# Introduction aux équations aux dérivées partielles

## 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 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 :

— cinq séances de cours de deux heures chacune.

— une séance de programmation de deux 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 $F$ est une fonction 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 les modèles présentés dans cette section sont classiques.*

## 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 et si l'on considère plusieurs dimensions d'espace, cette équation devient

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

où $\nabla\cdot$ et $\nabla$ sont respectivement les opérateurs divergence et gradient.

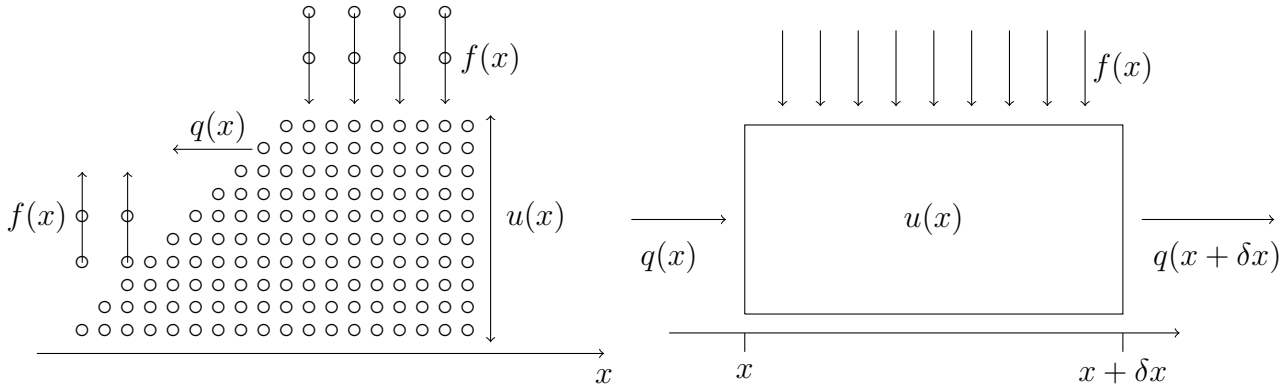


FIGURE 1 – Répartition de particules dans un domaine. Gauche : représentation du problème. Droite : équilibre des flux sur une portion infinitésimale du domaine.

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 donnons dans la partie gauche de la figure 1 une représentation du 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 les représentant. 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 pas du temps.

Pour établir les équations de ce problème, considérons une portion infinitésimale du domaine (de taille $\delta x$) 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

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

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 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, $f$

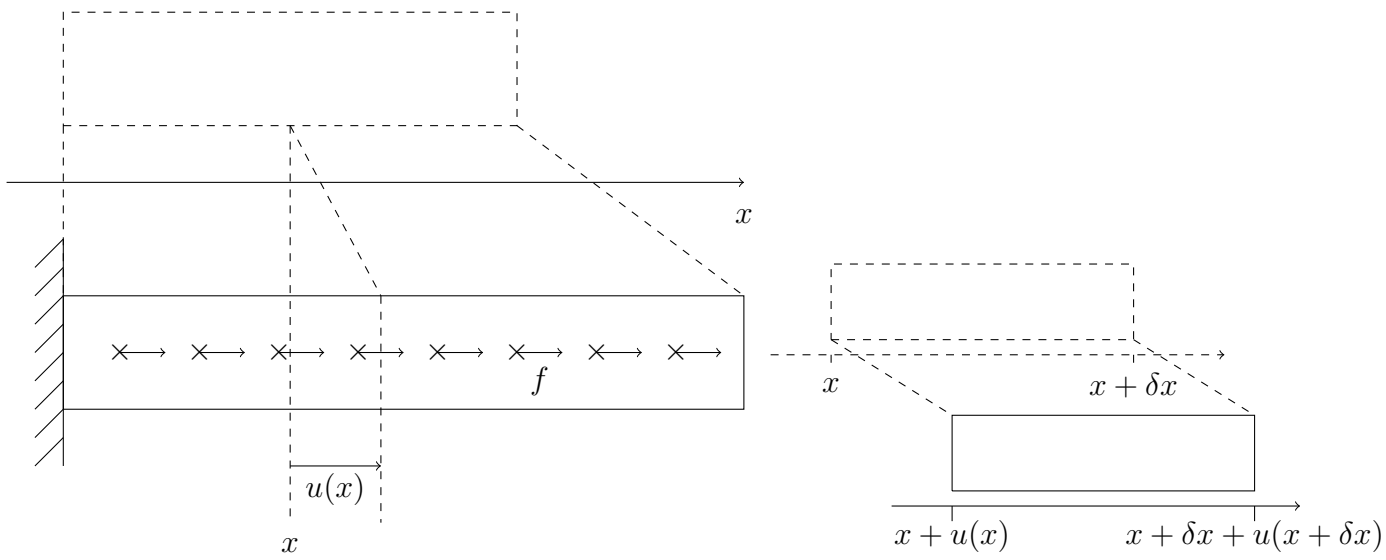


FIGURE 2 – Une barre élastique en équilibre. Gauche : représentation du problème. Droite : déformation d'un élément infinitésimal de matière. Le trait continu représente la configuration soumise à une charge $f$. Le trait discontinu représente la configuration au repos (sans $f$).

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 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 $\delta x$ 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 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. En général, la valeur $g$ sera 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 | déplacements de mol. |
| $k$ | conduction thermique | raideur | diffusion |
| $f$ | source | force externe | réactions chimiques |
| $u = g$ | temp. constante | déplacement imposé | concentration imposée |
| $q = 0$ | paroi adiabatique | frontière libre | frontière imperméable |

TABLE 1 – Résumé des problèmes physiques abordés.

**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".

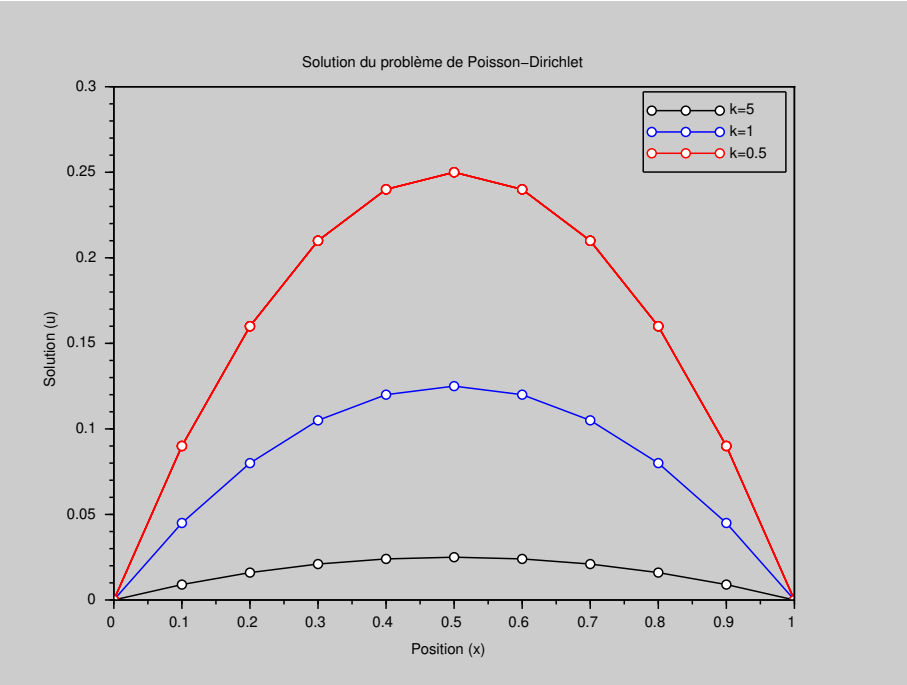


FIGURE 3 – Solution de l'équation elliptique 1D (3) pour différents $k$ (pour $f(x) = 1$).

## 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 $t$.

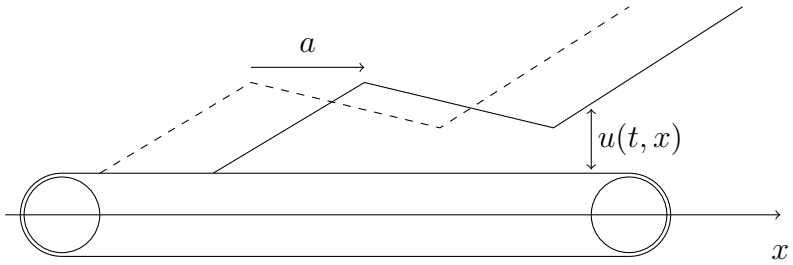


FIGURE 4 – Des tas de sable transportés sur un tapis roulant. La ligne discontinue représente les tas en $t = 0s$ et la ligne continue en $t = 1s$.

Si l'on considère un temps infinitésimal $\delta t$, 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

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

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.*

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.

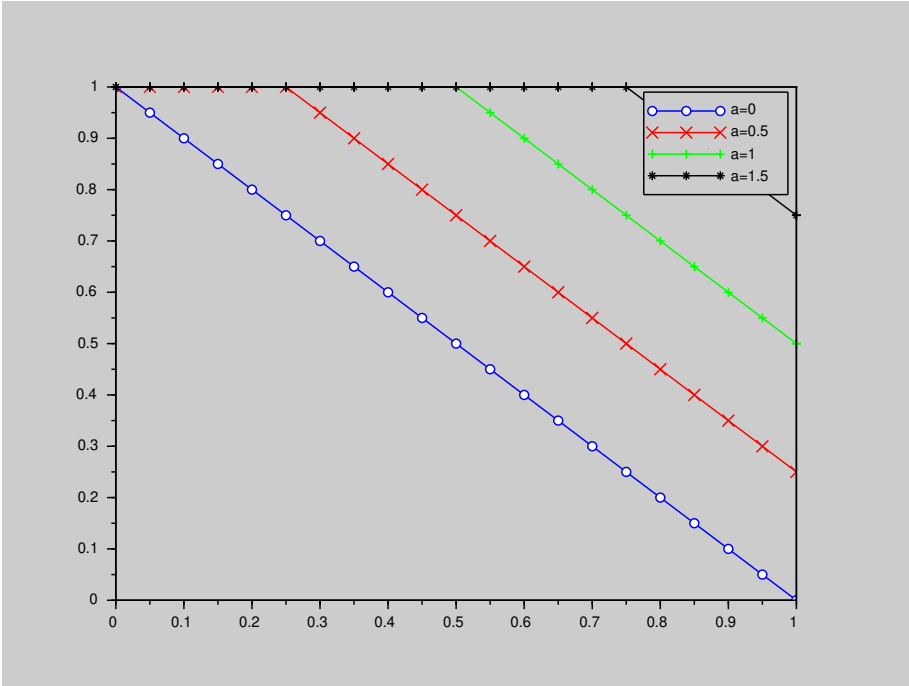


FIGURE 5 – Solution de l'équation de transport à $t = 0.5$ pour différents $a$. La donnée initiale correspond à la solution pour $a = 0$.

## 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 conduction 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 diffusion thermique.

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

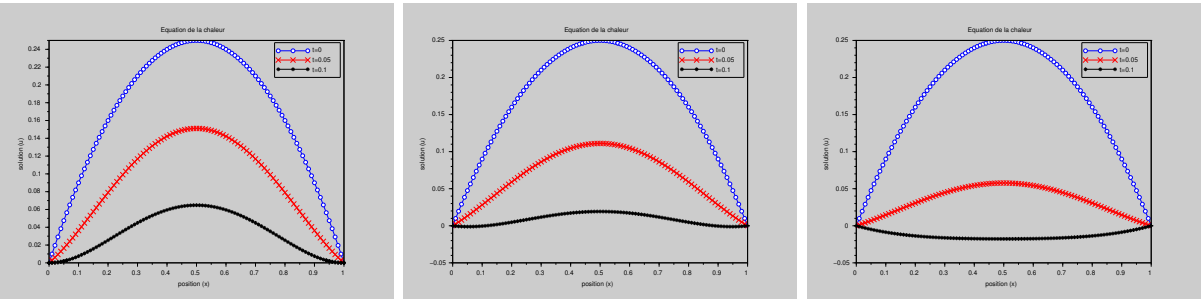


FIGURE 6 – Solution de l'équation de la chaleur pour $f(x) = -1$. Gauche : $\nu = 0.5$. Milieu : $\nu = 1$. Droite : $\nu = 2$. Les courbes représentent la solution à $t = 0$ (bleu), $t = 0.05$ (rouge) et $t = 0.1$ (noir).

Nous illustrons sur la figure 6 l'influence du paramètre $\nu$. Plus $\nu$ 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 $\delta x$ nous dit que l'accélération de la position de cette tranche $(x+u(t,x))$ 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 $\rho$ 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 $\rho$ 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.

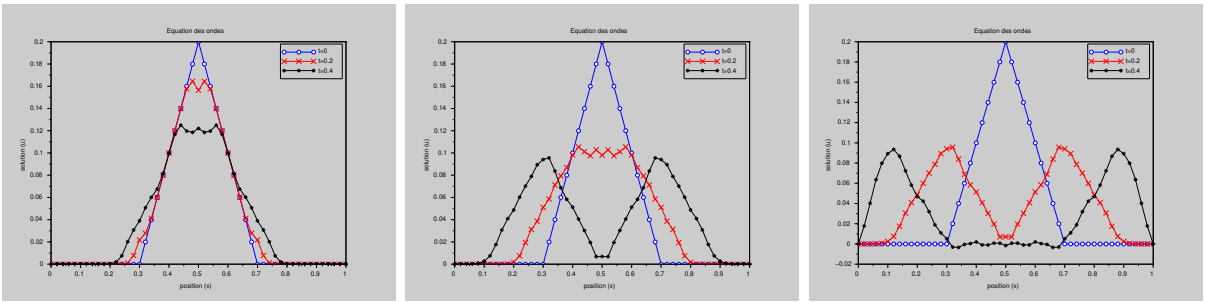


FIGURE 7 – Solution de l'équation des ondes pour $f=0$ et $v_0=0$. Gauche : $c=0.2$. Milieu : $c=0.5$. Droite : $c=1$. Les courbes représentent la solution à $t=0$ (bleu), $t=0.2$ (rouge) et $t=0.4$ (noir).

On 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 (on ne s'étendra pas davantage sur le sujet).

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 $\Omega$ 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

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

## 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 $C^\ell$ 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 $\Omega$, on dit que $u$ est une fonction harmonique.*

Il existe des liens entre fonctions harmoniques et fonctions holomorphes. En effet, si $\varphi$ 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 homogène

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

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 $\ell$ 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 :*

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

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

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

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

où $\nabla v$ 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*

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

**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 $\ell$ 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). *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 $H^1 = \{v \in L^2(\Omega) \mid \nabla u \in L^2(\Omega)\}$ et $H^1_\bullet = \{v \in H^1(\Omega) \mid \int_\Omega v \, \mathrm{d}x = 0\}$. 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*

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

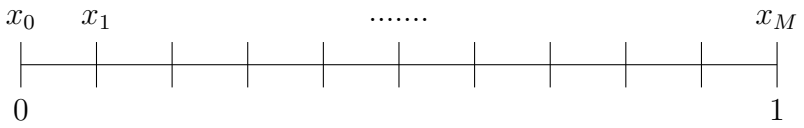


FIGURE 8 – Représentation de la discrétisation de l'intervalle $(0, 1)$.

## 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 aux dérivées ordinaires) ou EDP (équation aux dérivées partielles) par des valeurs censées représenter cette fonction en certains points. Dans toutes les discrétisations abordées dans 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 $AU = F$, avec*

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

*Démonstration.* La relation (20) est équivalente au système linéaire $AU = F$. 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 $AU = F$. La réciproque se prouve de manière similaire.

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

**Remarque 6.** *La matrice $A$ 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 $A$ 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 $\delta_j$, on a $\underline{\delta} = A^{-1}\underline{\epsilon}$ avec $(\underline{\epsilon})_j = O(h^2)$. En supposant que $\sup_{X \in \mathbb{R}^{M-1}\setminus\{0\}} \dfrac{\max_j |(A^{-1}X)_j|}{\max_j |X_j|}$ 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 $\log(\varepsilon) = 2\log(h) + \log(C)$.

**Exercice 7.** *Coder le schéma décrit précédemment et retrouver les résultats de la figure 9.*

**Exercice 8** (Coefficient de raideur)**.** *On peut décider d'ajouter un coefficient de raideur $k > 0$ constant. La nouvelle matrice devient alors $kA$ où $A$ 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*

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

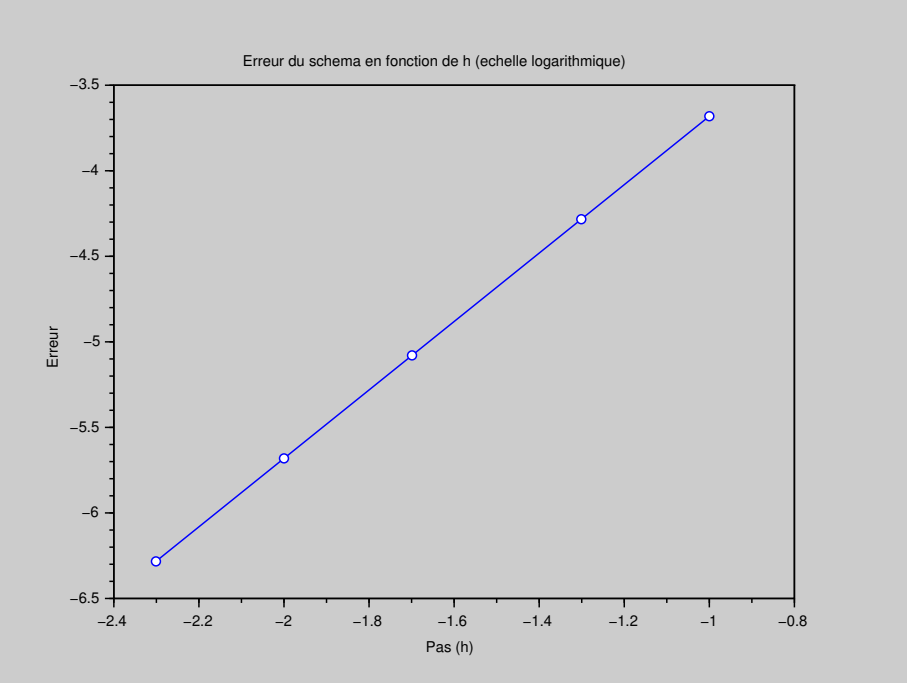


FIGURE 9 – Erreur en fonction de $h$ (échelle logarithmique) pour $f(x) = x^2$.

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 $\alpha$ ?*

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*

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

**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".

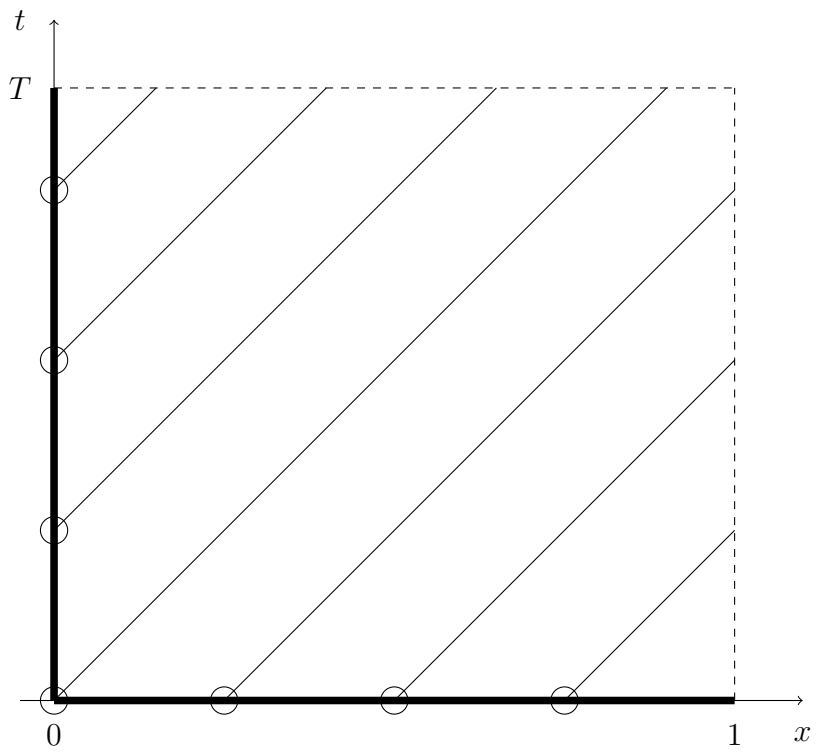


FIGURE 10 – Les droites caractéristiques de l'équation de transport ($a = 1$ et $T = 0.9$).

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

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

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

Cette propriété signifie que l'information est conservée au cours du temps : le comportement de l'équation ne dégrade pas cette information.

## 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. Nous allons maintenant 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 $u(t,x) = \alpha(t - x/a)$. Si $x \ge at$, on applique la proposition 8 avec $s = t$, avec la condition initiale on obtient $u(t,x) = u_0(x-at)$. 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 $0 \le f(x) < 1$.*

**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 $0 \le n \le N$, $t_n = nh_t$ 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 $0 \le j \le M$, $x_j = jh_x$ et $h_x = 1/M$. De plus, nous allons approcher la fonction $u(t,x)$ par une suite $(u_j^n)_{\substack{0\le n\le N\\ 0\le j\le M}}$. 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 au traitement sur le temps long de l'information issue de la donnée initiale. Pour cela, on peut essayer de suivre le déplacement de cette donnée initiale dans un domaine infini en considérant des conditions périodiques. Dans cette section, on s'intéresse 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 discrértisations de l'équation (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 $(N, M)$.*

### 3.3.2 Analyse numérique : définitions et théorèmes

Nous définissons maintenant des notions que nous utilisons ensuite pour analyser les deux schémas introduits précédemment. Pour cela, représentons la suite $(u_j^n)_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}}$ par un vecteur $U^n$ tel que $u_j^n$ soit égal à la $j$-ème coordonnée du vecteur $U^n$ généré par récurrence comme suit :

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

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

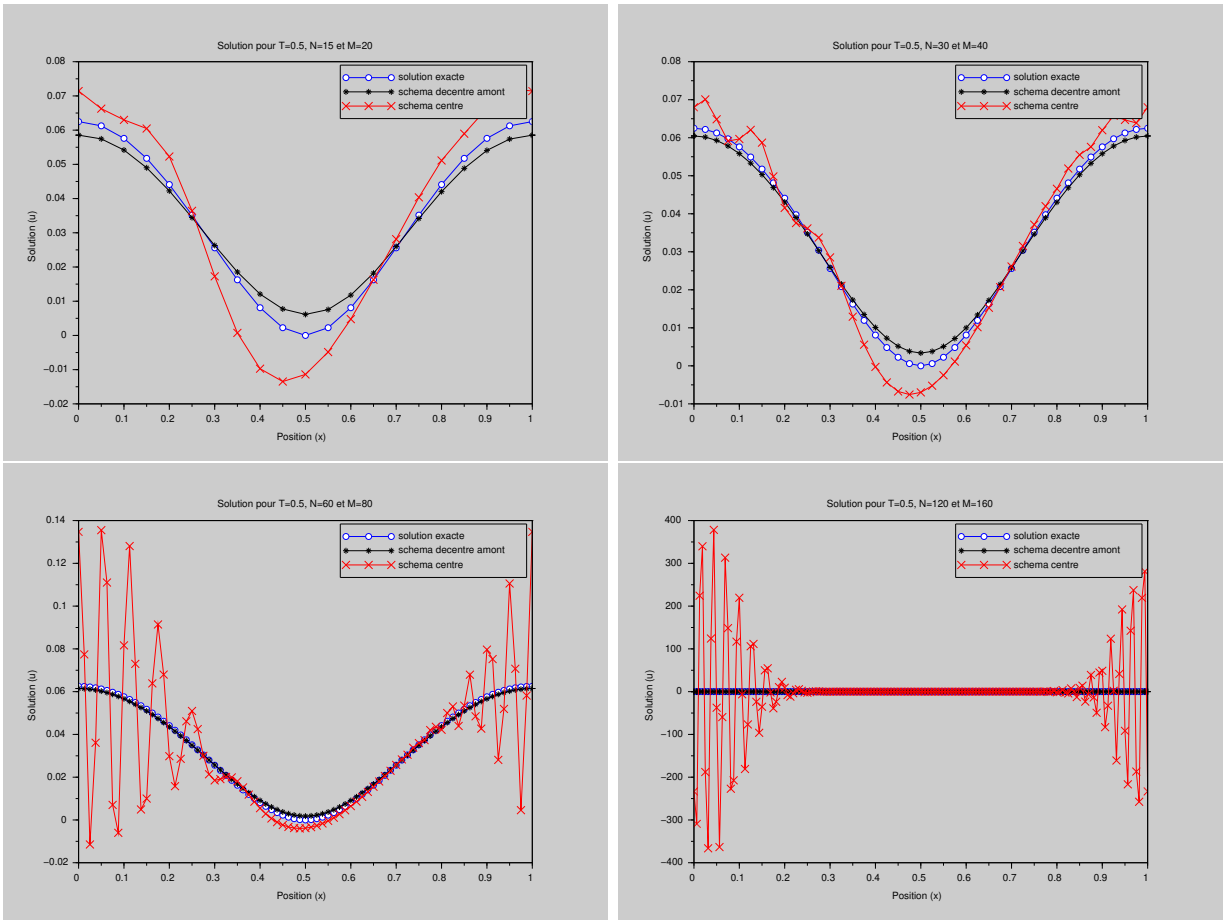


FIGURE 11 – Équation de transport avec conditions périodiques : solution exacte (à l'instant final) et solutions (à l'instant final) obtenues par les schémas décentré et centré pour $T = 0.5$, $a = 1$ et $(N, M) = (15, 20)$, $(30, 40)$, $(60, 80)$ et $(120, 160)$.

numérotera 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 ce cours.*

Définissons maintenant un certain nombre de notions utiles pour étudier le comportement du schéma (29).

**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 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*

$$\max_{\substack{0 \leq n \leq N \\ 1 \leq j \leq M}} |u_j^n| \leq C_1 \max_{1 \leq j \leq M} |(U_0)_j| + C_2 \max_{\substack{0 \leq n \leq N \\ 1 \leq j \leq M}} |F_j^n|, \tag{30}$$

*où $u_j^n = (U^n)_j$ avec $(U^n)$ l'unique suite vérifiant* (29).

*On dit également que le schéma* (29) *est inconditionnellement stable 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 du problème exact associé à* (29).

**Définition 4** (Consistance)**.** *On dit que le schéma numérique* (29) *est consistant si pour toute solution régulière $u$ du problème exact on a*

$$\lim_{\substack{N\to+\infty \\ M\to+\infty}} \max_{\substack{0\leq n\leq N \\ 1\leq j\leq M}} \frac{|\varepsilon_j^n|}{h_t} = 0, \tag{32}$$

*où $\varepsilon_j^n = (\varepsilon^n)_j$ 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 si pour toute solution régulière $u$ (du problème exact) il existe une constante $C > 0$ indépendante de $h_t$ et $h_x$ telle que*

$$\forall N, M \geq 2, \quad \max_{\substack{0\leq n\leq N \\ 1\leq j\leq M}} \frac{|\varepsilon_j^n|}{h_t} \leq C(h_t^p + h_x^q). \tag{33}$$

**Remarque 10.** *Notons que la constante $C$ dans la définition 4 peut dépendre de la solution $u$ du problème exact.*

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

$$\lim_{\substack{(h_t,h_x)\in\mathcal{S} \\ (h_t,h_x)\to(0,0)}} \max_{\substack{0\leq n\leq N \\ 1\leq j\leq M}} |u(t_n, x_j) - u_j^n| = 0, \tag{34}$$

*avec $Nh_t = T$ et $Mh_x = 1$ où $u$ est la solution du problème exact.*

*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 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*

$$\forall (h_t, h_x) \in \mathcal{S}, \quad \max_{\substack{0\leq n\leq N \\ 1\leq j\leq M}} |u(t_n, x_j) - u_j^n| \leq C(h_t^p + h_x^q). \tag{35}$$

*On dit enfin que le schéma est convergent (resp. convergent à l'ordre $p$ en temps et $q$ en espace) 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 sous la condition* $\mathcal{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$ *et consistant (respectivement consistant d'ordre* $p$ *en temps et* $q$ *en espace), alors le schéma* (29) *est convergent (respectivement convergent d'ordre* $p$ *en temps et* $q$ *en espace) 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.*

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), les quantités $\max_{\substack{0\leq n\leq N\\ 1\leq j\leq M}} |u_j^n|$, $\max_{1\leq j\leq M} |(U_0)_j|$ et $\max_{\substack{0\leq n\leq N\\ 1\leq j\leq M}} |F_j^n|$ sont des normes portant sur $(u_j^n)$, $U_0$ et $(F^n)$. 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.

On pourrait considérer d'autres normes pour exprimer cette stabilité. Par exemple, on pourrait considérer la norme $\max_{0\leq n\leq N} \sqrt{h_x}(\sum_{0\leq j\leq M}(u_j^n)^2)^{1/2}$ pour $(u_j^n)$. 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 $h_x$.

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. 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* $(h_t, h_x) \to (0,0)$*. En général, l'analyse numérique n'est menée que dans 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 $\max_{\substack{0\le n\le N\\ 1\le j\le M}} |w_j^n| \le C_2 \max_{\substack{0\le n\le N\\ 1\le j\le M}} \dfrac{|\varepsilon_j^n|}{h_t}$. 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 $\|V\| = \max\limits_{1\le j\le M} |V_j|$. 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 si et seulement si il existe* $C > 0$ *dépendant uniquement de* $T$ *tel que*

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

**Corrolaire 1.** *Si* $|||A||| \le 1$*, alors le schéma est stable.*

*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 $\forall V \in \mathbb{R}^M, \quad \|A^n V\| \le C\|V\|$ 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

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

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

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

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, 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. Dit autrement, la stabilité du schéma 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 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.

**Remarque 13.** *On peut prouver que la stabilité de Von Neumann correspond à la stabilité du schéma dans le sens $\|(u_j^n)\|_* \leq C_1 \|U_0\|_\# + C_2 \|(F^n)\|_*$, avec $C_1$ et $C_2$ ne dépendant que de $T$ et des normes $\|\cdot\|_*$ et $\|\cdot\|_\#$ différentes de celles de la définition 2.*

### 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 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. 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.*

*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 :

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

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.*

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

*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 (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 est instable pour $a > 0$ et qu'il faut changer le sens du décentrage si $a < 0$. En fait, il faut décentrer en utilisant les données qui correspondent à de l'information qui arrive et non de l'information qui part.

De plus, nous avons vu dans la proposition 12 que le schéma (12) 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 Courant–Friedrichs–Lewy. De la même façon on appelle $c = \frac{ah_t}{h_x}$ le nombre de CFL. 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 19).

!! étude de l'influence du nombre de cfl (nouvelle section ? ?) !! – DONNER DES FIGURES + EXO : RETROUVER LES FIGURES !!

### 3.3.4 Détermination pratique des ordres de convergence

!! expliquer comment ca marche + figures + exo : retrouver les figures !!

### 3.3.5 Conditions de Dirichlet

!! on s'intéresse au pb (...) !!

En appliquant (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

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

avec $\mathrm{c} = \frac{ah_t}{h_x}$. Remarquons que la condition de Dirichlet a été prise en compte dans le terme "second membre" $F^n$.

!! comparer avec matrice des conditions périodiques !!

On obtient les résultats (...)

!! est-ce qu'on ne ferait pas mieux de faire l'analyse de stabilité avec des conditions périodiques ? ? – Ca simplierait les choses en particulier pour les schémas instables !!

!! remplacer les dérivées continues par des taux d'accroissement assure la consistance mais pas forcément la stabilité !! – C'est tres interessant de faire ca sur l'équation de transport par ce que l'information ne se déplace que dans une direction !!

!! faire des sous-sections dans cette section !!

!! prouver que le schéma en intro n'est pas stable !!

!! donner un schéma stable sous CFL et faire les preuves + simus num !!

!! introduire nombre de CFL + sens physique + simus avec différentes CFL !!

!! donner un schéma inconditionnellement stable !!

!! comparer les résultats avec ceux de l'intro !!

!! diffusion numérique : comparer avec l'équation de la chaleur !!

!! discrétisation implicite en exercice !!

!! dessin représentant l'endroit où on va chercher l'information : schéma centré / décentré + effet de la cfl !!

!! faut-il parler des propriétés de positivité ? ? – OUI ET LE RELIER AVEC LA NOTION DE STABILITE !!

!! exercice sur les ordre de convergence : figure avec echelle logarithmique !!

**Exercice 19** (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 $u(t_n, 0) = \alpha(t_n)$, 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 $V$. En déduire que $|||A||| \leq 1$ et que ce schéma est inconditionnellement stable.*
4. *Montrer que ce schéma est consistant d'ordre 1 en temps et en espace. Que peut-on dire de la convergence de ce schéma ?*
5. *Coder le schéma proposé. On peut soit inverser la matrice à l'extérieur de la boucle en temps, soit résoudre un système linéaire à chaque pas de temps. Comparer les deux méthodes, laquelle est la plus efficace ?*

   *!! Coder ca !!*

   *NB : commandes scilab*
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. Pour $(N, M) = (...)$, que remarquez-vous en $(t, x) = (0.1, 0.5)$ ? Si vous l'avez codé, comparez avec le schéma explicite. Comment expliquez-vous ceci ?*
7. *Déterminer numériquement le taux de convergence de ce schéma. On prendra par exemple la solution du cas test précédent à $T = 1$ pour estimer la convergence du schéma. Calculer l'erreur et la tracer sur une figure à l'échelle logarithmique comme sur la figure (...). Quels ordres de convergence trouvez-vous numériquement ?*

# 4 Équation de la chaleur

!! parler des $\theta$-schémas – voir poly L. !!

# 5 Équation des ondes

# 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