \end{pmatrix}, \tag{39}$$

pour le schéma décentré amont (28). Les deux matrices précédentes sont dans $\mathbb{R}^{M\times M}$ et ont été écrites avec $\mathrm{c} = \frac{ah_t}{h_x}$.

Proposition 11. *Les schémas* (26) *et* (28) *sont tous les deux consistants pour la norme* $\|\cdot\|_\infty$*. De plus, le schéma* (26) *est consistant d'ordre 1 en temps et 2 en espace ; le schéma* (28) *est consistant d'ordre 1 en temps et 1 en espace (pour la norme* $\|\cdot\|_\infty$*).*

Démonstration. Prouvons que le schéma (26) est consistant d'ordre 1 en temps et 2 en espace. La preuve pour le schéma décentré amont est laissée en exercice (cf exercice 17).

On calcule l'erreur de troncature. Ceci revient à calculer une solution approchée au temps t_{n+1} par le schéma à partir de la solution exacte au temps t_n et de comparer le résultat obtenu à la solution exacte au temps t_{n+1}. On a ainsi

$$\varepsilon_j^{n+1} = u(t_{n+1}, x_j) - u(t_n, x_j) + \frac{ah_t}{2h_x}(u(t_n, x_{j+1}) - u(t_n, x_{j-1})),$$

où u est la solution de (25).

On considère ensuite les développements de Taylor suivants :

$$u(t_{n+1},x_j) = u(t_n,x_j) + h_t\frac{\partial u}{\partial t}(t_n,x_j) + O(h_t^2),$$
$$u(t_n,x_{j+1}) = u(t_n,x_j) + h_x\frac{\partial u}{\partial x}(t_n,x_j) + \frac{h_x^2}{2}\frac{\partial^2 u}{\partial x^2}(t_n,x_j) + O(h_x^3),$$
$$u(t_n,x_{j-1}) = u(t_n,x_j) - h_x\frac{\partial u}{\partial x}(t_n,x_j) + \frac{h_x^2}{2}\frac{\partial^2 u}{\partial x^2}(t_n,x_j) + O(h_x^3).$$

Nous obtenons donc

$$\frac{\varepsilon_j^{n+1}}{h_t} = \frac{\partial u}{\partial t}(t_n,x_j) + O(h_t) + a\frac{\partial u}{\partial x}(t_n,x_j) + O(h_x^2).$$

En utilisant le fait que u est solution de (25), on a $\dfrac{\partial u}{\partial t}(t_n,x_j) + a\dfrac{\partial u}{\partial x}(t_n,x_j) = 0$. Nous avons prouvé le fait que le schéma (26) est consistant d'ordre 2 en espace et d'ordre 1 en temps. □

Exercice 17. *Terminer la preuve de la proposition 11 : prouver que le schéma* (28) *est consistant d'ordre 1 en temps et en espace (pour la norme* $\|\cdot\|_\infty$*).*

Proposition 12. *Le schéma* (28) *est stable pour la norme* $\|\cdot\|_\infty$ *sous la condition* $c = \dfrac{ah_t}{h_x} \leq 1$*. De plus, ce schéma est instable si* $c > 1$ *(pour la norme* $\|\cdot\|_\infty$*).*

Démonstration. Supposons $c \leq 1$ et montrons que le schéma est stable. Soit $U_0 \in \mathbb{R}^M$, on calcule $U^1 = AU_0$. Pour $1 \leq j \leq M$, $u_j^1 = (1-c)u_j^0 + cu_{j-1}^0$. Ainsi, $|u_j^1| \leq |1-c||u_j^0| + |c||u_{j-1}^0|$. On obtient $\|U^1\| \leq (|1-c|+|c|)\|U_0\|$. Ceci étant valide pour tout $U_0 \in \mathbb{R}^M$, on a $|\!|\!|A|\!|\!| \leq |1-c|+|c|$. La condition $0 \leq c \leq 1$ donne $|\!|\!|A|\!|\!| \leq 1$ et le schéma est stable d'après le corrolaire 1.

Supposons maintenant que $c > 1$ et montrons que le schéma est instable. Pour cela, nous allons montrer qu'il est instable au sens de Von Neumann. Si $u_j^0 = e^{2i\pi k x_j}$, alors

$$u_j^1 = (1-c)u_j^0 + cu_{j-1}^0 = (1-c+ce^{-2i\pi k h_x})u_j^0 = \mathcal{A}(k)u_j^0.$$

Ici, nous avons $\mathcal{A}(k) = 1-c+ce^{-2i\pi k h_x}$. Pour $c > 1$ et $k = 1$, on a pour $h_x < 1$, $|\mathcal{A}| > 1$ (on montre que $\cos(2\pi k h_x) < 1$). Le schéma n'est pas stable au sens de Von Neumann donc, d'après la proposition 10, il n'est pas stable au sens de la définition 2. □

La mauvais comportement du schéma centré provient d'un défaut de stabilité. La preuve de ce résultat est laissée en exercice.

Exercice 18 (Instabilité du schéma explicite centré)**.** *Montrer que le schéma* (26) *est instable pour la norme* $\|\cdot\|_\infty$ *(c'est-à-dire n'est stable sous aucune condition* $\mathcal{S}$*).*

Nous pouvons tirer plusieurs conclusions de l'étude de ces deux schémas. Tout d'abord, le fait de décentrer la dérivée partielle en espace a permis de rendre le schéma stable. Notons que le sens dans le lequel on décentre cette dérivée est essentiel, on peut montrer que le schéma décentré aval issu de l'approximation $\frac{\partial u}{\partial x}(t_n,x_j) \simeq \frac{u(t_n,x_{j+1})-u(t_n,x_j)}{h_x}$ est instable pour $a > 0$ (toujours pour la norme $\|\cdot\|_\infty$) et qu'il faut changer le sens du décentrage si $a < 0$. Les numériciens commentent ce phénomène en disant qu'il faut décentrer la dérivée en espace en

utilisant les données qui correspondent à de l'information qui arrive et non à de l'information qui part. Cela signifie que si $a > 0$, l'information vient de la gauche, il faut donc décentrer la dérivée vers la gauche.

De plus, nous avons vu dans la proposition 12 que le schéma (28) est stable uniquement sous la condition $ah_t \leq h_x$ qui lie les pas de temps et d'espace. On appelle une telle condition "condition de CFL" pour les noms des trois scientifiques Courant, Friedrichs et Lewy. De la même façon on appelle $c = \frac{ah_t}{h_x}$ le nombre de CFL (parfois aussi appelé nombre de Courant). De telles conditions doivent être associées aux schémas explicites pour espérer les rendre stables. Seuls les schémas implicites peuvent se passer de ce type de condition (voir exercice 20).

Notons que remplacer les dérivées de l'EDP par des taux d'accroissements permet d'assurer la consistance d'un schéma mais pas sa stabilité comme nous venons de le voir. Il faut donc être très prudent quand on discrétise une EDP.

3.3.4 Discrétisation des conditions de Dirichlet

On s'intéresse dans cette section à la prise en compte des conditions de Dirichlet. On cherche donc à discrétiser le problème (22). Pour avoir les mêmes propriétés de stabilité que dans le cas des conditions aux limites périodiques on discrétise la dérivée en espace par une approximation décentrée amont, on utilise donc (27). On obtient le schéma numérique

$$\begin{cases} \forall 0 \leq j \leq M, u_j^0 = u_0(x_j), \\ \forall 1 \leq n \leq N, u_0^n = \alpha(t_n), \\ \forall 1 \leq n \leq N, \forall 1 \leq j \leq M, u_j^n = u_j^{n-1} - \dfrac{ah_t}{h_x}(u_j^{n-1} - u_{j-1}^{n-1}). \end{cases} \tag{40}$$

On a donc

$$A = \begin{pmatrix} 1-\mathrm{c} & 0 & & & \\ \mathrm{c} & 1-\mathrm{c} & \ddots & & \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & \mathrm{c} & 1-\mathrm{c} \end{pmatrix} \quad \text{et} \quad \forall n \geq 0, \quad h_t F^n = \mathrm{c} \begin{pmatrix} \alpha(t_n) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \tag{41}$$

avec $\mathrm{c} = \frac{ah_t}{h_x}$. Remarquons que la condition de Dirichlet a été prise en compte dans le terme "second membre" F^n. Si l'on compare la matrice A avec la matrice A_{AM} dans le cas de conditions périodiques (voir (39)), on voit que le terme c de la première ligne de la matrice A_{AM} a été retiré et remplacé par la présence d'un second membre. Ce terme c correspondait au fait que la valeur qui entrait en 0 dans le cas des conditions périodiques était celle de la valeur de la solution en 1. Notons que ce terme a été remplacé par un $c\alpha(t_n)$ dans le second membre (notons le coefficient c qui est le même que pour les conditions périodiques).

On peut, de la même façon que pour le cas des conditions périodiques, montrer que ce schéma est stable si et seulement la condition de CFL "$c \leq 1$" est vérifiée (vous pouvez chercher cette démonstration pour vous convaincre que vous avez compris la section précédente).

Proposition 13. *Pour un nombre de CFL $c \leq 1$, le schéma décentré amont conserve les bornes : s'il existe A et B réels tels que pour tout $j \in < 0, M >$, $A \leq u_0(x_j) \leq B$ et pour*

tout $n \in < 1, N >$, $A \leq \alpha(t_n) \leq B$, *alors pour tout* $n \in < 0, N >$ *et tout* $j \in < 0, M >$, *on a* $A \leq u_j^n \leq B$.

Démonstration. On a $u_j^{n+1} = u_j^n - \frac{ah_t}{h_x}(u_j^n - u_{j-1}^n) = (1-c)u_j^n + cu_{j-1}^n$ avec $c = \frac{ah_t}{h_x}$. Maintenant, si $A \leq u_j^n \leq B$ et $A \leq u_{j-1}^n \leq B$ alors en utilisant $c \geq 0$ et $1 - c \geq 0$, on a $A \leq u_j^{n+1} \leq B$.

On complète la preuve avec un argument de récurence. □

3.3.5 Diffusion numérique

On peut représenter le nombre de CFL comme le rapport entre les vitesses de propagation de l'information dans le cas continu et le cas discrétisé. En effet, si on considère une condition initiale constante et une perturbation arrivant en entrée du domaine, dans le problème continu la perturbation va progressivement chasser la condition initiale à vitesse a. Dans le problème discrétisé avec un schéma explicite comme (28), au premier pas de temps, le premier point en espace est modifié puis à chaque pas de temps un pas en espace est modifié. Le front de la pertubation de la condition initiale se propage donc à la vitesse h_x/h_t. Ainsi, quand $c \leq 1$, le front de la perturbation se propage plus vite dans le problème numérique que dans le problème continu et inversement quand $c \geq 1$.

Dans le cas où $c \leq 1$, le front de perturbation se propage plus vite dans le cas numérique et les valeurs ont tendance à s'applatir. On appelle ce phénomène la diffusion numérique : tout se passe comme s'il y avait un terme de diffusion supplémentaire dans l'équation. Quand $c \geq 1$, le schéma est instable et des phénomènes similaires à ceux du schéma centré apparaissent (voir figure 11).

Nous avons étudié le problème de transport–Dirichlet avec $\alpha(t) = 2$, $u_0(x) = 1$ et $a = 1$. La solution exacte est

$$u(t,x) = \begin{cases} 2 & \text{si } x \leq at, \\ 1 & \text{sinon.} \end{cases}$$

On calcule la solution numérique avec le schéma (40) pour différents nombres de CFL. L'influence de la CFL est ainsi illustrée sur la figure 12. On observe cet effet de diffusion numérique : un petit nombre de CFL tend à dissiper le saut comme le ferait un terme de diffusion (voir le chapitre sur l'équation de la chaleur). Bien entendu, l'effet de la diffusion numérique s'estompe quand $h \to 0$ (en supposant que l'on vérifie la condition de CFL).

Un dernier phénomène intéressant est le fait que le schéma numérique permet d'obtenir la solution exacte lorsque $c = 1$ (la démonstration est laissée en exercice). Cette valeur de CFL n'est en pratique pas utilisée car le schéma est alors proche d'être instable et de toutes façons on ne trouve la solution exacte que dans des cas simples tels que celui de l'équation de transport unidimensionnelle à vitesse constante. Dans le cas de problèmes plus complexes ne pouvant être résolus que par de la simulation numérique, il ne faut pas s'attendre à ce que le schéma fournisse la solution exacte.

Exercice 19. *Montrer que lorsque* $c = 1$, *l'erreur commise par le schéma* (40) *est nulle. La même démonstration peut être faite pour le cas des conditions périodiques* (28).

Application : la figure 5 représente les solutions exactes de l'équation de transport à $t = 0.5$ *pour* $M = 20$ *et différentes vitesses* a. *Elle a été obtenue en appliquant le schéma* (40). *Donner pour chaque valeur de* a *la valeur de* N *qui a été utilisée.*

NB : pour rappel, l'intervalle de temps est $(0, 0.5)$, *l'intervalle d'espace est* $(0, 1)$. *Vous pouvez essayer de retrouver la figure 5 pour vérifier vos résultats.*

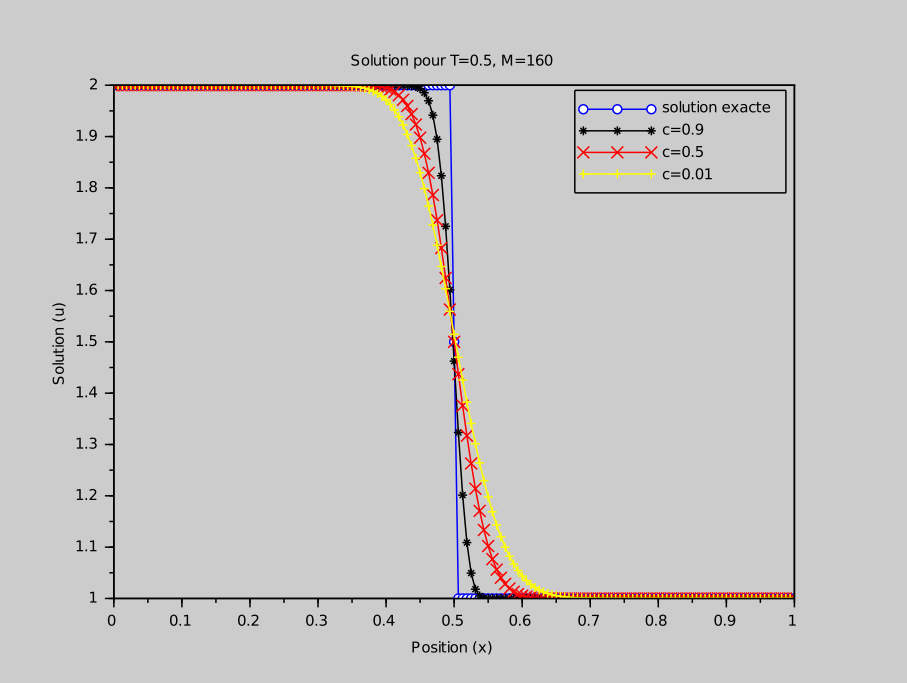

FIGURE 12 – Solution du problème de transport–Dirichlet pour différents nombres de CFL c ($t = 0.5$, $M = 160$).

3.3.6 Schéma implicite

Exercice 20 (Schéma implicite)**.** *Dans cet exercice, on s'intéresse à une discrétisation implicite du problème* (22)*. On approche les dérivées partielles en temps et espace par des approximations décentrées amont :*

$$\frac{\partial u}{\partial t}(t_n, x_j) \simeq \frac{u(t_n, x_j) - u(t_{n-1}, x_j)}{h_t}, \qquad \frac{\partial u}{\partial x}(t_n, x_j) \simeq \frac{u(t_n, x_j) - u(t_n, x_{j-1})}{h_x}. \tag{42}$$

1. *En considérant la condition de Dirichlet $u(t_n, 0) = \alpha(t_n)$ (donc pour $a > 0$), écrire le schéma numérique issu de* (42)*. Montrer qu'il peut se mettre sous la forme $\widetilde{A}U^n = U^{n-1} + h_t \widetilde{F}^n$. Donner $\widetilde{A}$ et $\widetilde{F}^n$.*
2. *Montrer que $\widetilde{A}$ est inversible. En déduire que ce schéma vérifie* (29) *en donnant A et F^n. On pourra donc utiliser les définitions de la section 3.3.2.*
3. *Soit $V \in \mathbb{R}^M$, on note $U = AV$. Calculer les coordonnées de U en fonction de celles de V. En déduire que $|||A||| \leq 1$ (pour $\|\cdot\|_\infty$) et que ce schéma est inconditionnellement stable pour la norme $\|\cdot\|_\infty$.*
4. *Montrer que ce schéma est consistant d'ordre 1 en temps et en espace pour la norme $\|\cdot\|_\infty$. Que peut-on dire de la convergence de ce schéma ?*
5. *Coder le schéma proposé. Tester plusieurs nombres de CFL. Vérifier que le schéma converge même quand $c > 1$.*
6. *On considère $a = 1$, $u_0 = 0$ et $\alpha(t) = t^2$. Quelle est la solution exacte du problème ? Approcher la solution grâce au schéma implicite. Que remarquez-vous en $(t, x) = (0.1, 0.5)$? Si vous l'avez codé, comparez avec le schéma explicite. Comment expliquez-vous ceci ?*

4 Équation de la chaleur

Nous considérons le domaine $\Omega = (0,1)$. Nous étudions dans ce chapitre les solutions au problème

$$\text{Trouver } u : (0,T)\times(0,1) \text{ vérifiant } \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = f & \text{dans } (0,T)\times(0,1), \\ u(0,.) = u_0 & \text{dans } (0,1), \\ u(t,0) = 0, \quad u(t,1) = 0 & \forall t \in (0,T), \end{cases} \tag{43}$$

où f et u_0 sont des fonctions données. Notons que l'on a imposé des conditions de Dirichlet homogènes sur tout le bord du domaine (ici $x = 0$ et $x = 1$).

4.1 Recherche de solutions régulières

Le but de cette section est de mettre en évidence la forme que peuvent avoir les solutions du problème (43). Pour cela, nous supposerons que les données f et u_0 sont suffisamment régulières. On traitera d'abord le cas du domaine $\Omega = (0,1)$, puis nous étudierons les solutions du problème sur la droite réelle $\Omega = \mathbb{R}$.

4.1.1 Cas du domaine borné $\Omega = (0,1)$

Le but de cette section est de construire une solution explicite au problème (43). Nous commençons tout d'abord par appliquer la méthode de séparation des variables. On cherche donc une solution de la forme $u(t,x) = \Phi(t)\Psi(x)$ où Φ et Ψ sont des fonctions à déterminer. Pour simplifier nous supposons que f est nul, on a alors $\Phi'(t)\Psi(x) = \Phi(t)\Psi''(x)$. On pourrait se passer de cette hypothèse (voir par exemple le texte B1 de la session 2006).

Tant que les fonctions Φ et Ψ ne s'annulent pas, on peut écrire $\frac{\Phi'(t)}{\Phi(t)} = \frac{\Psi''(x)}{\Psi(x)}$. Étant donné que le premier terme dépend uniquement de t et le second uniquement de x, on peut dire que ces deux termes sont constants. On suppose que cette constante est non nulle (sinon on a une solution identiquement nulle). Pour satisfaire les conditions de Dirichlet, on impose aussi $\Psi(0) = \Psi(1) = 0$.

Les fonctions Φ et Ψ vérifient donc

$$\begin{aligned} &\Phi'(t) = K\Phi(t), \quad 0 \le t \le T, \\ &\Psi''(x) = K\Psi(x), \quad 0 \le x \le 1, \quad \Psi(0) = 0, \quad \Psi(1) = 0. \end{aligned}$$

La fonction Φ vaut donc $\Phi(t) = e^{Kt}$. Discutons maintenant de Ψ et du signe de K.

Si $K > 0$, alors $\Psi(x) = ae^{\sqrt{K}x} + be^{-\sqrt{K}x}$. Les conditions aux limites donnent $a + b = 0$ et $a(e^{\sqrt{K}} - e^{-\sqrt{K}}) = 0$. La seule solution est donc la solution nulle.

On suppose maintenant $K < 0$. Dans ce cas, $\Psi(x) = ae^{i\sqrt{|K|}x} + be^{-i\sqrt{|K|}x}$. Les conditions aux limites donnent alors $a + b = 0$ et $a\sin(\sqrt{|K|}) = 0$. La seule façon d'avoir une solution non identiquement nulle est d'imposer $\sin(\sqrt{|K|}) = 0$ et donc de prendre $K = -\pi^2 k^2$ avec $k \in \mathbb{Z}^*$ (rappelons qu'on a supposé $K < 0$). On remarque ainsi que les fonctions de la forme $u(t,x) = \sin(k\pi x)e^{-k^2\pi^2 t}$ vérifient le problème (43) (avec $f = 0$ et $u_0(x) = \sin(k\pi x)$).

Nous venons de trouver des solutions pour des données initiales sous la forme particulière $u_0(x) = \sin(k\pi x)$. Pour généraliser ce raisonnement à d'autres données initiales, nous allons

maintenant chercher des solutions sous la forme d'une série de Fourier dont chaque terme sera de la forme de la solution particulière trouvée précédemment.

On suppose que la donnée initiale u_0 est $C^2([0,1])$ avec $u_0(0) = u_0(1) = 0$. On note U_0 la fonction définie sur $[-1,1]$ par

$$U_0(x) = \begin{cases} u_0(x) & \text{si } x \in [0,1], \\ -u_0(-x) & \text{sinon.} \end{cases}$$

On prolonge également U_0 sur $\mathbb{R}$ par 2-périodicité. La fonction obtenue est $C^1(\mathbb{R})$. D'après le théorème de Dirichlet, cette fonction est donc égale à sa série de Fourier (qui converge). De plus, comme U_0 est impaire, cette série de Fourier ne comprend que des sinus.

Donc, en particulier, sur $[0,1]$ on a

$$u_0(x) = \sum_{k \in \mathbb{N}^*} b_k \sin(k\pi x),$$

avec

$$b_k = \int_{-1}^{1} U_0(x) \sin(k\pi x) \, \mathrm{d}x = 2 \int_0^1 u_0(x) \sin(k\pi x) \, \mathrm{d}x.$$

On obtient une autre expression de b_k par deux intégrations par parties (et on utilise $u_0 \in C^2([0,1])$)

$$b_k = \frac{2}{k\pi} \int_0^1 u_0'(x) \cos(k\pi x) \, \mathrm{d}x = \frac{-2}{k^2\pi^2} \int_0^1 u_0''(x) \sin(k\pi x) \, \mathrm{d}x.$$

Cette dernière expression permet de donner la majoration suivante sur la valeur des coefficients b_k

$$|b_k| \leq \frac{C}{k^2},$$

avec $C > 0$ une constante indépendante de k. Pour $t \geq 0$, on a $e^{-k^2\pi^2 t} \leq 1$, la série $u(t,x) = \sum_{k \in \mathbb{N}^*} b_k \sin(k\pi x) e^{-k^2\pi^2 t}$ est donc normalement convergente sur $[0,T] \times [0,1]$.

De plus, pour $0 < t$, $e^{-k^2\pi^2 t} < 1$, la série $\sum_{k \in \mathbb{N}^*} -k^2\pi^2 b_k \sin(k\pi x) e^{-k^2\pi^2 t}$ est donc normalement convergente sur $[a,T] \times [0,1]$ pour tout $a > 0$. La fonction u est donc dérivable par rapport à la variable t et sa dérivée est donnée par

$$\frac{\partial u}{\partial t}(t,x) = -\sum_{k \in \mathbb{N}^*} k^2\pi^2 b_k \sin(k\pi x) e^{-k^2\pi^2 t}.$$

On peut également montrer que u est dérivable deux fois par rapport à la variable x et que cette dérivée seconde à la même valeur que la dérivée première en t. Cette série de Fourier vérifie donc l'équation de la chaleur.

Enfin, en utilisant des arguments présentés dans la section 4.2, on peut montrer que cette solution est unique.

Nous avons prouvé le théorème suivant.

Théorème 5. *Si $u_0 \in C^2([0,1])$ vérifie $u_0(0) = u_0(1) = 0$ et si $f = 0$, alors le problème (43) admet une unique solution. De plus, cette solution est donnée explicitement par*

$$u(t,x) = \sum_{k \in \mathbb{N}^*} b_k \sin(k\pi x) e^{-k^2\pi^2 t},$$

avec $b_k = 2 \int_0^1 u_0(x) \sin(k\pi x) \, \mathrm{d}x$.

4.1.2 Cas de la droite réelle $\Omega = \mathbb{R}$

On s'intéresse ici au problème

$$\text{Trouver } u : (0,T)\times\mathbb{R} \text{ vérifiant } \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = f & \text{dans } (0,T)\times\mathbb{R}, \\ u(0,.) = u_0 & \text{dans } \mathbb{R}, \end{cases} \tag{44}$$

où f et u_0 sont donnés. Notons que comme le domaine spatial $\mathbb{R}$ n'a pas de bord, le problème (44) n'a pas de conditions aux limites.

On peut utiliser la transformée de Fourier de u par rapport à x pour résoudre ce problème. On la note

$$\mathcal{F}(u)(t,\omega) := \hat{u}(t,\omega) := \int_{\mathbb{R}} e^{-i\omega x} u(t,x)\ \mathrm{d}x.$$

On notera la transformée de Fourier inverse

$$\mathcal{F}^{-1}(v)(t,x) := \frac{1}{2\pi}\int_{\mathbb{R}} e^{i\omega x} v(t,\omega)\ \mathrm{d}\omega.$$

Ces transformations peuvent être appliquées à des fonctions $L^1(\mathbb{R})$ en espace. On rappelle que si, de plus, u est dérivable par rapport à x et de dérivée intégrable alors $\mathcal{F}(\partial_x u)(t,\omega) = i\omega\mathcal{F}(u)(t,\omega)$.

La définition de la transformée de Fourier et de son inverse peut être étendue aux fonctions de $L^2(\mathbb{R})$. De plus, si $u \in L^2(\mathbb{R})$, alors $\hat{u} \in L^2(\mathbb{R})$.

Dans la suite, on suppose que $f(t,.) \in L^2(\mathbb{R})$ et que $u(t,.) \in L^2(\mathbb{R})$ est dérivable par rapport à x et vérifie $\partial_x u(t,.) \in L^2(\mathbb{R})$, on montre que $\hat{u}(t,\omega)$ est solution du problème

$$\frac{\partial \hat{u}}{\partial t}(t,\omega) + \omega^2 \hat{u}(t,\omega) = \hat{f}(t,\omega) \text{ dans } (0,T)\times\mathbb{R},$$
$$\hat{u}(0,\omega) = \hat{u}_0(\omega) \text{ dans } \mathbb{R}.$$

Il s'agit d'une famille d'EDO linéaires indexée par ω. Si f est continue en temps, alors, d'après le théorème de Cauchy–Lipschitz global, chacune de ces EDO admet une unique solution donnée par

$$\hat{u}(t,\omega) = \hat{u}_0(\omega)e^{-\omega^2 t} + \int_0^t \hat{f}(s,\omega)e^{-\omega^2(t-s)}\ \mathrm{d}s.$$

Étant donné que $u(t,.) \in L^2(\mathbb{R})$, on a $\hat{u}(t,.) \in L^2(\mathbb{R})$. Après quoi, on applique la transformée de Fourier inverse pour trouver une solution à l'équation (44). En utilisant les propriétés de la transformée de Fourier, on obtient donc

$$\begin{aligned} u(t,x) =& \mathcal{F}^{-1}(\hat{u}_0) * \mathcal{F}^{-1}(e^{-\omega^2 t}) + \mathcal{F}^{-1}\Big(\int_0^t \hat{f}(s,\omega)e^{-\omega^2(t-s)}\ \mathrm{d}s\Big) \\ =& (u_0 * G(t,.))(t,x) + \int_{\mathbb{R}}\int_0^t G(t-s,x-y)f(s,y)\ \mathrm{d}s\ \mathrm{d}y, \end{aligned}$$

avec $G(t,x) = \mathcal{F}^{-1}(e^{-\omega^2 t})(x) = \dfrac{1}{2\pi}\sqrt{\dfrac{\pi}{t}}e^{-\frac{x^2}{4t}}$ le noyau de la chaleur et $*$ le produit de convolution. Pour rappel $(u_0 * G(t,.))(t,x) := \int_{\mathbb{R}} u_0(y)G(t,x-y)\ \mathrm{d}y$. Pour le calcul de l'expression de $G(t,x)$, on pourra faire l'exercice 21.

On a donc établi le résultat suivant.

Théorème 6. *Supposons que $u \in L^2(0,T;H^2(\mathbb{R})) \cap H^1(0,T;L^2(\mathbb{R}))$ est solution du problème (44). Alors $u(t,x) = (u_0 * G(t,.))(t,x) + \int_{\mathbb{R}} \int_0^t G(t-s,x-y)f(s,y)\ \mathrm{d}s\ \mathrm{d}y$, avec $G(t,x) = \dfrac{1}{2\pi}\sqrt{\dfrac{\pi}{t}}e^{-\frac{x^2}{4t}}$.*

Remarque 14. *A $t > 0$ fixé, $G(t,.)$ est une Gaussienne. Quand $t \to 0^+$, le pic de la Gaussienne devient singulier et on retrouve une masse de Dirac.*

Exercice 21. *Pour $t > 0$, calculer $\mathcal{F}^{-1}(e^{-\omega^2 t})$.*

Indice : utiliser des intégrations par parties et les propriétés de la transformée de Fourier pour prouver que pour tout $t > 0$, $x \mapsto \mathcal{F}^{-1}(e^{-\omega^2 t})$ est solution d'une EDO d'ordre 2. Prouver que le problème de Cauchy vérifié par cette fonction admet une unique solution et vérifier que le noyau de la chaleur est cette unique solution.

Proposition 14. *Pour tout $t > 0$, pour tout $x \in \mathbb{R}$, on a $\dfrac{\partial G}{\partial t}(t,x) - \dfrac{\partial^2 G}{\partial x^2}(t,x) = 0$.*

Remarque 15. *En fait, on peut montrer que $\dfrac{\partial G}{\partial t} - \dfrac{\partial^2 G}{\partial x^2} = \delta_{(0,0)}$ au sens des distributions. La fonction G est donc aussi appelée solution élémentaire de la chaleur.*

Proposition 15 (Effet régularisant de la chaleur)**.** *Si $f = 0$ et $u_0 \in L^2(\mathbb{R})$, alors $u \in C^\infty([a,T] \times \mathbb{R})$ pour tout a vérifiant $0 < a < T$.*

Démonstration. C'est une conséquence de $u(t,x) = (G(t,.) * u_0)(x)$, $\dfrac{\mathrm{d}^n}{\mathrm{d}x^n}(G(t,.) * u_0) = \dfrac{\partial^n G}{\partial x^n}(t,.) * u_0$ et $G \in C^\infty([a,T] \times \mathbb{R})$. □

Remarque 16. *L'effet régularisant s'observe aussi sur un domaine borné.*

Proposition 16 (Propagation de la chaleur à vitesse infinie)**.** *Si $u_0 \geq 0$ (non uniformément nulle) a un support borné et $f = 0$, alors pour tous $t > 0$ et $x \in \mathbb{R}$, on a $u(t,x) > 0$.*

Démonstration. Utiliser l'expression $u(t,x) = (G(t,.) * u_0)(x)$. □

4.2 Propriétés générales

On considère à nouveau le domaine borné $\Omega := (0,1)$ et le problème (43). Dans cette section, on interprète l'équation de la chaleur au sens des distributions : les dérivées sont des dérivées faibles et les égalités sont des égalités entre distributions. Ceci permet de donner des résultats plus généraux notamment en utilisant les espaces de Sobolev $H^k(0,1)$.

On peut montrer qu'une solution dépend continûment des données u_0 et f.

Proposition 17. *Il existe $C > 0$ tel que toute fonction $u \in C^0([0,T];L^2(0,1)) \cap L^2(0,T;H_0^1(0,1))$ solution de (43) avec $u_0 \in L^2(0,1)$ et $f \in L^2(0,T;L^2(0,1))$ vérifie*

$$\sup_{t\in[0,T]} \|u(t,.)\|_{L^2(0,1)} + \|\partial_x u\|_{L^2(0,T;L^2(0,1))} \leq C(\|u_0\|_{L^2(0,1)} + \|f\|_{L^2(]0,T[\times]0,1[)}). \qquad (45)$$

Démonstration. On suppose tout d'abord la fonction u régulière, le résultat final sera obtenu par densité. Les dérivées de u sont donc considérées au sens usuel.

Multiplions la première ligne de (43) par u et intégrons en espace sur le domaine $\Omega = (0,1)$, on obtient

$$\int_\Omega \frac{\partial u}{\partial t} u = \int_\Omega \frac{\partial^2 u}{\partial x^2} u + \int_\Omega f u.$$

On utilise ensuite l'intégration par parties $\int_\Omega \frac{\partial^2 u}{\partial x^2} u = -\int_\Omega \left(\frac{\partial u}{\partial x}\right)^2$ (on rappelle que $u(t,0) = u(t,1) = 0$) ainsi que l'inégalité de Cauchy–Schwarz $\int_\Omega fu \leq \|f\|_{L^2(\Omega)} \|u\|_{L^2(\Omega)}$, que l'on combine avec l'inégalité de Young $\|f\|_{L^2(\Omega)}\|u\|_{L^2(\Omega)} \leq \frac{1}{2}(\|f\|^2_{L^2(\Omega)} + \|u\|^2_{L^2(\Omega)})$.

On obtient ainsi

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t,\cdot)\|^2_{L^2(\Omega)} = \int_\Omega \frac{\partial u}{\partial t} u = -\int_\Omega \left(\frac{\partial u}{\partial x}\right)^2 + \int_\Omega f u,$$

et

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|u(t,\cdot)\|^2_{L^2(\Omega)} \leq -\left\|\frac{\partial u}{\partial x}(t,\cdot)\right\|^2_{L^2(\Omega)} + \frac{1}{2}(\|f(t,\cdot)\|^2_{L^2(\Omega)} + \|u(t,\cdot)\|^2_{L^2(\Omega)}).$$

De plus, d'après l'inégalité de Poincaré, on a $\|u(t,\cdot)\|^2_{L^2(\Omega)} \leq \|\frac{\partial u}{\partial x}(t,\cdot)\|^2_{L^2(\Omega)}$. Notons que la constante de cette inégalité vaut 1 car on s'est placés dans le domaine $\Omega = (0,1)$. On a donc

$$\frac{\mathrm{d}}{\mathrm{d}t}\|u(t,\cdot)\|^2_{L^2(\Omega)} + \left\|\frac{\partial u}{\partial x}(t,\cdot)\right\|^2_{L^2(\Omega)} \leq \|f(t,\cdot)\|^2_{L^2(\Omega)}.$$

On intègre selon t, ce qui donne

$$\|u(t,\cdot)\|^2_{L^2(\Omega)} + \|\partial_x u\|^2_{L^2(0,t;L^2(0,1))} \leq \|u_0\|^2_{L^2(\Omega)} + \|f\|^2_{L^2((0,t)\times\Omega)}.$$

On obtient la borne sur $\sup_{t\in[0,T]} \|u(t,.)\|_{L^2(0,1)}$ en prenant le sup sur t et la borne sur $\|\partial_x u\|^2_{L^2(0,T;L^2(0,1))}$ en prenant $t = T$. □

On peut utiliser ce résultat pour montrer l'existence d'une unique solution faible au problème de Chaleur–Dirichlet.

Proposition 18. *Le problème* (43) *est bien posé : si $f \in L^2(0,T;L^2(0,1))$ et $u_0 \in L^2(0,1)$, alors il existe une unique solution u dans $C^0([0,T];L^2(0,1)) \cap L^2(0,T;H^1_0(0,1))$ et cette solution dépend continûment des données d'après l'estimée* (45).

L'unicité et la continuité par rapport aux données sont une conséquence directe de la Proposition 17 (pour l'unicité : considérer deux solutions, leur différence vérifie (43) avec des données nulles, d'après (45) cette différence est nulle). L'existence est plus délicate et on ne la présentera pas ici. Le lecteur currieux peut regarder la Section 6.3.

Notons que nous avons un gain de régularité de la solution dans $C^0([0,T];L^2(0,1)) \cap L^2(0,T;H^1_0(0,1))$ par rapport au second membre qui est dans $L^2(0,T;L^2(0,1))$ (ce gain

de régularité est plus important en espace qu'en temps). Notons également que le fait que la solution soit dans $C^0([0,T];L^2(0,1))$ est en accord avec la donnée initiale qui est dans $L^2(0,1)$.

On s'intéresse maintenant à la question de la réversibilité de l'équation de la chaleur. On considère une solution u de (43) et $\widetilde{u}(t,x) := u(T-t,x)$. La fonction $\widetilde{u}$ vérifie l'équation de la chaleur rétrograde :

$$\text{Trouver } \widetilde{u} : (0,T)\times(0,1) \text{ vérifiant } \begin{cases} -\dfrac{\partial \widetilde{u}}{\partial t} - \dfrac{\partial^2 \widetilde{u}}{\partial x^2} = \widetilde{f} & \text{dans } (0,T)\times(0,1), \\ \widetilde{u}(0,.) = u_T & \text{dans } (0,1), \\ \widetilde{u}(t,0) = 0, \quad \widetilde{u}(t,1) = 0 & \forall t \in (0,T), \end{cases} \tag{46}$$

où $\widetilde{f}(t,x) := f(T-t,x)$ et $u_T(x) := u(T,x)$. Ce problème est mal posé, on dit alors que l'équation de la chaleur n'est pas réversible.

Proposition 19 (Non réversibilité de l'équation de la chaleur)**.** *Le problème* (43) *n'est pas réversible : le problème* (46) *est mal posé.*

Démonstration. En utilisant les mêmes arguments que ceux évoqués précédemment, on doit pouvoir montrer l'existence d'une unique solution au problème (46). Son caractère mal posé vient de la dépendance non continue aux données. Pour prouver cela, on va utiliser les résutats de la Section 4.1.1.

On considère $\widetilde{f} = 0$ et on suppose par l'absurde qu'il existe une constante $C > 0$ telle que pour toute solution $\widetilde{u}$ de (46), on a

$$\sup_{0\le t\le T} \|\widetilde{u}(t,.)\|_{L^2(0,1)} \le C \|\widetilde{u}(0,.)\|_{L^2(0,1)}.$$

En adaptant le raisonnement de la Section 4.1.1, on peut montrer que les fonctions $\widetilde{u}_k(t,x) := \sin(k\pi x)e^{k^2\pi^2 t}$ sont solution pour tout $k \in \mathbb{N}^*$. On a $\sup_{0\le t\le T} \|\widetilde{u}_k(t,.)\|_{L^2(0,1)} = \frac{1}{\sqrt{2}} e^{k^2\pi^2 T}$ et $\|\widetilde{u}(0,.)\|_{L^2(0,1)} = \frac{1}{\sqrt{2}}$. On obtient une contradiction en faisant tendre k vers $+\infty$. □

Remarque 17. *La non-réversibilité de l'équation peut être interprétée comme l'impossibilité pratique de retrouver la solution à partir de la donnée finale. En effet, si cette donnée est bruitée, ne serait-ce qu'un peu, alors les bruits indésirables peuvent exploser aussi vite qu'on veut et polluer la solution (à cause de l'exponentielle croissante en temps).*

Exercice 22. *L'équation des ondes est-elle réversible ? La réponse est donnée dans le chapitre suivant.*

Proposition 20 (Principe du maximum)**.** *Si pour tous $t \in (0,T)$ et $x \in (0,1)$, on a $f(t,x) \le 0$. Alors $\sup_{(t,x)\in[0,T]\times[0,1]} u(t,x) = \max(0, \sup_{x\in[0,1]} u_0(x))$ avec u la solution de* (43)*.*

Notons que nous avons utilisé le fait que les conditions de Dirichlet étaient homogènes. Dans le cas de conditions non homogènes, il faut aussi prendre en compte les données de Dirichlet.

Remarque 18 (Conditions de Dirichlet non homogènes)**.** *Dans le problème* (43)*, on a utilisé des conditions de Dirichlet homogènes. Pour résoudre le problème*

$$\textit{Trouver } u : (0,T)\times(0,1) \textit{ vérifiant } \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = f & \textit{dans } (0,T)\times(0,1), \\ u(0,.) = u_0 & \textit{dans } (0,1), \\ u(t,0) = \alpha(t), \quad u(t,1) = \beta(t) & \forall t \in (0,T), \end{cases}$$

on peut utiliser un relèvement (comme dans la Section 2.1). Pour cela, il faut choisir une fonction g vérifiant $g(t,0)=\alpha(t)$ et $g(t,1)=\beta(t)$, puis résoudre le problème avec conditions de Dirichlet homogènes

$$\textit{Trouver } \widetilde{u}:(0,T)\times(0,1) \textit{ vérifiant } \begin{cases} \dfrac{\partial \widetilde{u}}{\partial t}-\dfrac{\partial^2 \widetilde{u}}{\partial x^2}=f-\dfrac{\partial g}{\partial t}+\dfrac{\partial^2 g}{\partial x^2} & \textit{dans } (0,T)\times(0,1),\\ \widetilde{u}(0,.)=u_0-g(0,.) & \textit{dans } (0,1),\\ \widetilde{u}(t,0)=0, \quad \widetilde{u}(t,1)=0 & \forall t\in(0,T). \end{cases}$$

La solution du premier problème sera alors donnée par $u(t,x)=\widetilde{u}(t,x)+g(t,x)$. Au passage, pour le domaine (1d) $\Omega=(0,1)$, on peut facilement choisir $g(t,x)=(1-x)\alpha(t)+x\beta(t)$.

Exercice 23. *Vérifier que $u(t,x)=\widetilde{u}(t,x)+g(t,x)$ est bien solution du premier problème.*

4.3 Approximation par différences finies

Nous nous plaçons ici dans le même cadre que celui utilisé pour l'équation de transport en Section 3.3. Nous décomposons les intervalles de temps et d'espace respectivement en N et M sous-intervalles. On note t_n $(0\le n\le N)$ et x_j $(0\le j\le M)$ les bornes de ces sous-intervalles. On rappelle $x_j=jh_x$, $h_x=1/M$, $t_n=nh_t$ et $h_t=T/N$. Nous cherchons à approcher la solution de (43) en t_n et x_j, i.e. $u(t_n,x_j)$, par le terme général u_j^n d'une suite (u_j^n). Nous proposons plusieurs schémas pour calculer cette suite. Leurs propriétés seront étudiées en utilisant les éléments de la Section 3.3.2.

4.3.1 Schéma explicite centré

Le schéma le plus intuitif à utiliser est un schéma explicite centré. Pour cela, approchons les dérivées de la façon suivante :

$$\frac{\partial u}{\partial t}(t_n,x_j)\simeq\frac{u(t_{n+1},x_j)-u(t_n,x_j)}{h_t}, \qquad \frac{\partial^2 u}{\partial x^2}(t_n,x_j)\simeq\frac{u(t_n,x_{j+1})-2u(t_n,x_j)+u(t_n,x_{j-1})}{h_x^2}.$$

Nous obtenons ainsi le schéma suivant

$$\begin{cases} u_j^0=u_0(x_j), & (1\le j\le M-1),\\ u_0^n=u_M^n=0, & (0\le n\le N),\\ u_j^{n+1}=u_j^n+\dfrac{h_t}{h_x^2}(u_{j+1}^n-2u_j^n+u_{j-1}^n)+h_t f(t_n,x_j), & (0\le n\le N-1, 1\le j\le M-1). \end{cases} \tag{47}$$

Exercice 24. *La figure 6 a été obtenue grâce au schéma (47) avec $f(t,x)=-1$ et $u_0(x)=x(1-x)$. Coder ce schéma et retrouver la figure du milieu ($\nu=1$).*

Nous allons maintenant étudier les propriétés de ce schéma. Nous pouvons prouver que ce schéma est consistant d'ordre 1 en temps et 2 en espace pour la norme $\|\cdot\|_\infty$. De plus, il est stable sous la condition de CFL $2h_t\le h_x^2$ et instable si cette condition n'est pas remplie (pour la norme $\|\cdot\|_\infty$). Ces résultats peuvent être prouvés en adaptant ce qui a été fait dans la Section 3.3.

Exercice 25. *Dans cet exercice, nous étudions le schéma (47) avec la norme $\|\cdot\|_\infty$.*

1. *Montrer que nous sommes bien dans le cadre de la Section 3.3.2, i.e. montrer que* $U^n = (u_j^n)_{j=1}^{M-1}$ *vérifie* (29) *(avec* U_0*,* A *et* F^n *à déterminer).*
2. *Prouver que ce schéma est consistant d'ordre 1 en temps et 2 en espace.*
3. *Prouver que ce schéma est stable sous la condition de CFL* $2h_t \leq h_x^2$*.*
4. *Prouver que ce schéma est instable si la condition de CFL précédente n'est pas vérifiée.*

NB : On pourra adapter ce qui a été fait dans la Section 3.3.3.

Remarque 19 (CFL parabolique)**.** *Ici, la condition de CFL associée à ce schéma explicite est* $2h_t \leq h_x^2$ *(à comparer avec la CFL de l'équation de transport* $ah_t \leq h_x$*). Le carré au dessus du pas d'espace rend cette condition plus restrictive que celle de l'équation de transport. Par exemple, pour rester à CFL constante, si on divise le pas d'espace par deux, il faut diviser le pas de temps par deux pour l'équation de transport et par quatre pour l'équation de la chaleur. Cette CFL plus restrictive est une des raisons pour lesquelles on préfère généralement utiliser un schéma implicite pour résoudre l'équation de la chaleur.*

Bien que, comme nous venons de le voir, on peut étudier le schéma explicite centré avec la norme $\|\cdot\|_\infty$, nous allons introduire une nouvelle norme : la norme $\|\cdot\|_{L^2}$. Cette norme sera utile pour la suite (en particulier pour le schéma implicite et pour l'équation des ondes).

Définition 7 (Norme $\|\cdot\|_{L^2}$ pour les suites)**.** *Pour* $V \in \mathbb{R}^{M-1}$*, on définit*

$$\|V\|_{L^2} := \Big(h_x \sum_{j=1}^{M-1} |V_j|^2\Big)^{\frac{1}{2}}. \tag{48}$$

Remarque 20. *Notons qu'il s'agit de la norme* $\|\cdot\|_{\ell^2}$ *classique multipliée par un facteur* $h_x^{\frac{1}{2}}$*. Ce facteur permet de donner à cette norme un comportement similaire à celui de la norme* $\|\cdot\|_{L^2}$ *des fonctions.*

Nous montrons maintenant que la norme $\|\cdot\|_\infty$ domine la norme $\|\cdot\|_{L^2}$.

Proposition 21. *La norme* $\|\cdot\|_\infty$ *domine la norme* $\|\cdot\|_{L^2}$ *: il existe* $C > 0$ *telle que*

$$\|V\|_{L^2} \leq C\|V\|_\infty \quad \forall V \in \mathbb{R}^{M-1}.$$

Démonstration. Par un calcul direct

$$\|V\|_{L^2}^2 = h_x \sum_{j=1}^{M-1} |V_j|^2 \leq h_x(M-1)\|V\|_\infty^2 \leq \|V\|_\infty^2.$$

Notons qu'ici on peut prendre $C = 1$ par ce que $\Omega = (0,1)$. Dans un cadre plus général, il faudrait prendre $C = |\Omega|^{\frac{1}{2}}$. □

Remarque 21. *Ceci sera utilisé en particulier pour prouver la consistance d'un schéma : on la prouvera pour la norme* $\|\cdot\|_\infty$ *et ce sera immédiatement valable pour la norme* $\|\cdot\|_{L^2}$*.*

Nous voulons maintenant étudier la convergence du schéma (47) pour la norme $\|\cdot\|_{L^2}$. Comme nous l'avons vu dans l'exercice 25, le schéma est consistant et stable pour la norme $\|\cdot\|_\infty$. D'après le théorème de Lax, il est donc convergent (d'ordre 1 en temps et 2 en espace) pour cette même norme (sous la condition de CFL $2h_t \leq h_x^2$). On peut donc très rapidement établir que ce schéma est convergent pour la norme $\|\cdot\|_{L^2}$ sous la même condition de CFL. Ceci vient du fait que la norme $\|\cdot\|_\infty$ domine la norme $\|\cdot\|_{L^2}$:

$$\max_{0\leq n\leq N} \|\widetilde{u}^n - U^n\|_{L^2} \leq C \max_{0\leq n\leq N} \|\widetilde{u}^n - U^n\|_\infty \leq C(h_t + h_x^2).$$

Cependant, pour de nombreux schémas, la convergence en norme $\|\cdot\|_\infty$ est plus difficile à établir qu'en norme $\|\cdot\|_{L^2}$. Nous allons donc chercher à établir la convergence en norme $\|\cdot\|_{L^2}$ sans utiliser la stabilité en norme $\|\cdot\|_\infty$ (en pratique c'est très souvent la stabilité qui est plus difficile à établir).

Comme nous l'avons vu dans l'exercice 25, le schéma (47) vérifie le cadre de la Section 3.3.2 et il est consistant pour la norme $\|\cdot\|_\infty$ d'ordre 1 en temps et 2 en espace (donc aussi consistant au même ordre pour la norme $\|\cdot\|_{L^2}$). Si nous prouvons la stabilité en norme $\|\cdot\|_{L^2}$ alors nous pourrons utiliser le Théorème de Lax. Comme nous allons le voir, l'étude de la stabilité pour la norme $\|\cdot\|_{L^2}$ passe par l'étude des valeurs propres de la matrice du schéma.

Proposition 22. *On se place dans le cadre de la Section 3.3.2. Si $A \in \mathbb{R}^{(M-1)^2}$ est symétrique alors*

$$|\!|\!| A |\!|\!|_{L^2} = \rho(A),$$

où $\rho(A) := \sup_{1\leq j\leq M-1} |\lambda_j|$ est le rayon spectral de la matrice A avec (λ_j) les valeurs propres de A.

Démonstration. Si A est symétrique à valeurs réelles, alors elle est diagonalisable dans $\mathbb{R}$. Il existe donc une base de vecteurs propres (V_j). Pour tout vecteur V, on peut écrire $V = \sum_{j=1}^{M-1} \alpha_j V_j$ et on a

$$|\!|\!| A |\!|\!|_{L^2} = \sup_{V\in\mathbb{R}^{M-1}\setminus\{0\}} \frac{\|AV\|_{L^2}}{\|V\|_{L^2}} = \sup_{\alpha\in\mathbb{R}^{M-1}\setminus\{0\}} \frac{(\sum_j (\alpha_j\lambda_j)^2 \|V_j\|^2)^{\frac{1}{2}}}{(\sum_j \alpha_j^2 \|V_j\|^2)^{\frac{1}{2}}} \leq \rho(A).$$

De plus, pour j tel que $\rho(A) = |\lambda_j|$, on a $\|AV_j\|_{L^2} = |\lambda_j| \|V_j\|_{L^2}$ et donc $\rho(A) \leq |\!|\!| A |\!|\!|_{L^2}$. On a montré $|\!|\!| A |\!|\!|_{L^2} = \rho(A)$. □

L'idée est donc de calculer les valeurs propres de A et d'utiliser le Corrolaire 1 pour prouver la stabilité du schéma. Pour le calcul des valeurs propres, nous pouvons utiliser le résultat suivant

Proposition 23. *Soit $A \in \mathbb{R}^{(M-1)^2}$ une matrice tridiagonale définie par*

$$A = \begin{pmatrix} a & b & & \\ b & \ddots & \ddots & \\ & \ddots & \ddots & b \\ & & b & a \end{pmatrix},$$

avec $a, b \in \mathbb{R}$. *Les valeurs propres de* A *sont les* $\lambda_k = a + 2b\cos\left(\dfrac{k\pi}{M}\right)$ *avec* $1 \leq k \leq M-1$. *Le vecteur propre associé* V_k *a pour composantes* $(V_k)_j = \sin\left(j\dfrac{k\pi}{M}\right)$.

NB : les vecteurs propres ne dépendent pas de a *et* b.

Démonstration. Pour une démonstration de cette proposition, le lecteur pourra se référer au Lemme 2.2 du livre de Brigitte Lucquin, "Équations aux dérivées partielles et leurs approximations", 2004. □

On peut montrer que nous sommes dans le cadre de la Section 3.3.2 avec $A = I - h_t A_\Delta$ où A_Δ est la matrice discrétisant le Laplacien donnée en (21). On applique donc la Proposition 23 avec $a = 1 - \frac{2h_t}{h_x^2}$ et $b = \frac{h_t}{h_x^2}$. Les valeurs propres de cette matrice sont donc les $\lambda_k = 1 - \frac{2h_t}{h_x^2}(1 - \cos(\frac{k\pi}{M}))$ $(1 \leq k \leq M-1)$. Sous la condition de CFL $2h_t \leq h_x^2$, on a $|\lambda_k| \leq 1$ et donc $\rho(A) \leq 1$. En appliquant la Proposition 22 et le Corrolaire 1, on établit le résultat suivant

Proposition 24. *Le schéma explicite centré* (47) *est stable pour la norme* $\|\cdot\|_{L^2}$ *sous la condition de CFL* $2h_t \leq h_x^2$.

Comme nous l'avons indiqué précédemment, ce schéma est consistant pour la norme $\|\cdot\|_{L^2}$. Il est donc convergent sous la condition de CFL $2h_t \leq h_x^2$ (ordre 1 en temps et 2 en espace).

Remarque 22. *Si on prenait en compte une diffusivité thermique* $\nu > 0$ *constante différente de* 1, *l'équation deviendrait* $\partial_t u - \nu\partial_{xx}^2 u = f$ *et la CFL associée pour le schéma explicite serait* $2\nu h_t \leq h_x^2$.

4.3.2 Schéma implicite centré

Comme évoqué dans la section précédente, l'utilisation de schémas explicites est très couteuse dans le cadre de l'équation de la chaleur (à cause de la condition de CFL parabolique). On s'intéresse donc à la possibilité d'utiliser des schémas qui seraient inconditionnellement stables. Un premier exemple d'un tel schéma est donné par le schéma implicite centré.

On approche les dérivées partielles de la manière suivante :

$$\frac{\partial u}{\partial t}(t_n, x_j) \simeq \frac{u(t_n, x_j) - u(t_{n-1}, x_j)}{h_t}, \qquad \frac{\partial^2 u}{\partial x^2}(t_n, x_j) \simeq \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{h_x^2}.$$

Le schéma implicite centré est donc défini comme

$$\begin{cases} u_j^0 = u_0(x_j), & 1 \leq j \leq M-1, \\ u_0^n = u_M^n = 0, & 0 \leq n \leq N, \\ u_j^n = u_j^{n-1} - \dfrac{h_t}{h_x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + h_t f(t_n, x_j), & 1 \leq n \leq N, 1 \leq j \leq M-1. \end{cases} \tag{49}$$

Le schéma peut donc être écrit sous la forme

$$\widetilde{A}U^n = U^{n-1} + h_t\widetilde{F}^n,$$

avec U^n le vecteur des u_j^n pour $1 \leq j \leq N-1$ et $(\widetilde{F}^n)_j = f(t_n, x_j)$. De plus, $\widetilde{A} = I - h_t A_\Delta$ où A_Δ est la matrice du problème de Laplace–Dirichlet (cf (21)). Pour rappel, cette matrice est exigible.

Pour montrer qu'il existe une unique solution à ce problème discret, il faut montrer que la matrice $\widetilde{A}$ est inversible. Cela a été fait par un calcul direct dans l'exercice 20 dans le cadre de l'équation de transport. Cependant ce calcul est bien plus difficile dans le cas présent. La méthode utilisée consiste alors à calculer les valeurs propres de cette matrice.

En utilisant la Proposition 23, les valeurs propres de la matrice $\widetilde{A}$ sont de la forme

$$\lambda_k = 1 + \frac{2h_t}{h_x^2}\left(1 - \cos\left(\frac{k\pi}{M}\right)\right) \qquad (1 \leq k \leq M-1).$$

Toutes ces valeurs propres sont strictement positives pour tous $h_t, h_x > 0$. La matrice $\widetilde{A}$ est donc inversible et le schéma est bien défini. De plus, nous avons

$$U^{n+1} = (\widetilde{A})^{-1}U^n + h_t(\widetilde{A})^{-1}\widetilde{F}^n.$$

Nous sommes donc bien dans le cadre de la Section 3.3.2. Nous pouvons ainsi appliquer tous les résultats de cette section (notamment le Théorème de Lax).

Nous allons étudier la stabilité de ce schéma en norme $\|\cdot\|_{L^2}$. Pour cela, nous devons donc étudier les valeurs propres de la matrice associée au schéma, i.e. la matrice $(\widetilde{A})^{-1}$.

Nous avons noté (λ_k) les valeurs propres de $\widetilde{A}$, notons (V_k) les vecteurs propres associés. Nous avons $V_k = (\widetilde{A})^{-1}\widetilde{A}V_k = \lambda_k(\widetilde{A})^{-1}V_k$ et donc les valeurs propres de $(\widetilde{A})^{-1}$ sont les (λ_k^{-1}). De plus, on a $\lambda_k \geq 1$ pour tout k. Donc $\rho((\widetilde{A})^{-1}) \leq 1$ et en appliquant la Proposition 22 et le Corrolaire 1, on établit la stabilité du schéma pour la norme $\|\cdot\|_{L^2}$. Notons que cette stabilité est inconditionnelle (il n'y a pas de condition de CFL à respecter), ce qui est attendu pour un schéma implicite.

On peut également montrer que ce schéma est consistant d'ordre 1 en temps et 2 en espace en norme $\|\cdot\|_\infty$ (et donc aussi en norme $\|\cdot\|_{L^2}$). D'après le théorème de Lax, ce schéma converge inconditionnellement avec un ordre 1 en temps et 2 en espace.

Proposition 25 (Convergence)**.** *Le schéma implicite centré est inconditionnellement stable pour la norme $\|\cdot\|_{L^2}$. De plus, il est consistant à l'ordre 1 en temps et 2 en espace. Il est donc inconditionnellement convergent à l'ordre 1 en temps et 2 en espace pour la norme $\|\cdot\|_{L^2}$.*

Remarque 23. *Le temps de résolution d'un pas de temps pour un schéma implicite est un peu plus long que pour un schéma explicite car il faut résoudre un système linéaire (notons qu'on ne calcule pas $(\widetilde{A})^{-1}$ mais qu'on résout plutôt le système linéaire associé). Ceci est compensé par le fait qu'il n'y a pas de restriction sur le pas de temps et on peut donc prendre des pas plus grand, impliquant moins de pas de temps.*

Remarque 24. *On pourrait prouver la stabilité de ce schéma en norme $\|\cdot\|_\infty$ mais ce serait plus difficile, c'est pour cela que l'on préfère la norme $\|\cdot\|_{L^2}$. Dans certains cas où la stabilité pour la norme $\|\cdot\|_{L^2}$ est aussi difficile à établir, on se contente de la stabilité au sens de Von Neumann (bien que ce soit moins rigoureux d'un point de vue mathématique).*

Exercice 26. *Dans cet exercice, nous comparons le comportement des schémas* (47) *et* (49)*.*

1. *Coder le schéma* (47). *Vérifier que ce schéma converge sous la condition* $2h_t \leq h_x^2$. *Observer l'apparition d'instabilités si cette condition n'est pas respectée.*
2. *Coder le schéma* (49). *Vérifier que ce schéma converge même si on ne respecte pas* $2h_t \leq h_x^2$.

4.4 Étude du problème de Chaleur–Neumann

Cette section constitue un exercice portant sur le problème de Chaleur–Neumann. Elle est facultative.

On s'intéresse au problème suivant :

$$\text{Trouver } u : (0,T)\times(0,1) \text{ vérifiant } \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0 & \text{dans } (0,T)\times(0,1), \\ u(0,.) = u_0 & \text{dans } (0,1), \\ \dfrac{\partial u}{\partial x}(t,0) = 0, \quad \dfrac{\partial u}{\partial x}(t,1) = 0 & \forall t \in (0,T), \end{cases} \tag{50}$$

où u_0 est une fonction donnée. Notez que, pour simplifier, nous n'avons pas considéré de terme source f sur l'équation.

Exercice 27. *On s'intéresse ici au problème* (50).

1. *Soit* u *une solution régulière de* (50). *On note* $\bar{u}(t) = \int_0^1 u(t,x)\,\mathrm{d}x$ *sa valeur moyenne sur le domaine. Montrer que cette valeur moyenne* $\bar{u}$ *se conserve au cours du temps.*
2. *Appliquer la méthode de séparation de variables pour trouver une solution particulière à ce problème lorsque* u_0 *a une forme particulière.*
3. *Pour* u_0 *régulière vérifiant* $\partial_x u_0(0) = \partial_x u_0(1) = 0$, *écrire* u_0 *sous la forme d'une série de Fourier. En déduire l'existence d'une solution* $u(t,x)$ *au problème* (50).
4. *Montrer que la solution* $u(t,x)$ *exhibée lors de la question précédente vérifie*

$$\|u(t,.) - \bar{u}\|_\infty \leq Ce^{-\pi^2 t},$$

avec une constante $C > 0$ *dépendant de* u_0 *uniquement. On a montré que la solution converge uniformément vers sa valeur moyenne lorsque* $t \to +\infty$. *De plus, cette convergence se fait à une vitesse exponentielle.*

5. *Adapter un des deux schémas précédents aux conditions de Neumann. Le coder et vérifier numériquement ce comportement lorsque* $t \to +\infty$.

5 Équation des ondes

On se place dans le domaine $\Omega = (0,1)$. On s'intéresse au problème de l'équation des ondes avec des conditions de Dirichlet homogènes

$$\text{Trouver } u : (0,T)\times(0,1) \text{ vérifiant } \begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2 \dfrac{\partial^2 u}{\partial x^2} = f & \text{dans } (0,T)\times(0,1), \\ u(0,.) = u_0 & \text{dans } (0,1), \\ \dfrac{\partial u}{\partial t}(0,.) = u_1 & \text{dans } (0,1), \\ u(t,0) = 0, \quad u(t,1) = 0 \quad \forall t \in (0,T), \end{cases} \tag{51}$$

avec $c > 0$ la célérité des ondes, f un terme source et u_0, u_1 des conditions initiales. Notons que si l'on compare avec l'équation de la chaleur, l'équation des ondes possède une dérivée seconde en temps. Cette dérivée seconde (au lieu de première) demande donc une condition aux limites (en temps) supplémentaire. On a donc besoin, en plus de connaître la valeur de la solution à l'instant initial (u_0), de connaître la valeur de la dérivée en temps de la solution à l'instant initial (u_1). On peut comparer avec des problèmes classiques 1d dérivant des lois de Newton (par exemple la chute d'un objet soumis uniquement à son propre poids) : on a une dérivée seconde en temps et il faut une donnée initiale sur la solution et sa dérivée.

5.1 Recherche de solutions régulières

Comme pour l'équation de la chaleur, on cherche ici à construire des solutions régulières à l'équation des ondes. On considère d'abord le cas du domaine borné $\Omega = (0,1)$ puis le cas de la droite réelle $\Omega = \mathbb{R}$.

5.1.1 Cas du domaine borné $\Omega = (0,1)$

Le but de cette section est de construire une solution explicite au problème (51). On adapte ce qui a été fait dans la Section 4.1.1 (le lecteur pourra comparer). Pour simplifier, on se restreint au cas particulier où $f = 0$. Notons que cette méthode peut être adaptée au cas où f n'est pas nul.

On utilise la méthode de séparation des variables. On cherche donc une solution de la forme $u(t,x) = \Phi(t)\Psi(x)$. En reportant cette expression dans l'équation, on s'aperçoit que ces fonctions doivent satisfaire la relation $\Phi''(t)\Psi(x) = c^2\Phi(t)\Psi''(x)$. On cherche donc des fonctions qui vérifient $\dfrac{\Phi''(t)}{\Phi(t)} = c^2\dfrac{\Psi''(x)}{\Psi(x)} = C$ avec C une constante (puisque ces deux expressions sont égales et dépendent du temps pour l'une et de l'espace pour l'autre). De plus, nous ne sommes pas intéressés par le cas où cette constante serait nulle car la solution associée serait alors triviale. Les fonctions Φ et Ψ vérifient donc les conditions

$$\Phi''(t) - C\Phi(t) = 0, \qquad c^2\Psi''(x) - C\Psi(x) = 0, \qquad \Psi(0) = \Psi(1) = 0.$$

Si $C > 0$, on peut montrer en utilisant les conditions de Dirichlet homogènes que la seule solution possible est la solution nulle. On s'intéresse donc uniquement au cas où $C < 0$. Dans ce cas, on a $\Phi(t) = A\sin(\sqrt{|C|}t) + B\cos(\sqrt{|C|}t)$ et $\Psi(x) = A'\sin(\dfrac{\sqrt{|C|}}{c}x) + B'\cos(\dfrac{\sqrt{|C|}}{c}x)$.

Avec les conditions de Dirichlet homogènes, on montre que $B' = 0$ et $C = -k^2c^2\pi^2$ avec $k \in \mathbb{Z}$. Ainsi, si les conditions initiales le permettent, on peut considérer des solutions de la forme $(t,x) \mapsto (A\sin(k\pi ct) + B\cos(k\pi ct))\sin(k\pi x)$.

Pour être plus général et ainsi vérifier les conditions initiales imposées, on va considérer des solutions sous la forme de séries de Fourier utilisant ces termes. Supposons que u_0 et u_1 sont dans $C^2([0,1])$ avec $u_0(0) = u_0(1) = 0$ et $u_1(0) = u_1(1) = 0$. On peut alors écrire

$$u_0(x) = \sum_{k=1}^{+\infty} b_k \sin(k\pi x) \qquad \text{et} \qquad u_1(x) = \sum_{k=1}^{+\infty} c_k \sin(k\pi x).$$

avec $b_k = 2\int_0^1 u_0(x)\sin(k\pi x)\,\mathrm{d}x$ et $c_k = 2\int_0^1 u_1(x)\sin(k\pi x)\,\mathrm{d}x$. On peut montrer que ces coefficients sont en $O(1/k^2)$ de sorte que les séries précédentes sont normalement convergentes (voir Section 4.1.1).

La fonction définie par

$$u(t,x) = \sum_{k=1}^{+\infty} (b_k\cos(k\pi ct) + \frac{c_k}{k\pi c}\sin(k\pi ct))\sin(k\pi x),$$

est alors continue et solution de l'équation des ondes au sens des distributions (les dérivées sont à considérer au sens des distributions).

Remarque 25. *La solution u est périodique en temps, on a $u(\frac{2}{c},x) = u_0(x)$ et $\partial_t u(\frac{2}{c},x) = u_1(x)$.*

5.1.2 Cas de la droite réelle

On considère le problème sur $\mathbb{R}$ entier ($\Omega = \mathbb{R}$) et on considère aussi $T = +\infty$

$$\text{Trouver } u : \mathbb{R}_+ \times \mathbb{R} \text{ vérifiant } \begin{cases} \dfrac{\partial^2 u}{\partial t^2} - c^2\dfrac{\partial^2 u}{\partial x^2} = f & \text{dans } \mathbb{R}_+ \times \mathbb{R}, \\ u(0,.) = u_0 & \text{dans } \mathbb{R}, \\ \dfrac{\partial u}{\partial t}(0,.) = u_1 & \text{dans } \mathbb{R}, \end{cases} \tag{52}$$

où, à nouveau, $c > 0$ est la célérité des ondes, f est le terme source et u_0 et u_1 sont les données initiales.

Théorème 7 (Formule de d'Alembert)**.** *Si $u_0 \in C^2(\mathbb{R})$, $u_1 \in C^1(\mathbb{R})$ et $f \in C^1(\mathbb{R}_+ \times \mathbb{R})$, alors le problème* (52) *admet une unique solution $u \in C^2(\mathbb{R}_+ \times \mathbb{R})$ donnée par*

$$u(t,x) = \frac{u_0(x+ct) + u_0(x-ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s + \frac{1}{2c}\int_0^t \int_{x-c(t-s)}^{x+c(t-s)} f(s,y)\,\mathrm{d}y\,\mathrm{d}s.$$

Démonstration. On prouve ce théorème par analyse et synthèse. On suppose que u est une solution de (52). L'idée de la preuve consiste à s'intéresser aux droites $(t,x) \mapsto x - ct$ et $(t,x) \mapsto x + ct$ selon lesquelles les ondes se déplacent. On introduit donc le changement de variables $y = x + ct$ et $z = x - ct$, on note $v(y,z) := u\left(\dfrac{y-z}{2c}, \dfrac{y+z}{2}\right)$ la représentation de u selon ces nouvelles variables.

On prouve ensuite que v vérifie l'équation

$$\frac{\partial^2 v}{\partial y \partial z} = -\frac{1}{4c^2} f\left(\frac{y-z}{2c}, \frac{y+z}{2}\right).$$

Cela se fait aisément en calculant cette dérivée croisée et en rappelant que u est solution de (52).

On peut ensuite calculer la fonction v en intégrant la relation précédente une fois selon y et une fois selon z. On obtient

$$v(y,z) = G(z) + F(y) - \frac{1}{4c^2}\int_0^z \int_0^y f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r,$$

où F et G sont des fonctions à déterminer. On peut donc écrire une première expression pour u

$$\begin{aligned} u(t,x) =& v(x+ct, x-ct) \\ =& G(x-ct) + F(x+ct) - \frac{1}{4c^2}\int_0^{x-ct} \int_0^{x+ct} f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r. \end{aligned}$$

Il faut maintenant déterminer l'expression de F et G. Pour cela, on évalue l'expression précédente en $t = 0$ et on utilise les conditions initiales, on obtient

$$u_0(x) = G(x) + F(x) - \frac{1}{4c^2}\int_0^x \int_0^x f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r. \tag{53}$$

Pour obtenir une deuxième relation, on suppose que F et G sont dérivables, on dérive l'expression de u et on évalue en $t = 0$, on obtient

$$u_1(x) = -cG'(x) + cF'(x) - \frac{1}{4c}\Big(\int_0^x f\left(\frac{x-r}{2c}, \frac{x+r}{2}\right) \,\mathrm{d}r - \int_0^x f\left(\frac{s-x}{2c}, \frac{s+x}{2}\right) \,\mathrm{d}s\Big). \tag{54}$$

Les fonctions F et G correspondent à des constantes d'intégration (chacune par rapport à une variable différente). On peut donc imposer la valeur de leur constante avec la relation $F(0) = G(0)$. En combinant les relations (53) et (54), on peut montrer que

$$\begin{aligned} F(x) = \frac{u_0(x)}{2} + \int_0^x \frac{u_1(s)}{2c}\,\mathrm{d}s + \frac{1}{8c^2}\Big(&\int_0^x\int_0^x f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r + \int_0^x\int_0^s f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}r\,\mathrm{d}s \\ &- \int_0^x\int_0^r f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r\Big), \\ G(x) = \frac{u_0(x)}{2} - \int_0^x \frac{u_1(s)}{2c}\,\mathrm{d}s + \frac{1}{8c^2}\Big(&\int_0^x\int_0^x f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r - \int_0^x\int_0^s f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}r\,\mathrm{d}s \\ &+ \int_0^x\int_0^r f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r\Big). \end{aligned}$$

On peut maintenant simplifier les intégrales pour obtenir

$$\begin{aligned} F(x) &= \frac{u_0(x)}{2} + \int_0^x \frac{u_1(s)}{2c}\,\mathrm{d}s + \frac{1}{4c^2}\int_0^x\int_0^s f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}r\,\mathrm{d}s, \\ G(x) &= \frac{u_0(x)}{2} - \int_0^x \frac{u_1(s)}{2c}\,\mathrm{d}s + \frac{1}{4c^2}\int_0^x\int_0^r f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right) \,\mathrm{d}s\,\mathrm{d}r. \end{aligned}$$

On peut maintenant injecter $F(x)$ et $G(x)$ dans l'expression de $u(t,x)$. Après simplification des intégrales, on obtient

$$u(t,x) = \frac{u_0(x+ct)+u_0(x-ct)}{2} + \frac{1}{2c}\int_{x-ct}^{x+ct} u_1(s)\,\mathrm{d}s + \frac{1}{4c^2}\int_{x-ct}^{x+ct}\int_{x-ct}^{s} f\left(\frac{s-r}{2c}, \frac{s+r}{2}\right)\,\mathrm{d}r\,\mathrm{d}s.$$

On termine cette preuve en remaniant le dernier terme par les changements de variables $\tilde{t} = \dfrac{s-r}{2c}$ et $\tilde{x} = s - c\tilde{t}$. □

Remarque 26 (Forme de la solution)**.** *Pour $f = 0$, on a $u(t,x) = F(x+ct)+G(x-ct)$. Sans terme source, la solution correspond donc à une onde se déplaçant vers la gauche (terme $F(x+ct)$) et une onde se déplaçant vers la droite (terme $G(x-ct)$), les deux à vitesse constante c. En fait, même dans le cas d'un domaine borné, comme $\Omega = (0,1)$, la solution peut se mettre sous cette forme tant que l'onde ne touche pas les bords du domaine (car alors des interactions plus compliquées apparaissent). On peut voir ici que l'équation des ondes se comporte (un peu) comme deux équation de transport. On peut pousser cette similitude en montrant l'écriture $\partial_t^2 - c^2\partial_x^2 = (\partial_t - c\partial_x)(\partial_t + c\partial_x)$. Notons également que la preuve précédente peut être vue comme une méthode des caractéristiques où on s'est intéressé aux droites $x - ct = a$ et $x + ct = b$ (pour a et b fixés) en introduisant les changements de variables $y = x + ct$ et $z = x - ct$.*

Remarque 27 (Propagation de l'information à vitesse finie)**.** *Une conséquence de la remarque précédente est que l'information se déplace dans les deux directions $+x$ et $-x$ à vitesse constante c. On a donc un comportement différent de celui de l'équation de la chaleur pour laquelle l'information se déplace à vitesse infinie.*

Exercice 28. *On suppose que $f = 0$. Retrouver le résultat du Théorème 7 en utilisant la méthode de la transformée de Fourier comme dans la Section 4.1.2.*

La première étape consiste à prouver que la transformée de Fourier de la solution vérifie $\hat{u}(t,\omega) = \hat{u}_0(\omega)\cos(c\omega t) + \frac{\hat{u}_1(\omega)}{c\omega}\sin(c\omega t)$. La deuxième étape consiste à calculer la transformée de Fourier inverse de $\cos(c\omega t)$ et $\sin(c\omega t)$ (au sens des distributions bien sûr). Enfin, on peut utiliser les propriétés de la transformée de Fourier pour retrouver le résultat de l'énoncé du Théorème 7.

5.2 Principales propriétés des solutions

On se place dans $\Omega = (0,1)$. On définit l'énergie du système

$$E(t) := \frac{1}{2}\int_0^1 \left(\frac{\partial u}{\partial t}\right)^2 \mathrm{d}x + \frac{c^2}{2}\int_0^1 \left(\frac{\partial u}{\partial x}\right)^2 \mathrm{d}x = \frac{1}{2}\|\partial_t u(t,.)\|^2_{L^2(0,1)} + \frac{c^2}{2}\|\partial_x u(t,.)\|^2_{L^2(0,1)}.$$

Nous rappelons que l'équation des ondes peut représenter un mouvement de corde. Le premier terme correspond alors à l'énergie cinétique de la corde et le second à l'énergie potentielle associée à la force de tension de la corde. Nous montrons que la dérivée de l'énergie mécanique totale $E(t)$ est égale à la puissance des efforts extérieurs f. En particulier, sans effort extérieur cette énergie se conserve. On peut aussi borner cette énergie.

Proposition 26 (Conservation et borne de l'énergie)**.** *Si $u \in H^1(0,T;L^2(0,1)) \cap L^2(0,T;H_0^1(0,1))$ est une solution de* (51)*, alors pour presque tout $t \in (0,T)$ on a*

$$E'(t) = \int_0^1 f \partial_t u \, \mathrm{d}x.$$

On peut de plus montrer que pour presque tout $t \in (0,T)$

$$E(t) \leq E(0)e^T + \frac{1}{2}\int_0^T \left[\int_0^1 (f(s,x))^2 \, \mathrm{d}x\right] e^{T-s} \, \mathrm{d}s. \tag{55}$$

Démonstration. En multipliant l'équation de la chaleur par $\dfrac{\partial u}{\partial t}$ et en intégrant sur $(0,1)$, on obtient

$$\begin{aligned}\int_0^1 f\partial_t u \, \mathrm{d}x &= \int_0^1 \frac{\partial^2 u}{\partial t^2}\frac{\partial u}{\partial t} \, \mathrm{d}x - c^2 \int_0^1 \frac{\partial u}{\partial t}\frac{\partial^2 u}{\partial x^2} \, \mathrm{d}x \\ &= \int_0^1 \frac{\partial^2 u}{\partial t^2}\frac{\partial u}{\partial t} \, \mathrm{d}x + c^2 \int_0^1 \frac{\partial^2 u}{\partial x \partial t}\frac{\partial u}{\partial x} \, \mathrm{d}x = E'(t).\end{aligned}$$

Notons qu'on a utilisé $\partial_t u(t,0) = \partial_t u(t,1)$ pour intégrer par parties.

En utilisant les inégalités de Cauchy–Schwarz et de Young, on obtient

$$\begin{aligned}|E'(t)| &\leq \|f(t,.)\|_{L^2(0,1)} \|\partial_t u(t,.)\|_{L^2(0,1)} \\ &\leq \|f(t,.)\|^2_{L^2(0,1)} + \|\partial_t u(t,.)\|^2_{L^2(0,1)} \\ &\leq \|f(t,.)\|^2_{L^2(0,1)} + E(t).\end{aligned}$$

Pour rappel, le Lemme de Grönwall nous dit que si une fonction η vérifie $\eta'(t) \leq \phi(t)\eta(t) + \psi(t)$ alors on a $\eta(t) \leq \exp(\int_0^t \phi(s) \, \mathrm{d}s)\eta(0) + \int_0^t \exp(\int_s^t \phi(u) \, \mathrm{d}u)\psi(s) \, \mathrm{d}s$.

On obtient le résultat en appliquant le Lemme de Grönwall ($\phi = 1$ et $\psi = \|f\|^2_{L^2(0,1)}$). □

On peut maintenant utiliser ce résultat sur l'énergie pour montrer que le problème des ondes avec conditions de Dirichlet est bien posé.

Proposition 27 (Caractère bien posé du problème)**.** *Pour $u_0 \in H_0^1(\Omega)$, $u_1 \in L^2(\Omega)$ et $f \in L^2(0,T;L^2(\Omega))$, la problème* (51) *admet une unique solution u dans $H^1(0,T;L^2(\Omega)) \cap L^2(0,T;H_0^1(\Omega))$. De plus, il existe une constante $C > 0$ (indépendante de u, u_0, u_1 et f) telle que*

$$\|u\|_{H^1(0,T;L^2(\Omega))} + \|u\|_{L^2(0,T;H_0^1(\Omega))} \leq C(\|u_0\|_{H_0^1(\Omega)} + \|u_1\|_{L^2(\Omega)} + \|f\|_{L^2(0,T;L^2(\Omega))}). \tag{56}$$

!! Est-ce qu'il s'agit des normes ou bien plutôt des semi-normes à gauche ? ?

Démonstration. L'existence utilise une analyse similaire à celle de la Section 5.1.1, mais dans un cadre plus général au sens des distributions. On ne la développera pas ici.

L'unicité est une conséquence de la linéarité du problème et de la dépendance aux données (considérer la différence entre deux solutions).

Il reste à prouver l'inégalité (56). On part de l'inégalité (55). On définit les semi-normes

$$|u|_{H^1(0,T;L^2(\Omega))} := \left(\int_0^T \int_0^1 \left(\frac{\partial u}{\partial t}(t,x) \right)^2 \mathrm{d}x \, \mathrm{d}t \right)^{\frac{1}{2}},$$
$$|u|_{L^2(0,T;H_0^1(\Omega))} := \left(\int_0^T \int_0^1 \left(\frac{\partial u}{\partial x}(t,x) \right)^2 \mathrm{d}x \, \mathrm{d}t \right)^{\frac{1}{2}}.$$

On a ainsi

$$\begin{aligned} |u|^2_{H^1(0,T;L^2(\Omega))} + |u|^2_{L^2(0,T;H_0^1(\Omega))} &\leq 2\max(1,c^{-2}) \int_0^T E(t) \, \mathrm{d}t \\ &\leq 2\max(1,c^{-2}) e^T T (E(0) + \|f\|^2_{L^2(0,T;L^2(\Omega))}). \end{aligned}$$

De plus, $E(0) \leq \frac{1}{2}\max(1,c)(\|u_0\|^2_{H_0^1(\Omega)} + \|u_1\|^2_{L^2(\Omega)})$.

On utilise ensuite l'inégalité de Poincaré

$$\|u\|_{L^2(0,T;L^2(\Omega))} \leq C |u|_{L^2(0,T;H_0^1(\Omega))},$$

et $\|u\|^2_{L^2(0,T;H_0^1(\Omega))} = |u|^2_{L^2(0,T;H_0^1(\Omega))} + \|u\|^2_{L^2(0,T;L^2(\Omega))}$ et de la même façon $\|u\|^2_{H^1(0,T;L^2(\Omega))} = |u|^2_{H^1(0,T;L^2(\Omega))} + \|u\|^2_{L^2(0,T;L^2(\Omega))}$.

On conclut avec l'inégalité de Young. □

On s'intéresse maintenant à la réversibilité de l'équation des ondes. Pour u une solution de (51), on pose $\widetilde{u}(t,x) := u(T-t,x)$. La fonction $\widetilde{u}$ vérifie le problème rétrograde

$$\begin{aligned} \frac{\partial^2 \widetilde{u}}{\partial t^2} - c^2 \frac{\partial^2 \widetilde{u}}{\partial x^2} &= \widetilde{f} \text{ dans } (0,T) \times (0,1), \\ \widetilde{u}(0,\cdot) &= u(T,\cdot) \text{ dans } (0,1), \\ \frac{\partial \widetilde{u}}{\partial t}(0,\cdot) &= -\frac{\partial u}{\partial t}(T,\cdot) \text{ dans } (0,1), \\ \widetilde{u}(t,0) = 0, \quad \widetilde{u}(t,1) &= 0 \qquad \forall t \in (0,T), \end{aligned} \tag{57}$$

avec $\widetilde{f}(t,x) := f(T-t,x)$. Il s'agit du problème (51) avec des données différentes. Ce problème est donc bien posé d'après la Proposition 27. L'équation des ondes est donc réversible.

Proposition 28 (Réversibilité de l'équation des ondes)**.** *L'équation des ondes est réversible : le problème rétrograde* (57) *est bien posé.*

Ceci signifie que retrouver la solution initiale à partir des données finales n'est pas particulièrement difficile pour cette équation.

Remarque 28 (Conditions de Dirichlet non homogènes)**.** *On peut gérer des conditions de Dirichlet non homogènes par un relèvement (voir la même remarque pour l'équation de Poisson et l'équation de la chaleur).*

5.3 Approximation par différences finies

!! Est-ce qu'on ne se concentre pas sur le schéma explicite centré ? ?

Nous étudions dans cette section différents schémas pour approcher la solution de l'équation. La discrétisation la plus naturelle que l'on puisse imaginer consiste à approcher les dérivées d'ordre deux par des approximations centrées. On utilise donc les approximations :

$$\frac{\partial^2 u}{\partial t^2}(t_n, x_j) \simeq \frac{u(t_{n+1}, x_j) - 2u(t_n, x_j) + u(t_{n-1}, x_j)}{h_t^2},$$
$$\frac{\partial^2 u}{\partial x^2}(t_n, x_j) \simeq \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{h_x^2}.$$

On obtient ainsi le schéma explicite suivant :

$$\forall 1 \leq n \leq N, \quad \forall 0 \leq j \leq M, \quad u_j^{n+1} = 2u_j^n - u_j^{n-1} + \frac{c^2 h_t^2}{h_x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + h_t^2 f(t_n, x_j).$$

Ce schéma est à deux pas de temps. C'est-à-dire que si on introduit un vecteur U^n représentant les u_j^n, pour calculer U^{n+1}, il faut U^n et U^{n-1}. Pour pouvoir calculer les U^n pour tout $n \geq 0$, il faut donc initialiser les deux premiers termes de la suite. Comme d'habitude, on initialise le premier pas de temps avec les conditions initiales. En plus de cela, il faut aussi initialiser le second pas de temps qu'on prendra comme $U^1 = U^0 + h_t V_0$ avec V_0 le vecteur des vitesses initiales. Cela signifie que l'on considère qu'au cours du premier pas de temps, la solution a été modifiée par la seule action de la vitesse initiale. Il s'agit d'une manière de procéder qui est classique. On obtient donc le schéma suivant

$$\begin{cases} \forall 1 \leq n \leq N, \quad \forall 0 \leq j \leq M, \quad u_j^{n+1} = 2u_j^n - u_j^{n-1} + \dfrac{c^2 h_t^2}{h_x^2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n) + h_t^2 f(t_n, x_j), \\ \forall 0 \leq n \leq N, \quad u_0^n = u_M^n = 0, \\ \forall 1 \leq j \leq M-1, \quad u_j^0 = u_0(x_j) \quad \text{et} \quad u_j^1 = u_0(x_j) + h_t u_1(x_j), \end{cases} \tag{58}$$

où u_0 et u_1 sont les données initiales du problème en position et en vitesse.

Exercice 29. *La figure (...) a été obtenue avec le schéma* (58) *pour les données (...). Coder ce schéma et retrouver la figure.*

!! RETROUVER LES PARAMETRES DE CETTE FIGURE !!

!! Completer cette remarque !!

Remarque 29. *Notons que la dérivée seconde présente dans l'équation des ondes rend très naturel le recours à un schéma à deux pas de temps. Ce type de schéma peut aussi être utilisé pour les équations vues précédemment (voir schéma saute-mouton). On peut aussi envisager d'utiliser un schéma à un pas de temps pour résoudre l'équation des ondes (voir ...).*

Sous forme matricielle, on peut écrire ce schéma comme

$$\begin{aligned} &U^{n+1} = (2I - c^2 h_t^2 A_\Delta)U^n - U^{n-1} + h_t^2 F^n, && 1 \leq n \leq N-1, \\ &(U^0)_j = u_0(x_j), && 1 \leq j \leq M-1, \\ &(U^1)_j = u_0(x_j) + h_t u_1(x_j), && 1 \leq j \leq M-1, \end{aligned}$$

avec I la matrice identité, A_Δ la matrice de Laplacien–Dirichlet définie en (21), U^n le vecteur d'état défini par $U_j^n := u_j^n$ et F^n le second membre défini par $F_j^n := f(t_n, x_j)$.

!! La matrice de Laplacien est définie avec des h au lieu de h_x, voir comment adapter cela!!

Cette écriture du schéma ne permet pas d'exploiter le cadre de la Section 3.3.2 (comparer avec (29)). On va essayer d'écrire ce même schéma sous une autre forme qui va permettre de se replacer dans le cadre des schémas à un pas. Pour cela, introduisons la variable

$$\Theta^n := \begin{pmatrix} V^n \\ U^n \end{pmatrix} := \begin{pmatrix} U^{n-1} \\ U^n \end{pmatrix}.$$

Cette variable vérifie le problème suivant

$$\Theta^{n+1} = \hat{A}\Theta^n + h_t^2 \hat{F}^n, \qquad 1 \leq n \leq N,$$

avec

$$\hat{A} := \begin{pmatrix} 0 & I \\ -I & 2I - c^2 h_t^2 A_\Delta \end{pmatrix}, \qquad \hat{F}^n := \begin{pmatrix} 0 \\ F^n \end{pmatrix}.$$

On s'est donc ramenés au cadre des schémas à un pas. La seule différence avec (29) est qu'il y a un h_t^2 au lieu d'un h_t devant $\hat{F}$. Notons que le second membre n'influence pas la stabilité (on peut donc prouver la stabilité du schéma en se plaçant dans le cadre de la Section 3.3.2). Pour la consistance, par contre, il faudra adapter.

Remarque 30. *Notons que cette 'astuce' d'introduire une nouvelle variable est classique dans le cadre des EDO. Par exemple, on ne peut pas utiliser directement le théorème de Cauchy–Lipschitz sur l'équation $\partial_t^2 u - Au = 0$ (avec des données de Cauchy adaptées) du fait de la dérivée seconde en temps. Par contre, on y arrivera sur l'équation satisfaite par la variable intermédiaire $v := \begin{pmatrix} u \\ \partial_t u \end{pmatrix}$. On pourra ensuite en déduire des propriétés de u.*

Nous prouvons maintenant la stabilité du schéma explicite centré pour la norme $\|\cdot\|_{L^2}$. Nous allons donc calculer les valeurs propres de $\hat{A}$ et s'assurer qu'elles sont toutes de module inférieur ou égal à un. On considère $\lambda \in \mathbb{C} \setminus \{0\}$ et on calcule

$$\begin{aligned} \left|\hat{A} - \lambda I\right| &= \begin{vmatrix} -\lambda I & I \\ -I & (2-\lambda)I - c^2 h_t^2 A_\Delta \end{vmatrix} \\ &= \begin{vmatrix} -\lambda I & 0 \\ -I & (2-\lambda-\frac{1}{\lambda})I - c^2 h_t^2 A_\Delta \end{vmatrix} \\ &= (-\lambda)^{M-1} \left|(2-\lambda-\frac{1}{\lambda})I - c^2 h_t^2 A_\Delta\right|. \end{aligned}$$

On note μ_j les valeurs propres de la matrice $2I - c^2 h_t^2 A_\Delta$, c'est-à-dire $|(2-\mu)I - c^2 h_t^2 A_\Delta| = \prod_{j=1}^{M-1}(\mu_j - \mu)$. On a donc $|\hat{A} - \lambda I| = (-\lambda)^{M-1} \prod_{j=1}^{M-1}(\mu_j - \lambda - \frac{1}{\lambda}) = \prod_{j=1}^{M-1}(\lambda^2 - \mu_j\lambda + 1)$. Le polynôme caractéristique étant continu, cette expression est valable également pour $\lambda = 0$. Les valeurs propres de la matrice $\hat{A}$ sont donc les racines des polynômes $\lambda^2 - \mu_j\lambda + 1$ (il faut considérer tous les j dans $< 1, M-1 >$). Le discriminant de ce polynôme vaut $\Delta_j = \mu_j^2 - 4$.

On rappelle que le schéma est stable pour la norme $\|\cdot\|_{L^2}$ si et seulement si toutes les valeurs propres sont de module inférieur ou égal à un.

• S'il y a un j tel que $\mu_j^2 > 4$, alors on a deux valeurs propres réelles qui vérifient $\lambda_1\lambda_2 = 1$ et $\lambda_1 \neq \lambda_2$. Donc soit $\lambda_1 > 1$ ou soit $\lambda_2 > 1$ et dans les deux cas le schéma n'est pas stable.

• Pour j tel que $\mu_j^2 = 4$, on a une valeur propre double qui vaut 1 et ce j n'empêche pas le schéma d'être stable.

• Pour j tel que $\mu_j^2 < 4$, on a deux valeurs propres conjuguées qui vérifient $|\lambda_1| = |\lambda_2| = 1$ et ce j n'empêche pas le schéma d'être stable.

Une condition nécessaire et suffisante pour que le schéma soit stable pour la norme $\|\cdot\|_{L^2}$ est donc d'avoir $-2 \leq \mu_j \leq 2$ pour tout $j \in < 1, M-1 >$.

On remarque que la matrice $2I - c^2h_t^2A_\Delta$ est de la forme des matrices étudiées dans la Proposition 23 avec $a = 2 - \frac{2c^2h_t^2}{h_x^2}$ et $b = \frac{c^2h_t^2}{h_x^2}$. On peut donc calculer

$$\mu_j = a + 2b\cos\left(\frac{j\pi}{M}\right) = 2 - 2\alpha^2\left(1 - \cos\left(\frac{j\pi}{M}\right)\right),$$

avec $\alpha = \dfrac{ch_t}{h_x}$.

On remarque que la condition $\mu_j \leq 2$ est acquise. La condition $-2 \leq \mu_j$ (pour tout $j \in < 1, M-1 >$) est valable si et seulement si $\alpha \leq 1$.

Le schéma est donc stable pour la norme $\|\cdot\|_{L^2}$ si et seulement si la condition de CFL suivante est vérifiée

$$\alpha = \frac{ch_t}{h_x} \leq 1.$$

!! Ajouter une remarque sur la CFL : comparaison avec l'équation de transport, donner un sens physique à cette condition de CFL, étendre ce sens au cas parabolique!!

!! Faire la consistance de manière propre (régler le problème du h_t^2)!!

!! Eclaircir la subtilité de la condition initiale (voir notes papier ou cours ENSTA)!!

!! Creuser la différence Von Neumann vs transformée de Fourier discrète!!

5.4 Étude du problème de Neumann

!! faire d'autres conditions limites (Neumann) ? ?

6 Pour aller plus loin

!! Ecrire tous les exercices au fur et à mesure et les reporter dans une section à la fin du poly!!

!! les elements de cette section sont hors-programme mais ne sont pas déconnectés de celui-ci au sens où ils peuvent être utilisés dans un texte (ils seront alors présentés)!!!! Ne travailler cette section qu'une fois que les autres sont parfaitement maitrisées!!

6.1 Équation de transport dans un domaine multi-dimensionnel

!! Ajouter une annexe sur les EDO + y faire reference dans la premiere section d'analyse numerique!!

6.2 Équation de Burgers

6.3 Recherche de solutions faibles à l'équation de la chaleur

!! AJOUTER UN CHAPEAU A CETTE SECTION !!

Cette section constitue une preuve à la proposition 18. On cherche une solution faible à l'équation de la chaleur. Cela signifie que pour presque tout $t \in (0,T)$, on considère les termes de l'égalité $\dfrac{\partial u}{\partial t}(t,\cdot) - \dfrac{\partial^2 u}{\partial x^2}(t,\cdot) = f(t,\cdot)$ comme des éléments de $H^{-1}(\Omega)$. On cherche donc une solution u qui vérifie

$$\forall v \in H_0^1(\Omega), \quad \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(t,x)v(x)\,\mathrm{d}x + \int_\Omega \frac{\partial u}{\partial x}(t,x)\frac{\partial v}{\partial x}(x)\,\mathrm{d}x = \int_\Omega f(t,x)v(x)\,\mathrm{d}x. \tag{59}$$

!! Est-ce qu'on peut prouver l'unicité de la solution dans un espace particulier ? ? (pas juste dans l'espace des solutions de cette forme) —- On a l'unicité avec la Proposition de stabilité !! —- Il suffit juste de spéficier la régularité de cette solution particulière !!

Cette démonstration est issue du livre de B. Lucquin. On note $\Omega = (0,1)$.

• Recherche d'une solution : on note $\mathcal{A}$ l'opérateur $-\Delta$ sur $H_0^1(\Omega)$, on peut le définir comme

$$\forall v, w \in H_0^1(\Omega), \quad \int_\Omega \mathcal{A}(v)w\,\mathrm{d}x := \int_\Omega v'(x)w'(x)\,\mathrm{d}x.$$

On utilise une base orthogonale de fonctions propres de l'opérateur $\mathcal{A}$. On note $w_k(x) = \sqrt{2}\sin(k\pi x)$ et $\lambda_k = (k\pi)^2$. On peut montrer que l'on a

$$\forall v \in H_0^1(\Omega), \quad \int_\Omega \mathcal{A}(w_k)v\,\mathrm{d}x = \int_\Omega w_k'(x)v'(x)\,\mathrm{d}x = \lambda_k\int_\Omega w_k(x)v(x)\,\mathrm{d}x.$$

En ce sens, pour $k \geq 1$, la fonction w_k est un vecteur propre de $\mathcal{A}$ associé à la valeur propre λ_k. Nous allons maintenant chercher une solution u au problème (43) sous la forme

$$u(t,x) = \sum_{\ell=1}^{+\infty} u_\ell(t)w_\ell(x), \tag{60}$$

avec pour $\ell \geq 1$, u_ℓ une fonction à déterminer.

Les fonctions w_k vérifient $\int_\Omega w_k(x)w_\ell(x)\,\mathrm{d}x = \delta_{k\ell}$ où $\delta_{k\ell}$ représente le symbole de Kronecker (faîtes ce calcul).

En cherchant u sous la forme (60) et en prenant $v = w_k$ dans (59), on obtient pour tout $k \geq 1$, $u_k'(t) + \lambda_k u_k(t) = \int_\Omega f(t,x)w_k(x)\,\mathrm{d}x$. La condition initiale donne $u_k(0) = \int_\Omega u_0(x)w_k(x)\,\mathrm{d}x$. Les fonctions u_k vérifient donc

$$\forall t \in [0,T], \quad u_k(t) = \int_\Omega u_0(x)w_k(x)\,\mathrm{d}x e^{-\lambda_k t} + \int_0^t\int_\Omega f(s,x)w_k(x)\,\mathrm{d}x e^{-\lambda_k(t-s)}\,\mathrm{d}s. \tag{61}$$

Il faut maintenant montrer que la série de fonctions (60) définie à partir de (61) converge dans les espaces $C^0([0,T];L^2(0,1))$ et $L^2(0,T;H_0^1(0,1))$. Pour cela, nous allons considérer la suite $(u^{(n)})_n$ avec $u^{(n)}(t,x) = \sum_{\ell=1}^n u_\ell(t)w_\ell(x)$. Ces deux espaces étant dans des espaces de Hilbert, nous allons montrer que cette suite est de Cauchy dans chacun de ces espaces.

• Convergence dans l'espace $C^0([0,T];L^2(0,1))$: en utilisant l'orthogonalité des w_k, nous avons pour $0 \le m \le n$,

$$\|u^{(n)}(t,\cdot) - u^{(m)}(t,\cdot)\|^2_{L^2(0,1)} = \|\sum_{\ell=m+1}^{n} u_\ell(t) w_\ell\|^2_{L^2(0,1)} = \sum_{k,\ell=m+1}^{n} u_\ell(t)u_k(t)(w_\ell,w_k)_{L^2(0,1)}$$
$$= \sum_{k,\ell=m+1}^{n} u_\ell(t)u_k(t)\delta_{k\ell} = \sum_{\ell=m+1}^{n} (u_\ell(t))^2.$$

On note

$$a_k(t) = (u_0,w_k)_{L^2(0,1)} e^{-\lambda_k t}, \qquad b_k(t) = \int_0^t (f(s,.),w_k)_{L^2(0,1)} e^{-\lambda_k(t-s)}\,\mathrm{d}s.$$

En utilisant (61) et l'inégalité de Young, on a $(u_k(t))^2 = (a_k(t)+b_k(t))^2 \le 2(a_k(t)^2+b_k(t)^2)$. Ainsi,

$$\|u^{(n)}(t,\cdot) - u^{(m)}(t,\cdot)\|^2_{L^2(0,1)} \le 2(A_{n,m}(t) + B_{n,m}(t)),$$

avec

$$A_{n,m}(t) = \sum_{\ell=m+1}^{n} (a_\ell(t))^2, \qquad B_{n,m}(t) = \sum_{\ell=m+1}^{n} (b_\ell(t))^2.$$

Pour $t \ge 0$, on a $A_{n,m}(t) \le \sum_{\ell=m+1}^{n} (u_0,w_\ell)^2_{L^2(0,1)}$. Étant donné que les w_k forment une base hilbertienne de $L^2(0,1)$ et que $u_0 \in L^2(0,1)$, on a $u_0 = \sum_{\ell=1}^{+\infty}(u_0,w_\ell)_{L^2(0,1)} w_\ell$ et avec l'égalité de Parseval, on a $\|u_0\|^2_{L^2(0,1)} = \sum_{\ell=1}^{+\infty}(u_0,w_\ell)^2_{L^2(0,1)}$. La série de terme général $(u_0,w_\ell)^2_{L^2(0,1)}$ converge donc et on a $\sup_{t\in[0,T]} |A_{n,m}(t)| \to 0$ quand n et $m \to +\infty$.

!! VERIFIER QU'ON A BIEN UNE BASE HILBERTIENNE DANS L^2 ET H^1_0 !! — FAIRE UN POINT SUR TOUTES LES PROPRIETES DE CETTE BASE !!

Montrons maintenant le même résultat sur $B_{n,m}(t)$. On a

$$\Big(\int_0^t (f(s,\cdot),w_k)_{L^2(0,1)} e^{-\lambda_k(t-s)}\,\mathrm{d}s\Big)^2 \le \Big(\int_0^t (f(s,\cdot),w_k)^2_{L^2(0,1)}\,\mathrm{d}s\Big)\Big(\int_0^t e^{-2\lambda_k(t-s)}\,\mathrm{d}s\Big)$$
$$\le \frac{1}{2\lambda_k}\Big(\int_0^t (f(s,\cdot),w_k)^2_{L^2(0,1)}\,\mathrm{d}s\Big)$$
$$\le \frac{1}{2\pi^2}\Big(\int_0^T (f(s,\cdot),w_k)^2_{L^2(0,1)}\,\mathrm{d}s\Big),$$

où on a utilisé l'inégalité de Cauchy–Schwarz et $\lambda_k = k^2\pi^2 \ge \pi^2$. De plus, comme $f \in L^2(0,T;L^2(0,1))$, pour presque tout $s \in [0,T]$, $f(s,\cdot)$ appartient à $L^2(0,1)$. D'après l'égalité de Parseval, pour presque tout $s \in [0,T]$, $\|f(s,\cdot)\|^2_{L^2(0,1)} = \sum_{\ell=1}^{+\infty}(f(s,\cdot),w_\ell)^2_{L^2(0,1)}$. Ainsi, $\|f\|^2_{L^2(0,T;L^2(0,1))} = \int_0^T \sum_{\ell=1}^{+\infty}(f(s,\cdot),w_\ell)^2_{L^2(0,1)}\,\mathrm{d}s$.

On peut donc montrer que la série de terme général $\int_0^T (f(s,\cdot),w_k)^2_{L^2(0,1)}\,\mathrm{d}s$ converge (en étant bornée par $\|f\|^2_{L^2(0,T;L^2(0,1))}$). Donc $\sup_{t\in[0,T]} |B_{n,m}(t)| \to 0$ quand n et $m \to +\infty$.

La suite $(u^{(n)})$ est donc de Cauchy dans $C^0([0,T];L^2(0,1))$ et converge donc.

• Convergence dans l'espace $L^2([0,T];H_0^1(0,1))$: La norme $|u|_{1,\Omega}^2 = \int_0^1 (u'(x))^2\,\mathrm{d}x = \mathcal{A}(u,u)$ est une norme sur $H_0^1(0,1)$ équivalente à la norme usuelle (par l'inégalité de Poincaré, voir (...)). On a

$$\begin{aligned}|(u^{(n)}-u^{(m)})(s,\cdot)|_{1,\Omega}^2 &= \mathcal{A}((u^{(n)}-u^{(m)})(s,\cdot),(u^{(n)}-u^{(m)})(s,\cdot))\\ &= \sum_{k,\ell=m+1}^{n} u_k(s)u_\ell(s)\mathcal{A}(w_k,w_\ell)\\ &= \sum_{k,\ell=m+1}^{n} u_k(s)u_\ell(s)\lambda_k\delta_{k\ell} = \sum_{k=m+1}^{n}\lambda_k(u_k(s))^2.\end{aligned}$$

On montre donc que $|(u^{(n)}-u^{(m)})(s,\cdot)|_{1,\Omega}^2 \le 2(C_{n,m}(t)+D_{n,m}(t))$ avec

$$C_{n,m}(t) := \sum_{\ell=m+1}^{n}\lambda_\ell(a_\ell(t))^2, \qquad D_{n,m}(t) := \sum_{\ell=m+1}^{n}\lambda_\ell(b_\ell(t))^2.$$

En raisonnant comme précédemment, on a

$$\int_0^T \lambda_k e^{-2\lambda_k s}(u_0,w_k)_{L^2(\Omega)}^2\,\mathrm{d}s \le \frac{1}{2}(u_0,w_k)_{L^2(\Omega)}^2,$$

et la série de terme général $(u_0,w_k)_{L^2(\Omega)}^2$ converge par l'égalité de Parseval. Donc $\int_0^T C_{n,m}(s)\,\mathrm{d}s \to 0$ quand n et $m \to +\infty$.

De la même façon,

$$\begin{aligned}\int_0^T \lambda_k\Big(\int_0^t (f(s,\cdot),w_k)_{L^2(0,1)}e^{-\lambda_k(t-s)}\,\mathrm{d}s\Big)^2\,\mathrm{d}t &\le \int_0^T \lambda_k\Big(\int_0^t (f(s,\cdot),w_k)_{L^2(0,1)}^2\,\mathrm{d}s\Big)\Big(\int_0^t e^{-2\lambda_k(t-s)}\,\mathrm{d}s\Big)\,\mathrm{d}t\\ &\le \frac{1}{2}\int_0^T\Big(\int_0^t (f(s,\cdot),w_k)_{L^2(0,1)}^2\,\mathrm{d}s\Big)\,\mathrm{d}t\\ &\le \frac{T}{2}\Big(\int_0^T (f(s,\cdot),w_k)_{L^2(0,1)}^2\,\mathrm{d}s\Big).\end{aligned}$$

Comme nous l'avons vu précédemment, la série de terme général $\int_0^T (f(s,\cdot),w_k)_{L^2(0,1)}^2\,\mathrm{d}s$ converge.

Nous avons donc montré que la suite $(u^{(n)})$ est de Cauchy et donc converge dans $L^2(0,T;H_0^1(0,1))$.

• Solution de l'équation : montrons maintenant que u définie par (60) avec (61) vérifie l'équation (43) au sens faible (c'est-à-dire (59)).

Soit $v \in H_0^1(\Omega)$. On note comme précédemment $u^{(N)}(t,x) := \sum_{\ell=1}^N u_\ell(t) w_\ell(x)$. On a

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}(u^{(N)}(t,\cdot),v)_{L^2(\Omega)} &= \sum_{k=1}^N u_k'(t)(w_k,v)_{L^2(\Omega)} \\
&= \sum_{k=1}^N -\lambda_k (u_0,w_k)_{L^2(\Omega)}(w_k,v)_{L^2(\Omega)} e^{-\lambda_k t} + (f(t,\cdot),w_k)_{L^2(\Omega)}(w_k,v)_{L^2(\Omega)} \\
&\qquad - \lambda_k \int_0^t (f(s,\cdot),w_k)_{L^2(\Omega)}(w_k,v)_{L^2(\Omega)} e^{-\lambda_k(t-s)}\,\mathrm{d}s \\
&= \sum_{k=1}^N -\Big((u_0,w_k)_{L^2(\Omega)} e^{-\lambda_k t} + \int_0^t (f(s,\cdot),w_k)_{L^2(\Omega)} e^{-\lambda_k(t-s)}\,\mathrm{d}s\Big)\mathcal{A}(w_k,v) \\
&\qquad + (f(t,\cdot),w_k)_{L^2(\Omega)}(w_k,v)_{L^2(\Omega)} \\
&= -\mathcal{A}(u^{(N)}(t,\cdot),v) + \sum_{k=1}^N (f(t,\cdot),w_k)_{L^2(\Omega)}(w_k,v)_{L^2(\Omega)}.
\end{aligned}$$

Quand $N \to +\infty$, $u^{(N)}$ converge vers u dans $L^2(0,T;H_0^1(\Omega))$. Ainsi, le terme $\dfrac{\mathrm{d}}{\mathrm{d}t}(u^{(N)}(t,\cdot),v)_{L^2(\Omega)}$ est correctement défini quand $N \to +\infty$ et sa limite vaut $\dfrac{\mathrm{d}}{\mathrm{d}t}(u(t,\cdot),v)_{L^2(\Omega)}$. De plus, $v \in L^2(\Omega)$ et (w_k) est une base Hilbertienne, $\sum_{k=1}^N (f(t,\cdot),w_k)_{L^2(\Omega)}(w_k,v)_{L^2(\Omega)}$ tend donc vers $(f(t,\cdot),v)_{L^2(\Omega)}$ quand $N \to +\infty$. La fonction u vérifie donc l'équation (59), ce qui conclut cette démonstration.

A Principales propriétés de la transformée de Fourier

$$\mathcal{F}(\frac{\partial u}{\partial x}) = i\omega \mathcal{F}(u), \tag{62}$$

$$\mathcal{F}(u \times v) = \mathcal{F}(u) * \mathcal{F}(v), \tag{63}$$

$$\mathcal{F}^{-1}(\mathcal{F}(u)) = u, \tag{64}$$

avec $u * v$ le produit de convolution. Les propriétés (63) sont aussi vérifiées par $\mathcal{F}^{-1}$, les preuves sont similaires et ne sont pas données dans ce document.