

# 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

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

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

$$ax^2 + bxy + cy^2 + dx + ey + f = A, \quad (2)$$

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.

## 1.1 Équations elliptiques

Nous présentons deux problèmes physiques qui font intervenir des équations elliptiques. 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



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

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

$$\frac{\partial q}{\partial x} = f. \quad (3)$$

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

$$q(x) = -k(x) \frac{\partial u}{\partial x}(x). \quad (4)$$

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

$$-\frac{\partial}{\partial x} \left( k \frac{\partial u}{\partial x} \right) = f. \quad (5)$$

Si l'on considère plusieurs dimensions d'espace, cette équation devient

$$-\nabla \cdot (k \nabla u) = f, \quad (6)$$

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

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



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  $f$ ).

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 (3).

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 (4) 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 (4), 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  $\delta x + u(x + \delta x) - u(x) \simeq \left(1 + \frac{\partial u}{\partial x}\right) \delta x$  et

la dérivée partielle en  $x$  est donc bien une élongation par unité de longueur.

D'autres problèmes physiques peuvent être 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 classes de problèmes exposées ici correspondent à des équations de nature différente. La première classe de problème est une équation parabolique en instationnaire, tandis que la deuxième classe est une équation hyperbolique.

Évoquons maintenant les conditions frontières 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 la 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 quoiqu'il arrive.

Une autre condition frontière 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 : quoiqu'il arrive aucun flux de chaleur ne traversera cette frontière (pensez par exemple à une bouteille isotherme).

Ces deux conditions frontières peuvent éventuellement être utilisées simultanément en des points différents du domaine (Dirichlet sur une partie de la frontière et Neumann sur le reste). On parle alors de conditions frontières 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 (6) est elliptique.

**Remarque 1.** Dans la plupart des applications,  $k$  est une constante. Prenons par exemple  $k = 1$ . L'équation (6) devient  $\Delta u = f$ . On appelle cette équation équation de Laplace ou équation de Poisson.

Pour finir cette section, nous illustrons le comportement de la solution de (5) en fonction de  $k$  (voir figure 3). Comme nous l'avons évoqué précédemment, une solution avec un  $k$  plus grand est plus "plate".

**Remarque 2.** 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.

A venir A venir A venir

FIGURE 3 – Solution de l'équation elliptique 1D (5) pour différents  $k$ .

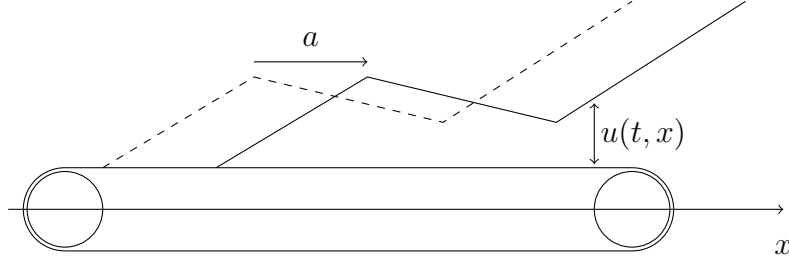


FIGURE 4 – Des tas de sable transportés sur un tapis roulant. La ligne discontinue représente les tas en  $t = 0s$  et la ligne continue en  $t = 1s$ .

## 1.2 Équation de transport

La deuxième équation que nous étudions est l'équation de transport (unidimensionnelle). 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  $\delta t$ , 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

$$\frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0. \quad (7)$$

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.

A venir A venir A venir

FIGURE 5 – Solution de l'équation de transport pour différents  $a$ .

### 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 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  $c(x)\delta x \frac{\partial u}{\partial t}(t, x) = f(t, x)\delta x + q(t, x) - q(t, x + \delta x)$ , ce qui donne

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

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 (4)). 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 (4) et (8), on obtient l'équation de la chaleur

$$\frac{\partial u}{\partial t} - \nu \frac{\partial^2 u}{\partial x^2} = f, \quad (9)$$

où  $\nu = k/c > 0$  est la diffusion thermique.

De manière similaire à ce qui a été expliqué dans la section 1.1, cette l'équation de la chaleur, en plus de représenter l'évolution de la température dans un milieu peut aussi représenter l'évolution de la densité d'une population ou encore l'évolution d'une concentration chimique d'une espèce au sein d'un diluant.

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

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'aplatir rapidement.

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



FIGURE 6 – Solution de l'équation de la chaleur pour différents  $\nu$ .

## 1.4 Équation des ondes

L'équation des ondes est 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,  $\rho(x)\delta x \frac{\partial^2 u}{\partial t^2} = f(t, x)\delta x + q(t, x) - q(t, x + \delta x)$ , ce qui nous donne l'équation

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

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 (4). 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 (4) et (10) on obtient l'équation des ondes

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = f, \quad (11)$$

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 différents  $c$ .

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.

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 frontières utilisables sont les mêmes que celles du cas stationnaire (voir tableau 1).

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

## 2 Équations elliptiques

Comme évoqué précédemment, on s'intéresse au problème

$$\begin{cases} -\nabla \cdot (k \nabla u) = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \quad (12)$$

avec  $0 < k_0 \leq k(x) \leq k_1$ .

### 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  $u \in C^{\ell+2}(\Omega)$ .*

!! commenter + donner lien vers preuve!!

**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).*

!! commenter + ref livre!!

**Remarque 4.** *D'autres conditions frontière sont possibles. Condition de Neumann, de Robin, conditions mixtes, conditions de Dirichlet non homogènes.*

!! expliquer le principe d'un relèvement!!

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

!! lien avec les fonctions holomorphes!!

### 2.2 Formulation variationnelle, théorème de Lax–Milgram

Le but de cette section est de présenter une preuve du théorème 1 en utilisant 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 :

$$\text{Trouver } u \in V, \text{ tel que } \forall v \in V, a(u, v) = \ell(v). \quad (13)$$

Ceci signifie que le problème (13) admet une unique solution  $u \in V$  et que celle-ci vérifie  $\|u\|_V \leq \frac{C_\ell}{\alpha}$ .

!! lien vers un livre !!

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

$$H_0^1(\Omega) := \{v \in L^2(\Omega) \mid \nabla v \in L^2(\Omega) \text{ et } v|_{\partial\Omega} = 0 \text{ sur } \partial\Omega\}, \quad (14)$$

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

$$\int_{\Omega} -\nabla \cdot (k \nabla u) v \, dx = \int_{\Omega} f v \, dx.$$

En intégrant par parties on obtient

$$\int_{\Omega} k \nabla u : \nabla v \, dx - \int_{\partial\Omega} k (\nabla u \cdot n) v \, ds = \int_{\Omega} f v \, dx.$$

Puisque  $v|_{\partial\Omega} = 0$ , le deuxième terme de cette expression est nul. Nous avons donc établi la proposition suivante (la réciproque utilise des arguments similaires).

**Proposition 2.** *Le problème (12) est équivalent au problème suivant.*

$$\text{Trouver } u \in H_0^1(\Omega), \text{ tel que } \forall v \in H_0^1(\Omega), \int_{\Omega} k \nabla v : \nabla v \, dx = \int_{\Omega} f v \, dx. \quad (15)$$

**Proposition 3.** *Le problème (15) est bien posé.*

*Démonstration.* Le problème (15) correspond au problème (13) avec  $V = H_0^1(\Omega)$ ,  $\forall v, w \in H_0^1(\Omega), a(v, w) = \int_{\Omega} k \nabla v : \nabla w \, dx$  et  $\ell(v) = \int_{\Omega} f v \, dx$ . 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  $\|v\|_{H_0^1(\Omega)} := (\int_{\Omega} v^2 + \nabla v : \nabla v \, dx)^{1/2}$  est un espace de Hilbert (se reporter à un cours sur les distributions). D'après l'inégalité de Cauchy–Schwarz, on a  $|\ell(v)| \leq (\int_{\Omega} f^2 \, dx)^{1/2} (\int_{\Omega} v^2 \, dx)^{1/2} \leq (\int_{\Omega} f^2 \, dx)^{1/2} \|v\|_{H_0^1(\Omega)}$ . 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  $\forall v \in H_0^1(\Omega), \int_{\Omega} v^2 \, dx \leq c \int_{\Omega} \nabla v : \nabla v \, dx$ . Avec cette inégalité, on peut prouver que  $a$  est coercive. Toutes les hypothèses du théorème de Lax–Milgram sont réunies, le problème (15) est donc bien posé.  $\square$

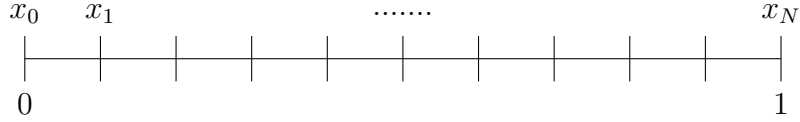


FIGURE 8 – Représentation de la discrétisation de l'intervalle  $(0, 1)$ .

## 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  $\Omega = (0, 1)$  et  $k = 1$ . Notre problème est donc

$$\begin{cases} -\frac{d^2u}{dx^2} = f \text{ dans } (0, 1), \\ u(0) = u(1) = 0. \end{cases} \quad (16)$$

Nous discrétisons l'intervalle  $(0, 1)$  en  $N > 0$  sous-intervalles. On définit donc les points  $x_j = jh$  ( $0 \leq j \leq N$ ) où  $h = 1/N$ . On a bien  $x_0 = 0$  et  $x_N = 1$  (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 dérivées discrètes. Pour cela, on rappelle qu'une dérivée est la limite d'un taux d'accroissement. Par exemple, on peut montrer que  $\frac{d^2u}{dx^2}(x) = \lim_{h \rightarrow 0} \frac{u(x+h) - 2u(x) + u(x-h))}{h^2}$ . Avec les notations introduites précédemment, cela signifie que, pour  $h$  suffisamment petit,

$$\frac{d^2u}{dx^2}(x_j) \simeq \frac{u(x_{j+1}) - 2u(x_j) + u(x_{j-1}))}{h^2}. \quad (17)$$

La méthode des différences finies consiste donc à définir une suite  $(u_j)_{0 \leq j \leq N}$  qui reprenne les éléments du problème (16). En particulier, il faut remplacer les dérivées continues par des dérivées discrètes. Il existe plusieurs façons de définir la suite  $(u_j)_{0 \leq j \leq N}$  et toutes n'ont pas les mêmes propriétés. La propriété la plus importante est la convergence : on veut que  $\lim_{N \rightarrow +\infty} \max_{0 \leq j \leq N} |u(x_j) - u_j| = 0$ . Nous aborderons plus en détails ces aspects dans la section 3.3.

Nous proposons maintenant de discrétiser le problème (16) en utilisant (17). La suite ainsi obtenue vérifie

$$\begin{cases} \forall 1 \leq j \leq N-1, -\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f(x_j), \\ u_0 = u_N = 0. \end{cases} \quad (18)$$

**Proposition 4.** *La relation (18) définit une unique suite  $(u_j)_{0 \leq j \leq N}$  dont les valeurs peuvent*

être déterminées en résolvant le système linéaire  $AU = F$ , avec

$$A = \frac{1}{h^2} \begin{pmatrix} 2 & -1 & 0 & \cdots & \\ -1 & 2 & -1 & 0 & \cdots \\ 0 & -1 & 2 & \ddots & \\ \vdots & & \ddots & \ddots & \ddots \\ & & & \ddots & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}, \quad U = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N-1} \end{pmatrix}, \quad F = \begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \end{pmatrix}. \quad (19)$$

*Démonstration.* La relation (18) est équivalente au système linéaire  $AU = F$ . Pour prouver cela, on écrit les équations discrètes

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

En prenant en compte  $u_0 = u_N = 0$ , 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.  $\square$

**Remarque 5.** 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.

!! Ajouter resultats de simulations!!

!! Ajouter exercice avec le problème de Neumann!!

!! La preuve des caracteristiques de la matrice  $A$  en exercice??

!! trouver une solution pour prouver la convergence de ce schéma sans introduire stabilité et consistance!!

!! exercice : prendre  $k$  constant et faire des simus avec — retrouver les résultats de simu de la section 1!!

### 3 Équation de transport

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

considérons le problème

$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \text{ dans } (0, T) \times (0, 1), \\ \forall t \in (0, T), u(t, 0) = \alpha(t), \\ \forall x \in (0, 1), u(0, x) = u_0(x), \end{cases} \quad (20)$$

où  $a \geq 0$  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.

### 3.1 Propriétés générales

Le problème (20) 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.** *Le problème (20) admet une unique solution donnée par (...)*

**Remarque 6.** *On n'a besoin d'une donnée de Dirichlet que d'un seul côté du domaine. Puisque l'on a choisi  $a \geq 0$ , il faut fixer la donnée de Dirichlet à gauche. On aurait aussi pu prendre  $a \leq 0$  et utiliser la condition de Dirichlet  $u(t, 1) = \alpha(t)$ .*

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, on peut reconstituer la solution initiale. Dans le cas présent, en utilisant la solution au temps final et la solution en sortie du domaine ( $x = 1$ ), on peut reconstituer la valeur de la solution à l'instant initial ainsi que la valeur de la donnée de Dirichlet  $\alpha(t)$  en tout temps  $t \in [0, T]$ .

**Proposition 6** (Réversibilité de l'équation de transport). (...)

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  $t = ax + b$  avec  $b \in \mathbb{R}$ . Ces droites sont donc les droites caractéristiques de notre problème, nous les représentons sur la figure 9.

!! equation des droites a montrer en exo!!

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 9 par un cercle). Nous retrouvons donc l'expression de la solution donnée en (...).

!! preuve de l'unicité??

Pour montrer la réversibilité de l'équation, il suffit de parcourir ces droites caractéristiques dans l'autre sens (...)

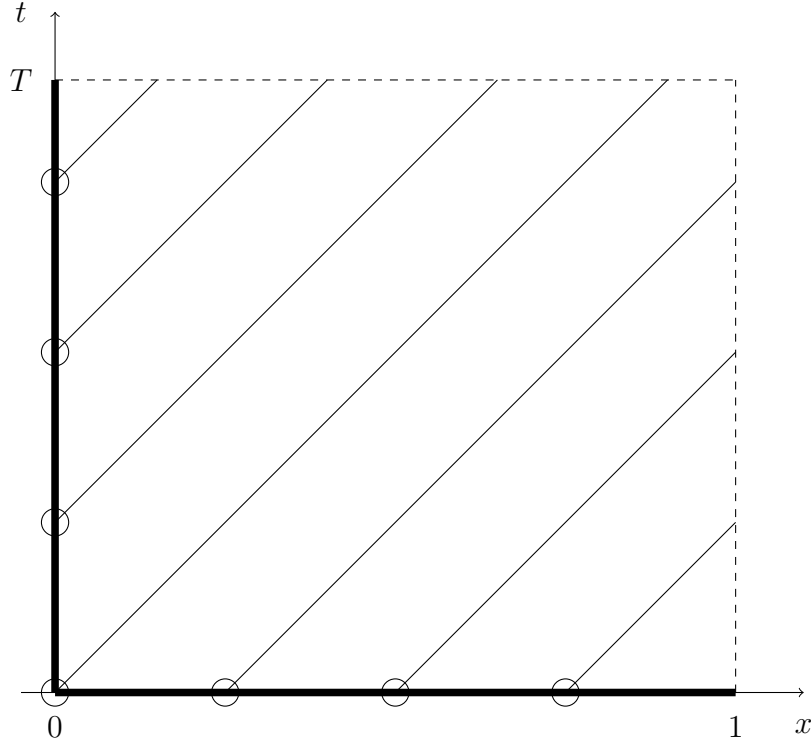


FIGURE 9 – Les droites caractéristiques de l'équation de transport ( $a = 1$  et  $T = 0.9$ ).

### 3.3 Discrétisation par les différences finies et analyse numérique

Nous cherchons maintenant à discrétiser le problème (20) 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 les 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,  $u_j^n$  devra être une approximation de  $u(t_n, x_j)$ .

#### 3.3.1 Conditions périodiques et motivation de l'analyse

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

$$\begin{cases} \frac{\partial u}{\partial t} + a \frac{\partial u}{\partial x} = 0 \text{ dans } (0, T) \times (0, 1), \\ \forall t \in (0, T), u(t, 0) = u(t, 1), \\ \forall x \in (0, 1), u(0, x) = u_0(x), \end{cases} \quad (21)$$

avec  $a > 0$  et  $u_0$  vérifie les conditions périodiques  $u_0(0) = u_0(1)$ . Avec ces conditions frontières, l'information qui sort en  $x = 1$  réentre immédiatement en  $x = 0$ .

En appliquant la méthode des caractéristiques, on peut montrer que la solution exacte de ce problème est

$$u(t, x) = u_0(f(x - at)), \quad (22)$$

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  $0 \leq f(x) < 1$ .

!! Calcul de la solution du pb périodique avec méthode des caractéristiques en exo??

Nous souhaitons maintenant discrétiser l'équation (21) et comparer la solution obtenue numériquement avec (22). La méthode consiste à approcher les dérivées partielles (en temps et en espace) par des taux d'accroissement. Nous proposons de comparer le comportement de deux discrétisations différentes. Nous proposons tout d'abord d'approcher la dérivée en temps par une approximation décentrée et la dérivée en espace par une approximation centrée comme suit

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

On obtient le schéma numérique

$$\begin{cases} \forall 0 \leq j \leq M, u_j^0 = u_0(x_j), \\ \forall 1 \leq n \leq N, u_0^n = u_M^n, \\ \forall 1 \leq n \leq N, \forall 0 \leq j \leq M, u_j^n = u_j^{n-1} - \frac{ah_t}{2h_x}(u_{j+1}^{n-1} - u_{j-1}^{n-1}), \end{cases} \quad (23)$$

où nous considérons  $u_0^n = u_M^n$ ,  $u_{M+1}^n = u_1^n$  et  $u_{-1}^n = u_{M-1}^n$ .

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

$$\frac{\partial u}{\partial t}(t_n, x_j) \simeq \frac{u(t_{n+1}, x_j) - u(t_n, x_j)}{h_t}, \quad \frac{\partial u}{\partial x}(t_n, x_j) \simeq \frac{u(t_n, x_j) - u(t_n, x_{j-1}))}{h_x}. \quad (24)$$

On obtient le schéma numérique

$$\begin{cases} \forall 0 \leq j \leq M, u_j^0 = u_0(x_j), \\ \forall 1 \leq n \leq N, u_0^n = u_M^n, \\ \forall 1 \leq n \leq N, \forall 0 \leq j \leq M, u_j^n = u_j^{n-1} - \frac{ah_t}{h_x}(u_j^{n-1} - u_{j-1}^{n-1}), \end{cases} \quad (25)$$

où nous considérons  $u_0^n = u_M^n$  et  $u_{-1}^n = u_{M-1}^n$ .

Les résultats numériques sont reportés dans la figure 10. Nous observons des comportements numériques très différents pour ces deux schémas qui sont pourtant similaires. Il ne suffit donc pas de remplacer les dérivées partielles par des taux d'accroissement pour obtenir des résultats convenables. Il faut aussi vérifier certaines propriétés que nous abordons dans la section suivante.

!! exercice : coder les deux schémas et retrouver les résultats de la figure 10!!

A venir A venir A venir

FIGURE 10 – Équation de transport avec conditions périodiques

### 3.3.2 Outils d'analyse numérique

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

$$U^0 = U_0 \quad \text{et} \quad \forall n \geq 0, \quad U^{n+1} = AU^n + h_t F^n, \quad (26)$$

où  $U_0$  est donné par  $(U_0)_j = u(t_0, x_j)$ . Notons également que nous considérons dans la suite  $U_0, F^n \in \mathbb{R}^{M+1}$  et  $A \in \mathbb{R}^{(M+1) \times (M+1)}$ . Si ces vecteurs et matrice ont une autre dimension, cela ne change pas la suite de notre propos.

**Remarque 7.** Dans le cas présent, nous considérons une suite dont la formule de récurrence ne nécessite que le pas de temps précédent. En pratique, certains schémas nécessitent plusieurs pas de temps précédents. Le schéma (26) peut être adapté à ce cas de figure. Nous verrons ceci plus loin dans ce cours.

**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 (26) est stable sous la condition  $\mathcal{S}$  s'il existe  $C_1, C_2 > 0$  qui ne dépendent que de  $T$  tels que  $\forall (h_x, h_t) \in \mathcal{S}, \forall U_0 \in \mathbb{R}^{M+1}, \forall 0 \leq n \leq N, \forall F^n \in \mathbb{R}^{M+1}$ , on a

$$\max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |u_j^n| \leq C_1 \max_{0 \leq j \leq M} |(U_0)_j| + C_2 \max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |F_j^n|, \quad (27)$$

où  $(u_j^n)$  est la suite générée par la récurrence (26).

On dit également que le schéma est inconditionnellement stable si la définition de stabilité conditionnelle s'applique avec  $\mathcal{S} = \mathbb{R}_+^* \times \mathbb{R}_+^*$ .

**Définition 3** (Erreur de troncature). L'erreur de troncature du schéma (26) au temps  $t_n$  est un vecteur  $\varepsilon^n \in \mathbb{R}^{M+1}$  défini par

$$\varepsilon^{n+1} := \tilde{u}^{n+1} - A\tilde{u}^n - h_t F^n, \quad (28)$$

où  $\tilde{u}^n \in \mathbb{R}^{M+1}$  est défini par  $(\tilde{u}^n)_j = u(t_n, x_j)$  avec  $u$  la solution du problème exact associé à (26).

**Définition 4** (Consistance). On dit que le schéma numérique (26) est consistant si pour toute solution régulière  $u$  du problème exact on a

$$\lim_{\substack{h_t \rightarrow 0 \\ h_x \rightarrow 0}} \max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} \frac{|\varepsilon_j^n|}{h_t} = 0, \quad (29)$$



où  $\varepsilon_j^n = (\varepsilon^n)_j$  l'erreur de troncature.

On dit de plus que le schéma (26) 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

$$\forall h_t, h_x > 0, \quad \max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} \frac{|\varepsilon_j^n|}{h_t} \leq C(h_t^p + h_x^q). \quad (30)$$

**Définition 5** (Convergence). On dit que le schéma numérique (26) converge (ou est convergent) sous la condition  $\mathcal{S} \subset \mathbb{R}_+^* \times \mathbb{R}_+^*$  si

$$\lim_{\substack{(h_t, h_x) \in \mathcal{S} \\ (h_t, h_x) \rightarrow (0,0)}} \max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |u(t_n, x_j) - u_j^n| = 0, \quad (31)$$

où  $u$  est la solution du problème exact.

De manière similaire, on dit que le schéma numérique (26) 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

$$\forall (h_t, h_x) \in \mathcal{S}, \quad \max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |u(t_n, x_j) - u_j^n| \leq C(h_t^p + h_x^q). \quad (32)$$

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 suivant.

**Théorème 4.** Si le schéma (26) 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 (26) est convergent (respectivement convergent d'ordre  $p$  en temps et  $q$  en espace) sous la condition  $\mathcal{S}$ .

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

Concernant la stabilité (cf définition 2), les quantités  $\max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |u_j^n|$ ,  $\max_{0 \leq j \leq M} |(U_0)_j|$  et  $\max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |F_j^n|$  sont des normes portant sur  $(u_j^n)$ ,  $U_0$  et  $(F^n)$ . On peut donner dire que la stabilité du schéma correspond à la continuité de l'application linéaire qui à  $U_0$  et  $(F^n)$  associe  $(u_j^n)$  en ajoutant que la constante associée à cette continuité ne doit pas dépendre de  $h_t$  et  $h_x$ . On pourrait considérer d'autres normes pour exprimer cette stabilité, il faudrait alors utiliser des poids adaptés en  $h_t$  et  $h_x$  pour avoir une définition équivalente (la constante  $C$  ne doit pas dépendre de  $h_t$  et  $h_x$ ). 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  $(u_j^n)$ . Pour ne pas avoir à considérer des poids en  $h_x$  et  $h_t$ , nous considérons les normes avec des "max". Cette continuité permet d'obtenir le fait qu'une "petite" modification de la donnée initiale ou des termes source ou frontière ne crée qu'une "petite" modification du résultat obtenu. 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 un tel exemple dans la suite de cette section.

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

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 lorsqu'on prend des pas de temps et d'espace de plus en plus petits. 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 les pas de temps et d'espace tendent vers 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.

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 soient 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 (on appelle ce résultat le théorème de Lax).

**Remarque 8.** *Les notions de consistance et de convergence sont différentes. Les jurys se plaignent des candidats qui confondent ces deux notions. Ne faites pas cette erreur.*

Nous allons maintenant prouver le théorème 4.

*Preuve du théorème 4.* Dans cette preuve, nous utilisons toutes les définitions précédentes. Tout d'abord, par définition de l'erreur de troncature,  $\tilde{u}^{n+1} = A\tilde{u}^n + h_t F^n + \varepsilon^{n+1}$ . On montre ainsi que  $w^n = \tilde{u}^n - U^n$  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  $\max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |w_j^n| \leq C_2 \max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} \frac{|\varepsilon_j^n|}{h_t}$ . On conclut en utilisant la définition de la consistance.  $\square$

Pour  $A \in \mathbb{R}^{(M+1) \times (M+1)}$ , on note

$$|||A||| = \sup_{V \in \mathbb{R}^{M+1} \setminus \{0\}} \frac{\|AV\|}{\|V\|},$$

où  $\|V\| = \max_{0 \leq j \leq M} |V_j|$ . Nous donnons maintenant une caractérisation de la stabilité du schéma.

**Proposition 7.** *Le schéma (26) est stable si et seulement si il existe  $C > 0$  dépendant uniquement de  $T$  tel que*

$$\forall n \in \mathbb{N}, \quad |||A^n||| \leq C. \quad (33)$$

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

*Preuve de la proposition 7.* Supposons tout d'abord que le schéma est stable. Dans ce cas, appliquons le schéma à  $U_0 = V \in \mathbb{R}^{M+1}$  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  $\forall V \in \mathbb{R}^{M+1}, \|A^n V\| \leq C \|V\|$  et donc (33).

Maintenant supposons que (33) est établie. Appliquons le schéma à  $U_0 \in \mathbb{R}^{M+1}$  et  $\forall n \in \mathbb{N}, F^n \in \mathbb{R}^{M+1}$ . Nous pouvons prouver que la solution obtenue est

$$U^n = A^n U_0 + h_t \sum_{\ell=0}^{n-1} A^{n-\ell-1} F^\ell. \quad (34)$$

Ainsi,  $\|U^n\| \leq \|A^n\| \|U_0\| + h_t \sum_{\ell=0}^{n-1} \|A^{n-\ell-1}\| \|F^\ell\|$ . En utilisant (33), on obtient  $\|U^n\| \leq C \|U_0\| + C h_t n \max_{0 \leq \ell \leq N} \|F^\ell\|$ . De plus,  $n h_t = t_n \leq T$ . On a donc établi la stabilité du schéma.  $\square$

!! Exercice : établir (34)!!

### 3.3.3 Analyse des deux schémas pour conditions périodiques

Nous allons montrer dans cette section que le schéma numérique proposé en (23) n'est pas stable contrairement au schéma (25).

On considère comme degrés de liberté  $U^n = \begin{pmatrix} u_0^n \\ \vdots \\ u_{M-1}^n \end{pmatrix} \in \mathbb{R}^M$ . Les deux schémas précédents peuvent être écrits sous la forme matricielle (26) avec  $F^n = 0$  et pour matrices

$$A_C = \frac{1}{2} \begin{pmatrix} 2 & -c & 0 & \dots & 0 & c \\ c & 2 & -c & & & 0 \\ 0 & \ddots & \ddots & \ddots & & \\ \vdots & & \ddots & \ddots & \ddots & \\ 0 & & & \ddots & \ddots & -c \\ -c & 0 & & & c & 2 \end{pmatrix}, \quad (35)$$

pour le schéma centré (23) et

$$A_{AM} = \begin{pmatrix} 1-c & 0 & \dots & 0 & c \\ c & \ddots & & & 0 \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & c \\ & & & & 1-c \end{pmatrix}, \quad (36)$$

pour le schéma décentré amont (25). Les deux matrices précédentes ont été écrites avec  $c = \frac{a h_t}{h_x}$ . Notons que les deux matrices précédentes sont dans  $\mathbb{R}^{M \times M}$ .

**Proposition 8.** *Les schémas (23) et (25) sont tous les deux consistants. De plus, le schéma (23) est consistant d'ordre 1 en temps et 2 en espace ; le schéma (25) est consistant d'ordre 1 en temps et 1 en espace.*

*Démonstration.* Prouvons que le schéma (23) est consistant d'ordre 1 en temps et 2 en espace. La preuve pour le schéma décentré amont est laissée en exercice.

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 (21).

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 (21), nous avons prouvé le fait que le schéma (23) est consistant d'ordre 2 en espace et d'ordre 1 en temps.  $\square$

!! exercice : consistance du schéma (25)!!

**Proposition 9.** *Le schéma (25) est stable sous la condition  $c = \frac{ah_t}{h_x} \leq 1$ .*

*Démonstration.* Soit  $U_0 \in \mathbb{R}^M$ , on calcule  $U^1 = AU_0$ . Pour  $0 \leq j \leq M-1$ ,  $u_j^1 = (1-c)u_j^0 + cu_{j-1}^0$ . 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 corollaire 1.  $\square$

Lorsque l'on veut prouver qu'un schéma de la forme (26) est instable, la façon standard de procéder est d'utiliser la stabilité 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 carrée de l'unité et  $k \in \mathbb{Z}$ . On étudie ensuite l'effet qu'a l'application du schéma sur ce nombre complexe. On montre que  $\forall n \in \mathbb{N}, u_j^{n+1} = \mathcal{A}u_j^n$  avec  $\mathcal{A} \in \mathbb{C}$ .

**Définition 6** (Stabilité au sens de von Neumann). *On dit qu'un schéma (26) est stable au sens de von Neumann si  $|\mathcal{A}| \leq 1$ .*

**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  $u_j^0 = \Re(e^{2i\pi k x_j})$  alors  $\forall 0 \leq n \leq N, u_j^n = \Re(\mathcal{A}^n e^{2i\pi k x_j})$ . Idem, si  $u_j^0 = \Im(e^{2i\pi k x_j})$ , alors  $\forall 0 \leq n \leq N, u_j^n = \Im(\mathcal{A}^n e^{2i\pi k x_j})$ .

Ainsi, si le schéma (26) est stable alors  $\max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |\Re(\mathcal{A}^n e^{2i\pi k x_j})| \leq C$  où  $C > 0$  ne dépend que de  $T$ . En faisant le même raisonnement sur la partie imaginaire, on obtient  $\max_{\substack{0 \leq n \leq N \\ 0 \leq j \leq M}} |\Im(\mathcal{A}^n e^{2i\pi k x_j})| \leq \sqrt{2}C$ .

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

Appliquons ce critère pour prouver le résultat suivant.

**Proposition 11.** *Le schéma (25) est instable si  $c = \frac{ah_t}{h_x} > 1$ .*

*Démonstration.* Si  $u_j^n = b_n(k)e^{2i\pi k x_j}$ , alors

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

Ici, nous avons  $u_j^{n+1} = \mathcal{A}u_j^n$  avec  $\mathcal{A} = 1 - c + ce^{-2i\pi k h_x}$ . Pour  $c > 1$ , on a  $|\mathcal{A}| > 1$  (si  $\cos(2\pi k h_x) < 1$ ). Le schéma n'est pas stable au sens de von Neumann donc il n'est pas stable au sens de la définition 2.  $\square$

**Remarque 9.** *On peut même prouver l'équivalence des deux notions de stabilité en utilisant le fait que les exponentielles complexes constituent une base Hilbertienne de l'espace des fonctions de carré intégrable sur  $(0, 1)$ .*

!! exercice le schéma (23) est instable.!!

!! expliquer qu'il faut aller chercher l'information qui vient pas celle qui part!!

!! exercice : consistance + stabilité d'un schéma implicite!!

### 3.3.4 Conditions de Dirichlet

!! on s'intéresse au pb (...)!!

En appliquant (24), on obtient le schéma numérique

$$\begin{cases} \forall 0 \leq j \leq M, u_j^0 = u_0(x_j), \\ \forall 1 \leq n \leq N, u_0^n = \alpha(t_n), \\ \forall 1 \leq n \leq N, \forall 1 \leq j \leq M, u_j^n = u_j^{n-1} - \frac{ah_t}{h_x}(u_j^{n-1} - u_{j-1}^{n-1}). \end{cases} \quad (37)$$

On a donc

$$A = \begin{pmatrix} 1-c & 0 & & & \\ c & 1-c & \ddots & & \\ 0 & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 0 \\ 0 & & & c & 1-c \end{pmatrix} \quad \text{et} \quad \forall n \geq 0, \quad h_t F^n = c \begin{pmatrix} \alpha(t_n) \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (38)$$

avec  $c = \frac{ah_t}{h_x}$ . 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 simplifierait 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 très intéressant de faire ça 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??

## 4 Équation de la chaleur

## 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 éléments 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 maîtrisées!!

### 6.1 Équation de transport dans un domaine multi-dimensionnel