MÉTHODES NUMÉRIQUES POUR LES EDP

UNIVERSITÉ DE CERGY

Contents

1. Approximation des dérivées par différence finie 1
1.1. Méthode générale 2
1.2. Méthodes des différences finies classiques 2
2. Résolution numérique des équations différentielles ordinaires 7
2.1. Le problème de Cauchy 7
2.2. Méthodes numériques à un pas 9
2.3. Analyse des méthodes à un pas 11
2.4. Méthode d'Euler implicite 15
2.5. Stabilité 17
2.6. Méthode de Runge-Kutta 19
3. Approximation numérique d'équations aux dérivées partielles 21
3.1. Schéma θ pour l'équation de la chaleur 21
3.2. Théorème de Lax 25
3.3. Le cas multidimensionnel 26
3.4. Exercices 27
4. Méthode des volumes finis 27
4.1. Équation de transport non linéaire 28
4.2. Forme intégrale et quantité conservée 28
4.3. Maillage, volumes de contrôle et moyennes de cellule 28
4.4. Schéma de volumes finis : origine du schéma 29
4.5. Schéma discret d'Euler 29
4.6. Conservation discrète 29
4.7. Consistance 30
4.8. Stabilité 32
4.9. Convergence 33
4.10. Exercices 34
References 34

1. Approximation des dérivées par différence finie

Un problème qu'on rencontre souvent en analyse numérique est l'approximation de la dérivée d'une fonction $f : [a, b] \to \mathbb{R}$ sur un intervalle donné $[a, b]$.

1.1. **Méthode générale.** Une approche naturelle consiste à introduire $n+1$ nœuds $x_k \in [a,b]$ uniformément répartis, c'est-à-dire tels que

$$x_0 = a, \qquad x_n = b, \qquad x_{k+1} = x_k + h, \quad \forall k \in \{0,\dots,n-1\},$$

où

$$h := \frac{b-a}{n}.$$

On approche alors $f'(x_i)$ en utilisant les valeurs nodales $f(x_k)$, dont on considère avoir l'accès. On note u'_i l'approximation de $f'(x_i)$, donc

$$u'_i \simeq f'(x_i).$$

De manière générale, on définit les u'_i via l'équation

$$h \sum_{k=-m}^{m} \alpha_k \, u'_{i-k} = \sum_{k=-m'}^{m'} \beta_k \, f(x_{i-k}), \tag{1}$$

où $\{\alpha_k\}, \{\beta_k\} \in \mathbb{R}$ sont $m+m'+1$ coefficients à déterminer, et où on peut utiliser la convention $u'_j = 0$ et $f(x_j) = 0$ pour tout $j \notin \{0,\dots,n\}$. Cette équation déterminant une approximation est appelée schéma.

Le coût du calcul est un critère important dans le choix du schéma, il faut par exemple noter que si $m \neq 0$, la détermination des quantités u'_i requiert la résolution d'un système linéaire.

Definition 1.1 (Stencil)**.** *L'ensemble des nœuds impliqués dans la construction de la dérivée de y en un nœud donné est appelé* stencil.

1.2. **Méthodes des différences finies classiques.**

1.2.1. *Méthode "forward".* Le moyen le plus simple pour construire une formule du type (1) consiste à revenir à la définition de la dérivée. Si $f'(x_i)$ existe, alors

$$f'(x_i) = \lim_{h \to 0^+} \frac{f(x_i+h) - f(x_i)}{h}. \tag{2}$$

Definition 1.2 (Différence finie progressive)**.** *En remplaçant la limite par le taux d'accroissement, avec h fini, on obtient l'approximation*

$$u'_{i,FD} = \frac{f(x_{i+1}) - f(x_i)}{h}, \qquad \forall i \in \{0,\dots,n-1\}. \tag{3}$$

Cette relation est un cas particulier de (1) où $m=0$, $\alpha_0 = 1$, $m'=1$, $\beta_{-1} = 1$, $\beta_0 = -1$, $\beta_1 = 0$. Le second membre de (3) est appelé différence finie progressive, ou "en avant".

L'approximation que l'on fait revient à remplacer $f'(x_i)$ par la pente de la droite passant par les points $(x_i, f(x_i))$ et $(x_{i+1}, f(x_{i+1}))$.

Pour estimer l'erreur commise, il suffit d'écrire le développement de Taylor de f (qui sera toujours supposée assez régulière). En effet, par le théorème de Taylor-Lagrange, il existe $\beta_i \in]x_i, x_{i+1}[$ tel que

$$f(x_{i+1}) = f(x_i) + h f'(x_i) + \frac{h^2}{2} f''(\beta_i).$$

Ainsi,

$$f'(x_i) - u'_{i,FD} = -\frac{h}{2} f''(\beta_i).$$

1.2.2. *Méthode centrée.* Au lieu de (3), on aurait pu utiliser un taux d'accroissement centré, obtenant alors l'approximation suivante.

Definition 1.3 (Différence finie centrée)**.**

$$u'_{i,CD} = \frac{f(x_{i+1}) - f(x_{i-1})}{2h}, \qquad \forall i \in \{1, \dots, n-1\}. \tag{4}$$

Le schéma (4) est un cas particulier de (1) où $m = 0$, $\alpha_0 = 1$, $m' = 1$, $\beta_{-1} = \frac{1}{2}$, $\beta_0 = 0$, $\beta_1 = -\frac{1}{2}$. Le second membre de (4) est appelé différence finie centrée. Géométriquement, l'approximation revient à remplacer $f'(x_i)$ par la pente de la droite passant par les points $(x_{i-1}, f(x_{i-1}))$ et $(x_{i+1}, f(x_{i+1}))$.

Lemma 1.4. *Il existe $\beta_i \in [x_{i-1}, x_{i+1}]$ tel que*

$$f'(x_i) - u'_{i,CD} = -\frac{h^2}{6} f^{(3)}(\beta_i).$$

Démonstration. On utilise le développement de Taylor autour de x_i aux points $x_{i+1} = x_i + h$ et $x_{i-1} = x_i - h$ et le théorème de Taylor-Lagrange, on obtient

$$f(x_i + h) = f(x_i) + hf'(x_i) + \frac{h^2}{2} f''(x_i) + \frac{h^3}{6} f^{(3)}(\beta_1),$$
$$f(x_i - h) = f(x_i) - hf'(x_i) + \frac{h^2}{2} f''(x_i) - \frac{h^3}{6} f^{(3)}(\beta_2),$$

où $\beta_1 \in]x_i, x_i + h[$ et $\beta_2 \in]x_i - h, x_i[$. Ainsi,

$$f'(x_i) - u'_{i,CD} = -\frac{h^2}{12}\big(f^{(3)}(\beta_1) + f^{(3)}(\beta_2)\big).$$

Puisque $f^{(3)}$ est continue sur $]x_i - h, x_i + h[$, la moyenne

$$\frac{f^{(3)}(\beta_1) + f^{(3)}(\beta_2)}{2}$$

est une valeur intermédiaire de $f^{(3)}$ sur cet intervalle. Par théorème des valeurs intermédiaires, il existe $\beta_i \in]x_i - h, x_i + h[$ tel que

$$f^{(3)}(\beta_i) = \frac{f^{(3)}(\beta_1) + f^{(3)}(\beta_2)}{2}.$$

□

La formule (4) fournit donc une approximation de $f'(x_i)$ qui est du second ordre par rapport à h.

1.2.3. *Méthode "backward".* Enfin, on peut définir de manière analogue un troisième schéma.

Definition 1.5 (Différence finie rétrograde)**.**

$$u'_{i,BD} = \frac{f(x_i) - f(x_{i-1})}{h}, \qquad \forall i \in \{1, \dots, n\}. \tag{5}$$

L'erreur suivante lui correspond

$$f'(x_i) - u'_{i,BD} = \frac{h}{2} f''(\beta_i),$$

pour un certain $\beta_i \in]x_{i-1}, x_i[$. Les valeurs des paramètres dans (5) sont $m = 0$, $\alpha_0 = 1$, $m' = 1$ et $\beta_{-1} = 0$, $\beta_0 = 1$, $\beta_1 = -1$.

1.2.4. *Approximation de dérivées d'ordres supérieurs.* Des schémas d'ordre élevé, ou encore des approximations par différences finies de dérivées de f d'ordre supérieur, peuvent être construits en augmentant l'ordre des développements de Taylor. Voici un exemple concernant l'approximation de f''. Si $f \in C^4([a, b])$, on obtient

$$f''(x_i) = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2} - \frac{h^2}{24}\Big(f^{(4)}(x_i + \theta_i h) + f^{(4)}(x_i - \omega_i h)\Big),$$

où $0 < \theta_i, \omega_i < 1$, d'où on déduit le schéma aux différences finies centrées

$$u''_i = \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2}, \qquad \forall i \in \{1, \dots, n-1\}. \tag{6}$$

L'erreur correspondante est

$$f''(x_i) - u''_i = -\frac{h^2}{24}\Big(f^{(4)}(x_i + \theta_i h) + f^{(4)}(x_i - \omega_i h)\Big).$$

La formule (6) fournit donc une approximation de $f''(x_i)$ du second ordre par rapport à h.

1.2.5. *Différences finies compactes.* Pour abréger on note $f_i^{(k)} = f^{(k)}(x_i)$ et $f_i := f(x_i)$. Des approximations plus précises de f' sont données par les formules suivantes

Definition 1.6 (Différences finies compactes)**.** *On définit u'_i via les équations*

$$\alpha u'_{i-1} + u'_i + \alpha u'_{i+1} = \frac{\beta}{2h}\big(f_{i+1} - f_{i-1}\big) + \frac{\gamma}{4h}\big(f_{i+2} - f_{i-2}\big), \tag{7}$$

où $i \in \{2, \dots, n-2\}$.

Les coefficients α, β et γ doivent être déterminés de manière à ce que les relations (7) conduisent à des valeurs de u_i qui approchent $f'(x_i)$ à l'ordre le plus élevé par rapport à h. Pour cela, on choisit des coefficients qui minimisent l'erreur de consistance

$$\sigma_i := \alpha f'_{i-1} + f'_i + \alpha f'_{i+1} - \left[\frac{\beta}{2h}(f_{i+1} - f_{i-1}) + \frac{\gamma}{4h}(f_{i+2} - f_{i-2})\right]. \tag{8}$$

Nous pouvons donner une définition non rigoureuse mais générale des erreurs de consistance.

Definition 1.7 (Erreur de consistance)**.** *L'erreur de consistance d'un schéma consiste à considérer le schéma, à y remplacer la grandeur approximée par la grandeur exacte, et à regarder l'erreur qui y est faite.*

Definition 1.8 (Erreur de convergence)**.** *L'erreur de convergence est l'erreur entre une quantité exacte et son approximation.*

On considère une norme $\|\cdot\|$ sur $\mathbb{R}^{N+1}$ quelconque.

Lemma 1.9 (Consistance implique convergence)**.** *Considérons un schéma de différences finies compactes pour approcher f', écrit sous forme matricielle*

$$A\,u' = BF,$$

où

$$u' := (u'_i)_{i=0}^N, \qquad F' := \big(f'(x_i)\big)_{i=0}^N, \qquad F := \big(f(x_i)\big)_{i=0}^N,$$

et où on a $u'_i \simeq f'(x_i)$. B peut dépendre de h mais pas A, et A est inversible. Supposons qu'il existe $C > 0$ et $n \in \mathbb{N}$ tels que pour tout $h > 0$ dans un voisinage de 0,

$$\big\|AF' - BF\big\| \leqslant Ch^n,$$

qui est l'erreur de consistance. Alors l'erreur de convergence est

$$\big\|u' - F'\big\| \leqslant C\,\big\|A^{-1}\big\|\,h^n.$$

Démonstration. On a $Au' = BF$ et on définit l'erreur de consistance $\sigma = (\sigma_i)_{i=0}^N$ par $\sigma := AF' - BF$. En soustrayant ces deux relations, on obtient $A(u' - F') = -\sigma$ et donc $u' - F' = -A^{-1}\sigma$. □

Autrement dit, l'ordre de convergence global est égal à l'ordre de consistance n. Dans (7) on a $N = 3$, et A est une matrice ayant 1 sur sa diagonale et α en-dessous et au-dessus de sa diagonale. Plus explicitement,

$$A = \begin{pmatrix} 1 & \alpha & 0 & \cdots & 0 \\ \alpha & 1 & \alpha & \ddots & \vdots \\ 0 & \alpha & 1 & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \alpha \\ 0 & \cdots & 0 & \alpha & 1 \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}$$

$$B = \frac{1}{h}\begin{pmatrix} 0 & \frac{\beta}{2} & 0 & -\frac{\gamma}{4} & \cdots & 0 \\ -\frac{\beta}{2} & 0 & \frac{\beta}{2} & 0 & -\frac{\gamma}{4} & \vdots \\ 0 & -\frac{\beta}{2} & 0 & \frac{\beta}{2} & 0 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & -\frac{\beta}{2} & 0 & \frac{\beta}{2} \\ 0 & \cdots & \frac{\gamma}{4} & 0 & -\frac{\beta}{2} & 0 \end{pmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}.$$

Lemma 1.10 (Ordre 6 des différences finies compactes)**.** *Dans le cas de* (7)*, il existe un unique schéma d'ordre* 6 *et il correspond aux paramètres*

$$\alpha = \frac{1}{3}, \qquad \beta = \frac{14}{9}, \qquad \gamma = \frac{1}{9}. \tag{9}$$

Démonstration. En supposant que $f \in C^5([a,b])$ et en écrivant le développement de Taylor en x_i, on trouve

$$f_{i\pm1} = f_i \pm hf_i' + \frac{h^2}{2}f_i^{(2)} \pm \frac{h^3}{6}f_i^{(3)} + \frac{h^4}{24}f_i^{(4)} \pm \frac{h^5}{120}f_i^{(5)} + \frac{h^6}{6!}f_i^{(6)} + O(h^7),$$
$$f_{i\pm2} = f_i \pm 2hf_i' + 2h^2f_i^{(2)} \pm \frac{4}{3}h^3f_i^{(3)} + \frac{2}{3}h^4f_i^{(4)} \pm \frac{4}{15}h^5f_i^{(5)} + \frac{2^6h^6}{6!}f_i^{(6)} + O(h^7),$$
$$f_{i\pm1}' = f_i' \pm hf_i^{(2)} + \frac{h^2}{2}f_i^{(3)} \pm \frac{h^3}{6}f_i^{(4)} + \frac{h^4}{24}f_i^{(5)} \pm \frac{h^5}{120}f_i^{(5)} + O(h^6).$$

Ainsi

$$\alpha f_{i-1}' + f_i' + \alpha f_{i+1}' = (2\alpha+1)f_i' + \alpha h^2 f_i^{(3)} + \alpha\frac{h^4}{12}f_i^{(5)} + O(h^6).$$

On calcule ensuite

$$f_{i+1} - f_{i-1} = 2hf_i' + \frac{h^3}{3}f_i^{(3)} + \frac{h^5}{60}f_i^{(5)} + O(h^7),$$

et

$$f_{i+2} - f_{i-2} = 4hf_i' + \frac{8}{3}h^3f_i^{(3)} + \frac{8}{15}h^5f_i^{(5)} + O(h^7).$$

Par conséquent, le second membre vaut

$$\frac{\beta}{2h}\big(f_{i+1} - f_{i-1}\big) + \frac{\gamma}{4h}\big(f_{i+2} - f_{i-2}\big)$$
$$= (\beta+\gamma)f_i' + \Big(\frac{\beta}{6} + \frac{2\gamma}{3}\Big)h^2f_i^{(3)} + \Big(\frac{\beta}{120} + \frac{2\gamma}{15}\Big)h^4f_i^{(5)} + O(h^6).$$

Par substitution dans (8), on obtient

$$\sigma_i = (2\alpha+1)f_i' + \alpha\frac{h^2}{2}f_i^{(3)} + \alpha\frac{h^4}{12}f_i^{(5)} - (\beta+\gamma)f_i'$$
$$- \frac{h^2}{2}\left(\frac{\beta}{6} + \frac{2\gamma}{3}\right)f_i^{(3)} - \frac{h^4}{60}\left(\frac{\beta}{2} + 8\gamma\right)f_i^{(5)} + O(h^6).$$

On construit des schémas du second ordre en annulant le coefficient de f_i', c'est-à-dire en imposant

$$2\alpha + 1 = \beta + \gamma,$$

des schémas d'ordre 4 en annulant aussi le coefficient de $f_i^{(3)}$,

$$6\alpha = \beta + 4\gamma,$$

et des schémas d'ordre 6 en annulant aussi le coefficient de $f_i^{(5)}$,

$$10\alpha = \beta + 16\gamma.$$

Le système linéaire formé par ces trois dernières relations est non singulier et a une unique solution (9).

Par le Lemme 1.9, l'erreur de convergence est la même que l'erreur de consistance. □

Il y a une seule méthode d'ordre 6 mais il existe en revanche une infinité de méthodes du second et du quatrième ordre. Parmi celles-ci, citons un schéma très utilisé qui correspond aux coefficients

$$\alpha = \frac{1}{4}, \qquad \beta = \frac{3}{2}, \qquad \gamma = 0.$$

Des schémas d'ordre plus élevé peuvent être construits au prix d'un accroissement supplémentaire du stencil.

1.2.6. *Conditions de bord.* Les schémas aux différences finies traditionnels correspondent au choix $\alpha = 0$ et permettent de calculer de manière explicite l'approximation de la dérivée première de f en un nœud, contrairement aux schémas compacts qui nécessitent dans tous les cas la résolution d'un système linéaire de la forme $Au = BF$.

Pour pouvoir résoudre le système, il est nécessaire de se donner les valeurs des variables u_i pour $i < 0$ et $i > n$. On est dans une situation simple quand f est une fonction périodique de période $b - a$, auquel cas

$$u_{i+n} = u_i \qquad \forall i \in \mathbb{Z}.$$

Dans le cas non périodique, le système (7) doit être complété par des relations aux nœuds voisins des extrémités de l'intervalle d'approximation. Par exemple, la dérivée première en x_0 peut être calculée en utilisant la relation

$$u_0' + \alpha u_1' = \frac{1}{h}\big(Af_1 + Bf_2 + Cf_3 + Df_4\big),$$

et en imposant

$$A = \frac{-3 + \alpha + 2D}{2}, \qquad B = 2 + 3D, \qquad C = \frac{-1 - \alpha + 6D}{2},$$

afin que le schéma soit au moins précis à l'ordre deux. Dans ce document, nous essaierons le plus possible d'éviter les problématiques liées aux conditions de bord.

2. Résolution numérique des équations différentielles ordinaires

2.1. **Le problème de Cauchy.** Soit $d \in \mathbb{N}$, I désigne un intervalle de $\mathbb{R}$, $t_0 \in I$, le problème de Cauchy associé à une EDO du premier ordre s'écrit de la manière suivante. Il faut trouver une fonction réelle $y \in C^1(I, \mathbb{R}^d)$ telle que

$$\begin{cases} y'(t) = f(t, y(t)) & \text{si } t \in I \\ y(t_0) = y_0 \end{cases} \tag{10}$$

où $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ est continue par rapport aux deux variables. Si f ne dépend pas explicitement de t, l'équation différentielle est dite autonome. Le cas scalaire correspond à $d = 1$.

2.1.1. *Forme intégrale.* En intégrant (10) entre t_0 et t, on obtient

$$y(t) - y_0 = \int_{t_0}^{t} f(\tau, y(\tau))\, d\tau. \tag{11}$$

La solution de (10) est donc nécessairement de classe C^1 sur I et satisfait l'équation intégrale (11). Inversement, si y est définie par (11), alors elle est continue sur I et $y(t_0) = y_0$. De plus, en tant que primitive de la fonction continue $f(\cdot, y(\cdot))$, on a $y \in C^1(I)$ et elle satisfait l'équation différentielle :

$$y'(t) = f(t, y(t)).$$

Ainsi, si f est continue, le problème de Cauchy (10) est équivalent à l'équation intégrale (11). Nous verrons plus loin comment tirer parti de cette équivalence pour les méthodes numériques.

2.1.2. *Existence locale et unicité.* Rappelons maintenant deux résultats d'existence et d'unicité pour (10). On supposera $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ localement lipschitzienne en (t_0, y_0) par rapport à y, ce qui signifie qu'il existe une boule ouverte $J \subseteq I$ centrée en t_0 de rayon r_J, une boule ouverte Σ centrée en y_0 de rayon r_Σ et une constante $L > 0$ telles que :

$$|f(t, y_1) - f(t, y_2)| \leqslant L|y_1 - y_2| \qquad \forall t \in J,\ \forall y_1, y_2 \in \Sigma.$$

Cette condition est automatiquement vérifiée si la dérivée de f par rapport à y est continue. En effet, dans ce cas, il suffit de prendre

$$L = \max_{(t,y)\in \overline{J\times\Sigma}} \left|\partial_y f(t, y)\right|.$$

Lemma 2.1 (Rappel sur l'existence de la solution locale)**.** *Soit $f : I \times \mathbb{R}^d \to \mathbb{R}^d$ localement lipschitzienne en (t_0, y_0) par rapport à y. Alors le problème de Cauchy* (10) *admet une unique solution dans une boule ouverte de centre t_0 et de rayon $r_0 > 0$.*

Cette solution est appelée solution locale.

2.1.3. *Existence globale et unicité.*

Lemma 2.2 (Rappel sur l'existence d'une solution globale)**.** *Le problème de Cauchy admet une solution globale unique si f est uniformément lipschitzienne par rapport à y, c'est-à-dire si on peut prendre $J = I$, $\Sigma = \mathbb{R}$.*

2.1.4. *Stabilité sous perturbation.* En vue de l'analyse de stabilité du problème de Cauchy, on considère le problème suivant :

$$\begin{cases} \dot{z}(t) = f(t, z(t)) + \delta(t), & t \in I, \\ z(t_0) = y_0 + \delta_0, \end{cases} \tag{12}$$

où $\delta_0 \in \mathbb{R}$ et où δ est une fonction continue sur I. Le problème (12) est déduit de (10) en perturbant la donnée initiale y_0 par δ_0 et la fonction f par δ. Caractérisons à présent la sensibilité de la solution z par rapport à ces perturbations. Intuitivement, la stabilité correspond au fait que si l'EDO est parturbée, alors la solution change d'une manière "continue".

Definition 2.3 (Problème de Cauchy stable)**.** *Soit I un ensemble borné. Le problème de Cauchy* (10) *est dit stable sur I si, pour toute perturbation $(\delta_0, \delta(t))$ satisfaisant*

$$|\delta_0| \leqslant \varepsilon, \qquad |\delta(t)| \leqslant \varepsilon \quad \forall t \in I,$$

avec $\varepsilon > 0$ assez petit pour garantir l'existence de la solution du problème perturbé (12)*, alors*

$$\exists C > 0 \quad \textit{tel que} \quad |y(t) - z(t)| \leqslant C\varepsilon \quad \forall t \in I. \tag{13}$$

La constante C dépend en général de t_0, y et f, mais pas de ε.

Quand I n'est pas borné supérieurement, on dit que (10) *est asymptotiquement stable si, en plus de* (13)*, on a*

$$|\delta(t)| \underset{t\to+\infty}{\longrightarrow} 0 \qquad \Longrightarrow \qquad |y(t) - z(t)| \underset{t\to+\infty}{\longrightarrow} 0.$$

2.1.5. *Grönwall.* Rappelons le lemme de Grönwall pour le problème de Cauchy.

Lemma 2.4 (Grönwall)**.** *Soit p une fonction positive intégrable sur l'intervalle $]t_0, t_0 + T[$, et soient g et φ deux fonctions continues sur $[t_0, t_0 + T]$, avec g croissante. Si φ satisfait*

$$\varphi(t) \leqslant g(t) + \int_{t_0}^{t} p(\tau)\,\varphi(\tau)\,d\tau \qquad \forall t \in [t_0, t_0 + T],$$

alors

$$\varphi(t) \leqslant g(t) \exp\left(\int_{t_0}^{t} p(\tau)\,d\tau\right) \qquad \forall t \in [t_0, t_0 + T].$$

2.1.6. *Utilité du numérique.* On ne sait intégrer qu'un très petit nombre d'EDO non linéaires. De plus, même quand c'est possible, il n'est pas toujours facile d'exprimer explicitement la solution ; considérer par exemple l'équation très simple :

$$y' = \frac{y - t}{y + t},$$

dont la solution n'est définie que de manière implicite par la relation :

$$\frac{1}{2}\log(t^2 + y^2) + \arctan\left(\frac{y}{t}\right) = C,$$

où C est une constante dépendant de la condition initiale.

Pour cette raison, nous sommes conduits à considérer des méthodes numériques. Celles-ci peuvent en effet être appliquées à n'importe quelle EDO, sous la seule condition qu'elle admette une unique solution.

2.2. **Méthodes numériques à un pas.** Abordons à présent l'approximation numérique du problème de Cauchy (10). On fixe $0 < T < +\infty$ et on note $I =]t_0, t_0 + T[$ l'intervalle d'intégration. Pour $h > 0$, soit

$$t_n = t_0 + nh, \qquad n = 0, 1, 2, \ldots, N_h,$$

une suite de nœuds de I induisant une discrétisation de I en sous-intervalles $I_n := [\,t_n, t_{n+1}\,]$.

La longueur h de ces sous-intervalles est appelée pas de discrétisation. Le nombre N_h est le plus grand entier tel que

$$t_{N_h} \leqslant t_0 + T.$$

On a donc $hN_h \simeq T$.

Soit u_j l'approximation au nœud t_j de la solution exacte $y(t_j) =: y_j$,

$$u_j \simeq y_j.$$

De même, $f_j := f(t_j, u_j)$. On pose naturellement

$$u_0 = y_0.$$

Definition 2.5 (Méthode à un pas, méthode multipas)**.** *Une méthode numérique pour l'approximation du problème* (10) *est dite à un pas si* $\forall n \geqslant 0$*, le schéma définissant* u_{n+1} *ne dépend que de* u_n*. Autrement, on dit que le schéma est une méthode multi-pas (ou à pas multiples).*

Une méthode multipas est par exemple quand u_{n+1} dépend de u_n et u_{n-1}. Pour l'instant, nous concentrons notre attention sur les méthodes à un pas. En voici quelques-unes.

Definition 2.6 (Méthode d'Euler explicite)**.**

$$u_{n+1} = u_n + hf(t_n, u_n).$$

Definition 2.7 (Méthode d'Euler implicite)**.**

$$u_{n+1} = u_n + hf(t_{n+1}, u_{n+1}).$$

Dans les deux cas, y' est approchée par un schéma aux différences finies (resp. progressif puis rétrograde). Puisque ces deux schémas sont des approximations au premier ordre par rapport à h de la dérivée première de y, on s'attend à obtenir une approximation d'autant plus précise que le pas du maillage h est petit.

Definition 2.8 (Méthode du trapèze, ou de Crank–Nicolson)**.**

$$u_{n+1} = u_n + \frac{h}{2}(f(t_n, u_n) + f(t_{n+1}, u_{n+1})).$$

Cette méthode provient de l'approximation de l'intégrale (11) par la formule de quadrature du trapèze.

Definition 2.9 (Méthode de Heun)**.**

$$u_{n+1} = u_n + \frac{h}{2}\left(f(t_n, u_n) + f(t_{n+1}, u_n + hf_n)\right).$$

Definition 2.10 (Méthode explicite, implicite)**.** *Une méthode est dite explicite si la valeur* u_{n+1} *peut être calculée directement à l'aide des valeurs précédentes* $(u_k)_{k\leqslant n}$ *(ou d'une partie d'entre elles). Une méthode est dite implicite si* u_{n+1} *n'est défini que par une relation implicite faisant intervenir la fonction* f*.*

Ainsi, la substitution opérée dans la méthode de Heun a pour effet de transformer la méthode implicite du trapèze en une méthode explicite. La méthode d'Euler explicite est explicite, tandis que celle d'Euler implicite est implicite. Noter que les méthodes implicites nécessitent à chaque pas de temps la résolution d'un problème non linéaire (si f dépend non linéairement de la seconde variable).

Pour les méthodes implicites, il faut à chaque itération résoudre un problème consistant à trouver le zéro d'une fonction. Pour Euler implicite, afin de

déterminer u_{n+1} à partir de u_n et t_{n+1} (auxquels on a accès), il faut résoudre l'équation

$$F(x) = 0,$$

où $F(x) := x - u_n - hf(t_{n+1}, x)$. On trouve donc le nombre $x = u_{n+1}$, comme solution.

2.3. **Analyse des méthodes à un pas.**

2.3.1. *Convergence.* Comme en Définition 1.7, la consistance mesure à quel point le schéma numérique reproduit l'équation originale quand le pas tend vers 0. Par ailleurs, la convergence dit quelque chose au niveau de la solution.

On rappelle que le max est une norme

$$\|(u_n)_{0\leqslant n\leqslant j}\|_{\ell^\infty} := \max_{n\in\{0,\dots,j\}} |u_n| .$$

Definition 2.11 (Méthode convergente et ordre de convergence)**.** *Une méthode est dite convergente si*

$$\max_{0\leqslant n\leqslant N} |u_n - y_n| \leqslant C(h)$$

où $C(h) \underset{h\to 0}{\longrightarrow} 0$. *On dit que l'ordre de convergence est* $p > 0$ *s'il existe* $c > 0$ *tel que* $C(h) = ch^p$.

2.3.2. *Grönwall discret.*

Lemma 2.12 (Grönwall discret)**.** *Soit* $(k_n)_{n\in\mathbb{N}}$ *et* $(A_n)_{n\in\mathbb{N}}$ *des suites de réels positifs et* $(\phi_n)_{n\in\mathbb{N}}$ *une suite telle que pour tout* $n \in \mathbb{N}$,

$$\phi_n \leqslant A_n + \sum_{s=0}^{n-1} k_s \phi_s,$$

Si (A_n) *est croissante pour tout* $n \geqslant 0$, *alors pour tout* $n \in \mathbb{N}$,

$$\phi_n \leqslant A_n \exp\left(\sum_{s=0}^{n-1} k_s\right).$$

Démonstration. L'idée de la preuve est d'éliminer les termes récurrents de type ϕ_s dans la somme, en les remplaçant par leur majorant inductif. Nous allons montrer par récurrence sur n que

$$\phi_n \leqslant A_n \exp\left(\sum_{s=0}^{n-1} k_s\right).$$

- Initialisation. On a $\phi_0 \leqslant A_0$, et comme $\sum_{s=0}^{-1} k_s = 0$, alors

$$\phi_0 \leqslant A_0 = A_0 e^0 = A_0 \exp\left(\sum_{s=0}^{-1} k_s\right).$$

- Hérédité. Supposons le résultat vrai pour tout $s < n$, c'est-à-dire

$$\phi_s \leqslant A_s \exp\left(\sum_{i=0}^{s-1} k_i\right), \qquad \forall s < n.$$

En partant de l'inégalité fondamentale,

$$\phi_n \leqslant A_n + \sum_{s=0}^{n-1} k_s \phi_s,$$

nous remplaçons chaque ϕ_s dans la somme par sa borne inductive :

$$\phi_n \leqslant A_n + \sum_{s=0}^{n-1} k_s\, A_s \exp\left(\sum_{i=0}^{s-1} k_i\right) \underset{A_s \text{ croissante}}{\leqslant} A_n \left(1 + \sum_{s=0}^{n-1} k_s \exp\left(\sum_{i=0}^{s-1} k_i\right)\right).$$

On reconnaît maintenant la forme discrète de l'intégrale exponentielle.

Lemma 2.13. *Nous voulons montrer que, si $k_s \geqslant 0$ pour tout s, alors*

$$1 + \sum_{s=0}^{n-1} k_s \exp\left(\sum_{i=0}^{s-1} k_i\right) \leqslant \exp\left(\sum_{s=0}^{n-1} k_s\right). \tag{14}$$

Démonstration. Pour cela, on introduit les notations $S_0 := 0$,

$$S_s := \sum_{i=0}^{s-1} k_i, \qquad s \geqslant 0, \qquad B_n := 1 + \sum_{s=0}^{n-1} k_s e^{S_s}, \qquad n \geqslant 0.$$

L'inégalité (14) s'écrit donc simplement $B_n \leqslant e^{S_n}$. Nous allons le prouver par récurrence sur n.

Pour $n = 0$, on a $B_0 = 1$, $S_0 = 0$, donc $B_0 = 1 = e^{S_0}$.

Supposons que, pour un certain $n \geqslant 0$, on ait $B_n \leqslant e^{S_n}$. Nous allons montrer que cela implique $B_{n+1} \leqslant e^{S_{n+1}}$. Par définition de B_{n+1}, on a

$$B_{n+1} = B_n + k_n e^{S_n} \leqslant e^{S_n} + k_n e^{S_n} = (1 + k_n)\, e^{S_n} \underset{1+x \leqslant e^x}{\leqslant} e^{k_n + S_n} = e^{S_{n+1}}.$$

□

Nous obtenons donc finalement

$$\phi_n \leqslant A_n \exp\left(\sum_{s=0}^{n-1} k_s\right),$$

ce qui conclut l'hérédité et prouve le Lemme 2.12. □

Corollary 2.14. *Soit (a_n) une suite positive. Si pour tout $n \in \{0, \dots, N_h\}$,*

$$a_{n+1} \leqslant (1 + ch)a_n + Ch^{p+1},$$

alors $a_n \leqslant (a_0 + CTh^p)\, e^{cT}$.

Le lemme de Grönwall n'est pas nécessaire dans ce cas mais on va l'utiliser.

Démonstration. On a

$$a_{n+1} - a_n \leqslant ch\, a_n + Ch^{p+1}.$$

En sommant cette inégalité de $n = 0$ à $n = m - 1$ (avec $m \geqslant 1$ arbitraire), on obtient

$$a_m - a_0 = \sum_{n=0}^{m-1} (a_{n+1} - a_n) \leqslant \sum_{s=0}^{m-1} \left(ch\, a_s + Ch^{p+1}\right) = ch \sum_{s=0}^{m-1} a_s + Ch^{p+1}\, m.$$

On passe a_0 à droite et en utilisant le lemme de Grönwall discret, on obtient

$$a_m \leqslant \big(a_0 + Ch^{p+1}m\big)\, e^{cmh} \leqslant (a_0 + CTh^p)\, e^{cT}.$$

□

2.3.3. *Consistance implique convergence.* Considérons un schéma du type

$$u_{n+1} = \Phi\big(t_n, u_n, h\big). \tag{15}$$

On constate que ces schémas sont explicites. Par exemple, Euler explicite et la méthode de Heun se mettent sous cette forme, on a

- $\Phi\big(t, y, h\big) = y + hf(t, y)$ pour Euler explicite
- pour la méthode de Heun,

$$\Phi(t, y, h) = y + \frac{h}{2}\Big(f(t, y) + f\big(t + h,\, y + hf(t, y)\big)\Big).$$

On définit l'erreur de troncature locale

$$\tau_{n+1} := y_{n+1} - \Phi\left(t_n, y_n, h\right).$$

Definition 2.15 (Consistance d'un schéma)**.** *Une méthode est dite consistante si*

$$\max_{0 \leqslant n \leqslant N_h - 1} |\tau_n| \underset{h \to 0}{\longrightarrow} 0.$$

Proposition 2.16. *Prenons un schéma du type* (15)*. Supposons que*

$$|\Phi(t, y, h) - \Phi(t, z, h)| \leqslant (1 + Ch)|y - z|.$$

Si $|\tau_n| \leqslant Ch^{p+1}$ pour un $C > 0$ indépendant de h et de n (i.e. si la méthode est consistante d'ordre $p + 1$), alors la méthode est convergente d'ordre p, c'est-à-dire

$$\|u_n - y_n\| \leqslant ch^p$$

pour un $c > 0$ indépendant de h et de n.

Démonstration. On considère l'erreur $e_n := u_n - y_n$. On a

$$e_{n+1} = e_n + \Phi(t_n, u_n, h) - \Phi(t_n, y_n, h) - \tau_{n+1},$$

donc

$$|e_{n+1}| \leqslant (1 + Ch)|e_n| + Ch^{p+1}.$$

On termine en appliquant le Corollaire 2.14. □

2.3.4. *Méthode d'Euler explicite.*

Theorem 2.17 (Ordre de la méthode d'Euler explicite)**.** *Supposons que f est Lipschitzienne en sa seconde variable. La méthode d'Euler explicite est convergente d'ordre 1, c'est-à-dire que*

$$\max_{0 \leqslant n \leqslant N_h} |u_n - y_n| \leqslant Ch.$$

Démonstration. On utilise la formule de Taylor sur $y(t)$ autour de t_n, via Taylor-Lagrange

$$y_{n+1} = y(t_n + h) = y_n + hy'(t_n) + \frac{h^2}{2}y''(\beta_n),$$

pour un certain $\beta_n \in]t_n, t_{n+1}[$. Mais comme $y'(t) = f(t, y(t))$, cela donne

$$y_{n+1} = y_n + hf(t_n, y_n) + \frac{h^2}{2}y''(\beta_n).$$

L'erreur de consistance est

$$\sigma_{n+1} := y_{n+1} - \Phi\left(t_n, y_n, h\right) = \frac{h^2}{2}y''(\beta_n).$$

Sous l'hypothèse que y'' est bornée sur $[0, T]$, il existe $M > 0$ tel que $|y''(t)| \leqslant M$, et ainsi, pour tout n,

$$|\sigma_{n+1}| \leqslant \frac{M}{2}h^2 =: C_0 h^2,$$

et on voit que la méthode est consistante.

De plus, f est L-Lipschitzienne par rapport à sa seconde variable, donc

$$|\Phi(t, y, h) - \Phi(t, z, h)| \leqslant (1 + Lh)|y - z|.$$

On termine en appliquant la Proposition 2.16. □

2.3.5. *Méthode de Heun.*

Theorem 2.18 (Ordre de la méthode de Heun)**.** *Supposons que f est Lipschitzienne en sa seconde variable. La méthode de Heun est convergente d'ordre 2, c'est-à-dire que*

$$\max_{0 \leqslant n \leqslant N_h} |u_n - y_n| \leqslant Ch^2.$$

Démonstration. On a

$$y'(t_n) = f(t_n, y_n), \qquad y''(t_n) = \frac{\partial f}{\partial t}(t_n, y_n) + \frac{\partial f}{\partial y}(t_n, y_n)\, f(t_n, y_n).$$

Développons

$$\begin{aligned} y_{n+1} = y(t_n + h) &= y_n + h\, y'(t_n) + \frac{h^2}{2}y''(t_n) + O(h^3) \\ &= y_n + hf(t_n, y_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_n, y_n) + \frac{\partial f}{\partial y}(t_n, y_n) f(t_n, y_n)\right) + O(h^3). \end{aligned}$$

Développons maintenant

$$\begin{aligned} &f\big(t_n + h,\, y_n + hf(t_n, y_n)\big) \\ &\quad = f(t_n, y_n) + h\,\frac{\partial f}{\partial t}(t_n, y_n) + h\,\frac{\partial f}{\partial y}(t_n, y_n)\, f(t_n, y_n) + O(h^2). \end{aligned}$$

On en déduit

$$\begin{aligned} \Phi(t_n, y_n, h) &= y_n + \frac{h}{2}\Big(f(t_n, y_n) + f\big(t_n + h,\, y_n + hf(t_n, y_n)\big)\Big) \\ &= y_n + hf(t_n, y_n) + \frac{h^2}{2}\left(\frac{\partial f}{\partial t}(t_n, y_n) + \frac{\partial f}{\partial y}(t_n, y_n) f(t_n, y_n)\right) + O(h^3). \end{aligned}$$

On a donc

$$\sigma_{n+1} = y_{n+1} - \Phi(t_n, y_n, h) = O(h^3).$$

Par ailleurs,

$$\Phi(t,y,h) - \Phi(t,z,h) = y - z + \frac{h}{2}\left(f(t,y) - f(t,z)\right) + \frac{h}{2}\left(f\left(t+h, y+hf(t,y)\right) - f\left(t+h, z+hf(t,z)\right)\right)$$

et

$$\begin{aligned}
&\left|f\left(t+h, y+hf(t,y)\right) - f\left(t+h, z+hf(t,z)\right)\right| \\
&\quad \leqslant L\left|y+hf(t,y) - (z+hf(t,z))\right| \leqslant L\left(|y-z| + h\left|f(t,y) - f(t,z)\right|\right) \\
&\quad \leqslant L(1+Lh)\left|y-z\right|.
\end{aligned}$$

Enfin,

$$|\Phi(t,y,h) - \Phi(t,z,h)| \leqslant (1 + Lh\,(1+Lh/2))\,|y-z|.$$

On termine en appliquant la Proposition 2.16. □

2.3.6. *Milne-Simpson.* Le schéma de Milne-Simpson est défini par

$$u_{i+1} = u_{i-1} + \frac{h}{3}\Big(f(t_{i-1}, u_{i-1}) + 4f(t_i, u_i) + f(t_{i+1}, u_{i+1})\Big).$$

Exercice 2.19. *Écrire l'erreur de consistance en utilisant y la solution exacte de $y'(t) = f(t, y(t))$. Montrer que l'erreur de consistance de Milne-Simpson est d'ordre 4. Quel est l'ordre de convergence ?*

2.4. **Méthode d'Euler implicite.**

Theorem 2.20 (Ordre de la méthode d'Euler implicite)**.** *On suppose que f est L-Lipschitzienne en sa seconde variable. Pour $h < 1/L$, la méthode d'Euler implicite est convergente d'ordre 1.*

Démonstration. On définit d'abord l'erreur de consistance

$$\sigma_{n+1} := y_{n+1} - \big(y_n + hf(t_{n+1}, y_{n+1})\big).$$

Un développement de Taylor de y autour de t_{n+1} donne

$$y_n = y_{n+1} - hy'(t_{n+1}) + \frac{h^2}{2}y''(\beta_{n+1}),$$

d'où

$$\sigma_{n+1} = -\frac{h^2}{2}y''(\beta_{n+1}), \qquad |\sigma_{n+1}| \leqslant Ch^2.$$

On introduit l'erreur $e_n := u_n - y_n$. En soustrayant l'identité vérifiée par y et le schéma numérique, on obtient :

$$e_{n+1} = e_n + h\big(f(t_{n+1}, u_{n+1}) - f(t_{n+1}, y_{n+1})\big) - \sigma_{n+1}.$$

Par hypothèse que f est L-Lipschitz,

$$|e_{n+1}| \leqslant |e_n| + hL\,|e_{n+1}| + |\sigma_{n+1}|,$$

donc

$$(1 - hL)|e_{n+1}| \leqslant |e_n| + |\sigma_{n+1}|.$$

Pour h assez petit tel que $1 - hL > 0$, on obtient

$$|e_{n+1}| \leqslant \frac{1}{1-hL}\big(|e_n| + |\sigma_{n+1}|\big) \leqslant (1 + C_1 h)|e_n| + C_2 h^2,$$

avec des constantes C_1, C_2 indépendantes de h. On termine en appliquant le Corollaire 2.14. □

Proposition 2.21 (Ordre de la méthode de Crank–Nicolson)**.** *Supposons que $f \in \mathcal{C}^2$. La méthode de Crank–Nicolson est convergente d'ordre 2.*

Démonstration. • On commence par prouver la consistance. On définit l'erreur de consistance

$$\sigma_{n+1} := y_{n+1} - \Big(y_n + \frac{h}{2}\big(f(t_n, y_n) + f(t_{n+1}, y_{n+1})\big)\Big),$$

et on veut connaître son comportement quand h est petit.

On part de la formulation intégrale de l'EDO,

$$y_{n+1} = y_n + \int_{t_n}^{t_{n+1}} f\big(s, y(s)\big)\, ds.$$

La formule du trapèze appliquée à l'intégrale donne en fait l'erreur de consistance

$$\int_{t_n}^{t_{n+1}} f\big(s, y(s)\big)\, ds = \frac{h}{2}\Big(f\big(t_n, y_n\big) + f\big(t_{n+1}, y_{n+1}\big)\Big) + \sigma_{n+1},$$

et on veut connaître l'ordre de σ_{n+1} en h. On pose

$$g(s) := f\big(s, y(s)\big).$$

On a $g \in \mathcal{C}^2$ sur l'intervalle considéré. On rappelle que $t_{n+1} = t_n + h$. On a

$$\sigma_{n+1} = \int_{t_n}^{t_{n+1}} g(s)\mathrm{d}s - \frac{h}{2}\Big(g(t_n) + g(t_{n+1})\Big). \tag{16}$$

Pour tout $s \in [t_n, t_{n+1}]$, il existe $\beta_s \in [t_n, t_{n+1}]$ tel que

$$g(s) = g(t_n) + (s - t_n)g'(t_n) + \frac{(s-t_n)^2}{2} g''(\beta_s).$$

On intègre de t_n à t_{n+1}, en posant $\nu = s - t_n$, on obtient

$$\begin{aligned}\int_{t_n}^{t_{n+1}} g(s)\, ds &= \int_0^h \Big[g(t_n) + \nu g'(t_n) + \frac{\nu^2}{2} g''(\beta_{t_n+\nu})\Big]\, d\nu \\ &= h\, g(t_n) + \frac{h^2}{2} g'(t_n) + \frac{1}{2}\int_0^h \nu^2 g''(\beta_{t_n+\nu})\, d\nu.\end{aligned}$$

On applique Taylor-Lagrange en t_n, il existe $\eta_n \in [t_n, t_{n+1}]$ tel que

$$g(t_{n+1}) = g(t_n) + h g'(t_n) + \frac{h^2}{2} g''(\eta_n),$$

et alors

$$\frac{h}{2}\big(g(t_n) + g(t_{n+1})\big) = h\, g(t_n) + \frac{h^2}{2} g'(t_n) + \frac{h^3}{4} g''(\eta_n).$$

En reformant (16) on a

$$\sigma_{n+1} = \frac{1}{2}\int_0^h \tau^2 g''(\beta_{t_n+\tau})\, d\tau - \frac{h^3}{4} g''(\eta_n).$$

Or, g'' est bornée sur $[0,T]$, donc $\left|g''(s)\right| \leqslant M$ pour tout $s \in [0,T]$. Alors

$$\left|\frac{1}{2}\int_0^h \tau^2 g''(\beta_{t_n+\tau})\,d\tau\right| \leqslant \frac{1}{2}M\int_0^h \tau^2\,d\tau = \frac{Mh^3}{6},$$

et M est indépendant de T et de n. On en déduit qu'il existe une constante $C_T > 0$, indépendante de h et de n, telle que

$$|\tau_{n+1}| \leqslant Ch^3.$$

• Prouvons maintenant la convergence. On introduit l'erreur $e_n := u_n - y_n$. En soustrayant l'identité vérifiée par y et le schéma numérique, on obtient

$$e_{n+1} = e_n + \frac{h}{2}\big(f(t_n,u_n) - f(t_n,y_n) + f(t_{n+1},u_{n+1}) - f(t_{n+1},y_{n+1})\big) - \tau_{n+1}.$$

Par hypothèse que f est L-Lipschitz,

$$|e_{n+1}| \leqslant |e_n|\left(1+\frac{h}{2}L\right) + \frac{h}{2}L|e_{n+1}| + |\tau_{n+1}|.$$

et

$$\left(1-\frac{h}{2}L\right)|e_{n+1}| \leqslant |e_n|\left(1+\frac{h}{2}L\right) + |\tau_{n+1}|.$$

Pour $h < 2/L$, on a

$$|e_{n+1}| \leqslant |e_n|\frac{1+\frac{h}{2}L}{1-\frac{h}{2}L} + |\tau_{n+1}| \leqslant |e_n|\,(1+C_1h) + Ch^3,$$

on termine en utilisant Grönwall discret de la même manière que pour la méthode d'Euler implicite. □

2.5. **Stabilité.**

2.5.1. *Définition.* Les erreurs viennent de plusieurs sources

- erreur d'arrondi données par la précision machine, qui n'est pas strictement nulle
- erreur sur la donnée initiale
- erreurs de troncature du fait du pas fini

Un schéma numérique est stable si ces pertubations ne produisent pas une divergence de la solution numérique à temps long.

Definition 2.22 (Stabilité)**.** *Une méthode numérique est dite absolument stable si, à pas h fixé,*

$$|u_n| \underset{t_n\to+\infty}{\longrightarrow} 0. \tag{17}$$

Une méthode numérique est dite stable si pour tout $T > 0$ il existe C_T tel que pour tout $n \in \{0,\dots,N_h\}$,

$$|u_n| \leqslant C_T\,|u_0|\,, \tag{18}$$

où C_T est indépendant de et de u_0.

Dans (17), on remarque que $t_n \to +\infty$ est équivalent à $n \to +\infty$.

2.5.2. *Stabilité pour $y' = \lambda y$.* On va appliquer cette notion au problème de Cauchy

$$\begin{cases} y'(t) = \lambda y(t), & t > 0 \\ y(0) = 1, \end{cases} \tag{19}$$

où $\lambda \in \mathbb{C}$. On sait que la solution exacte est $y(t) = e^{\lambda t}$ et que $y(t) \underset{t\to+\infty}{\longrightarrow} 0$ si et seulement si $\operatorname{Re}\lambda < 0$. Et dans ce cas, le problème est stable au sens de la Définition 2.3.

Proposition 2.23. *On considère le cas* (19) *où $\lambda \in \mathbb{C}$, et $h > 0$. On a que*

- *Euler explicite est absolument stable si et seulement si*

$$\operatorname{Re}\lambda < 0, \qquad et \qquad h < -\frac{2\operatorname{Re}\lambda}{|\lambda|^2}. \tag{20}$$

- *Euler implicite est absolument stable si et seulement si*

$$\operatorname{Re}\lambda < 0.$$

- *la méthode du trapèze est absolument stable si et seulement si*

$$\operatorname{Re}\lambda < 0.$$

- *pour $\lambda \in \mathbb{R}$, la méthode de Heun est absolument stable si et seulement si*

$$\lambda < 0, \qquad h < -\frac{2}{\lambda}.$$

Pour $\lambda \in \mathbb{R}$, les régions de stabilité d'Euler explicite et de Heun sont les mêmes. On voit qu'il semble falloir ajouter des conditions pour la stabilité des schémas explicites, alors qu'il y a besoin de moins de conditions pour la stabilité des schémas implicites.

Démonstration. Le schéma d'Euler explicite s'écrit

$$u_{n+1} = u_n + h\lambda u_n = (1 + h\lambda)\, u_n.$$

La solution numérique est donc

$$u_n = (1 + h\lambda)^n u_0,$$

On définit $z := h\lambda$, le facteur d'amplification est $R(z) = 1 + z$. La condition de stabilité est donc $|R(z)| < 1$. Or, en passant au carré, on calcule

$$|R(z)| = 1 + 2h\operatorname{Re}\lambda + h^2\,|\lambda|^2$$

et on voit que la condition est équivalente à

$$h\left(h\,|\lambda|^2 + 2\operatorname{Re}\lambda\right) < 0,$$

ce qui est équivalent à (20).

Pour le schéma implicite

$$u_{n+1} = u_n + h\lambda u_{n+1}, \qquad \text{donc} \qquad u_{n+1} = \frac{1}{1 - h\lambda}\, u_n$$

et la solution numérique est

$$u_n = \left(\frac{1}{1 - h\lambda}\right)^n u_0.$$

Le facteur d'amplification est $R(z) = \frac{1}{1-z}$, la condition de stabilité s'écrit $|R(z)| \leqslant 1$. Comme précédemment, on calcule que $|R(h\lambda)| < 1$ si et seulement si $h > \frac{2\,\mathrm{Re}\,\lambda}{|\lambda|^2}$. La méthode d'Euler implicite est ainsi absolument stable pour tout $h > 0$ dès que $\mathrm{Re}\,\lambda < 0$.

Pour la méthode du trapèze,

$$u_{n+1} = u_n + \frac{h}{2}\big(f(t_n, u_n) + f(t_{n+1}, u_{n+1})\big) = u_n + \frac{h\lambda}{2}\big(u_n + u_{n+1}\big).$$

On obtient ainsi

$$\left(1 - \frac{\lambda h}{2}\right) u_{n+1} = \left(1 + \frac{\lambda h}{2}\right) u_n, \qquad u_{n+1} = \frac{1 + \frac{\lambda h}{2}}{1 - \frac{\lambda h}{2}}\, u_n.$$

et

$$u_n = \left(\frac{1 + \frac{\lambda h}{2}}{1 - \frac{\lambda h}{2}}\right)^n u_0.$$

Le facteur d'amplification est

$$R(z) = \frac{1 + \frac{z}{2}}{1 - \frac{z}{2}},$$

et on calcule que $|R(\lambda h)| < 1$ si et seulement si $\mathrm{Re}\,\lambda < 0$.

Pour la méthode de Heun,

$$\begin{aligned} u_{n+1} &= u_n + \frac{h}{2}\big(2\lambda u_n + h\lambda^2 u_n\big) = u_n + h\lambda u_n + \frac{h^2\lambda^2}{2} u_n \\ &= \Big(1 + h\lambda + \tfrac{1}{2}(h\lambda)^2\Big) u_n. \end{aligned}$$

La solution numérique peut donc s'écrire

$$u_n = R(h\lambda)^n u_0.$$

où le facteur d'amplification est $R(z) = 1 + z + \frac{z^2}{2}$. Dans le cas où $\lambda \in \mathbb{R}$, on a que $|R(\lambda h)| < 1$ si et seulement si

$$-1 < 1 + h\lambda + \frac{(h\lambda)^2}{2} < 1$$

ce qui est équivalent à

$$-4 < h^2\lambda^2 + 2h\lambda < 0,$$

il faut résoudre deux inégalités quadratiques. L'inégalité de droite est vérifiée si et seulement si $h < -\frac{2}{\lambda}$ et l'inégalité de gauche est toujours vérifiée. □

2.6. **Méthode de Runge-Kutta.** Le but est d'introduire une classe de méthodes plus précises que les méthodes d'Euler explicite et implicite, à un ordre plus élevé.

Le principe est alors de construire des valeurs approchées u_k au temps t_n pour chaque $0 \leq n \leq N_h$ suivant le schéma à un pas

$$u_0 = y_0, \quad \text{et} \quad u_{n+1} = \Phi(t_n, u_n, h),$$

dans laquelle la fonction Φ caractérise la méthode considérée. Pour la méthode d'Euler explicite, cette fonction est donnée par $\Phi(t,y,h) = y + hf(t,y)$. La méthode de Runge-Kutta d'ordre deux est définie par

$$\Phi(t,y,h) = y + hf\left(t+\tfrac{h}{2}, y+\tfrac{h}{2}f(y,t)\right),$$

tandis que la méthode de Runge-Kutta d'ordre quatre est donnée par

$$\Phi(t,y,h) = \frac{h}{6}\big(\boldsymbol{n}_1 + 2\boldsymbol{n}_2 + 2\boldsymbol{n}_3 + \boldsymbol{n}_4\big),$$

où

$$\begin{aligned}
\boldsymbol{n}_1 &= f(y,t),\\
\boldsymbol{n}_2 &= f\big(t+\tfrac{h}{2}, y+\tfrac{h}{2}\boldsymbol{n}_1\big),\\
\boldsymbol{n}_3 &= f\big(t+\tfrac{h}{2}, y+\tfrac{h}{2}\boldsymbol{n}_2\big),\\
\boldsymbol{n}_4 &= f(t+h, y+h\boldsymbol{n}_3).
\end{aligned}$$

Plus généralement, une méthode de Runge-Kutta d'ordre s est donnée par les formules

$$\Phi(t,y,h) = h\sum_{i=1}^{s} b_i \boldsymbol{n}_i,$$

où

$$\begin{aligned}
&\boldsymbol{n}_1 = f(t,y),\\
&\boldsymbol{n}_2 = f(t+c_2h, y+ha_{21}\boldsymbol{n}_1),\\
&\boldsymbol{n}_3 = f(t+c_3h, y+h(a_{31}\boldsymbol{n}_1 + a_{32}\boldsymbol{n}_2)),\\
&\ldots = \ldots,\\
&\boldsymbol{n}_s = f(t+c_sh, y+h(a_{s1}\boldsymbol{n}_1 + a_{s2}\boldsymbol{n}_2 + \ldots + a_{s,s-1}\boldsymbol{n}_{s-1})).
\end{aligned}$$

Les coefficients $(a_{ij})_{1\le j<i\le s}$, $(c_i)_{2\le i\le s}$, et $(b_i)_{1\le i\le s}$ sont souvent représentés par un tableau dit de Butcher

$$\begin{array}{c|ccccc}
0 & & & & & \\
c_2 & a_{21} & & & & \\
c_3 & a_{31} & a_{32} & & & \\
\vdots & \vdots & & \ddots & & \\
c_s & a_{s1} & a_{s2} & \cdots & a_{s,s-1} & \\
\hline
 & b_1 & b_2 & \cdots & b_{s-1} & b_s
\end{array}$$

Par exemple, les tableaux de Butcher des méthodes d'ordre 2 est

$$\begin{array}{c|cc}
0 & & \\
\frac{1}{2} & \frac{1}{2} & \\
\hline
 & 0 & 1
\end{array}$$

et celle d'ordre 4 est

$$\begin{array}{c|cccc}
0 & & & & \\
\frac{1}{2} & \frac{1}{2} & & & \\
\frac{1}{2} & 0 & \frac{1}{2} & & \\
1 & 0 & 0 & 1 & \\
\hline
 & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6}
\end{array}$$

De plus, la méthode de Heun est une méthode de Runge-Kutta avec tableau

$$\begin{array}{c|cc} 0 & & \\ 1 & 1 & \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}$$

3. Approximation numérique d'équations aux dérivées partielles

3.1. Schéma θ pour l'équation de la chaleur.

3.1.1. *Présentation.* On étudie l'équation de la chaleur unidimensionnelle sur $\mathbb{R}_+ \times [0,1]$. Elle s'écrit, pour un paramètre $\nu > 0$ appelé coefficient de diffusion,

$$\frac{\partial y}{\partial t} - \nu \frac{\partial^2 y}{\partial x^2} = 0, \qquad y(\cdot,0) = y(\cdot,1) = 0, \qquad y(0,x) \underset{\forall x \in [0,1]}{=} y_0(x). \quad (21)$$

On suppose connue l'existence de la solution classique et on peut montrer qu'elle est $\mathcal{C}^\infty$. Notre but est d'approcher numériquement la solution avec un schéma aux différences finies appelé *θ-schéma*.

On se donne une discrétisation en temps

$$t_n = n\Delta t, \quad n \in \mathbb{N}$$

et en espace

$$x_j = j\Delta x, \qquad j \in \{0, \dots, J+1\}, \qquad \Delta x = \frac{1}{J+1}$$

On note y la solution exacte de (21), on définit $y_j^n := y(t_n, x_j)$, et on note

$$u_j^n \simeq y(t_n, x_j)$$

une approximation de y_j^n. On aura donc les conditions de bord

$$u_0^n = u_{J+1}^n = 0.$$

On note $u^n := (u_j^n)_{1 \leqslant j \leqslant J} \in \mathbb{R}^J$ le vecteur contenant toute l'approximation pour un temps donné.

3.1.2. *Définition du θ-schéma.* Pour $\theta \in [0,1]$, on définit

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} = \nu\theta \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{(\Delta x)^2} + \nu(1-\theta)\frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2}. \quad (22)$$

On voit que

- le cas $\theta = 0$ est un schéma explicite, à partir de u^n on peut calculer u^{n+1},
- les cas $\theta \in]0,1]$ sont des schémas implicites,
- le cas $\theta = 1$ correspond à Euler implicite,
- le cas $\theta = \frac{1}{2}$ est le cas Crank-Nicolson pour la discrétisation temporelle.

Definition 3.1 (Stencil)**.** *Le stencil pour (n,j) est l'ensemble des (m,k) qu'il faut connaître pour pouvoir calculer u_j^n.*

On rappelle la définition de la norme euclidienne ℓ^2 dans $\mathbb{R}^n$,

$$\|v\|_{\ell^2} := \sqrt{\sum_{j=1}^{n} |v_j|^2}.$$

Theorem 3.2. *Soit $T > 0$, on considère l'équation différentielle sur $t \in [0,T]$ et $n \in \{0,\dots,N\}$ où $N\Delta t \leqslant T \leqslant (N+1)\Delta t$. Pour la norme ℓ^2, le θ-schéma est convergent*

- *si $\theta \geqslant \frac{1}{2}$,*
- *si $\theta < \frac{1}{2}$ sous la condition CFL*

$$(1-2\theta)\frac{2\nu\Delta t}{(\Delta x)^2} \leqslant 1. \tag{23}$$

Il est d'ordre 2 en espace. Il est d'ordre 1 en temps si $\theta \neq \frac{1}{2}$, et d'ordre 2 en temps si $\theta = \frac{1}{2}$. Plus précisément, sous la condition CFL, on a

$$\max_{0\leqslant n\leqslant N} \|u^n - y^n\|_{\ell^2} \leqslant C_T(\Delta x)^2 + C_T \begin{cases} \Delta t & \text{si } \theta \neq \frac{1}{2} \\ (\Delta t)^2 & \text{si } \theta = \frac{1}{2}, \end{cases} \tag{24}$$

où C_T ne dépend pas de n ni de Δt ni de Δx, mais dépend de T. De plus, il est stable au sens où pour tout $n \leqslant N$,

$$\|u^n\|_{\ell^2} \leqslant K_T \left\|u^0\right\|_{\ell^2},$$

où K_T ne dépend pas de n ni de Δt ni de Δx ni de u^0, mais dépend de T.

3.1.3. *Le schéma est bien défini.* Définissons

$$A := \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix} \in \mathbb{R}^{J\times J}, \qquad \beta := \frac{\nu\Delta t}{(\Delta x)^2}.$$

Pour tout n, j, on a $(Au^n)_j = u^n_{j+1} - 2u^n_j + u^n_{j-1}$. Le schéma (22) se réécrit

$$(I + \theta\beta A)\, u^{n+1} = (I - (1-\theta)\beta A)\, u^n.$$

Le schéma est bien défini si $I+\theta\beta A$ est inversible, car alors u^{n+1} est calculable par

$$u^{n+1} = (I + \theta\beta A)^{-1} (I - (1-\theta)\beta A)\, u^n.$$

Montrons que $I + \theta\beta A$ est inversible.

Lemma 3.3. *Pour $p \in \{1,\dots,J\}$, le vecteur*

$$V^p := \left(\sin\left(\tfrac{jp\pi}{J+1}\right)\right)_{1\leqslant j\leqslant J}$$

est vecteur propre de A pour la valeur propre

$$\lambda_p = 4\left(\sin\left(\tfrac{p\pi}{2(J+1)}\right)\right)^2. \tag{25}$$

Démonstration. Par définition de A,

$$\begin{aligned}(AV^p)_j &= 2(V^p)_j - (V^p)_{j-1} - (V^p)_{j+1}\\ &= 2\sin\Big(\tfrac{jp\pi}{J+1}\Big) - \sin\Big(\tfrac{(j-1)p\pi}{J+1}\Big) - \sin\Big(\tfrac{(j+1)p\pi}{J+1}\Big).\end{aligned}$$

On utilise que pour tout $x \in \mathbb{R}$,

$$\sin((j+1)x) + \sin((j-1)x) = 2\sin(jx)\cos(x).$$

En prenant $x = \frac{p\pi}{J+1}$, on obtient

$$(AV^p)_j = 2\Big(1 - \cos\Big(\tfrac{p\pi}{J+1}\Big)\Big)\sin\Big(\tfrac{jp\pi}{J+1}\Big) = \lambda_p (V^p)_j$$

où $\lambda_p = 2\Big(1 - \cos\Big(\frac{p\pi}{J+1}\Big)\Big)$, car $1 - \cos x = 2\left(\sin\Big(\frac{x}{2}\Big)\right)^2$. □

Ainsi, en notant $\sigma(B)$ le spectre de B pour toute matrice B, on a

$$\begin{aligned}\min \sigma\left(I + \theta\beta A\right) &= I + \theta\beta \min \sigma\left(A\right)\\ &= 1 + 4\theta\beta\left(\sin\Big(\tfrac{\pi}{2(J+1)}\Big)\right)^2 \geqslant 1 > 0,\end{aligned}$$

donc la matrice est inversible et le schéma bien défini.

3.1.4. *Consistance du schéma.* On remarque d'abord que

$$\frac{y_j^{n+1} - y_j^n}{\Delta t} = \frac{\partial y}{\partial t}(t_n, x_j) + \frac{\Delta t}{2}\frac{\partial^2 y}{\partial t^2}(t_n, x_j) + O((\Delta t)^2).$$

Puis

$$\begin{aligned}&\frac{y_{j+1}^n - 2y_j^n + y_{j-1}^n}{(\Delta x)^2} = \frac{y_{j+1}^n - y_j^n - \Big(y_j^n - y_{j-1}^n\Big)}{(\Delta x)^2}\\ &= \frac{1}{\Delta x}\left(\frac{\partial y}{\partial x}(t_n, x_j) + \frac{\Delta x}{2}\frac{\partial^2 y}{\partial x^2}(t_n, x_j) + \frac{(\Delta x)^2}{3!}\frac{\partial^3 y}{\partial x^3}(t_n, x_j) + O((\Delta x)^3)\right)\\ &\quad - \frac{1}{\Delta x}\left(\frac{\partial y}{\partial x}(t_n, x_j) + \frac{\Delta x}{2}\frac{\partial^2 y}{\partial x^2}(t_n, x_j) - \frac{(\Delta x)^2}{3!}\frac{\partial^3 y}{\partial x^3}(t_n, x_j) + O((\Delta x)^3)\right)\\ &= \frac{\partial^2 y}{\partial x^2}(t_n, x_j) + O((\Delta x)^2).\end{aligned}$$

De même,

$$\begin{aligned}\frac{y_{j+1}^{n+1} - 2y_j^{n+1} + y_{j-1}^{n+1}}{(\Delta x)^2} &= \frac{\partial^2 y}{\partial x^2}(t_{n+1}, x_j) + O((\Delta x)^2)\\ &= \frac{1}{\nu}\frac{\partial y}{\partial t}(t_{n+1}, x_j) + O((\Delta x)^2)\\ &= \frac{1}{\nu}\left(\frac{\partial y}{\partial t}(t_n, x_j) + \Delta t\frac{\partial^2 y}{\partial t^2}(t_n, x_j)\right) + O((\Delta x)^2 + (\Delta t)^2)\end{aligned}$$

On note $y^n := (y_j^n)_{1\leqslant j\leqslant J}$. En remplaçant tous ces termes dans le schéma, on obtient l'erreur de consistance en (t_n, x_j).

$$\begin{aligned}
\sigma_j^n &:= \frac{1}{\Delta t}\left((I+\theta\beta A)\, y^{n+1} - (I-(1-\theta)\beta A)\, y^n\right)_j \\
&= \frac{y_j^{n+1}-y_j^n}{\Delta t} - \left(\nu\theta\frac{y_{j+1}^{n+1}-2y_j^{n+1}+y_{j-1}^{n+1}}{(\Delta x)^2} + \nu(1-\theta)\frac{y_{j+1}^n-2y_j^n+y_{j-1}^n}{(\Delta x)^2}\right) \\
&= \frac{\partial y}{\partial t}(t_n,x_j) + \frac{\Delta t}{2}\frac{\partial^2 y}{\partial t^2}(t_n,x_j) - \theta\frac{\partial y}{\partial t}(t_n,x_j) - \theta\Delta t\frac{\partial^2 y}{\partial t^2}(t_n,x_j) \\
&\quad - \nu(1-\theta)\frac{\partial^2 y}{\partial x^2}(t_n,x_j) + O((\Delta x)^2+(\Delta t)^2) \\
&= (1-\theta)\left(\frac{\partial y}{\partial t} - \nu\frac{\partial^2 y}{\partial x^2}\right)(t_n,x_j) + \Delta t\left(\tfrac{1}{2}-\theta\right)\frac{\partial^2 y}{\partial t^2}(t_n,x_j) + O((\Delta x)^2+(\Delta t)^2) \\
&= \Delta t\left(\tfrac{1}{2}-\theta\right)\frac{\partial^2 y}{\partial t^2}(t_n,x_j) + O((\Delta x)^2+(\Delta t)^2).
\end{aligned}$$

On a

$$\left|\sigma_j^n\right| = O((\Delta x)^2) + \begin{cases} O(\Delta t) & \text{si } \theta \neq \frac{1}{2} \\ O((\Delta t)^2) & \text{si } \theta = \frac{1}{2} \end{cases}$$

et donc

$$\max_{0\leqslant n\leqslant N}\|\sigma^n\| = O((\Delta x)^2) + \begin{cases} O(\Delta t) & \text{si } \theta \neq \frac{1}{2} \\ O((\Delta t)^2) & \text{si } \theta = \frac{1}{2} \end{cases} \tag{26}$$

donc le schéma est consistant, d'ordre donné par (26).

3.1.5. *Stabilité du schéma.* On rappelle que pour toute matrice M symmétrique et tout vecteur v, on a

$$\|Mv\|_{\ell^2} \leqslant \|v\|_{\ell^2}\max\{|\lambda| \mid \lambda\in\sigma(M)\}.$$

On définit

$$B := (I+\theta\beta A)^{-1}(I-(1-\theta)\beta A), \qquad \text{on a} \qquad u^{n+1} = Bu^n.$$

Comme A est symétrique, pour toute fonction f lisse en les λ_p, $p\in\{1,\dots,J\}$, les valeurs propres de $f(A)$ sont les $f(\lambda_p)$. En se rappelant la définition (25) de λ_p, on a que les J valeurs propres de B sont, pour $j\in\{1,\dots,J\}$,

$$\mu_p := \frac{1-(1-\theta)\beta\lambda_p}{1+\theta\beta\lambda_p}, \qquad \mu := \max_{1\leqslant p\leqslant J}|\mu_p|.$$

Comme $\theta\geqslant 0$, alors $1+\theta\beta\lambda_p \geqslant 1 > 0$. La méthode est stable dans la norme euclidienne ℓ^2 si et seulement si

$$\forall p\in\{1,\dots,J\}, \qquad |\mu_p|\leqslant 1. \tag{27}$$

On rappelle la définition de la norme d'opérateur pour ℓ^2, pour toute matrice M,

$$\|M\|_{\ell^2\to\ell^2} := \sup_{v\in\mathbb{R}^{J+1}}\frac{\|Mv\|_{\ell^2}}{\|v\|_{\ell^2}}.$$

Elle respecte, pour toutes matrices M et P, $\|MP\|_{\ell^2\to\ell^2} \leqslant \|M\|_{\ell^2\to\ell^2}\|P\|_{\ell^2\to\ell^2}$. On a toujours $u^n = B^n u^0$ et donc dans ce cas,

$$\|u^n\|_{\ell^2} \leqslant \|B^n\|_{\ell^2\to\ell^2}\left\|u^0\right\| \leqslant \|B\|^n_{\ell^2\to\ell^2}\left\|u^0\right\|,$$

donc $\|u^n\|_{\ell^2}$ reste borné si et seulement si $\|B\|_{\ell^2\to\ell^2}$, ce qui est équivalent à (27). Or, (27) équivaut à

$$(1+\theta\beta\lambda_p)^2 - (1-(1-\theta)\beta\lambda_p)^2 \geqslant 0,$$

On calcule donc

$$(1+\theta\beta\lambda_p)^2 - (1-(1-\theta)\beta\lambda_p)^2 = \beta\lambda_p\left(2-(1-2\theta)\beta\lambda_p\right).$$

Comme $\beta\lambda_p \geqslant 0$, la stabilité équivaut à $2 \geqslant (1-2\theta)\beta\lambda_p$.

Si $\theta \geqslant \frac{1}{2}$, alors l'inégalité est toujours vérifiée. Si $\theta < \frac{1}{2}$, alors $1-2\theta > 0$ et la condition devient

$$\beta\lambda_p < \frac{2}{1-2\theta}.$$

On a que

$$\max_{1\leqslant p\leqslant J}\lambda_p = 4\sin^2\Big(\frac{J\pi}{2(J+1)}\Big) < 4.$$

Ainsi, si on a la condition CFL

$$4\beta \leqslant \frac{2}{1-2\theta},$$

alors

$$\beta\lambda_p \leqslant \frac{\lambda_p}{4}\frac{2}{1-2\theta} < \frac{2}{1-2\theta}.$$

3.1.6. *Convergence du schéma.* On utilisera pour ça le théorème 3.4.

3.2. **Théorème de Lax.** Le théorème de Lax montre que

- stabilité (le schéma ne crée pas d'oscillations rapides)
- et consistance (au niveau de l'EDP discrète, l'erreur entre l'application du schéma à u^n et y^n tend vers 0)

implique convergence.

Theorem 3.4 (Lax : stabilité + consistance $\implies$ convergence)**.** *Soit y la solution suffisamment régulière de l'équation de la chaleur* (21)*. Soit u_j^n la solution numérique discrète obtenue par un schéma de différences finies avec la donnée initiale $u_j^0 = y_0(x_j)$. On prend la norme euclidienne $\|\cdot\|_{\ell^2}$. On suppose que le schéma est*

- *linéaire à deux niveaux*
- *consistant d'ordre p en espace et à l'ordre q en temps pour $\|\cdot\|_{\ell^2}$, où l'erreur de consistance est*

$$\sigma^n := \frac{1}{\Delta t}\left(y^{n+1} - By^n\right)$$

- *stable pour $\|\cdot\|_{\ell^2}$.*

On définit $e^n := u_j^n - y_j^n$. Alors pour tout temps $T > 0$ il existe une constante $C_T > 0$ indépendante de Δx et Δt telle que

$$\max_{0\leqslant t_n\leqslant T}\|e^n\|_{\ell^2} \leqslant C_T\big((\Delta x)^p + (\Delta t)^q\big). \tag{28}$$

On remarque que l'estimation (28) est indépendante du nombre de points de discrétisation J.

Démonstration. Un schéma linéaire à deux niveaux peut s'écrire sous la forme condensée c'est-à-dire

$$u^{n+1} = Bu^n,$$

où B est la matrice d'itération (carrée de taille J). On note $y^n = (y_j^n)_{1\leqslant j\leqslant J}$ avec $y_j^n = y(t_n, x_j)$. Par hypothèse sur la consistance, il existe un vecteur σ^n tel que

$$y^{n+1} = By^n + \Delta t\, \sigma^n$$

avec

$$\|\sigma^n\|_{\ell^2} \leqslant C\big((\Delta x)^p + (\Delta t)^q\big).$$

On obtient

$$e^{n+1} = Be^n - \Delta t\, \sigma^n,$$

d'où, par récurrence,

$$e^n = B^n e^0 - \Delta t \sum_{k=1}^{n} B^{n-k}\sigma^{k-1}.$$

Or, la stabilité du schéma veut dire que

$$\|u^n\|_{\ell^2} = \left\|B^n u^0\right\|_{\ell^2} \leqslant K \left\|u^0\right\|_{\ell^2}$$

pour toute donnée initiale, c'est-à-dire que $\|B^n\|_{\ell^2\to\ell^2} \leqslant K$ où la constante K ne dépend pas de n. D'autre part, $e^0 = 0$, donc la relation précédente donne

$$\|e^n\|_{\ell^2} \leqslant \Delta t \sum_{k=1}^{n} \left\|B^{n-k}\right\|_{\ell^2} \left\|\sigma^{k-1}\right\|_{\ell^2} \leqslant \Delta t\, n\, KC\big((\Delta x)^p + (\Delta t)^q\big),$$

ce qui fournit l'inégalité voulue avec la constante $C_T = TKC$ (puisque $n\Delta t \leqslant T$). □

Le Théorème de Lax est valable pour toute EDP, pas seulement pour l'équation de la chaleur. Il admet une réciproque au sens où un schéma linéaire consistant à deux niveaux qui converge est nécessairement stable, mais nous ne préciserons pas ce sujet.

3.3. **Le cas multidimensionnel.** Nous donnons ici simplement un aperçu des méthodes multidimensionnelles, sans rentrer dans les détails. Nous pouvons facilement adapter le cas unidimensionel en espace au cas multidimensionel en espace. Considérons $\Omega = (0,1)\times(0,L)$ avec des conditions aux limites de Dirichlet pour le problème exact suivant

$$\begin{cases} \dfrac{\partial y}{\partial t} - \nu\dfrac{\partial^2 y}{\partial x^2} - \nu\dfrac{\partial^2 y}{\partial y^2} = 0, & (x,y,t)\in\Omega\times\mathbb{R}_+^*, \\ y(t=0,x,y) = y_0(x,y), & (x,y)\in\Omega, \\ y(t,x,y) = 0, & t\in\mathbb{R}_+^*,\ (x,y)\in\partial\Omega. \end{cases} \tag{29}$$

On introduit deux discrétisations en espace $\Delta x = 1/(N_x+1) > 0$ et $\Delta y = L/(N_y+1) > 0$, où $N_x, N_y \in \mathbb{N}$. Le pas de temps sera Δt, et les coordonnées sont donc, pour $n \geqslant 0,\ 0 \leqslant j \leqslant N_x+1,\ 0 \leqslant k \leqslant N_y+1$,

$$(t_n, x_j, y_k) = (n\Delta t,\, j\Delta x,\, k\Delta y). \tag{30}$$

On note y la solution exacte de (29), et $u^n_{j,k}$ les valeurs d'une solution approchée. Les conditions aux limites de Dirichlet se traduisent, pour $n > 0$, en

$$u^n_{0,k} = u^n_{N_x+1,k} = 0,\ \forall k, \qquad u^n_{j,0} = u^n_{j,N_y+1} = 0,\ \forall j. \tag{31}$$

La donnée initiale est discrétisée en $u^0_{j,k} = y_0(x_j, y_k) \quad \forall j,k$.

La généralisation au cas bidimensionnel du schéma explicite est évidente

$$\frac{u^{n+1}_{j,k} - u^n_{j,k}}{\Delta t} + \nu\,\frac{-u^n_{j-1,k} + 2u^n_{j,k} - u^n_{j+1,k}}{(\Delta x)^2} + \nu\,\frac{-u^n_{j,k-1} + 2u^n_{j,k} - u^n_{j,k+1}}{(\Delta y)^2} = 0, \tag{32}$$

pour $n \geqslant 0$, $j \in \{1,\dots,N_x\}$ et $k \in \{1,\dots,N_y\}$. La seule différence notable avec le cas unidimensionnel est le caractère deux fois plus sévère de la condition CFL.

3.4. **Exercices.**

3.4.1. *Advection.* On considère l'équation d'advection linéaire à vitesse constante

$$\partial_t y + a\,\partial_x y = 0, \qquad a \in \mathbb{R},$$

et le schéma explicite centré associé

$$u^{n+1}_i = u^n_i - \lambda\big(u^n_{i+1} - u^n_{i-1}\big), \qquad \lambda := \frac{a\Delta t}{\Delta x}.$$

On définit l'erreur de consistance

$$\sigma^n_i := \frac{y^{n+1}_i - y^n_i}{\Delta t} + a\,\frac{y^n_{i+1} - y^n_{i-1}}{2\Delta x}.$$

Montrer que le schéma est consistant d'ordre 1 en temps et 2 en espace (on ne demande pas ceci au niveau de la convergence de la solution mais au niveau de la consistance).

3.4.2. *Schéma de Gear.* On considère l'équation de la chaleur (21) et le schéma de Gear

$$\frac{3u^{n+1}_i - 4u^n_i + u^{n-1}_i}{2\Delta t} + \nu\frac{-u^{n+1}_{i-1} + 2u^{n+1}_i - u^{n+1}_{i+1}}{(\Delta x)^2} = 0.$$

Montrer qu'il est d'ordre 2 en espace et en temps.

4. Méthode des volumes finis

La méthode des volumes finis est utilisée quand il existe une quantité conservée et lorsqu'on veut que cette propriété soit exactement respectée par le schéma numérique. On montrera un tel schéma sur l'exemple suivant.

4.1. **Équation de transport non linéaire.** Soit $y_0 \in \mathcal{C}^\infty(\mathbb{R}, \mathbb{R})$ à support compact, c'est-à-dire que

$$\exists r \geqslant 0,\ \forall x \in]-\infty, -r] \cup [r, +\infty[, \quad y_0(x) = 0.$$

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction de flux (au moins $\mathcal{C}^1$). On considère l'équation de transport linéaire à vitesse constante

$$\begin{cases} \partial_t y(t,x) + \partial_x \left(f\left(y(t,x)\right)\right) = 0, & x \in \mathbb{R},\ t > 0 \\ y(x,0) = y_0(x). \end{cases} \tag{33}$$

Quand pour tout $x \in \mathbb{R}$, $f(x) = ax$, où $a \in \mathbb{R}$, on a l'équation de transport linéaire et la solution exacte est $y(t,x) = y_0(at - x)$.

4.2. **Forme intégrale et quantité conservée.** Dans l'équation exacte (33), la quantité conservée est la masse totale

$$M(t) := \int_{\mathbb{R}} y(t,x)\mathrm{d}x \in \mathbb{R}.$$

Lemma 4.1 (Conservation de la masse pour la solution exacte)**.** *Pour tout $t \geqslant 0$, $M(t) = M(0)$.*

Démonstration. Soit $x_1, x_2 \in \mathbb{R}$ tels que $x_1 < x_2$. En intégrant l'équation (33) sur $[x_1, x_2]$, on a

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} \int_{x_1}^{x_2} y(t,x)\mathrm{d}x = \int_{x_1}^{x_2} \partial_t y(t,x)\,\mathrm{d}x = -\int_{x_1}^{x_2} \partial_x f(y(t,x))\mathrm{d}x \\ = f(y(t,x_1)) - f(y(t,x_2)). \end{aligned}$$

Or, comme y_0 est à support compact, $y(t,\cdot)$ est à support compact pour tout $t \in \mathbb{R}_+$. Donc $y(t,x) \to 0$ quand $x \to \pm\infty$. Faire $x_1 \to -\infty$ et $x_2 \to +\infty$ donne que $\frac{\mathrm{d}}{\mathrm{d}t} M(t) = f(0) - f(0) = 0$. □

4.3. **Maillage, volumes de contrôle et moyennes de cellule.** On introduit un maillage (éventuellement non uniforme) donné par des interfaces

$$\cdots < x_{i-\frac{1}{2}} < x_{i+\frac{1}{2}} < x_{i+\frac{3}{2}} < \cdots,$$

et les cellules (volumes de contrôle)

$$I_i = [x_{i-\frac{1}{2}}, x_{i+\frac{1}{2}}], \qquad \Delta x_i = x_{i+\frac{1}{2}} - x_{i-\frac{1}{2}}.$$

On définit la moyenne sur chaque cellule, qui sera l'inconnue de la méthode

$$y_i(t) := \frac{1}{\Delta x_i} \int_{I_i} y(t,x)\,\mathrm{d}x, \tag{34}$$

et quand tous les Δx_i sont petits, on a bien sûr $y_i(t) \simeq y(t, x_i)$. On remarque aussi que

$$M(t) = \sum_{i \in \mathbb{Z}} y_i(t) \Delta x_i. \tag{35}$$

4.4. **Schéma de volumes finis : origine du schéma.** En intégrant (33) sur I_i, on obtient

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_{I_i} y(t,x)\,\mathrm{d}x + f(y(x_{i+\frac{1}{2}},t)) - f(y(x_{i-\frac{1}{2}},t)) = 0,$$

et donc

$$\frac{\mathrm{d}}{\mathrm{d}t}y_i(t) = -\frac{1}{\Delta x_i}\Big(f(y(x_{i+\frac{1}{2}},t)) - f(y(x_{i-\frac{1}{2}},t))\Big).$$

La quantité $y_i(t)$ représente la quantité de masse dans le domaine I_i. Sa dérivée représente la variation de masse, et elle est donnée par le flux, qui est la somme entre la masse entrante par la gauche de I_i, $\frac{1}{\Delta x_i}f(y(x_{i-\frac{1}{2}},t))$ et la masse entrante par la droite $\frac{1}{\Delta x_i}f(y(x_{i+\frac{1}{2}},t))$. L'idée des volumes finis est d'approximer les flux exacts aux interfaces par un *flux numérique*.

On introduit un flux numérique

$$F : \mathbb{R}\times\mathbb{R}\to\mathbb{R},$$

qui approxime le flux à l'interface lorsque la solution est approximée par des valeurs constantes à gauche et à droite.

Par exemple le flux numérique de Lax–Friedrichs est défini par

$$F(u_L,u_R) = \frac{1}{2}\big(f(u_L)+f(u_R)\big) - \frac{\alpha}{2}(u_R-u_L), \tag{36}$$

où $\alpha > 0$ est tel que $\alpha \geqslant \max_{u\in\mathcal{U}}|f'(u)|$, et $\mathcal{U}$ est un intervalle contenant les valeurs de la solution.

4.5. **Schéma discret d'Euler.** Pour $\Delta t > 0$, on note u_i^n une approximation de $y_i(n\Delta t)$, donc

$$u_i^n \simeq y_i(n\Delta t) \simeq y_i^n$$

et on pose

$$F^n_{i+\frac{1}{2}} := F(u_i^n, u_{i+1}^n).$$

Le schéma d'Euler explicite est

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x_i}\Big(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\Big). \tag{37}$$

Le schéma d'Euler implicite est

$$u_i^{n+1} = u_i^n - \frac{\Delta t}{\Delta x_i}\big(F^{n+1}_{i+\frac{1}{2}} - F^{n+1}_{i-\frac{1}{2}}\big).$$

Nous allons faire l'analyse d'Euler explicite.

4.6. **Conservation discrète.** Définissons la version discrète de la masse

$$m^n := \sum_{i\in\mathbb{Z}} u_i^n \Delta x_i.$$

Comme $u_i^n \simeq y_i(n\Delta t)$, et par (35), on a

$$m^n \simeq M(n\Delta t).$$

Le schéma a été choisi de manière à ce que cette masse approximée soit conservée.

Proposition 4.2 (Conservation de la masse dans le schéma discret)**.** *Le schéma* (37) *de volumes finis est conservatif au sens où pour tout* $n \in \mathbb{N}$*,*

$$m^n = m^0 = M(0).$$

Démonstration. Soit $J \in \mathbb{N}$. En multipliant le schéma par Δx_i et en sommant sur i, on obtient

$$\begin{aligned}\sum_{-J\leqslant i\leqslant J} u_i^{n+1}\Delta x_i &= \sum_{-J\leqslant i\leqslant J} u_i^n \Delta x_i - \Delta t \sum_{-J\leqslant i\leqslant J} \left(F^n_{i+\frac{1}{2}} - F^n_{i-\frac{1}{2}}\right) \\ &= \sum_{-J\leqslant i\leqslant J} u_i^n \Delta x_i + \Delta t \left(F^n_{-J-\frac{1}{2}} - F^n_{J+\frac{1}{2}}\right). \end{aligned} \tag{38}$$

Or, $F^n_{J+\frac{1}{2}} = F(u_J^n, u_{J+1}^n)$ mais comme y_0 est à support compact, on a que quel que soit $s \in \mathbb{N}$, $u_i^s \to 0$ quand $i \to \pm\infty$. Ainsi, $F^n_{J+\frac{1}{2}} \to F(0,0)$ quand $J \to +\infty$. On a de même $F^n_{-J-\frac{1}{2}} \to F(0,0)$ quand $J \to +\infty$. En faisant $J \to +\infty$ dans (38), on obtient

$$m^{n+1} = m^n.$$

Par ailleurs $m^0 = M(0)$ car initialement $u_i^0 = \int_{I_i} y_0$ pour tout $i \in \mathbb{Z}$. □

4.7. **Consistance.** Nous le faisons sur un maillage uniforme $\Delta x_i = \Delta x$ car le cas non uniforme est similaire.

On définit les valeurs exactes échantillonnées

$$y_i^n := y(t_n, x_i), \qquad x_i = i\Delta x, \quad t_n = n\Delta t.$$

Nous définissons l'erreur de troncature locale

$$\sigma_i^n := \frac{y_i^{n+1} - y_i^n}{\Delta t} + \frac{1}{\Delta x}\left(F(y_i^n, y_{i+1}^n) - F(y_{i-1}^n, y_i^n)\right)$$

On suppose que $f \in \mathcal{C}^2(\mathbb{R})$, le flux numérique $F : \mathbb{R}^2 \to \mathbb{R}$ est *consistant* avec f, c'est-à-dire que

$$F(v, v) = f(v), \qquad \forall v \in \mathbb{R}. \tag{39}$$

On peut vérifier que le flux (36) respecte cette propriété.

Proposition 4.3 (Consistance)**.** *Supposons* (39)*, et* F *est* $\mathcal{C}^1$ *au voisinage de la diagonale* $\{(v,v), v \in \mathbb{R}\}$*. Soit* $y \in \mathcal{C}^2(\mathbb{R}\times[0,T])$ *une solution* régulière *de l'équation exacte* (33)*. Alors pour le schéma Euler explicite* (37)*, on a que pour tout* $T > 0$*, il existe* $C_T > 0$*, indépendant de* Δt*, de* Δx*, de* i *et de* $n \leqslant N$*, tel que*

$$|\sigma_i^n| \leqslant C_T(\Delta t + \Delta x).$$

Démonstration. On a

$$y_i^{n+1} = y_i^n + \Delta t\,(\partial_t y)_i^n + \frac{\Delta t^2}{2}\,(\partial_t^2 y)_i^n + O(\Delta t^3).$$

Donc

$$\frac{y_i^{n+1} - y_i^n}{\Delta t} = (\partial_t y)_i^n + \frac{\Delta t}{2}(\partial_t^2 y)_i^n + O(\Delta t^2).$$

De même,

$$y_{i+1}^n = y_i^n + \Delta x\,(\partial_x y)_i^n + \frac{(\Delta x)^2}{2}(\partial_x^2 y)_i^n + O(\Delta x^3)$$
$$y_{i-1}^n = y_i^n - \Delta x\,(\partial_x y)_i^n + \frac{(\Delta x)^2}{2}(\partial_x^2 y)_i^n + O(\Delta x^3).$$

En particulier,

$$y_{i+1}^n - y_i^n = \Delta x\,(\partial_x y)_i^n + O((\Delta x)^2), \quad y_i^n - y_{i-1}^n = \Delta x\,(\partial_x y)_i^n + O((\Delta x)^2). \tag{40}$$

Développons le flux numérique au voisinage de la diagonale. Comme F est C^1 et que (y_i^n, y_{i+1}^n) et (y_{i-1}^n, y_i^n) sont proches de (y_i^n, y_i^n), et en utilisant la consistance $F(y_i^n, y_i^n) = f(y_i^n)$, on peut écrire des développements de Taylor à l'ordre 1

$$\begin{aligned} F(y_i^n, y_{i+1}^n) &= F(y_i^n, y_i^n) + \partial_2 F(y_i^n, y_i^n)\,(y_{i+1}^n - y_i^n) + O\big((y_{i+1}^n - y_i^n)^2\big), \\ &= f(y_i^n) + \partial_2 F(y_i^n, y_i^n)\,(y_{i+1}^n - y_i^n) + O((\Delta x)^2) \end{aligned}$$

où on a utilisé que $y_{i+1}^n - y_i^n = O(\Delta x)$. De même,

$$\begin{aligned} F(y_{i-1}^n, y_i^n) &= F(y_i^n, y_i^n) + \partial_1 F(y_i^n, y_i^n)\,(y_{i-1}^n - y_i^n) + O\big((y_{i-1}^n - y_i^n)^2\big) \\ &= f(y_i^n) + \partial_1 F(y_i^n, y_i^n)\,(y_{i-1}^n - y_i^n) + O((\Delta x)^2), \end{aligned}$$

Ainsi

$$\begin{aligned} &F(y_i^n, y_{i+1}^n) - F(y_{i-1}^n, y_i^n) \\ &\quad = \partial_2 F(y_i^n, y_i^n)\,(y_{i+1}^n - y_i^n) - \partial_1 F(y_i^n, y_i^n)\,(y_{i-1}^n - y_i^n) + O((\Delta x)^2) \\ &\quad \underset{(40)}{=} \Delta x\big(\partial_1 F(y_i^n, y_i^n) + \partial_2 F(y_i^n, y_i^n)\big)\,(\partial_x y)_i^n + O((\Delta x)^2). \end{aligned}$$

Puisque F est $\mathcal{C}^1$ et vérifie $F(y, y) = f(y)$, alors

$$f'(y) = \frac{\mathrm{d}}{\mathrm{d}y} F(y, y) = \partial_1 F(y, y) + \partial_2 F(y, y)$$

et donc

$$\frac{1}{\Delta x}\Big(F(y_i^n, y_{i+1}^n) - F(y_{i-1}^n, y_i^n)\Big) = f'(y_i^n)\,(\partial_x y)_i^n + O(\Delta x).$$

On a obtenu

$$\sigma_i^n = \Big((\partial_t y)_i^n + O(\Delta t)\Big) + \Big(f'(y_i^n)\,(\partial_x y)_i^n + O(\Delta x)\Big).$$

Or, comme y est solution exacte régulière, on a

$$(\partial_t y)_i^n + (\partial_x f(y))(t_n, x_i) = 0.$$

Mais $(\partial_x f(y))(t_n, x_i) = f'(y)\,(\partial_x y)_i^n$, donc

$$(\partial_t y)_i^n + f'(y_i^n)\,(\partial_x y)_i^n = 0.$$

□

4.8. **Stabilité.** On pose

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

Dans le cas général (flux non linéaire), une notion de stabilité utile en volumes finis est souvent la stabilité ℓ^∞ (principe du maximum).

On suppose que le flux numérique F vérifie :

$$\begin{cases} \partial_1 F(u,v) \geqslant 0, \qquad \partial_2 F(u,v) \leqslant 0 \\ \exists L \geqslant 0 \, \forall u, v \in \mathbb{R}, j \in \{1,2\}, |\partial_j F(u,v)| \leqslant L. \end{cases} \tag{41}$$

Par exemple le flux (36) vérifie ces propriétés.

Proposition 4.4 (Stabilité, principe du maximum discret)**.** *Nous supposons* (39), (41)*, que F est $\mathcal{C}^1$, et la condition CFL*

$$2\lambda L \leqslant 1. \tag{42}$$

Alors pour tout $n \in \mathbb{N}$,

$$\left\| u^{n+1} \right\|_{\ell^\infty} \leqslant \left\| u^n \right\|_{\ell^\infty}.$$

On en déduit facilement la relation de stabilité $\|u^n\|_{\ell^\infty} \leqslant \left\|u^0\right\|_{\ell^\infty}$ pour tout $n \in \mathbb{N}$.

Démonstration. On fixe n et i. On considère la fonction

$$\Phi(\alpha, \beta, \gamma) := \beta - \lambda\Big(F(\beta,\gamma) - F(\alpha,\beta)\Big),$$

le schéma se réécrit

$$u_i^{n+1} = \Phi(u_{i-1}^n, u_i^n, u_{i+1}^n).$$

- Prouvons que Φ est croissante en chacun de ses arguments. On calcule

$$\frac{\partial \Phi}{\partial \alpha} = \lambda \, \partial_1 F(\alpha,\beta) \geqslant 0, \qquad \frac{\partial \Phi}{\partial \gamma} = -\lambda \, \partial_2 F(\beta,\gamma) \geqslant 0,$$
$$\frac{\partial \Phi}{\partial \beta} = 1 - \lambda\Big(\partial_1 F(\beta,\gamma) - \partial_2 F(\alpha,\beta)\Big),$$

Avec (41),

$$\partial_1 F(\beta,\gamma) - \partial_2 F(\alpha,\beta) \leqslant |\partial_1 F(\beta,\gamma)| + |\partial_2 F(\alpha,\beta)| \leqslant 2L.$$

Cela donne

$$\frac{\partial \Phi}{\partial \beta} \geqslant 1 - 2\lambda L \underset{(42)}{\geqslant} 0.$$

- Soient

$$a^n := \min_{j \in \mathbb{Z}} u_j^n, \qquad b^n := \max_{j \in \mathbb{Z}} u_j^n.$$

Alors pour tout $i \in \mathbb{Z}$ et pour tout $n \in \mathbb{N}$,

$$a^n \leqslant u_{i-1}^n, \ u_i^n, \ u_{i+1}^n \leqslant b^n.$$

Comme Φ est croissante en chacun de ses arguments,

$$\Phi(a^n, a^n, a^n) \leqslant \Phi(u_{i-1}^n, u_i^n, u_{i+1}^n) \leqslant \Phi(b^n, b^n, b^n).$$

Mais, par (39), on a $\Phi(c,c,c)=c$, donc

$$a^n \leqslant u_i^{n+1} \leqslant b^n.$$

On en déduit la conclusion. □

4.9. **Convergence.** On se place dans le cas du transport linéaire

$$f(x)=ax$$

avec $a\in\mathbb{R}$ constant.

On suppose une nouvelle condition CFL

$$|a|\lambda \leqslant 1. \tag{43}$$

On note $y(\cdot,t)$ la solution exacte, donnée explicitement par

$$y(x,t)=y_0(x-at).$$

On se restreint également au flux numérique de Lax-Friedrichs (36), avec $\alpha=|a|$, donc

$$F(u_L,u_R)=\frac{a+|a|}{2}u_L+\frac{a-|a|}{2}u_R.$$

On peut prouver la convergence d'ordre 1 en ℓ^∞ en temps fini. On rappelle que

$$\|u^n\|_{\ell^\infty}=\sup_{i\in\mathbb{Z}}\,|u_i^n|\,.$$

Theorem 4.5 (Convergence)**.** *Soit $T>0$ fixé. On fait les hypothèses des Proposition 4.3 et 4.4, et on suppose la condition CFL* (43)*. Alors il existe une constante $C_T>0$ telle que, pour tout n tel que $t_n\leqslant T$,*

$$\|u^n-y^n\|_{\ell^\infty}\leqslant C_T\,(\Delta t+\Delta x).$$

C_T est indépendante de n, de Δt et de Δx.

Démonstration. On a stabilité et consistance du schéma par les résultats précédents. On définit $\nu:=a\lambda$. Pour $a>0$, on a $F(u_L,u_R)=au_L$ et le schéma s'écrit

$$u_i^{n+1}=(1-\nu)\,u_i^n+\nu\,u_{i-1}^n.$$

Pour $a<0$, on a $F(u_L,u_R)=au_R$ et le schéma s'écrit

$$u_i^{n+1}=(1-|\nu|)\,u_i^n+|\nu|\,u_{i+1}^n.$$

On donne seulement la preuve pour $a>0$ puisque le cas $a<0$ est identique en échangeant $i-1$ et $i+1$. On définit l'erreur $e_i^n:=u_i^n-y_i^n$. La consistance se réécrit

$$y_i^{n+1}=(1-\nu)\,y_i^n+\nu\,y_{i-1}^n+\Delta t\,\sigma_i^n,$$

où σ_i^n est bornée uniformément par $|\sigma_i^n|\leqslant C(\Delta t+\Delta x)$, pour $t_n\leqslant T$, grâce à la Proposition (4.3). On obtient l'équation de propagation de l'erreur

$$e_i^{n+1}=(1-\nu)\,e_i^n+\nu\,e_{i-1}^n-\Delta t\,\sigma_i^n.$$

Par la condition CFL (43), on a $0\leqslant\nu\leqslant 1$, donc

$$|e_i^{n+1}|\leqslant(1-\nu)|e_i^n|+\nu|e_{i-1}^n|+\Delta t\,|\sigma_i^n|.$$

En prenant le supremum sur i, on obtient

$$\left\|e^{n+1}\right\|_{\ell^\infty} \leqslant (1-\nu)\left\|e^n\right\|_{\ell^\infty} + \nu\left\|e^n\right\|_{\ell^\infty} + \Delta t\,\left\|\sigma^n\right\|_{\ell^\infty} = \left\|e^n\right\|_{\ell^\infty} + \Delta t\,\left\|\sigma^n\right\|_{\ell^\infty}.$$

On pourrait maintenant utiliser le lemme de Grönwall discret, mais on peut aussi directement itérer . On rappelle que N est tel que $t_N \leqslant T \leqslant t_{N+1}$. On obtient

$$\begin{aligned}\left\|e^n\right\|_{\ell^\infty} &\leqslant \left\|e^0\right\|_{\ell^\infty} + \Delta t \sum_{k=0}^{n-1}\left\|\sigma^k\right\|_{\ell^\infty} = \Delta t\sum_{k=0}^{n-1}\left\|\sigma^k\right\|_{\ell^\infty} \leqslant n\Delta t \max_{0\leqslant k\leqslant N}\left\|\sigma^k\right\|_{\ell^\infty}\\ &\underset{t_n\leqslant T}{\leqslant} T\max_{0\leqslant k\leqslant N}\left\|\sigma^k\right\|_{\ell^\infty} \underset{\substack{\text{Prop}\\(4.3)}}{\leqslant} CT(\Delta t+\Delta x).\end{aligned}$$

□

4.10. **Exercices.**

4.10.1. *Schéma conservant l'énergie.* On considère l'équation d'advection (33) avec $f(y) = ay$, $a > 0$. On suppose que la donnée initiale y_0 est lisse et à support compact. On définit l'énergie

$$E(t) := \frac{1}{2}\int_{\mathbb{R}} y(t,x)^2\,dx.$$

Montrer que cette quantité est conservée, c'est-à-dire que $E(t)$ ne dépend pas de t.

On note les moyennes de cellule $y_i(t)$ comme en (34). On définit

$$u_i^{n+1/2} := \frac{u_i^{n+1}+u_i^n}{2},$$

on choisit un pas spatial $\Delta x > 0$ et on écrit un schéma de volumes finis

$$\frac{u_i^{n+1}-u_i^n}{\Delta t} + a\,\frac{u_{i+1}^{n+1/2}-u_{i-1}^{n+1/2}}{2\Delta x} = 0.$$

On définit l'énergie discrète

$$E^n := \frac{\Delta x}{2}\sum_{i\in\mathbb{Z}}\left|u_i^n\right|^2.$$

Montrer que le schéma donné conserve l'énergie discrète.

References