Chapitre 3

Différences finies pour les problèmes d'évolution

3.1 Approximation de l'équation de la chaleur

On considère le problème suivant :

$$\begin{cases} \partial_t u - \mu \partial_{xx} u = f(x,t) & x \in]0,1[,\ 0 < t \leq T, \\ u(0,t) = 0, \quad u(1,t) = 0, & 0 \leq t \leq T, \\ u(x,0) = u_0(x), & x \in]0,1[, \end{cases} \tag{3.1}$$

c'est l'équation de la chaleur, ou équation de diffusion. La fonction f et la condition initiale u_0 sont données et $\mu > 0$ est une constante donnée. On suppose que la donnée initiale et les données au bord sont compatibles, c'est à dire que la donnée initiale vérifie les conditions aux limites $u_0(0) = u_0(1) = 0$. On ne cherche pas ici à étudier cette edp dans le cadre général (avec des données générales, par exemple des conditions aux limites non constantes, i.e. dépendant du temps, $u(0,t) = \alpha(t), u(1,t) = \beta(t)$) mais à comprendre comment construire un schéma aux différences finies pour en approcher la solution. On va se placer dans un cas particulier où il est assez simple de construire une solution. Pour pouvoir faire l'analyse numérique d'une méthode, il est en effet utile de savoir que la solution qu'elle est censée approcher existe dans un certain espace, et est unique. Dans le cas $f = 0$, on peut effectivement monter le résultat d'existence et d'unicité suivant (qui utilise un développement en série de Fourier).

Proposition 3.1.1 *Soit $u_0 \in \mathcal{C}^2([0,1])$ vérifiant $u_0(0) = u_0(1) = 0$. Le problème (3.1) avec $f = 0$ a une solution $u \in \mathcal{C}^0([0,1]\times[0,T]) \cap \mathcal{C}^1(]0,1[\times]0,T])$, $\partial_{xx} u \in \mathcal{C}^0([0,1]\times]0,T])$ et une soution ayant une telle régularité est unique.*

Pour l'approximation par différences finies de la solution de (3.1), en plus de la grille uniforme en espace, on introduit une grille en temps : on se donne un entier $M > 0$ et on pose $\Delta t = \frac{T}{M}$, Δt est le pas de temps, et $t_n = n\Delta t$, $0 \leq n \leq M$. Par analogier avec Δt, on utilisera la notation

$$\Delta x = h = \frac{1}{N+1}, \quad x_i = ih = i\Delta x, \quad i = 0, \ldots, N+1.$$

Les points de la grille espace-temps sont donc les points (x_j, t_n), $0 \leq j \leq N+1, 0 \leq n \leq M$. On cherche à calculer des valeurs u_j^n qui approchent les valeurs exactes $u(x_j, t_n)$ en ces points

de grille grâce à un schéma aux différences finies. On garde la même discrétisation pour le "Laplacien" $\partial_{xx}u$, mais suivant la formule utilisée pour approcher la dérivée en temps, on obtient des schémas différents. Le premier schéma est le suivant :

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{\Delta t} - \mu \dfrac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{\Delta x^2} = f_j^n, \qquad 1 \le j \le N, \quad 0 \le n \le M, \\ u_0^n = 0, \quad u_{N+1}^n = 0, \qquad 0 \le n \le M, \\ \\ u_j^0 = u_0(x_j), \qquad 0 \le j \le N+1, \end{cases} \tag{3.2}$$

où $f_j^n = f(x_j,t_n)$. Il est obtenu en prenant l'équation exacte en (x_j,t_n), en discrétisant le Laplacien par différences finies comme on l'a vu au début de ce chapitre et en remplaçant la dérivée en temps par une différence finie progressive

$$\partial_t u(x_j,t_n) \sim \frac{u(x_j,t_{n+1}) - u(x_j,t_n)}{\Delta t}.$$

C'est un schéma explicite en temps : si on pose

$$r = \mu \frac{\Delta t}{\Delta x^2}, \tag{3.3}$$

l'équation aux différences du schéma (3.2) s'écrit

$$u_j^{n+1} = (1-2r)u_j^n + r(u_{j+1}^n + u_{j-1}^n) + \Delta t f_j^n, \tag{3.4}$$

qui est une formule *explicite* permettant de calculer u_j^{n+1} à partir des valeurs $u_{j+1}^n, u_j^n, u_{j-1}^n$ connues au temps t_n, sans nécessité d'inverser une fonction ou un système.

Posons pour $0 \le n \le M$, $U_h^n = (u_1^n, u_2^n, ..., u_N^n)^T \in \mathbb{R}^N$, $F_h^n = (f_1^n, f_2^n, ..., f_N^n)^T \in \mathbb{R}^N$ et définissons la matrice tridiagonale

$$Q_1 = Q_1(r) = \begin{pmatrix} 1-2r & r & 0 & \dots & \dots & 0 \\ r & 1-2r & r & 0 & & \vdots \\ 0 & r & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & r & 1-2r & r \\ 0 & \dots & \dots & 0 & r & 1-2r \end{pmatrix}$$

la formule (3.4) écrite pour $j = 1, ..., N$ donne la relation

$$U_h^{n+1} = Q_1 U_h^n + \Delta t F_h^n,$$

et permet le calcul des vecteurs U_h^n pour $0 < n \le M$ à partir d'une donnée initiale qu'on a prise exacte $U_h^0 = \overline{U}_h^0$ vecteur de $\mathbb{R}^N$ de composantes $(u_0(x_j))_{j=1,...,N}$.

Si au contraire, on considère une différence finie régressive (qu'on écrit alors au temps t_{n+1})

$$\partial_t u(x_j,t_{n+1}) \sim \frac{u(x_j,t_{n+1}) - u(x_j,t_n)}{\Delta t},$$

en prenant l'équation exacte en (x_j, t_{n+1}), on obtient le deuxième schéma

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{\Delta t} - \mu \dfrac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{\Delta x^2} = f_j^{n+1}, \quad 1 \le j \le N, \qquad 0 \le n \le M, \\ u_0^n = 0, \quad u_{N+1}^n = 0, \qquad 0 \le n \le M \\ \\ u_j^0 = u_0(x_j), \qquad 0 \le j \le N+1, \end{cases} \tag{3.5}$$

qui, lui, est un schéma *implicite*. On peut écrire l'équation aux différences du schéma (3.5) avec la notation (3.3) sous la forme

$$(1+2r)u_j^{n+1} - r(u_{j+1}^{n+1} + u_{j-1}^{n+1}) = u_j^n + \Delta t f_j^{n+1}, 1 \le j \le N, \tag{3.6}$$

et on a un système de matrice tridiagonale à résoudre pour calculer les N composantes du vecteur U_h^{n+1}. On définit la matrice $N \times N$

$$Q_2 = Q_2(r) = \begin{pmatrix} 1+2r & -r & 0 & \dots & \dots & 0 \\ -r & 1+2r & -r & 0 & & \vdots \\ 0 & r & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & -r & 1+2r & -r \\ 0 & \dots & \dots & 0 & -r & 1+2r \end{pmatrix}.$$

La formule (3.6) écrite pour $j = 1, ..., N$ donne le système linéaire

$$Q_2(r) U_h^{n+1} = U_h^n + \Delta t F_h^{n+1}.$$

On peut prendre une différence finie centrée,

$$\partial_t u(x_j, t_{n+1/2}) \sim \frac{u(x_j, t_{n+1}) - u(x_j, t_n)}{\Delta t},$$

on prend alors l'équation exacte en $(x_j, t_{n+1/2})$ puis on approche les valeurs $u(x_j, t_{n+1/2})$ dans

$$\partial_{xx} u(x_j, t_{n+1/2}) \sim \frac{u(x_{j+1}, t_{n+1/2}) - 2u(x_j, t_{n+1/2}) + u(x_{j-1}, t_{n+1/2})}{\Delta x^2}$$

par des moyennes $u(x_j, t_{n+1/2}) \sim \frac{1}{2}(u(x(j, t_n) + u(x_j, t_{n+1}))$ et on obtient un troisième schéma

$$\begin{cases} \dfrac{u_j^{n+1}-u_j^n}{\Delta t} - \mu \dfrac{(u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}) + (u_{j+1}^n - 2u_j^n + u_{j-1}^n)}{2\Delta x^2} = f_j^{n+1/2}, \quad 1 \le j \le N, \quad 0 \le n \le M \\ u_0^n = 0, \qquad u_{N+1}^n = 0, \quad 0 \le n \le M \\ \\ u_j^0 = u_0(x_j), \qquad 0 \le j \le N+1, \end{cases} \tag{3.7}$$

où $f_j^{n+1/2} = f(x_j, (n+1/2)\Delta t)$ ou on pourrait aussi prendre l'approximation $\frac{1}{2}(f_j^n + f_j^{n+1})$. Ce schéma (appelé schéma de Crank-Nicolson) est également implicite. En introduisant encore deux matrices $N \times N$ tridiagonales, disons Q_3, Q_4 (le faire), on obtient une relation de la forme

$$Q_3 U_h^{n+1} = Q_4 U_h^n + \Delta t F_h^{n+1/2}.$$

Remarque 3.1.1 *Dans les trois exemples, on vérifie que les schémas obtenus correspondent à avoir utilisé le schéma d'Euler (explicite ou implicite) ou celui de Crank-Nicolson pour l' EDO*

$$\begin{cases} \dfrac{dU_h}{dt}(t)+\mu A_{0h}U_h(t) = f_h(t) \quad 0 < t \leq T \\ \\ U_h(0) = \overline{U}_h^0 \end{cases} \tag{3.8}$$

où A_{0h} est définie par (2.7), les N composantes de $U_h(t)$ approchent les $u(x_i,t)$ et $\overline{U}_h^0$ est le vecteur de composantes $u_0(x_j)$. Ce système d'équations différentielles (3.8) est celui qui résulte de la semi-discrétisation en espace par la méthode des différences finies en espace de la section 2.1, le temps t étant considéré comme un "paramètre", on parle aussi de "méthode des lignes". On discrétise ensuite en temps, en calculant un vecteur U_h^n qui approche le vecteur $U_h(t_n)$.

Remarque 3.1.2 *On pourrait aussi commencer par semi-discrétiser le problème en temps, x étant considéré comme un "paramètre". Par exemple si on utilise la méthode d'Euler implicite, on obtient*

$$\begin{cases} \dfrac{u^{n+1}(x)-u^n(x)}{\Delta t} - \mu \dfrac{d^2u^{n+1}}{dx^2}(x) = f^{n+1}(x) \quad x \in]0,1[, \, 0 \leq n \leq M-1 \\ u^n(0) = u^n(1) = 0, \quad 0 \leq n \leq M \\ u^0(x) = u_0(x), \quad x \in]0,1[, \end{cases} \tag{3.9}$$

où $f^n(x) = f(x,t_n)$ et la fonction $u^n(x)$ approche $u(x,t_n)$. On utilise ensuite la méthode des différences finies en espace pour (3.9), elle consiste à prendre l'équation aux points x_j, en remplaçant la dérivée exacte par une différence finie comme à la section 2.1. On obtient alors le schéma 2 (3.5) où U_h^n approche le vecteur de coordonées $u^n(x_i)$.

Ces remarques vont aider à comprendre comment on pourra relier l'ordre en temps et en espace de la méthode résultant des deux discrétisations, aux ordres respectifs de chaque méthode de semi discrétisation. Cela permet aussi de comprendre comment on peut généraliser en utilisant d'autres méthodes de discrétisation en temps (par exemple une méthode de Runge-Kutta) ou en espace (par exemple une méthode d'éléments finis).

Pour les schémas 2 et 3, on vérifie que les matrices Q_2 et Q_3 sont des M-matrices et par conséquent inversibles. Les trois schémas précédents peuvent donc tous se mettre sous la forme générale

$$U_h^{n+1} = QU_h^n + \Delta t G_h^n, \; 0 \leq n \leq M-1, \tag{3.10}$$

où $Q = Q(r)$ est une matrice $N \times N$ connue, et G_h^n un vecteur de $\mathbb{R}^N$ connu à partir des données du problème, et $M\Delta t \leq T$. La condition initiale est $U_h^0 = u_h^0 = (u_0(x_1), \ldots, u_0(x_N))^T$. Il s'agit d'un schéma à deux niveaux en temps (seules les valeurs en t_n interviennent pour le calcul au temps t_{n+1}), on dit aussi schéma à un pas de temps.

Notons que si les conditions aux limites dans (3.1) ne sont pas 0 mais $u(0,t) = \alpha(t), u(1,t) = \beta(t)$, ces valeurs une fois discrétisées en $(\alpha(t_n), \beta(t_n))$ interviendront dans les composantes du terme au second membre G_h^n.

Introduisons pour ces schémas les notions de consistance, stabilité et convergence. Pour cela, on considère les valeurs exactes $u(x_j,t_n)$ aux points de la grille et le vecteur associé

$$\overline{U}_h^n = (u(x_1,t_n),u(x_2,t_n),...,u(x_N,t_n))^T \in \mathbb{R}^N.$$

Alors $\overline{U}_h^n$ ne vérifie pas le schéma (sinon on saurait calculer la solution exacte en n'importe quel point) mais on peut écrire

$$\overline{U}_h^{n+1} = Q\overline{U}_h^n + \Delta t G_h^n + \Delta t K_h^n,$$

ce qui définit le vecteur K_h^n, c'est l'*erreur de consistance* au temps t_n

$$K_h^n = \frac{1}{\Delta t}(\overline{U}_h^{n+1} - Q\overline{U}_h^n - \Delta t G_h^n).$$

Enfin, on considère une norme vectorielle $\| .\|$ sur $\mathbb{R}^N$ et la norme matricielle associée $\|.\|$.

Définition 3.1.1 *On dit que le schéma (3.10) est*
i) consistant avec l'edp (3.1) si pour toute solution u de (3.1)

$$\sup_{m;m\Delta t\leq T} \|K_h^m\| \to 0 \text{ quand } \Delta t,\Delta x \to 0.$$

ii) La méthode est d'ordre (p,q) (p en espace et q en temps) si, pour toute solution u suffisamment régulière de (3.1), il existe une constante $C = C(u) > 0$ telle que

$$\sup_{m;m\Delta t\leq T} \|K_h^m\| \leq C(\Delta x^p + \Delta t^q).$$

iii) Le schéma est stable (pour la norme $\| .\|$) s'il existe une constante C_0 (qui peut dépendre de T) telle que

$$\sup_{m;m\Delta t\leq T} \|Q^m\| \leq C_0.$$

iv) Le schéma est convergent si

$$\sup_{m;m\Delta t\leq T} \|\overline{U}_h^m - U_h^m\| \to 0 \text{ quand } \Delta t,\Delta x \to 0.$$

iv) Le schéma est convergent à l'ordre (p,q) (p en espace et q en temps) si, pour toute solution u suffisamment régulière de (3.1), il existe une constante $C = C(u) > 0$ telle que

$$\sup_{m;m\Delta t\leq T} \|\overline{U}_h^m - U_h^m\| \leq C(\Delta x^p + \Delta t^q).$$

Remarques. i) Attention aux *notations* : dans K_h^n, U_h^n, qui sont des vecteurs de $\mathbb{R}^N$, n est un indice de temps ; dans Q^m, où $Q = Q(r)$ est une matrice $N \times N$, pour m il s'agit de la puissance $Q^2 = QQ$, ... Si on utilisait une méthode à pas variable, dans laquelle $\Delta t_n \equiv t_{n+1} - t_n$ n'est pas nécessairement constant, la matrice Q dans (3.10) dépendrait de n, soit $Q^{(n)}$, et on aurait alors un produit $Q^{(m)}Q^{(m-1)}...Q^{(1)}$ à la place de la puissance Q^m.

ii) La régularité demandée pour la consistance est celle de la proposition 3.1.1 ; pour estimer l'ordre, on a besoin de supposer que la solution est plus régulière pour pouvoir faire des développements limités, voir la proposition 3.1.2 ci-dessous. Bien qu'on n'ait pas donné de résultat

général de régularité de la solution par rapport aux données (qui serait un analogue du théorème 1.4.3 dans le cas de l'équation de la chaleur), cette régularité peut effectivement être atteinte et découler d'hypothèses de régularité sur les données.

iii) Même si on ne l'a pas rappelé, la norme utilisée dans $\mathbb{R}^N$ doit être telle que les quantités continuent à avoir un sens quand $N \to \infty$, par exemple la norme ℓ^∞ ou la norme ℓ^2 discrète $\|\cdot\|_{2,\Delta}$ introduite dans la définition 2.2.3, qui s'écrit ici

$$\|U_h^j\|_{2,\Delta}^2 = h \sum_{j=1}^{N} |u_j^n|^2,$$

puisque les valeurs aux extrémités u_0^n et u_{N+1}^n sont nulles. Notons que le facteur h change la valeur de la norme mais pas la norme matricielle associée, et que la norme $\|\cdot\|_{2,\Delta}$ est majorée par la norme $\|\cdot\|_\infty$. □

Voici à présent un résultat fondamental, parfois appelé Théorème de Lax, qui relie les concepts de consistance, stabilité et convergence.

Théorème 3.1.1 *Si le schéma est consistant et stable, alors il est convergent.*

Démonstration. On introduit l'erreur $E_h^n = \overline{U}_h^n - U_h^n, 0 \le n \le M = T/\Delta t$. Alors $e_h^0 = 0$ par choix de la condition initiale, et des deux relations

$$\overline{U}_h^{n+1} = Q\overline{U}_h^n + \Delta t G_h^n + \Delta t K_h^n, \quad U_h^{n+1} = QU_h^n + \Delta t G_h^n$$

on obtient par différence, pour $n \le M-1$,

$$E_h^{n+1} = QE_h^n + \Delta t K_h^n,$$

et en itérant le processus, puisque $E_h^0 = 0$, on obtient

$$E_h^{n+1} = \Delta t \sum_{k=0}^{n} Q^k K_h^{n-k}.$$

On déduit

$$\|E_h^{n+1}\| \le \Delta t \sum_{k=0}^{n} \|Q^k\| \, \|K_h^{n-k}\|.$$

Si la méthode est stable

$$\|E_h^{n+1}\| \le C_0 (n+1)\Delta t \sup_{0\le k\le n} \|K_h^{n-k}\|,$$

et donc, comme cela est valable pour tout $n \le M-1$, avec $M\Delta t = T$

$$\sup_{m;m\Delta t\le T} \|E_h^m\| \le C_0 T \sup_{n;n\Delta t\le T} \|K_h^n\|,$$

et si la méthode est consistante, le second membre tend vers 0 quand $\Delta t, \Delta x \to 0$. □

Remarque 3.1.3 *i) En calquant la démonstration précédente on montrerait que si on a une perturbation P_h^0 sur la donnée initiale et que le schéma calcule une suite de valeurs $\tilde{u}_h^n$ solution d'un schéma perturbé*

$$\tilde{U}_h^{n+1} = Q\tilde{U}_h^n + \Delta t G_h^n + \Delta t P_h^n,$$

avec $\tilde{U}_h^0 = U_h^0 + P_h^0$, si le schéma est stable, la différence $\|\tilde{U}_h^n - U_h^n\|$ reste bornée en fonction des perturbations $\|P_h^0\|$ et $\max_{n\leq M}\|P_h^n\|$, *ce qui explique le terme de* stabilité.

ii) Le principe "stabilité + consistance implique convergence" est très général dans les méthodes de discrétisation.

Appliquons ce résultat aux schémas 1 et 2. Il faut vérifier la consistance et la stabilité ; commençons par étudier le **schéma 1** explicite. Pour ce schéma, l'erreur de consistance est donnée par l'équation

$$u(x_j,t_{n+1}) = (1-2r)u(x_j,t_n) + r(u(x_{j+1},t_n) + u(x_{j-1},t_n)) + \Delta t f_j^n + \Delta t \tau_j^n.$$

Proposition 3.1.2 *Supposons que la solution u du problème (3.1) vérifie : $u \in \mathcal{C}^0([0,1]\times[0,T]\cap\mathcal{C}^1(]0,1[\times]0,T])$, et $\frac{\partial^4 u}{\partial x^4}, \frac{\partial^2 u}{\partial t^2} \in \mathcal{C}^0([0,1]\times]0,T])$. Alors, pour le schéma (3.2), l'erreur de consistance satisfait*

$$\sup_{m;m\Delta t\leq T} \|K_h^m\|_\infty \leq C(\Delta x^2 + \Delta t)$$

où la constante C dépend de $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^4 u}{\partial x^4}|$, $\max_{x\in[0,1],t\in[0,T]} |\frac{\partial^2 u}{\partial t^2}|$.

Démonstration. L'erreur de consistance κ_j^n vérifie

$$\kappa_j^n = \frac{1}{\Delta t}\big(u(x_j,t_{n+1}) - u(x_j,t_n)\big) - \mu\frac{1}{\Delta x^2}\Big(u(x_{j+1},t_n) - 2u(x_j,t_n) + u(x_{j-1},t_n)\Big) - f(x_j,t_n).$$

On utilise de nouveau la formule de Taylor-Lagrange. Pour une fonction ϕ supposée de classe $\mathcal{C}^2$ sur $[0,T]$, on peut écrire : pour tout $n \in \{0,\dots,M-1\}$, il existe un nombre $\theta^{(n)} \in]0,1[$ tel que

$$\phi(t_{n+1}) = \phi(t_n) + \Delta t\phi'(t_n) + \frac{\Delta t^2}{2}\phi''(t_n + \theta^{(n)}\Delta t),$$

on applique le résultat à $\phi(t) = u(x_j,t)$, et alors $\phi'(t_n) = \partial_t u(x_j,t_n)$. De même on applique le théorème 2.1.1 à $\varphi(x) = u(x,t_n)$ et alors $\varphi''(x_j) = \partial_{xx}u(x_j,t_n)$. Comme u est solution de (3.1),

$$\partial_t u(x_j,t_n) - \mu\partial_{xx}u(x_j,t_n) = f(x_j,t_n)$$

donc

$$\kappa_j^n = \frac{\Delta t}{2}\frac{\partial^2 u}{\partial t^2}(x_j,t_n+\theta^{(n)}\Delta t) - \mu\frac{\Delta x^2}{12}\frac{\partial^4 u}{\partial x^4}(x_j+\theta_j\Delta_j,t_n),$$

pour des $\theta_j, \theta^{(n)} \in]-1,1[$, d'où le résultat. On a donc montré que le schéma est d'ordre 2 en espace et 1 en temps comme on s'y attendait puisqu'il résulte de l'utilisation d'un schéma de différences finies en espace d'ordre 2 (associé au Laplacien discret) et d'un schéma de différences finies en temps d'ordre 1 (la méthode d'Euler). □

On note que, puisque la norme $\|\cdot\|_\infty$ majore la norme $\|\cdot\|_{2,\Delta}$, on a aussi

$$\sup_{m;m\Delta t\le T}\|K_h^m\|_{2,\Delta}\le C(\Delta x^2+\Delta t)$$

Pour la stabilité, on a le résultat suivant.

Proposition 3.1.3 *Supposons $0<r\le 1/2$. Alors le schéma est stable pour la norme $\|.\|_\infty$ et pour $\|.\|_2$. Plus précisément $\|Q_1(r)\|_\infty=1$ et $\|Q_1(r)\|_2<1$.*

Démonstration. Considérons d'abord la norme $\|.\|_\infty$. Si $0<r\le 1/2$, alors les coefficients de $Q_1(r)$ sont positifs, et par le théorème 1.5.1, $\|Q_1(r)\|_\infty=1$, ce qui entraîne la stabilité pour cette norme. Pour la norme $\|.\|_2$, comme Q_1 est symétrique, le même théorème donne $\|Q_1\|_2=\max_j|\lambda_j(Q_1)|$. On vérifie que les vecteurs propres de Q_1 sont donnés par

$$s_k=\Big(\sin\Big(\frac{kj\pi}{(N+1)}\Big)\Big)_{j=1,\dots,N},\quad k=1,\dots,N,$$

qui ne sont rien d'autres que l'échantillonage aux points x_j des fonctions $x\mapsto\sin(k\pi x)$, c'est à dire les fonctions propre de l'opérateur $u\mapsto u''$ s'annulant aux bord de Ω. Les valeurs propres λ_k de Q_1 correspondantes sont données par

$$\lambda_k(Q_1)=1-4r\sin^2\Big(\frac{k\pi}{2(N+1)}\Big),\ 1\le k\le N.$$

Si $0<r\le 1/2$, on a $|\lambda_k(Q_1)|<1$ pour tout k, et la méthode est stable. □

La condition de stabilité $0<r\le 1/2$ s'exprime comme une contrainte sur le pas de temps

$$0<\Delta t\le\Delta x^2/2\mu. \tag{3.11}$$

Sous cette condition, la méthode est donc convergente et

$$\sup_{m;m\Delta t\le T}\|\overline{U}_h^m-U_h^m\|_p\le c(\Delta x^2+\Delta t),$$

pour $p=2$ et $p=\infty$, où c ne dépend que de u.

D'après l'estimation de l'erreur de consistance, on voit que la méthode est d'ordre 2 (en espace) et 1 (en temps). donc l'erreur est finalement en $O(\Delta x^2)$ et n'est pas détériorée par l'ordre 1 en temps (on raisonne avec $\mu>0$ fixé, qui n'est a priori pas très petit devant Δx). Cette estimation suppose cependant une condition qui limite le pas de temps, ici (3.11), on dit que la méthode est *conditionnellement stable*, c'est la conséquence de son caractère explicite.

Passons à l'étude du **Schéma 2** implicite. On écrit maintenant que la solution exacte vérifie

$$(1+2r)u(x_j,t_{n+1})-r(u(x_{j+1},t_{n+1})+u(x_{j-1},t_{n+1}))=u(x_j,t_n)+\Delta t f_j^{n+1}+\Delta t\overline{\kappa}_j^n,$$

ce qui définit l'erreur de consistance $\overline{\kappa}_j^n$ et on vérifie que l'on a encore

$$|\overline{\kappa}_j^n|\le C(\Delta x^2+\Delta t),$$

où la constante C dépend de $\max_{x\in[0,1],t\in[0,T]}|\frac{\partial^4 u}{\partial x^4}|$, $\max_{x\in[0,1],t\in[0,T]}|\frac{\partial^2 u}{\partial t^2}|$. Si on définit alors la matrice Q par $Q=Q_2^{-1}(r)$, et si on pose $K_h^n=Q\overline{K}_h^n$ et $G_h^n=QF_h^{n+1}$, le schéma peut se mettre sous la forme générale (3.10).

La stabilité résulte de la proposition suivante.

Proposition 3.1.4 *Le schéma (3.5) est stable pour la norme* $\|.\|_\infty$ *et pour* $\|.\|_2$. *Plus précisément* $\|Q\|_\infty \leq 1$ *et* $\|Q\|_2 < 1$.

Démonstration. Considérons d'abord la norme $\|.\|_\infty$. On voit que Q_2 est une M-matrice et que la somme sur chaque ligne est supérieure à 1. D'après la Proposition 2.2.2, on a donc

$$\|Q\|_\infty = \|Q_2^{-1}\|_\infty \leq 1.$$

ce qui implique la stabilité pour cette norme. Pour la norme $\|.\|_2$, on trouve que Q_2 admet les même vecteurs propres que Q_1, et on peut aussi calculer ses valeurs propres

$$\lambda_k(Q_2) = 1 + 4r\sin^2(\frac{k\pi}{2(N+1)}),\ 1 \leq k \leq N.$$

Puisque Q_2 est symmétrique, il en est de même pour Q et on a donc $\|Q\|_2 = \max_j |\lambda_k(Q)|$, or les valeurs propres de Q sont les inverses des $\lambda_k(Q_2)$, donc les valeurs

$$\lambda_k(Q) = (1 + 4r\sin^2(\frac{j\pi}{2(N+1)}))^{-1},\ 1 \leq k \leq N.$$

On a ainsi $|\lambda_k(Q)| < 1$, donc $\|Q\|_2 \leq 1$ et la méthode est stable pour cette norme. □

Le schéma 2 est donc inconditionnellement stable, il est aussi consistant, donc il est convergent.

3.2 Approximation de l'équation de transport

On s'intéresse à présent à l'équation de transport appelée aussi équation d'advection ou équation de convection

$$\frac{\partial u}{\partial t} + a\frac{\partial u}{\partial x} = 0, \quad x \in]0,1[, \quad t > 0, \tag{E}$$

avec diverses conditions aux limites possibles suivant le signe de a comme on l'a vu dans le Chapitre 1.

On note Δx le pas d'espace et Δt le pas de temps, $x_j = j\Delta x$ et $t_n = n\Delta t$ avec $0 \leq j \leq N+1$ et $0 \leq n \leq M = T/\Delta t$. On pose

$$\lambda = \frac{\Delta t}{\Delta x}.$$

On note u_j^n l'approximation de $u(x_j, t_n)$ que l'on veut calculer, et comme pour l'équation de la chaleur on note U_h^n le vecteur qui contient ces valeurs. Dans le cas de conditions périodiques on prendra plus précisément $U_h^n = (u_0^n, \ldots, u_N^n)^T$ puisque la valeur u_{N+1}^n est imposée égale à u_0^n. De même, on prendra $U_h^n = (u_1^n, \ldots, u_{N+1}^n)^T$ dans le cas de conditions aux limites à gauche si $a > 0$ puisque u_0^n est imposé, et $U_h^n = (u_0^n, \ldots, u_N^n)^T$ dans le cas de conditions aux limites à droite si $a < 0$.

Les schémas, pour résoudre notre EDP, diffèrent par la façon dont les dérivées en temps et en espace sont approchées. S'il est naturel de faire l'approximation

$$\frac{\partial u}{\partial t}(x_j, t_n) \simeq \frac{u_j^{n+1} - u_j^n}{\Delta t},$$

pour la dérivée en temps, il n'a pas de raison, a priori, de privilégier l'une ou l'autre des approximations suivantes de la dérivée en espace :

1. approximation centrée :
$$\frac{\partial u}{\partial x}(x_j,t_n) \simeq \frac{u^n_{j+1}-u^n_{j-1}}{2\Delta x},$$
2. approximation décentrée à droite :
$$\frac{\partial u}{\partial x}(x_j,t_n) \simeq \frac{u^n_{j+1}-u^n_{j}}{\Delta x},$$
3. approximation décentrée à gauche :
$$\frac{\partial u}{\partial x}(x_j,t_n) \simeq \frac{u^n_{j}-u^n_{j-1}}{\Delta x}.$$

Sans parler, bien sûr, de la multitude d'autres manières d'approcher une dérivée première. Par exemple, en utilisant plus de deux points. Il faut bien entendu faire attention aux conditions aux limites qui peuvent être périodiques ou avec des valeurs imposées en $j = 0$ ou $j = N+1$, comme on l'a expliqué dans la section précédente.

Commencons par quelques exemples de schémas explicites pour l'équation de transport.

1. **Schéma centré**
   Commençons par un schéma assez naturel. Nous verrons plus tard qu'il n'est pas stable et donc pas utilisé en pratique
$$u^{n+1}_j = u^n_j - \frac{\lambda a}{2}(u^n_{j+1}-u^n_{j-1}).$$
   Ce schéma est obtenu en utilisant une approximation centrée de la dérivée spatiale.
2. **Schémas décentrés**
   (a) **Schéma décentré (à gauche)**
$$u^{n+1}_j = u^n_j - \lambda a(u^n_j-u^n_{j-1}).$$
   Il est obtenu en utilisant une approximation décentrée (à gauche) de la dérivée spatiale.
   (b) **Schéma décentré (à droite)**
$$u^{n+1}_j = u^n_j - \lambda a(u^n_{j+1}-u^n_{j}).$$
   Il est obtenu en utilisant une approximation décentrée (à droite) de la dérivée spatiale.
3. **Schéma de Lax-Friedrichs**
$$u^{n+1}_j = \frac{u^n_{j+1}+u^n_{j-1}}{2} - \frac{\lambda a}{2}(u^n_{j+1}-u^n_{j-1}).$$
   On reconnaît une modification du schéma centré dans lequel u^n_j est remplacé par la moyenne de u^n_{j-1} et u^n_{j+1}.

4. **Schéma de Lax-Wendroff**

$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda^2 a^2}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$

On reconnaît une modification du schéma centré dans lequel on a rajouté un terme étrange qui contient une une discrétisation de la dérivée seconde en espace alors qu'il n'y a pas de dérivée seconde dans l'EDP considérée.

On associe à chaque schéma un "stencil", le stencil associé au point x_j est l'ensemble des points qui servent à calculer u_j^{n+1}, plus précisément, c'est l'ensemble des points x_k, tels que $u_j^{n+1} = \sum_k c_k u_k^n$. Voir la figure 3.1.

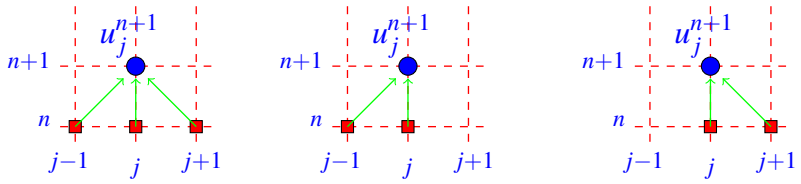

FIGURE 3.1 – Stencils de quelques schémas : de gauche à droite : schéma centré, schéma décentré à gauche et schéma décentré à droite. Les valeurs (à l'instant t_n) indiquées par des carrés permettent de calculer la valeur (à l'instant t_{n+1}) indiquée par un cercle.

Tous les schémas présentés sont des schémas à 3 points, c'est-à-dire qu'ils n'utilisent qu'au plus, les valeurs aux trois points x_{j-1}, x_j et x_{j+1} pour avancer en temps. On peut les écrire sous la forme

$$u_j^{n+1} = H(u_{j-1}^n, u_j^n, u_{j+1}^n), \tag{3.12}$$

où on a introduit une fonction H appelée "solution discrète". Dans le cas linéaire, qui nous intéresse ici, la fonction H est souvent une combinaison linéaire des valeurs u_k^n. Plus généralement, on peut étudier les schémas à $2L+1$ points. Ces schémas s'écrivent sous la forme générique

$$u_j^{n+1} = H(u_{j-L}^n, \cdots, u_{j+L}^n) \tag{3.13}$$

avec une fonction "solution discrète"

$$H(u_{j-L}^n, \cdots, u_{j+L}^n) = \sum_{l=-L}^{+L} c_l u_{j+l}^n \tag{3.14}$$

avec des coefficients c_ℓ qui dépendent de λ et a. Pour $L = 1$, on retrouve les schémas à 3 points, pour $L = 2$, on obtient des schémas à 5 points, ...

Voici les coefficients c_ℓ de la formule (3.14) pour les schémas à trois points cités plus haut.

1. Schéma centré
$$c_{-1} = \frac{\lambda a}{2}, \quad c_0 = 1, \quad c_1 = -\frac{\lambda a}{2}.$$
2. Schémas décentrés

(a) Schéma décentré à gauche

$$c_{-1} = \lambda a, \quad c_0 = 1 - \lambda a.$$

(b) Schéma décentré à droite

$$c_0 = 1 + \lambda a, \quad c_1 = -\lambda a.$$

3. Schéma de Lax-Friedrichs

$$c_{-1} = \frac{1+\lambda a}{2}, \quad c_1 = \frac{1-\lambda a}{2}.$$

4. Schéma de Lax-Wendroff

$$c_{-1} = \frac{\lambda a}{2}(\lambda a + 1), \quad c_0 = 1 - \lambda^2 a^2, \quad c_1 = \frac{\lambda a}{2}(\lambda a - 1).$$

La fonction H doit vérifier la condition

$$H(v, \cdots, v) = v \tag{3.15}$$

qui correspond au fait que les états constants sont propagés exactement par l'opérateur "solution exacte". Autrement dit, si la solution est constante à l'instant t_n, elle doit le rester à l'instant t_{n+1}. La relation (3.15) signifie pour un schéma linéaire que

$$\sum_{\ell=-L}^{L} c_\ell = 1, \tag{3.16}$$

ce qu'on supposera toujours par la suite.

Notons que pour un schéma linéaire (ce qui est le cas pour tous les schémas présentés ici) la relation (3.13) s'écrit aussi

$$U_h^{n+1} = QU_h^n, \tag{3.17}$$

où Q est une matrice, dont les diagonales constantes sont données par les coefficients $c_{-L}, \ldots, c_L$ et dont les première et dernière lignes doivent prendre en compte les conditions aux limites imposées en $j = 0$ et $j = N + 1$.

Considérons à présent le cas général d'une équation linéaire ou non linéaire sous forme conservative

$$\begin{cases} \dfrac{\partial u}{\partial t} + \dfrac{\partial}{\partial x}[f(u)] = 0, \quad x \in \mathbb{R}, \quad t > 0 \\ u(x,0) = u_0(x), \quad x \in \mathbb{R}, \end{cases}$$

où f est la fonction de flux.

En introduisant les points milieux $x_{j+1/2} = \frac{1}{2}(x_j + x_{j+1})$ et les cellules $C_i =]x_{i-1/2}, x_{i+1/2}[$ on remarque que lorsque Δx est petit les valeurs $u(x_j, t_n)$ sont proches des valeurs moyennes

$$m_j^n := \frac{1}{\Delta x} \int_{C_j} u(x, t_n) dx.$$

D'après l'équation de bilan, ces valeurs évoluent exactement suivant la relation

$$m_j^{n+1} - m_j^n = \frac{1}{\Delta x}\int_{t_n}^{t_{n+1}} [f(u(x_{j+1/2},t) - f(u(x_{j-1/2},t)]dt.$$

Le second membre n'est pas calculable mais peut être approché par

$$\frac{\Delta t}{\Delta x}(f(u(x_{j+1/2},t_n) - f(u(x_{j-1/2},t_n)) \approx \lambda(g(u_j^n,u_{j+1}^n) - g(u_{j-1}^n,u_j^n)),$$

où g est une fonction de deux variables appelée flux numérique. On exige que ce flux soit consistant avec le flux f de l'EDP au sens où $g(v,v) = f(v)$.

Ceci conduit au schéma dit de *volumes finis*

$$u_j^{n+1} = u_j^n - \lambda[g(u_j^n,u_{j+1}^n) - g(u_{j-1}^n,u_j^n)], \tag{3.18}$$

qui correspond au choix particulier $H(u,v,w) = v - \lambda[g(v,w) - g(u,v)]$. On note que ce choix assure automatiquement la relation $H(v,v,v) = v$ ainsi que la conservation de la masse totale $\sum_j u_j^{n+1} = \sum_j u_j^n$.

Dans le cas linéaire $f(u) = au$, tous les schémas de différences finies que nous avons cité peuvent s'interprêter comme des schémas de volumes finis en identifiant la fonction de flux numérique correspondante :

1. Schéma centré : $g_C(v,w) = a(v+w)/2$.
2. Schéma décentré : $g_D(v,w) = av$ dans le cas $a > 0$ ($= aw$ si $a < 0$).
3. Schéma de Lax-Friedrichs : $g_{LF}(v,w) = \frac{a}{2}(w+v) - \frac{1}{2\lambda}(w-v)$.
4. Déterminer le flux numérique associé au schéma de Lax-Wendroff (exercice).

3.3 Analyse des schémas

L'analyse numérique des schémas a pour objectif d'étudier leur convergence, dans une norme appropriée, et lorsqu'ils convergent, d'estimer l'erreur, calculée (souvent) dans la même norme, entre la solution exacte et la solution approchée. Comme nous l'avons déjà vu pour l'équation de la chaleur, deux propriétés importantes entrent en jeu : la stabilité et la consistance, dont on déduit la convergence.

Commençons par l'étude de consistance. Comme pour l'équation de la chaleur, on adopte la notation suivante.

Définition 3.3.1 *On appelle erreur de consistance du schéma (3.13), au point x_j et à l'instant t_n, le réel*

$$\kappa_j^n = \frac{1}{\Delta t}\left(u(x_j,t_{n+1}) - H(u(x_{j-L},t_n),\cdots,u(x_{j+L},t_n))\right).$$

L'erreur de consistance du schéma à l'instant t_n est le vecteur K_h^n dont les composantes sont les κ_j^n, et qui vérifie donc

$$\overline{U}_h^{n+1} = Q\overline{U}_h^n + \Delta t K_h^n$$

Pour une norme vectorielle donnée $\|\cdot\|$, le schéma est dit consistant si son erreur de consistance tend vers 0 *quand les pas de discrétisation Δt et Δx tendent vers* 0 *et consistant à l'ordre p en espace et q en temps si et seulement si il existe une constante $C > 0$ telle que*

$$\sup_{n\Delta t \le T} \|K_h^n\| \le C(\Delta x^p + \Delta t^q),$$

pour toute solution u suffisament régulière.

À partir des trois développement de Taylor ci-dessous, on peut estimer les erreurs de consistance des schémas présentés précédemment pour la norme ℓ^∞, et par conséquent pour la norme ℓ^2_Δ qui est majorée par celle-ci.

1. Pour le schéma centré, on a

$$\begin{aligned}\kappa_j^n &= \tfrac{1}{\Delta t}(u(x_j,t_{n+1}) - u(x_j,t_n)) + \tfrac{a}{2\Delta x}(u(x_{j+1},t_n) - u(x_{j-1},t_n)) \\ &= \Big(\partial_t u(x_j,t_n) + O(\Delta t)\Big) + a\Big(\partial_x u(x_j,t_n) + O((\Delta x)^2)\Big).\end{aligned}$$

   Comme u est solution de l'EDP, l'erreur de consistance est

$$\kappa_j^n = O(\Delta t) + O((\Delta x)^2).$$

   Le schéma est consistant, d'ordre 1 en temps et d'ordre 2 en espace.

2. Pour le schéma décentré à gauche, on a

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

   Le schéma est consistant, d'ordre 1 en temps et d'ordre 1 en espace. Il en est de même pour le schéma décentré à droite.

3. Schéma de Lax-Friedrichs.
   On montre (exercice) que schéma est consistant sous la condition que $\frac{(\Delta x)^2}{\Delta t}$ tende vers 0. Autrement dit, que Δx tende vers 0 plus vite que $\sqrt{\Delta t}$.

4. Schéma de Lax-Wendroff.
   On montre (exercice) que l'erreur de consistance du schéma de Lax-Wendroff vérifie

$$\kappa_j^n = O((\Delta t)^2) + O((\Delta x)^2).$$

   C'est donc un schéma consistant, d'ordre 2 en temps et d'ordre 2 en espace.

Passons à présent à l'étude de la stabilité.

Définition 3.3.2 *On dit qu'un schéma est stable pour une norme vectorielle $\|.\|$ si et seulement si il existe une constante C_0 (qui peut dépendre de T), telle que*

$$\sup_{n\Delta t \le T} \|Q^n\| \le C_0.$$

Bien entendu, le Théorème 3.1.1 de Lax déjà énoncé pour l'équation de la chaleur et indiquant que "consistance et stabilité implique convergence" reste valable pour l'équation de transport avec la même démonstration. On étudie d'abord la stabilité ℓ^∞, en présentant (dans le cas de conditions aux limites périodiques) des conditions suffisantes dont la vérification est très simple.

Proposition 3.3.1 *Un schéma linéaire (3.13)-(3.15) dont les coefficients sont positifs est stable dans ℓ^∞.*

Démonstration. Chaque ligne de la matrice Q fait apparaître les coefficients $c_{-L},\dots,c_L$, ce qui montre que

$$\|Q\|_\infty = \sum_{l=-L}^{L} |c_l|.$$

Si les coefficients sont positifs on a donc

$$\|Q\|_\infty = \sum_{l=-L}^{L} c_l = 1,$$

d'après la relation (3.16). □

Remarque 3.3.1 *Une fonction H vérifiant les propriétés énoncées à la Proposition 3.3.1 est croissante en chacun de ses arguments. Par conséquent, si deux conditions initiales V_h^0 et W_h^0 vérifient $V_h^0 \geq W_h^0$ au sens $v_j^0 \geq w_j^0$ pour tout j, alors $V_h^n \geq W_h^n$ pour tout n. C'est la traduction au niveau discret de la propriété de monotonie*

$$v_0 \leq w_0 \implies v(\cdot,t) \leq w(\cdot,t), \quad t>0,$$

vérifiée par les solutions de l'EDP de transport. On dit ainsi que le schéma est monotone.

Applications de la Proposition 3.3.1. On rappelle que $\lambda = \Delta t/\Delta x$.

1. Schéma centré. On ne peut pas utiliser la Proposition 3.3.1 car les coefficients de la combinaison linéaire ne peuvent pas être tous positifs. De fait, ce schema s'avère numériquement instable.
2. Schémas décentrés
   (a) Schéma décentré à gauche. Pour $a>0$ et sous la condition
   $$a\frac{\Delta t}{\Delta x} \leq 1,$$
   le schéma décentré à gauche est stable en norme L^∞.
   (b) Schéma décentré à droite. Pour $a<0$ et sous la condition $-a\lambda \leq 1$, le schéma décentré à droite est stable en norme ℓ^∞.

(c) Noter l'incohérence dans le choix du schéma décentré à gauche dans la cas d'une vitesse $a < 0$: la caractéristique issue du point (x_j, t_{n+1}) n'intersecte pas le stencil de ce schéma (voir la figure 3.1). Remarque analogue pour le schéma décentré à droite. On retient que le schéma décentré à gauche pour une vitesse positive et à droite pour une vitesse négative (pour tenir compte de la direction du vent) est stable, en norme L^∞ sous la condition

$$|a|\frac{\Delta t}{\Delta x} \leq 1. \tag{3.19}$$

appelée condition CFL [1]. Ce schéma est dit "upwind" : on y tient compte de la direction du vent. Si on ne précise pas le signe de a, on peut écrire

$$u_j^{n+1} = u_j^n - \lambda(a_-(u_{j+1}^n - u_j^n) + a_+(u_j^n - u_{j-1}^n)),$$

où $a_- = \min(a,0)$ et $a_+ = \max(a,0)$. Notant que $a = a_+ + a_-$ et $|a| = a_+ - a_-$, on peut écrire le schéma upwind sous la forme

$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{\lambda|a|}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n).$$

3. Schéma de Lax-Friedrichs. Montrer que le schéma de Lax-Friedrichs est stable, en norme ℓ^∞, sous la condition (3.19).
4. Schéma de Lax-Wendroff. On ne peut pas utiliser la Proposition 3.3.1 car les coefficients de la combinaison linéaire ne peuvent pas être tous positifs.

Le schéma de Lax-Wendroff n'est pas stable en norme ∞, c'est pourtant un schéma "prometteur" car le seul, parmi ceux que nous avons présentés, qui soit d'ordre 2 en espace et en temps. Pour ce schéma (comme d'autres !), il faut mesurer la stabilité dans une autre norme que la norme $\|.\|_\infty$. La norme ℓ^2 est un bon candidat.

Pour un schéma linéaire, la stabilité ℓ^2 peut se caractériser à l'aide de la transformation de Fourier discrète. On considère là aussi des conditions aux limites périodiques

$$u(x+1,t) = u(x,t).$$

En particulier

$$u(x_{j+N+1},t) = u(x_j,t), \quad j \in \mathbb{Z}.$$

Pour traduire cette périodicité sur le schéma on prend donc des vecteurs U_h^n qui ont pour composantes $(u_0^n, \ldots, u_N^n)$ et qu'on peut prolonger en définissant les valeurs u_j^n pour tout $j \in \mathbb{Z}$ par la relation de périodicité

$$u_{j+N+1}^n = u_j^n.$$

Ceci permet de donner un sens précis à la relation

$$u_j^{n+1} = \sum_{l=-L}^{L} c_l u_{j+l}^n, \quad j = 0, \ldots, N.$$

1. Initiales de trois mathématiciens ayant travaillé sur les EDP et/ou leurs approximations numériques : Richard Courant, Kurt Friedrichs, et Hans Lewy.

On voit en particulier que la matrice Q telle que $U_h^{n+1} = QU_h^n$ est une matrice circulante dont les coefficients sont donnés par

$$q_{i,j} = c_{(j-i)\%(N+1)}.$$

Pour tout vecteur $V = (v_0, \dots, v_N)^T \in \mathbb{R}^{N+1}$ on définit sa transformée de Fourier discrète $\hat{V} = (\hat{v}_0, \dots, \hat{v}_N)$ par

$$\hat{v}_k = \frac{1}{\sqrt{N+1}} \sum_{j=0}^{N} v_j \exp\Big(-i2\pi \frac{kj}{N+1}\Big) = \frac{1}{\sqrt{N+1}} \sum_{j=0}^{N} v_j \exp(-i2\pi kj\Delta x) = \langle V, F_k \rangle,$$

où F_k est le vecteur de coordonnées $\big(\frac{1}{\sqrt{N+1}} \exp(-i2\pi kj\Delta x)\big)_{j=0,\dots N}$. Notons que le vecteur F_k se prolonge naturellement par périodicité à tout $\mathbb{Z}$, et que les $\hat{v}_k$ sont en général des nombres complexes. Il est facile de vérifier (exercice) que $(F_0, \dots, F_N)$ est une base orthonormée de $\mathbb{C}^{N+1}$ muni du produit scalaire hilbertien usuel, et par conséquent

$$V = \sum_{k=0}^{N} \hat{v}_k F_k,$$

ce qui se lit aussi comme la formule de transformation inverse de Fourier

$$v_j = \frac{1}{\sqrt{N+1}} \sum_{k=0}^{n} \hat{v}_k \exp(i2\pi kj\Delta x).$$

On a en particulier la formule de Parseval pour la norme $\|\cdot\| = \|\cdot\|_2^2$:

$$\|V\|^2 = \sum_{j=0}^{N} |v_j|^2 = \sum_{k=0}^{N} |\hat{v}_k|^2 = \|\hat{V}\|^2.$$

L'action de Q sur un vecteur V prolongé par périodicité s'écrit

$$(QV)_j = \sum_{l=-L}^{L} c_l v_{j+l},$$

ce qui donne en particulier

$$\begin{aligned}(QF_k)_j &= \tfrac{1}{\sqrt{N+1}} \textstyle\sum_{l=-L}^{L} c_l \exp(i2\pi k(j+l)\Delta x) \\ &= \Big(\textstyle\sum_{l=-L}^{L} c_l \exp(i2\pi kl\Delta x)\Big) \exp(i2\pi kj\Delta x) \\ &= \alpha_k (F_k)_j,\end{aligned}$$

avec

$$\alpha_k := \sum_{l=-L}^{L} c_l \exp(i2\pi kl\Delta x).$$

Ceci traduit le fait que les F_k sont les fonctions propres des matrices circulantes : on a $QF_k = \alpha_k F_k$. Cela entraîne en particulier que

$$\hat{u}_k^{n+1} = \alpha_k \hat{u}_k^n.$$

On dit que α_k est le coefficient d'amplification du mode k de Fourier pour le schéma considéré. Grâce à la formule de Parseval, on obtient

$$\|Q\|_2 = \max_{k=0,\dots,N} |\alpha_k|,$$

ainsi que

$$\|Q^m\|_2 = \max_{k=0,\dots,N} |\alpha_k|^m.$$

Ceci nous montre que le schéma est stable pour la norme ℓ^2 si et seulement si

$$|\alpha_k| \leq 1, \quad k = 0, \dots, N.$$

Cette condition, qui assure le contrôle des coefficients de Fourier par

$$|\hat{u}_k^n| \leq |\hat{u}_k^0|, \quad n \in \mathbb{N},$$

est appelée *condition de stabilité de Von Neumann*.

Exemple 3 *Reprenons la discrétisation par un schéma explicite de l'équation de la chaleur étudiée dans la première section, en supposant des conditions périodiques. Sans terme source, le schéma s'écrit sous la forme*

$$u_j^{n+1} = (1-2r)u_j^n + ru_{j-1}^n + ru_{j+1}^n, \quad r = \mu \frac{\Delta t}{\Delta x^2}.$$

On en déduit que

$$\alpha_k = 1 - 2r\Big(1 - \frac{1}{2}(\exp(i2\pi k\Delta x) + \exp(-i2\pi k\Delta x))\Big) = 1 - 2r(1-\cos(2\pi k\Delta x)) = 1 - 4r\sin^2(\pi k\Delta x)).$$

On voit ainsi que la condition de stabilité de Von Neumann sera assurée si et seulement si $0 < r \leq 1/2$, ce qui est le même résultat que celui qu'on avait obtenu dans le cas des conditions homogènes de Dirichlet. Pour le schéma implicite (toujours pour l'équation de la chaleur avec conditions périodique), on montre que

$$\alpha_k = \frac{1}{1 + 4r\sin^2(k\pi\Delta x)}.$$

Ce schéma est donc inconditionnellement stable en norme ℓ^2, i.e. sans condition sur les pas de discrétisations Δt et Δx. Ce que nous avions déjà vu dans le cas des conditions homogènes de Dirichlet.

Revenons à l'équation de transport et au schéma (3.13), (3.14), (3.16).

1. Pour le schéma centré, on a

$$\alpha_k = 1 + \frac{\lambda a}{2}(e^{-i\theta} - e^{i\theta}) = 1 - i\lambda a \sin\theta,$$

en notant $\theta = 2k\pi\Delta x$. On en déduit que $|\alpha_k| > 1$, pour les k tels que $\sin\theta \neq 0$. Le schéma n'est donc pas ℓ^2-stable.

2. Pour le schéma upwind (décentré à gauche si $a > 0$ et à droite si $a < 0$), on trouve dans le cas $a > 0$
$$\alpha_k = 1 - \lambda a + \lambda a \cos\theta - i\lambda a \sin\theta.$$
Montrer que le schéma upwind est ℓ^2-stable sous la condition CFL
$$|a|\frac{\Delta t}{\Delta x} \leq 1. \tag{3.20}$$

3. Pour le schéma de Lax-Friedrichs, on trouve
$$\alpha_k = \cos\theta - i\lambda a \sin\theta.$$
Montrer que le schéma de Lax-Friedrichs est L^2-stable sous la même condition CFL.

4. Pour le schéma de Lax-Wendroff, on trouve
$$\alpha_k = 1 - 2(\lambda a)^2 \sin^2\theta - i\lambda a \sin(2\theta).$$
en notant $\theta = k\pi\Delta x$. Montrer que le schéma de Lax-Wendroff est ℓ^2-stable sous la même condition CFL.

Remarque 3.3.2 (À propos de la condition CFL (3.20)) *La quantité $|a|\Delta t$ est la distance parcourue, à la vitesse a, pendant un temps Δt. Comme la solution exacte vérifie $u(x_j, t_{n+1}) = u(x_j - a\Delta t, t_n)$, la condition CFL exprime le fait que la caractéristique issue du point x_j à l'instant t_{n+1} doit couper l'axe $t = t_n$ en un point contenu dans le stencil du schéma. Voir la figure 3.2.*

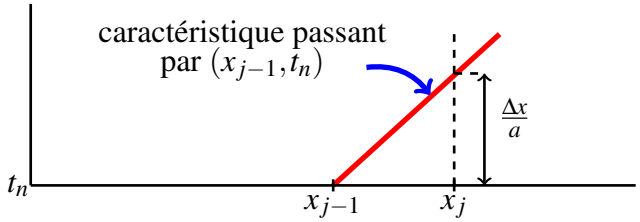

FIGURE 3.2 – Interprétation de la condition CFL (3.20). Cas $a > 0$. Si Δt est plus petit que $(\Delta x)/a$, alors la caractéristique passant par (x_j, t_{n+1}) coupe l'axe $t = t_n$ en un point situé entre x_{j-1} et x_j.

Nous terminons en revenant sur la notion d'ordre de précision des schéma. On suppose par la suite que le rapport
$$\lambda = \frac{\Delta t}{\Delta x}$$
est constant et donc que le pas de discrétisation en espace et le pas de discrétisation en temps tendent à la même vitesse vers 0.

Définition 3.3.3 *On dit qu'un schéma est d'ordre p, si pour toute solution assez régulière de l'équation (1.4), on a*

$$\kappa_j^n = O((\Delta t)^p) = O((\Delta t)^p + (\Delta x)^p). \tag{3.21}$$

Écrivons des développement de Taylor de la solution, supposée assez régulière. On utilise ici les notations : $\overline{u}_j^n = u(x_j,t_n)$, $\partial_x \overline{u}_j^n = \partial_x u(x_j,t_n)$, etc. On a d'une part

$$H(\overline{u}_{j-L}^n,\cdots,\overline{u}_{j+L}^n) = \left[\sum_{\ell=-L}^{L} c_\ell\right]\overline{u}_j^n + \left[\sum_{\ell=-L}^{L} \ell\Delta x c_\ell\right]\partial_x\overline{u}_j^n + \left[\sum_{\ell=-L}^{L} \frac{(\ell\Delta x)^2}{2}c_\ell\right]\partial_{xx}\overline{u}_j^n + \cdots$$

D'autre part, tenant compte des relations $u_t = -au_x$ et $u_{tt} = a^2 u_{xx}$, on a

$$\overline{u}_j^{n+1} = \overline{u}_j^n - a(\Delta t)\partial_x\overline{u}_j^n + \frac{(a\Delta t)^2}{2}\partial_{xx}\overline{u}_j^n + \cdots$$

D'où le développement de Taylor de l'erreur de consistance (tenant compte de (3.16)

$$\kappa_j^n = -\left[a + \frac{1}{\lambda}\sum_{\ell=-L}^{L}\ell c_\ell\right]\partial_x\overline{u}_j^n + \left[\frac{a^2\Delta t}{2} - \frac{1}{\lambda}\sum_{\ell=-L}^{L}\frac{\ell^2\Delta x}{2}c_\ell\right]\partial_{xx}\overline{u}_j^n + \cdots$$

Pour un rapport $\lambda = \frac{\Delta t}{\Delta x}$ constant, on peut ainsi écrire

$$\kappa_j^n = -\left[a + \frac{1}{\lambda}\sum_{\ell=-L}^{L}\ell c_\ell\right]\partial_x\overline{u}_j^n + \frac{\Delta t}{2}\left[a^2 - \frac{1}{\lambda^2}\sum_{\ell=-L}^{L}\ell^2 c_\ell\right]\partial_{xx}\overline{u}_j^n + O((\Delta t)^2).$$

On en déduit le résultat suivant :

Proposition 3.3.2 *Si les coefficients c_ℓ d'un schéma (3.13), (3.14), (3.16) vérifient $\sum_{\ell=-L}^{L}\ell c_\ell = -\lambda a$ alors le schéma est d'ordre au moins égal à 1. Si de plus $\sum_{\ell=-L}^{L}\ell^2 c_\ell = (\lambda a)^2$, il est d'ordre au moins égal à 2.*

Proposition 3.3.3 *Un schéma linéaire à trois points, conservatif, (c'est à dire s'écrivant sous la forme (3.18)), associé à un flux numérique consistant est d'ordre au moins égal à 1.*

Démonstration. Nous allons montrer que $\sum_{\ell=-L}^{L}\ell c_\ell = -\lambda a$, c'est-à-dire pour un schéma à trois points que $c_1 - c_{-1} = -\lambda a$. Si $g(v,w) = \alpha v + \beta w$ est le flux numérique associé au schéma, ses coefficients sont alors $c_{-1} = \lambda\alpha$, $c_0 = 1-\lambda(\alpha-\beta)$ et $c_1 = -\lambda\beta$. La relation recherchée découle de la consistance du flux g. □

Proposition 3.3.4 *Il existe un seul schéma linéaire à trois points, conservatif, dont le flux numérique est consistant et qui est d'ordre deux. C'est le schéma de Lax-Wendroff.*

Démonstration. Si le flux numérique est $g(v,w) = \alpha v + \beta w$, alors le schéma s'écrit

$$u_j^{n+1} = c_{-1}u_{j-1}^n + c_0 u_j^n + c_1 u_{j+1}^n$$

avec $c_{-1} = \alpha\lambda, c_0 = 1 - \alpha\lambda + \beta\lambda$ et $c_1 = -\beta\lambda$. On a $c_{-1} + c_0 + c_1 = 1$ et comme le flux est consistant, $c_{-1} - c_1 = \lambda a$. Un tel schéma peut donc s'écrire en fonction d'un seul paramètre qu'on prend égal à $q = c_{-1} + c_1 = 1 - c_0$. Le schéma s'écrit alors

$$u_j^{n+1} = u_j^n - \frac{\lambda a}{2}(u_{j+1}^n - u_{j-1}^n) + \frac{q}{2}(u_{j+1}^n - 2u_j^n + u_{j-1}^n). \tag{3.22}$$

Par la proposition 3.3.2, il est d'ordre deux si $c_{-1} + c_1 = (\lambda a)^2$, c'est à dire si $q = \lambda^2 a^2$. On reconnait alors le schéma de Lax-Wendroff. On pourra vérifier que ce schéma n'est pas d'ordre supérieur à deux. □

Considérons des schémas de la forme (3.22) où le réel q est un paramètre appelé *coefficient de viscosité* du schéma (pourquoi ?). Le schéma obtenu pour $q = (\lambda a)^2$ est donc le schéma de Lax-Wendroff.

Proposition 3.3.5 *Un schéma linéaire (3.13) à trois points conservatif et dont le flux numérique est consistant est stable dans ℓ^2 si et seulement si son coefficient de viscosité q vérifie*

$$(\lambda a)^2 \leq q \leq 1. \tag{3.23}$$

Démonstration. Le coefficient d'amplification du schéma (3.22) est donné par (on pose $c = \lambda a$)

$$\alpha_k = 1 - \frac{c}{2}(\exp(i2\pi k\Delta x) - \exp(-i2\pi k\Delta x)) + \frac{q}{2}(\exp(i2\pi k\Delta x) - 2 + \exp(-i2\pi k\Delta x))$$

$$= 1 - q + q\cos(2\pi k\Delta x) - ic\sin 2\pi k\Delta x$$

d'où,

$$\alpha_k = 1 - q(1 - \cos(2\pi k\Delta x)) - 2ic\cos(2\pi k\Delta x/2)\sin(2\pi k\Delta x/2).$$

Posant $y = (\sin(2\pi k\Delta x/2))^2$, on a

$$|\alpha_k|^2 = (1 - 2qy)^2 + 4c^2 y(1-y).$$

La condition de Von Neumann est donc vérifiée si et seulement si

$$0 \leq (1 - 2qy)^2 + 4c^2 y(1-y) \leq 1, \qquad \forall y \in [0,1].$$

Ces inégalités soient vérifiées si

$$-qy + q^2 y^2 + c^2 y(1-y) \leq 0, \qquad \forall y \in [0,1].$$

Soit puisque $y \in]0,1]$, si

$$-q + q^2 y + c^2(1-y) \leq 0.$$

On reconnaît une fonction affine de y, qui est donc négative sur un intervalle si et seulement elle est négative aux deux extrémités de cet intervalle. D'où le résultat. □