MÉTHODES NUMÉRIQUES POUR LES EDP

UNIVERSITÉ DE CERGY

Contents

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

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

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

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

Par substitution dans (8), on obtient

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

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 perturbée, 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.**

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

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

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,

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

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, en supposant que $h \leqslant 1$,

$$\begin{aligned}|\Phi(t,y,h) - \Phi(t,z,h)| &\leqslant (1 + Lh\,(1 + Lh/2))\,|y-z| \\ &\leqslant (1 + Lh\,(1 + L/2))\,|y-z|.\end{aligned}$$

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

$$|e_{n+1}| \leqslant \frac{1}{1-hL}\big(|e_n|+|\sigma_{n+1}|\big) \leqslant (1+C_1h)|e_n| + C_2h^2,$$

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

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

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

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

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 n 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 $|R(z)| < 1$. 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 $0 \le n \le N_h$ suivant le schéma à un pas

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

dans lequel 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

$$\begin{aligned}
&\boldsymbol{n}_1 = f(t, y),\\
&\boldsymbol{n}_2 = f(t + c_2 h, y + h a_{21} \boldsymbol{n}_1),\\
&\boldsymbol{n}_3 = f(t + c_3 h, y + h(a_{31}\boldsymbol{n}_1 + a_{32}\boldsymbol{n}_2)),\\
&\ldots = \ldots,\\
&\boldsymbol{n}_s = f(t + c_s h, y + h(a_{s1}\boldsymbol{n}_1 + a_{s2}\boldsymbol{n}_2 + \ldots + a_{s,s-1}\boldsymbol{n}_{s-1})).
\end{aligned}$$

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, le tableau 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 celui 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

$$F(u_L,u_R)=\frac{a+|a|}{2}u_L+\frac{a-|a|}{2}u_R=a\left(\delta_{a>0}u_L+\delta_{a<0}u_R\right).$$

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

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

Montrer que le schéma donné conserve l'énergie discrète.

5. Méthode variationnelle

5.1. **Le problème de Dirichlet.**

5.1.1. *Formulation classique.* Soit Ω un ouvert borné de $\mathbb{R}^d$, $d \geqslant 1$. On considère le problème

$$\begin{cases} -\Delta u = f, & \text{dans } \Omega, \\ u = 0, & \text{sur } \partial\Omega, \end{cases} \tag{44}$$

où $f \in C(\overline{\Omega})$ et $\Delta u = \partial_1^2 u + \partial_2^2 u$, où l'on désigne par $\partial_i^2 u$ la dérivée partielle d'ordre 2 par rapport à la i-ème variable.

Definition 5.1. *On appelle solution classique de* (44) *une fonction* $u \in C^2(\overline{\Omega})$ *qui vérifie* (44).

5.1.2. *Formulation faible.* On rappelle que l'espace $H_0^1(\Omega)$ est défini comme l'adhérence de $C_c^\infty(\Omega)$ dans

$$H^1(\Omega) = \{u \in L^2(\Omega);\ Du \in L^2(\Omega)\},$$

où Du désigne la dérivée faible de u. Par exemple $|\cdot|$ en dimension 1 a une dérivée faible mais n'est pas dérivable en 0. Les conctions discontinues en dimension 1 n'ont pas de dérivée faible, les dérivées distributionnelles sont des Dirac aux points de discontinuité. On rappelle que l'espace $H^1(\Omega)$ muni du produit scalaire

$$(u, v)_{H^1} := \int_\Omega uv + \sum_{i=1}^d \int_\Omega D_i u\, D_i v \tag{45}$$

est un espace de Hilbert. Les espaces $H^1(\Omega)$ et $H_0^1(\Omega)$ font partie des espaces dits "de Sobolev".

L'introduction de solutions plus générales permet de s'affranchir de la régularité C^2, on les appellera "solutions faibles".

Definition 5.2 (Formulation faible)**.** *Soit $f \in L^2(\Omega)$, on dit que u est solution faible de* (44) *si u est solution de*

$$\begin{cases} u \in H_0^1(\Omega), \\ \displaystyle\sum_{i=1}^d \int_\Omega D_i u\, D_i \varphi \quad = \int_\Omega f\, \varphi, \quad \forall \varphi \in H_0^1(\Omega). \end{cases} \tag{46}$$

5.1.3. *Formulation variationnelle.* On définit

$$J(v) := \frac{1}{2} \int_\Omega \nabla v \cdot \nabla v - \int_\Omega f\, v.$$

où on note

$$\int_\Omega \nabla u \cdot \nabla \varphi = \sum_{i=1}^d \int_\Omega D_i u\, D_i \varphi.$$

Definition 5.3 (Formulation variationnelle)**.** *Soit $f \in L^2(\Omega)$. On dit que u est solution variationnelle de* (44) *si u est solution du problème de minimisation*

$$\begin{cases} u \in H_0^1(\Omega), \\ J(u) \leqslant J(v) \quad \forall v \in H_0^1(\Omega). \end{cases} \tag{47}$$

5.1.4. *Classique implique faible.*

Lemma 5.4 (Une solution classique est une solution faible)**.** *Soit u une solution classique de* (44)*. Alors $u \in H_0^1(\Omega)$ et pour tout fonction $\varphi \in H_0^1(\Omega)$, on a*

$$\int_\Omega \nabla u \cdot \nabla \varphi = \int_\Omega f\, \varphi. \tag{48}$$

Proof. Soit $u \in C^2(\overline{\Omega})$ une solution classique de (44), et soit $\varphi \in C_c^\infty(\Omega)$, où $C_c^\infty(\Omega)$ désigne l'ensemble des fonctions de classe C^∞ à support compact

dans Ω. On multiplie (44) par φ et on intègre sur Ω (on appellera par la suite φ "fonction test"), on a donc

$$\int_\Omega -\varphi \Delta u = \int_\Omega f\,\varphi.$$

Notons que ces intégrales sont bien définies, puisque $\Delta u \in C(\Omega)$ et $f \in C(\Omega)$. Par intégration par parties (formule de Green), on a :

$$\begin{aligned}\int_\Omega -\Delta u\,\varphi &= -\sum_{i=1}^d \int_\Omega \partial_i^2 u\,\varphi = \sum_{i=1}^d \int_\Omega \partial_i u\,\partial_i\varphi - \int_{\partial\Omega} \operatorname{div}(\varphi \nabla u) \\ &\underset{\text{Green}}{=} \sum_{i=1}^d \int_\Omega \partial_i u\,\partial_i\varphi - \int_{\partial\Omega} (n\cdot\nabla u)(s)\,\varphi(s)\,d\gamma(s),\end{aligned}$$

où n_i désigne la i-ème composante du vecteur unitaire normal à la frontière $\partial\Omega$ de Ω, et extérieur à Ω, et $d\gamma$ désigne le symbole d'intégration sur $\partial\Omega$. Comme φ est nulle sur $\partial\Omega$, on obtient :

$$\sum_{i=1}^d \int_\Omega \partial_i u\,\partial_i\varphi = \int_\Omega f\,\varphi,$$

Prenons maintenant comme fonction test φ, non plus une fonction de $C_c^\infty(\Omega)$, mais une fonction de $H_0^1(\Omega)$. Comme $\varphi \in H_0^1(\Omega)$, par définition, il existe $(\varphi_n)_{n\in\mathbb{N}} \subset C_c^\infty(\Omega)$ telle que

$$\varphi_n \to \varphi \text{ dans } H^1 \text{ lorsque } n \to +\infty,$$

soit encore

$$\|\varphi_n - \varphi\|_{H^1} = \|\varphi_n - \varphi\|_{L^2}^2 + \sum_i \|D_i\varphi_n - D_i\varphi\|_{L^2}^2 \underset{n\to+\infty}{\longrightarrow} 0$$

Pour chaque fonction $\varphi_n \in C_c^\infty(\Omega)$ on a par (48) :

$$\sum_{i=1}^d \int_\Omega \partial_i u\,\partial_i\varphi_n = \int_\Omega f\,\varphi_n, \quad \forall n \in \mathbb{N}.$$

Or la i-ème dérivée partielle $\partial_i\varphi_n$ converge vers $D_i\varphi$ dans L^2 (donc dans L^2 faible) lorsque $n \to \infty$, et φ_n tend vers φ dans $L^2(\Omega)$. On a donc :

$$\int_\Omega \partial_i u\,\partial_i\varphi_n\,dx \underset{n\to+\infty}{\longrightarrow} \int_\Omega \partial_i u\,D_i\varphi\,dx$$

et

$$\int_\Omega f\,\varphi_n\,dx \underset{n\to+\infty}{\longrightarrow} \int_\Omega f\,\varphi\,dx$$

L'égalité est donc vérifiée pour toute fonction $\varphi \in H_0^1(\Omega)$.

Montrons maintenant que si u est solution classique de (44) alors $u \in H_0^1(\Omega)$. En effet, si $u \in C^2(\Omega)$, alors $u \in C(\overline{\Omega})$ donc $u \in L^2(\Omega)$; de plus $\partial_i u \in C(\overline{\Omega})$ donc $\partial_i u \in L^2(\Omega)$. On a donc bien $u \in H^1(\Omega)$. Il reste à montrer que $u \in H_0^1(\Omega)$.

Pour cela on rappelle (ou on admet) les théorèmes de trace suivants.

Theorem 5.5 (Existence de l'opérateur trace)**.** *Soit Ω un ouvert (borné ou non borné) de $\mathbb{R}^d$, $d \geqslant 1$, de frontière $\partial\Omega$ lipschitzienne, alors $C_c^\infty(\overline{\Omega})$ est dense dans $H^1(\Omega)$. On peut donc définir par continuité l'application "trace", linéaire continue de $H^1(\Omega)$ dans $L^2(\partial\Omega)$, définie par :*

$$\gamma(u) = u|_{\partial\Omega} \quad \textit{si } u \in C_c^\infty(\overline{\Omega}),$$

et par

$$\gamma(u) = \lim_{n\to+\infty} \gamma(u_n)$$

si $u \in H^1(\Omega)$, $u = \lim_{n\to+\infty} u_n$, $(u_n)_{n\in\mathbb{N}} \subset C_c^\infty(\overline{\Omega})$.

Dire que l'application (linéaire) γ est continue est équivalent à dire qu'il existe $C \in \mathbb{R}_+$ tel que

$$\|\gamma(u)\|_{L^2(\partial\Omega)} \leqslant C\|u\|_{H^1(\Omega)} \quad \text{pour tout } u \in H^1(\Omega). \tag{3.4}$$

Notons que $\gamma(H^1(\Omega)) \subset L^2(\partial\Omega)$, mais $\gamma(H^1(\Omega)) \neq L^2(\partial\Omega)$. On note $H^{1/2}(\partial\Omega) = \gamma(H^1(\Omega))$.

Remarquons que si Ω est un ouvert borné, alors $\overline{\Omega}$ est compact et donc toutes les fonctions C^∞ sont à support compact dans $\overline{\Omega}$.

Theorem 5.6 (Noyau de l'opérateur trace)**.** *Soit Ω un ouvert borné de $\mathbb{R}^d$ de frontière $\partial\Omega$ lipschitzienne, et γ l'opérateur trace défini ci-dessus. Alors*

$$\ker\gamma = H_0^1(\Omega).$$

Si $u \in C^2(\overline{\Omega})$ est une solution classique de (44), alors $\gamma(u) = u|_{\partial\Omega} = 0$ donc $u \in \ker\gamma$, et par le théorème précédent ceci prouve que $u \in H_0^1(\Omega)$.

Nous avons ainsi montré que toute solution classique de (44) vérifie $u \in H_0^1(\Omega)$ et l'égalité (48). □

5.1.5. *Existence et unicité des formulations faible et variationnelle.* On cherche à montrer l'existence et l'unicité de la solution de (46) et (47). Pour cela, on utilise le théorème de Lax–Milgram, qu'on rappelle ici.

Soit H un espace de Hilbert et a une forme bilinéaire sur H. On définit $\mathcal{J} : H \to \mathbb{R}$ par

$$\mathcal{J}(v) = \frac{1}{2}\,a(v,v) - T(v). \tag{49}$$

Theorem 5.7 (Lax–Milgram)**.** *Soit H un espace de Hilbert, soit a une forme bilinéaire continue coercive sur H et $T \in H'$. Il existe un unique élément u tel que*

$$\begin{cases} u \in H, \\ a(u,v) = T(v), \quad \forall v \in H. \end{cases} \tag{50}$$

De plus, si a est symétrique, u est l'unique solution du problème de minimisation

$$\begin{cases} u \in H, \\ \mathcal{J}(u) \leqslant \mathcal{J}(v), \qquad \forall v \in H. \end{cases} \tag{51}$$

Ici, a coercive signifie qu'il existe $C > 0$ tel que pour tout $v \in H_0^1(\Omega)$, $a(v,v) \geqslant C\,\|v\|_{H^1}$.

Montrons qu'on peut appliquer le théorème de Lax–Milgram pour les problèmes (46) et (47).

Proposition 5.8 (Existence et unicité de la solution de (44))**.** *Si $f \in L^2(\Omega)$, il existe un unique $u \in H_0^1(\Omega)$ solution de* (46) *et* (47)*.*

Proof. Montrons que les hypothèses du théorème de Lax–Milgram sont vérifiées. L'espace $H = H_0^1(\Omega)$ est un espace de Hilbert. La forme bilinéaire a est définie par

$$a(u,v) = \int_\Omega \nabla u \cdot \nabla v = \sum_{i=1}^{d} \int_\Omega D_i u \, D_i v,$$

et la forme linéaire T par

$$T(v) = \int_\Omega fv.$$

Montrons que $T \in H'$. En effet,

$$|T(v)| \leqslant \|f\|_{L^2} \|v\|_{L^2} \leqslant \|f\|_{L^2} \|v\|_{H^1}.$$

On en déduit que T est continue sur $H_0^1(\Omega)$, ce qui est équivalent à dire que $T \in H^{-1}(\Omega)$ (dual topologique de $H_0^1(\Omega)$).

Montrons que a est bilinéaire, continue et symétrique. La continuité s'obtient par

$$|a(u,v)| = \left| \int_\Omega \nabla u \cdot \nabla v \right| \leqslant \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} \leqslant \|u\|_{H^1} \|v\|_{H^1}.$$

Les caractères bilinéaire et symétrique sont évidents.

Montrons que a est coercitive. En effet,

$$a(v,v) = \int_\Omega |\nabla v|^2 = \sum_{i=1}^{d} \int_\Omega |D_i v|^2 \geqslant \frac{1}{\operatorname{diam}(\Omega)^2 + 1} \|v\|_{H^1}^2,$$

par l'inégalité de Poincaré, qui dit qu'il existe $C > 0$ tel que pour tout $v \in H_0^1$,

$$\|v\|_{L^2} \leqslant \operatorname{diam}(\Omega) \|\nabla v\|_{L^2}.$$

Comme $T \in H'$ et a est bilinéaire, continue, coercitive, le théorème de Lax–Milgram s'applique : il existe une unique fonction $u \in H_0^1(\Omega)$ solution de (46), et comme a est symétrique, u est l'unique solution du problème de minimisation associé. □

5.1.6. *Formulation forte.*

Definition 5.9 (Solution forte dans H^2)**.** *Soit $f \in L^2(\Omega)$, on dit que u est solution forte de* (44) *dans $H^2(\Omega)$ si*

$$u \in H^2(\Omega) \cap H_0^1(\Omega) \quad \text{et} \quad -\Delta u = f \text{ dans } L^2(\Omega).$$

Remarquons que si u est solution forte C^2 de (44), alors u est solution forte H^2. De même, si u est solution forte H^2 de (44) alors u est solution faible de (44). Les réciproques sont fausses.

On admettra le théorème de régularité suivant.

Theorem 5.10 (Régularité)**.** *Soit Ω un ouvert borné de $\mathbb{R}^d$. On suppose que Ω a une frontière de classe C^2, ou que Ω est convexe à frontière lipschitzienne. Si $f \in L^2(\Omega)$ et si $u \in H_0^1(\Omega)$ est solution faible de* (44)*, alors $u \in H^2(\Omega)$. De plus, si $f \in H^m(\Omega)$ alors $u \in H^{m+2}(\Omega)$.*

Remark 5.11 (Différences entre les méthodes de discrétisation)**.** *Lorsqu'on adopte une discrétisation par différences finies, on a directement le problème* (44)*. Lorsqu'on adopte une méthode de volumes finis, on discrétise le "bilan" obtenu en intégrant* (44) *sur chaque maille. Lorsqu'on utilise une méthode variationnelle, on discrétise la formulation variationnelle* (47) *dans le cas de la méthode de Ritz, la formulation faible* (46) *dans le cas de la méthode de Galerkin.*

Remarquons également que dans la formulation faible (46), les conditions aux limites de Dirichlet homogènes $u = 0$ sont prises en compte dans l'espace $u \in H_0^1(\Omega)$, et donc également dans l'espace d'approximation H_N. Pour le problème de Neumann homogène, les conditions aux limites ne sont pas explicites dans l'espace fonctionnel.

5.2. **Problème de Dirichlet non homogène.** On se place ici en dimension 1 d'espace, $d = 1$, et on considère :

$$\begin{cases} -u'' = f \quad \text{sur } (0,1), \\ u(0) = a, \\ u(1) = b, \end{cases} \tag{3.10}$$

où a et b sont des réels donnés. Ces conditions aux limites sont dites de type Dirichlet non homogène ; comme a et b ne sont pas forcément nuls, on cherche une solution dans $H^1(\Omega)$ et non plus dans $H_0^1(\Omega)$.

Cependant, pour se ramener à l'espace $H_0^1(\Omega)$ (en particulier pour obtenir que le problème est bien posé grâce au théorème de Lax–Milgram et à la coercivité de la forme bilinéaire $a(u,v) = \int_\Omega \nabla u \nabla v$ sur $H_0^1(\Omega)$), on va utiliser une technique dite de "relèvement".

On pose $u = u_0 + u_e$ où u_0 est définie par :

$$u_0(x) = a + (b-a)x.$$

On a en particulier $u_0(0) = a$ et $u_0(1) = b$. On a alors $u_e(0) = 0$ et $u_e(1) = 0$. La fonction u_e vérifie donc :

$$\begin{cases} -u_e'' = f, \\ u_e(0) = 0, \\ u_e(1) = 0, \end{cases}$$

dont on connaît la formulation faible et dont on sait qu'il est bien posé. Donc il existe un unique $u \in H^1(\Omega)$ vérifiant $u = u_0 + u_e$, où $u_e \in H_0^1(\Omega)$ est l'unique solution du problème

$$\int_0^1 u_e' \, v' = \int_0^1 f \, v \quad \forall v \in H_0^1((0,1)).$$

De manière plus générale, soit un relèvement

$$u_1 \in H_{a,b}^1((0,1)) = \{v \in H^1;\ v(0) = a \text{ et } v(1) = b\},$$

et soit $\overline{u} \in H_0^1((0,1))$ l'unique solution faible du problème :

$$\begin{cases} -\overline{u}'' = u_1'' + f, \\ \overline{u}(0) = 0, \\ \overline{u}(1) = 0. \end{cases}$$

Alors $\overline{u} + u_1$ est l'unique solution faible de (3.10), c'est-à-dire la solution du problème

$$\begin{cases} u \in H^1_{a,b}((0,1)), \\ \displaystyle\int_0^1 u'v' = \int_0^1 fv, \quad \forall v \in H^1_0((0,1)). \end{cases}$$

On pourrait montrer que u ne dépend pas du relèvement.

Considérons maintenant le cas de la dimension 2 d'espace : $d = 2$. Soit Ω un ouvert borné de $\mathbb{R}^d$, considérons le problème :

$$\begin{cases} -\Delta u = f \quad \text{dans } \Omega, \\ \quad u = g \quad \text{sur } \partial\Omega. \end{cases} \tag{3.11}$$

Pour se ramener au problème de Dirichlet homogène, on veut construire un relèvement, c'est-à-dire une fonction $u_0 \in H^1(\Omega)$ telle que $\gamma(u_0) = g$, où γ est l'application trace. On ne peut plus le faire de manière explicite comme en dimension 1. En particulier, on rappelle qu'en dimension 2, l'espace $H^1(\Omega)$ n'est pas inclus dans $C(\overline{\Omega})$, contrairement au cas de la dimension 1.

Mais si $g \in H^{1/2}(\partial\Omega)$, on sait qu'il existe $u_0 \in H^1(\Omega)$ tel que $g = \gamma(u_0)$. On cherche donc u sous la forme $u = u_e + u_0$ avec $u_e \in H^1_0(\Omega)$ et $u_0 \in H^1(\Omega)$ telle que $\gamma(u_0) = g$.

Soit $v \in H^1_0(\Omega)$; on multiplie (3.11) par v et on intègre sur Ω :

$$\int_\Omega -\Delta u\, v = \int_\Omega f\, v,$$

c'est-à-dire :

$$\int_\Omega \nabla u\, \nabla v = \int_\Omega f\, v.$$

Comme $u = u_0 + u_e$, on a donc :

$$\begin{cases} u_e \in H^1_0(\Omega), \\ \displaystyle\int_\Omega \nabla u_e\, \nabla v = \int_\Omega f\, v - \int_\Omega \nabla u_0\, \nabla v, \quad \forall v \in H^1_0(\Omega). \end{cases} \tag{3.12}$$

En dimension 2, il n'est pas toujours facile de construire le relèvement u_0. Il est donc usuel, dans la mise en œuvre des méthodes d'approximation (par exemple par éléments finis), de se servir de la formulation suivante, équivalente à (3.12) :

$$\begin{cases} u \in \{v \in H^1(\Omega);\ \gamma(v) = g \text{ sur } \partial\Omega\}, \\ \displaystyle\int_\Omega \nabla u\, \nabla v = \int_\Omega f\, v, \quad \forall v \in H^1_0(\Omega). \end{cases} \tag{3.13}$$

5.3. **Condition de Neumann.** Considérons maintenant le problème

$$\begin{cases} -\Delta u = f, \quad \text{dans } \Omega, \\ \dfrac{\partial u}{\partial n} = 0 \quad \text{sur } \partial\Omega, \end{cases} \tag{52}$$

où

$$\frac{\partial u}{\partial n} = n \cdot \nabla u,$$

n étant le vecteur normal à $\partial\Omega$ pointant vers l'extérieur. On appelle ce problème problème de Dirichlet avec conditions de Neumann homogènes. En intégrant la première équation du système, on voit que

$$\int_\Omega -\Delta u = \int_{\partial\Omega} \frac{\partial u}{\partial n} = \int_\Omega f = 0.$$

donc une condition nécessaire d'existence d'une solution est que $\int_\Omega f = 0$.

On remarque que la solution de (52) n'est pas unique, puisque si u est solution alors $u + c$ est aussi solution, pour tout $c \in \mathbb{R}$. Pour éviter ce problème on va chercher les solutions de (52) à moyenne nulle. On cherche donc à (52) résoudre dans l'espace

$$H = \left\{ v \in H^1(\Omega)\,;\ \int_\Omega v = 0 \right\}.$$

Maintenant a est coercive sur H grâce à l'inégalité suivante, qui sera admise.

Lemma 5.12 (Poincaré–Wirtinger)**.** *Soit Ω un ouvert borné de $\mathbb{R}^d$ de frontière lipschitzienne, alors il existe $C \in \mathbb{R}_+^*$, ne dépendant que de Ω, tel que pour tout $u \in H^1(\Omega)$, on a*

$$\left\| u - \frac{1}{|\Omega|} \int_\Omega u \right\|_{L^2(\Omega)} \leqslant C \left\| \nabla u \right\|_{L^2(\Omega)}. \tag{53}$$

On a alors $a(u,u) = \|\nabla u\|^2_{L^2(\Omega)}$ et

$$\|u\|^2_{H^1} = \|u\|^2_{L^2} + a(u,u) \underset{\substack{\int_\Omega u = 0 \\ (53)}}{\leqslant} \left(1 + C^2\right) a(u,u),$$

donc la constante de coercivité est $\alpha = (1 + C^2)^{-1}$.

Le problème

$$\begin{cases} u \in H, \\ a(u,v) = \displaystyle\int fv \qquad \forall v \in H, \end{cases}$$

admet donc une unique solution.

5.4. **Formulation faible et formulation variationnelle.** Nous donnons ici un exemple de problème pour lequel on peut établir une formulation faible, mais pas variationnelle. On se place en une dimension d'espace $d = 1$, et on considère $\Omega =]0,1[$ et $f \in L^2(]0,1[)$. On s'intéresse au problème suivant d'advection diffusion

$$\begin{cases} -u'' + u' = f, & \text{dans }]0,1[, \\ u(0) = u(1) = 0. \end{cases} \tag{54}$$

Cherchons une formulation faible. On choisit $v \in H^1_0(\Omega)$, on multiplie (54) par v et on intègre par parties :

$$\int_\Omega u'v' + \int_\Omega u'v = \int_\Omega fv.$$

Il est donc naturel de poser :

$$a(u,v)=\int_\Omega u'v'+\int_\Omega u'v,\qquad T(v)=\int_\Omega fv.$$

Il est évident que T est une forme linéaire continue sur $H_0^1(\Omega)$ (c'est à dire $T\in H^{-1}(\Omega)$) et que la forme a est bilinéaire continue, mais pas symétrique, donc on n'a pas l'existence du minimum dans (51). De plus elle est coercive. En effet,

$$a(u,u)=\int_\Omega u'^2+\int_\Omega u'u=\int_\Omega u'^2+\int_\Omega \frac{1}{2}(u^2)'.$$

Or, comme $u\in H_0^1(\Omega)$, on a $u=0$ sur $\partial\Omega$ et donc

$$\int_\Omega (u^2)'=u^2(1)-u^2(0)=0.$$

On en déduit que :

$$a(u,u)=\int_0^1 (u')^2,$$

et par l'inégalité de Poincaré, on conclut que a est coercive sur $H_0^1(\Omega)$. On en déduit par le théorème de Lax–Milgram, l'existence et l'unicité de u solution du problème

$$\begin{cases} u\in H_0^1(]0,1[),\\ \displaystyle\int_0^1 \big(u'v'+u'v\big)=\int_0^1 fv. \end{cases}$$

6. Méthodes de Ritz et Galerkin

6.1. **Principe général de la méthode de Ritz.** On se place sous les hypothèses suivantes :

$$\begin{cases} H \text{ est un espace Hilbert,}\\ a \text{ est une forme bilinéaire continue coercitive et symétrique,}\\ T\in H'. \end{cases} \tag{55}$$

On cherche à calculer $u\in H$ telle que :

$$a(u,v)=T(v),\qquad \forall v\in H, \tag{56}$$

ce qui revient à calculer $u\in H$ solution du problème de minimisation (3.8), avec J définie par (3.9).

L'idée de la méthode de Ritz est de remplacer H par un espace $H_N\subset H$ de dimension finie (où $\dim H_N=N$), et de calculer U solution de

$$\begin{cases} U\in H_N,\\ J(U)\leqslant J(v),\qquad \forall v\in H_N, \end{cases} \tag{57}$$

en espérant que U soit "proche" (en un sens à définir) de u.

Theorem 6.1. *Sous les hypothèses* (55)*, si H_N est un sous-espace vectoriel de H et $\dim H_N<+\infty$ alors le problème* (57) *admet une unique solution.*

Proof. Puisque H_N est un espace de dimension finie inclus dans H, c'est donc aussi un Hilbert. On peut donc appliquer le théorème de Lax–Milgram, et on en déduit l'existence et l'unicité de $U \in H_N$ solution de (57), qui est aussi solution de :

$$\begin{cases} U \in H_N, \\ a(U,v) = T(v), \qquad \forall v \in H_N. \end{cases}$$

□

Nous allons maintenant exposer une autre méthode de démonstration du théorème 6.1, qui a l'avantage d'être constructive, et qui nous permet d'introduire les idées principales des méthodes numériques envisagées plus loin. Comme l'espace H_N considéré dans le théorème est de dimension N, il existe une base $(\varphi_1, \dots, \varphi_N)$ de H_N. Si $u \in H_N$, on peut donc développer

$$u = \sum_{i=1}^{N} u_i \varphi_i.$$

On note

$$U = (u_1, \dots, u_N) \in \mathbb{R}^N.$$

L'application ξ qui à u associe U est une bijection de H_N dans $\mathbb{R}^N$. Posons $j = J \circ \xi^{-1}$. On a donc :

$$\begin{aligned} j(U) = J(u) &= \frac{1}{2} a\left(\sum_{i=1}^{N} u_i \varphi_i,\ \sum_{i=1}^{N} u_i \varphi_i \right) - T\left(\sum_{i=1}^{N} u_i \varphi_i \right) \\ &= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} u_i u_j a(\varphi_i, \varphi_j) - \sum_{i=1}^{N} u_i T(\varphi_i) = \frac{1}{2} U^t K U - U^t G \end{aligned}$$

où $K \in M_{N,N}(\mathbb{R})$ est définie par $K_{ij} = a(\varphi_i, \varphi_j)$, et où $G_i = T(\varphi_i)$. Chercher u_N solution de (57) est donc équivalent à chercher U solution de :

$$\begin{cases} U \in \mathbb{R}^N, \\ j(U) \leqslant j(V), \qquad \forall V \in \mathbb{R}^N, \end{cases} \tag{58}$$

où

$$j(V) := \frac{1}{2} V^t K V - V^t G. \qquad (3.24)$$

Il est facile de vérifier que la matrice K est symétrique car a l'est. De même, pour tout $U \in \mathbb{R}^N$,

$$\begin{aligned} \langle U, KU \rangle &= \sum_{1 \leqslant i,j \leqslant N} U_i K_{ij} U_j = \sum_{1 \leqslant i,j \leqslant N} U_i U_j a(\varphi_i, \varphi_j) \\ &= a\left(\sum_{j=1}^{N} U_j \varphi_j, \sum_{j=1}^{N} U_j \varphi_j \right) \underset{a \text{ coercive}}{\geqslant} \alpha \left\| \sum_{j=1}^{N} U_j \varphi_j \right\|^2 \underset{(\varphi_j)_j \text{ base}}{=} \alpha \|U\|^2 \end{aligned}$$

donc K est définie positive par coercivité de a. Donc j est une fonctionnelle quadratique sur $\mathbb{R}^N$, et on a donc existence et unicité de $U \in \mathbb{R}^N$ tel que

$j(U) \leqslant j(V)\ \forall V \in \mathbb{R}^N$. L'unique solution du problème de minimisation (58) est aussi la solution du système linéaire

$$KU = G,$$

on appelle souvent K la matrice de rigidité.

6.2. **Résumé sur la technique de Ritz.**

(1) On se donne $H_N \subset H$.
(2) On trouve une base de H_N.
(3) On calcule la matrice de rigidité K et le second membre G
(4) On minimise j par la résolution de $KU = G$.
(5) On calcule la solution approchée :

$$u^{(N)} = \sum_{i=1}^{N} u_i \varphi_i.$$

On appelle H_N l'espace d'approximation. Le choix de cet espace sera fondamental pour le développement de la méthode d'approximation. Le choix de H_N est formellement équivalent au choix de la base $(\varphi_i)_{i=1\ldots N}$. Pourtant, le choix de cette base est capital même si $u^{(N)}$ ne dépend que du choix de H_N et pas de la base.

6.3. **Choix de la base.** Un premier choix consiste à choisir des bases indépendantes de N c'est à dire

$$\{\text{base de } H_{N+1}\} = \{\text{base de } H_N\} \cup \{\varphi_{N+1}\}.$$

Les bases sont donc emboîtées les unes dans les autres. Considérons par exemple $H = H^1(]0,1[)$, et l'espace d'approximation :

$$H_N = \text{Vect}\{1, X, \ldots, X^{N-1}\}.$$

Les fonctions de base sont donc $\varphi_i = X^{i-1}$, $i = 1, \ldots, N$. On peut remarquer que ce choix de base amène à une méthode d'approximation qui donne des matrices pleines. Or, on veut justement éviter les matrices pleines, car les systèmes linéaires associés sont coûteux (en temps et mémoire) à résoudre.

Le choix idéal serait de choisir une base $(\varphi_i)_{i=1,\ldots,N}$ qui diagonalise a, c'est-à-dire telle que

$$a(\varphi_i, \varphi_j) = \lambda_i \delta_{ij},$$

où

$$\delta_{ij} = \begin{cases} 1 & \text{si } i = j, \\ 0 & \text{sinon.} \end{cases} \qquad (3.25)$$

On a alors $K = \text{diag}(\lambda_1, \ldots, \lambda_N)$, et explicitement

$$u^{(N)} = \sum_{i=1}^{N} \frac{T(\varphi_i)}{a(\varphi_i, \varphi_i)} \varphi_i.$$

Considérons par exemple le problème de Dirichlet (46), où $a(\varphi, \phi) = \int_\Omega \nabla\varphi \cdot \nabla\phi$. Si φ_i est la i-ème fonction propre de l'opération $-\Delta$ avec conditions aux limites de Dirichlet associée à λ_i, on obtient bien la propriété souhaitée. Malheureusement, il est rare que l'on puisse connaître explicitement les fonctions de base φ_i.

Un deuxième choix consiste à choisir des bases dépendantes de N. Mais dans ce cas, la base de H_N n'est pas incluse dans celle de H_{N+1}. La technique des éléments finis qu'on verra au chapitre suivant, est un exemple de ce choix. Dans la matrice K obtenue est creuse (c'est à dire qu'un grand nombre de ses coefficients sont nuls). Par exemple, pour des éléments finis appliqués à un opérateur du second ordre, on peut avoir un nombre de coefficients non nuls de l'ordre de $O(N)$.

6.4. **Convergence de l'approximation de Ritz.** Une fois qu'on a calculé u_N solution de (58), il faut se préocupper de savoir si $u^{(N)}$ est une bonne approximation de u solution de (56), c'est à dire de savoir si

$$u^{(N)} \underset{N\to+\infty}{\longrightarrow} u.$$

Pour vérifier cette convergence, on va se servir de la notion de consistance.

Definition 6.2 (Consistance)**.** *On dit que l'approximation de Ritz définie par l'espace $H_N \subset H$ avec $\dim H_N = N < +\infty$ est consistante si*

$$d(u, H_N) \underset{n\to+\infty}{\longrightarrow} 0, \qquad \forall u \in H, \tag{59}$$

La condition (59) est équivalente à

$$\inf_{v\in H_N} \|u - v\| \underset{n\to+\infty}{\longrightarrow} 0, \qquad \forall u \in H.$$

L'autre notion fondamentale pour prouver la convergence est la stabilité, elle même obtenue grâce à la propriété de coercivité de a. Par stabilité, on entend estimation a priori sur la solution approchée $u^{(N)}$ (avant même de savoir si elle existe), où $u^{(N)}$ est solution de (58) ou encore de :

$$\begin{cases} u^{(N)} \in H_N \\ a(u^{(N)}, v) = T(v) \qquad \forall v \in H_N. \end{cases} \tag{60}$$

On a l'estimation a priori suivante sur u_N.

Proposition 6.3 (Stabilité)**.** *Sous les hypothèses du théorème 6.4, on a*

$$\left\|u^{(N)}\right\|_H \leqslant \frac{\|T\|_{H'}}{\alpha}.$$

Proof. On a

$$\alpha\|u^{(N)}\|^2 \underset{\substack{a \\ \text{coercive}}}{\leqslant} a(u^{(N)}, u^{(N)}) \underset{(60)}{=} T(u^{(N)}) \underset{\substack{T \\ \text{continue}}}{\leqslant} \|T\|_{H'}\, \|u^{(N)}\|_H.$$

□

Theorem 6.4 (Céa)**.** *Soit H un espace de Hilbert réel. Soit a une forme bilinéaire continue symétrique coercive, soit $M > 0$ et $\alpha > 0$ tels que $a(u, v) \leqslant M\|u\|_H\|v\|_H$ et $a(u, u) \geqslant \alpha\|u\|_H^2$. Soit $T \in H'$ une forme linéaire continue. Soit $u \in H$ l'unique solution du problème*

$$\begin{cases} u \in H, \\ a(u, v) = T(v), \qquad \forall v \in H. \end{cases} \tag{61}$$

Soit $H_N \subset H$ tel que $\dim H_N = N$, et soit $u^{(N)} \in H_N$ l'unique solution de

$$\begin{cases} u^{(N)} \in H_N, \\ a(u^{(N)}, v) = T(v), \qquad \forall v \in H_N. \end{cases} \tag{62}$$

Alors

$$\left\| u - u^{(N)} \right\|_H \leqslant \sqrt{\frac{M}{\alpha}}\, d(u, H_N). \tag{63}$$

Proof. • On va montrer que $u^{(N)}$ est la projection de u sur H_N pour le produit scalaire $(\cdot,\cdot)_a$ induit par a, défini de $H \times H$ par $(u,v)_a = a(u,v)$. On note $\|u\|_a = \sqrt{a(u,u)}$, la norme induite par le produit scalaire a. La norme $\|\cdot\|_a$ est équivalente à la norme $\|\cdot\|_H$, en effet, grâce à la coercivité et la continuité de a,

$$\alpha \|u\|_H^2 \leqslant \|u\|_a^2 \leqslant M \|u\|_H^2.$$

Donc $(H, \|\cdot\|_a)$ est un espace de Hilbert. Soit u la solution de (61), et soit $v := P_{H_N} u$ la projection orthogonale de u sur H_N relative au produit scalaire $a(\cdot,\cdot)$. Par définition de la projection orthogonale, on a donc

$$P_{H_N} u - u = -P_{H_N}^{\perp} u \in H_N^{\perp},$$

soit encore

$$a(P_{H_N} u - u, w) = 0, \qquad \forall w \in H_N.$$

En soustrayant (61) et (62), on obtient la condition d'orthogonalité de Galerkin

$$a(u - u^{(N)}, w) = 0, \qquad \forall w \in H_N.$$

En combinant ces deux relations, il vient

$$a(P_{H_N} u - u^{(N)}, w) = 0, \qquad \forall w \in H_N.$$

Or $P_{H_N} u - u^{(N)} \in H_N$. En prenant $w = P_{H_N} u - u^{(N)}$, on obtient

$$a(P_{H_N} u - u^{(N)}, P_{H_N} u - u^{(N)}) = 0.$$

La coercivité de a implique alors $P_{H_N} u - u^{(N)} = 0$, donc

$$u^{(N)} = P_{H_N} u.$$

• Par définition de P_{H_N}, on a :

$$\|u - P_{H_N} u\|_a^2 \leqslant \|u - v\|_a^2, \qquad \forall v \in H_N,$$

ce qui s'écrit (puisque $P_{H_N} u = u^{(N)}$) :

$$a(u - u^{(N)}, u - u^{(N)}) \leqslant a(u - v, u - v), \qquad \forall v \in H_N.$$

Par coercivité et continuité de la forme bilinéaire a, on a donc $\forall v \in H_N$,

$$\alpha \|u - u^{(N)}\|_H^2 \leqslant a(u - u^{(N)}, u - u^{(N)}) \leqslant a(u - v, u - v) \leqslant M \|u - v\|_H^2.$$

On en déduit que :

$$\|u - u^{(N)}\|_H \leqslant \sqrt{\frac{M}{\alpha}}\, \|u - v\|_H, \qquad \forall v \in H_N.$$

En passant à l'inf sur v, on obtient alors (63). □

6.5. **Méthode de Galerkin.** On se place maintenant sous les hypothèses suivantes :

$$\begin{cases} H \text{ espace de Hilbert,} \\ a : \text{ forme bilinéaire continue et coercive,} \\ T \in H'. \end{cases} \tag{3.30}$$

Remarquons que maintenant, a n'est pas nécessairement symétrique, les hypothèses (3.30) sont donc plus générales que les hypothèses (3.21). On considère le problème

$$\begin{cases} u \in H, \\ a(u,v) = T(v), \qquad v \in H. \end{cases} \tag{64}$$

Par le théorème de Lax–Milgram, il y a existence et unicité de $u \in H$ solution de (64).

Le principe de la méthode de Galerkin est similaire à celui de la méthode de Ritz. On se donne $H_N \subset H$, tel que $\dim H_N < +\infty$, et on cherche à résoudre le problème approché :

$$(P_N) \quad \begin{cases} u^{(N)} \in H_N, \\ a(u^{(N)},v) = T(v), \qquad \forall v \in H_N. \end{cases} \tag{65}$$

Par le théorème de Lax–Milgram, on a immédiatement :

Theorem 6.5. *Sous les hypothèses, si $H_N \subset H$ et $\dim H_N = N$, il existe un unique $u^{(N)} \in H_N$ solution de* (65)*.*

Comme dans le cas de la méthode de Ritz, on va donner une autre méthode, constructive, de démonstration de l'existence et unicité de u_N qui permettra d'introduire la méthode de Galerkin. Comme $\dim H_N = N$, il existe une base $(\varphi_1, \dots, \varphi_N)$ de H_N. Soit $v \in H_N$, on peut donc développer v sur la base

$$v = \sum_{i=1}^{N} v_i \varphi_i,$$

et identifier v au vecteur $(v_1, \dots, v_N) \in \mathbb{R}^N$. En écrivant que $u^{(N)}$ satisfait (65) pour tout $v = \varphi_i$, $i = 1, \dots, N$:

$$a(u, \varphi_i) = T(\varphi_i), \qquad \forall i = 1, \dots, N,$$

et en développant u sur la base $(\varphi_i)_{i=1,\dots,N}$, on obtient :

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i) u_j^{(N)} = T(\varphi_i), \qquad \forall i = 1, \dots, N.$$

On peut écrire cette dernière égalité sous forme d'un système linéaire

$$KU = G, \tag{66}$$

où $U := \left(u_1^{(N)}, \dots, u_N^{(N)}\right)^T$, $K_{ij} = a(\varphi_j, \varphi_i)$ et $G_i = T(\varphi_i)$, pour $i, j = 1, \dots, N$. La matrice K n'est pas en général symétrique.

Proposition 6.6. *Sous les hypothèses du Théorème* (6.5)*, le système linéaire* (66) *admet une unique solution.*

Proof. On va montrer que K est inversible en vérifiant que son noyau est réduit à $\{0\}$. Soit $w \in \mathbb{R}^N$ tel que $Kw = 0$. Décomposons w sur le N base $(\varphi_1, \ldots, \varphi_N)$ de H_N. On a

$$\sum_{j=1}^{N} a(\varphi_j, \varphi_i) w_j = 0.$$

Multiplions cette relation par w_i et sommons pour $i = 1$ à N, on obtient

$$\sum_{i=1}^{N} \sum_{j=1}^{N} a(\varphi_j, \varphi_i) w_j w_i = 0.$$

Soit encore : $a(w, w) = 0$ où $w := (w_1, \ldots, w_N)^T$. Par coercivité de a, ceci entraine que $w = 0$. On en déduit que $w_i = 0$, $\forall i = 1, \ldots, N$, ce qui achève la preuve. □

Remark 6.7. *Si a est symétrique, la méthode de Galerkin est équivalente à celle de Ritz.*

En résumé, la méthode de Galerkin comporte les mêmes étapes que la méthode de Ritz, c'est à dire :

(1) On se donne $H_N \subset H$.
(2) On trouve une base de H_N.
(3) On calcule K et G.
(4) On résout $KU = G$.
(5) On écrit $u^{(N)} = \sum_{i=1}^{N} u_i \varphi_i$.

La seule différence est que l'étape 4 n'est pas issue d'un problème de minimisation. Comme pour la méthode de Ritz, il faut se poser la question du choix du sous espace H_N et de sa base, ainsi que de la convergence de l'approximation de u solution de (64) par $u^{(N)}$ obtenue par la technique de Galerkin. En ce qui concerne le choix de la base $\{\varphi_1, \ldots, \varphi_N\}$, les possibilités sont les mêmes que pour la méthode de Ritz, voir paragraphe 3.2.1. De même, la notion de consistance est identique à celle donnée pour la méthode de Ritz (voir définition 3.19) et la démonstration de stabilité est identique à celles effectuée pour la méthode de Ritz ; voir proposition 3.20 page 109. On peut alors établir le théorème de convergence :

Theorem 6.8. *Sous les hypothèses du théorème* (57)*, si u est la solution de* (64) *et u_N la solution de* (65)*, alors*

$$\|u - u^{(N)}\|_H \leqslant \frac{M}{\alpha}\, d(u, H_N). \tag{67}$$

Ici encore, M et α sont tels que : $\alpha\|v\|^2 \leqslant a(v, u) \leqslant M\|v\|^2$ pour tout v dans H (les réels M et α existent en vertu de la continuité et de la coercivité de a).

Proof. Comme la forme bilinéaire a est coercive de constante α, on a :

$$\alpha\|u - u^{(N)}\|_H^2 \leqslant a(u - u^{(N)}, u - u^{(N)}).$$

On a donc, pour tout $v \in H$:

$$\alpha\|u - u^{(N)}\|_H^2 \leqslant a(u - u^{(N)}, u - v) + a(u - u^{(N)}, v - u^{(N)}).$$

Or

$$a(u-u^{(N)},v-u^{(N)})=a(u,v-u^{(N)})-a(u^{(N)},v-u^{(N)}),$$

et par définition de u et $u^{(N)}$, on a :

$$\begin{aligned} a(u,v-u^{(N)}) &= T(v-u^{(N)}), \qquad \forall v\in H \\ a(u^{(N)},v-u^{(N)}) &= T(v-u^{(N)}) \qquad \forall v\in H_N. \end{aligned}$$

On en déduit que :

$$\alpha\|u-u^{(N)}\|_H^2 \leqslant a(u-u^{(N)},u-v), \qquad \forall v\in H_N,$$

et donc, par continuité de la forme bilinéaire a :

$$\alpha\|u-u^{(N)}\|_H^2 \leqslant M\|u-u^{(N)}\|_H\,\|u-v\|_H.$$

On obtient donc :

$$\|u-u^{(N)}\|_H \leqslant \frac{M}{\alpha}\,\|u-v\|_H, \qquad \forall v\in H_N,$$

ce qui entraine (67). □

Remark 6.9. *On peut remarquer que l'estimation* (67) *obtenue dans le cadre de la metode de Galerkin est moins bonne que l'estimation* (63) *obtenue dans le cadre de la methode de Ritz. Ceci est normal, puisque la methode de Ritz est un cas particulier de la methode de Galerkin.*

Grâce au théorème (6.8), on peut remarquer que $u^{(N)}$ converge vers u dans H lorsque N tend vers $+\infty$ dès que $d(u,H_N)\to 0$ lorsque $N\to+\infty$. C'est donc là encore une propriété de consistance dont nous avons besoin.

La propriété de consistance n'est pas toujours facile à montrer directement. On utilise alors la caractérisation suivante :

Proposition 6.10 (Caractérisation de la consistance)**.** *Soit V un sous espace vectoriel de H dense dans H. On suppose qu'il existe une fonction $r_N : V\to H_N$ telle que pour tout $v\in V$,*

$$\|v-r_N(v)\|_H \underset{N\to+\infty}{\longrightarrow} 0,$$

alors

$$d(u,H_N) \underset{N\to+\infty}{\longrightarrow} 0.$$

Proof. Soit $v\in V$, et $w=r_N(v)$. Par définition, on a

$$d(u,H_N)\leqslant\|u-r_N(v)\|_H\leqslant\|u-v\|_H+\|v-r_N(v)\|_H.$$

Comme V est dense dans H, pour tout $\varepsilon>0$, il existe $v\in V$, tel que $\|u-v\|_H\leqslant\varepsilon$. Choisissons v qui vérifie cette dernière inégalité. Par hypothèse sur r_N :

$$\forall\varepsilon>0,\ \exists N_0 \text{ tel que } N\geqslant N_0\Rightarrow\|v-r_N(v)\|\leqslant\varepsilon.$$

Donc si $N\geqslant N_0$, on a $d(u,H_N)\leqslant 2\varepsilon$. On en déduit que $d(u,H_N)\to 0$ quand $N\to+\infty$. □

7. La méthode des éléments finis

La méthode des éléments finis est une façon de choisir les bases des espaces d'approximation pour les méthodes de Ritz et Galerkin.

7.1. **Principe de la méthode.** On se limitera dans le cadre de ce cours à des problèmes du second ordre. L'exemple type sera le problème de Dirichlet qu'on rappelle ici :

$$\begin{cases} -\Delta u = f & \text{dans } \Omega, \\ u = 0 & \text{sur } \partial\Omega, \end{cases} \tag{68}$$

et l'espace de Hilbert sera l'espace de Sobolev $H^1(\Omega)$ ou $H_0^1(\Omega)$.

On se limitera à un certain type d'éléments finis, dits "de Lagrange". Donnons les principes généraux de la méthode.

Éléments finis de Lagrange. Soit $\Omega \subset \mathbb{R}^2$ (ou $\mathbb{R}^3$). Soit H l'espace fonctionnel dans lequel on recherche la solution (par exemple $H_0^1(\Omega)$ s'il s'agit du problème de Dirichlet (3.1)). On cherche $H_N \subset H = H_0^1(\Omega)$ et les fonctions de base $\varphi_1, \ldots, \varphi_N$. On va déterminer ces fonctions de base à partir d'un découpage de Ω en un nombre fini de cellules, appelés "éléments". La procédure est la suivante :

(1) On construit un "maillage" $\mathcal{T}$ de Ω (en triangles ou rectangles) que l'on appelle éléments K.
(2) Dans chaque élément, on se donne des points que l'on appelle "noeuds".
(3) On définit H_N par :
$$H_N = \left\{ u : \Omega \to \mathbb{R} \ / \ u|_K \in \mathbb{P}_k, \ \forall K \in \mathcal{T} \right\} \cap H,$$
où $\mathbb{P}_k$ désigne l'ensemble des polynômes de degré inférieur ou égal à k. Le degré des polynômes est choisi de manière à ce que u soit entièrement déterminée par ses valeurs aux noeuds. Pour une méthode d'éléments finis de type Lagrange, les valeurs aux noeuds sont également les "degrés de liberté", c.à.d. les valeurs qui déterminent entièrement la fonction recherchée.
(4) On construit une base $\{\varphi_i, \ldots, \varphi_N\}$ de H_N tel que le support de φ_i soit "le plus petit possible". Les fonctions φ_i sont aussi appelées fonctions de forme.

7.1.1. *Exemple en dimension 1.* Soit $\Omega =]0,1[\subset \mathbb{R}$ et soit $H = H_0^1([0,1[)$. On cherche un espace H_N d'approximation de H. Pour cela, on divise l'intervalle $]0,1[$ en N intervalles de longueur

$$h = \frac{1}{N+1}.$$

On pose $x_i = i$, $i = 0, \ldots, N+1$.

Les étapes 1. à 4. décrites précédemment donnent dans ce cas :

(1) **Construction des éléments.** On a construit $n+1$ éléments $K_i =]x_i, x_{i+1}[$, $i = 0, \ldots, N$.
(2) **Noeuds.** On a deux noeuds par élément, (x_i et x_{i+1} sont les noeuds de K_i, $i = 0, \ldots, N$). Le fait que $H_N \subset H_0^1(]0,1[)$ impose que les fonctions de H_N soient nulles en $x_0 = 0$ et $x_{N+1} = 1$. On appelle $x_1, \ldots, x_N$ les noeuds libres et x_0, x_{N+1} les noeuds liés. Les degrés

de liberté sont donc les valeurs de u en $x_1, \dots, x_N$. Aux noeuds liés, on a $u(x_0) = u(x_{N+1}) = 0$.

(3) **Choix de l'espace.** On choisit comme espace de polynôme :

$$\mathbb{P}_1 = \{ax + b,\ a, b \in \mathbb{R}\}$$

et on pose :

$$H_N = \Big\{u : \Omega \to \mathbb{R} \ \Big|\ u|_{K_i} \in \mathbb{P}_1,\ \forall i \in \{1, \dots, N\}, u \in \mathcal{C}^0(\Omega),\ u(0) = u(1) = 0\Big\}.$$

Rappelons que $H = H_0^1(]0, 1[) \subset C([0, 1])$. Avec le choix de H_N, on a bien $H_N \subset H$.

(4) **Choix de la base de H_N.**

On peut définir φ_i pour $i = 1$ à N par :

$$\begin{cases} \varphi_i : \text{ affine par morceaux, continue,} \\ \text{supp}(\varphi_i) = [x_{i-1}, x_{i+1}], \\ \varphi_i(x_i) = 1, \\ \varphi_i(x_{i-1}) = \varphi_i(x_{i+1}) = 0, \end{cases}$$

Il est facile de voir que $\varphi_i \in H_N$ et que $\{\varphi_1, \dots, \varphi_N\}$ engendre H_N, c'est à dire que pour tout $u \in H_N$, il existe $(u_1, \dots, u_N) \in \mathbb{R}^N$ tel que

$$u = \sum_{i=1}^{N} u_i \varphi_i.$$

7.1.2. *Exemple en dimension 2.* Soit Ω un ouvert polygonal de $\mathbb{R}^2$, et $H = H_0^1(\Omega)$. Les étapes de construction de la méthode des éléments finis sont encore les mêmes.

(1) **Éléments :** on choisit des triangles.

(2) **Noeuds :** on les place aux sommets des triangles. Les noeuds $x_i \in \Omega$ (intérieurs à Ω) sont libres, et les noeuds $x_i \in \partial\Omega$ (sur la frontière de Ω) sont liés. On notera Σ l'ensemble des noeuds libres, Σ_F l'ensemble des noeuds liés, et $\Sigma = \Sigma_I \cup \Sigma_F$.

(3) **Espace d'approximation.** L'espace des polynômes est l'ensemble des fonctions affines, noté $\mathbb{P}_1$. Une fonction $p \in \mathbb{P}_1$ est de la forme :

$$p : \mathbb{R}^2 \to \mathbb{R}, \qquad x = (x_1, x_2) \mapsto a_1 x_1 + a_2 x_2 + b,$$

avec $(a_1, a_2, b) \in \mathbb{R}^3$. L'espace d'approximation H_N est donc défini par :

$$H_N = \Big\{u \in \overline{C(\Omega)};\ u|_K \in \mathbb{P}_1,\ \forall K, \text{ et } u(x_i) = 0,\ \forall x_i \in \Sigma_F\Big\}.$$

(4) **Base de H_N :** On choisit comme base de H_N la famille de fonctions $\{\varphi_i\}_{i=1,\dots,N}$, où $N = \text{card}(\Sigma_I)$, où φ_i est définie, pour $i = 1$ à N, par :

$$\begin{cases} \varphi_i \text{ est affine par morceaux,} \\ \varphi_i(x_i) = 1, \\ \varphi_i(x_j) = 0, \qquad \forall j \neq i. \end{cases}$$

En résumé. Les questions à se poser pour construire une méthode d'éléments finis sont donc :

(1) La construction du maillage.
(2) Un choix cohérent entre éléments, noeuds et espace des polynômes.
(3) La construction de l'espace d'approximation H_N et de sa base $\{\varphi_i\}_{i=1\dots N}$.
(4) La construction de la matrice de rigidité K et du second membre G.
(5) L'évaluation de $d(u, H_N)$ en vue de l'analyse de convergence.

Pour construire les éléments, il faut éviter les angles trop grands ou trop petits. Il faut mettre beaucoup d'éléments là où u varie rapidement (ceci ne peut se faire que si on connait a priori les zones de de variation rapide, ou si on a les moyens d'évaluer l'erreur entre la solution exacte du problème et la solution calculée et de remailler les zones ou celle–ci est jugée trop grande.

On a vu aux paragraphes précédents que l'erreur entre la solution exacte u recherchée et la solution $u^{(N)}$ obtenue par la méthode de Ritz ou de Galerkin est majorée par une constante fois la distance entre H et H_N. On a donc intérêt à ce que cette distance soit petite. Pour ce faire, il paraît raisonnable d'augmenter la dimension de l'espace H_N. Pour cela, on a deux possibilités

- augmenter le nombre d'éléments : on augmente alors aussi le nombre global de noeuds, mais pas le nombre local.
- augmenter le degré des polynômes : on augmente alors le nombre de noeuds local, donc on augmente aussi le nombre global de noeuds, mais pas le nombre d'éléments. Ce deuxième choix (augmentation du degré des polynômes) ne peut se faire que si la solution est suffisamment régulière ; si la solution n'est pas régulière, on n'arrivera pas à diminuer $d(H, H_N)$ en augmentant le degré des polynômes.

7.2. **Convergence des éléments finis $\mathbb{P}_1$ en dimension 1.** On note l'espace des éléments finis $\mathbb{P}_1$

$$V_h := \left\{ v \in C([0,1]) \text{ tel que } v\big|_{[x_j, x_{j+1}]} \in \mathbb{P}_1 \text{ pour tout } 0 \leqslant j \leqslant n \right\}$$

et le sous-espace des fonctions s'annulant aux bords

$$V_h^0 := \{ v \in V_h \mid v(0) = v(1) = 0 \}.$$

7.2.1. *Énoncé du théorème de convergence.* La méthode des éléments finis est une méthode de Galerkin où l'espace variationnel est V_h^0. On se place en dimension 1, et $\Omega =]0,1[$. Le problème de Galerkin définit $u_h \in V_h^0$ via

$$\int_0^1 u_h' v_h' = \int_0^1 f v_h, \qquad \forall v_h \in V_h^0. \tag{69}$$

Theorem 7.1 (Convergence de la méthode $\mathbb{P}_1$)**.** *Soit $u \in C^2([0,1])$ la solution de* (68) *et soit $u_h \in V_h^0$ la solution de* (69)*. La méthode des éléments finis $\mathbb{P}_1$ converge, c'est-à-dire qu'il existe une constante C indépendante de h et de f telle que*

$$\|(u - u_h)'\|_{L^2(0,1)} \leqslant Ch \, \|f\|_{L^2(0,1)}.$$

7.2.2. *Preuve du théorème 7.1.*

Definition 7.2 (Opérateur d'interpolation $\mathbb{P}_1$)**.** *On appelle opérateur d'interpolation $\mathbb{P}_1$ l'application linéaire r_h de $C([0,1])$ dans V_h définie, pour tout $v \in C([0,1])$, par*

$$(r_h v)(x) := \sum_{j=0}^{n+1} v(x_j)\, \varphi_j(x)$$

Par un dessin, on voit que $r_h v$ est la fonction affine par morceaux qui coïncide avec v sur les sommets du maillage x_j.

On commence par montrer le lemme technique suivant.

Lemma 7.3. *Il existe une constante C indépendante de h telle que, pour tout $v \in C^2([0,1])$,*

$$\|v - r_h v\|_{L^2(0,1)} \leqslant h^2 \, \|v''\|_{L^2(0,1)}, \tag{70}$$

et

$$\|v' - (r_h v)'\|_{L^2(0,1)} \leqslant h \, \|v''\|_{L^2(0,1)}. \tag{71}$$

Proof. Soit $v \in C^2([0,1])$. Par définition, l'interpolée $r_h v$ est une fonction affine. Pour tout $x \in]x_j, x_{j+1}[$, on a

$$(r_h v)(x) = v(x_j) + \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j}\,(x - x_j). \tag{72}$$

donc

$$\begin{aligned} v(x) - r_h v(x) &\underset{(72)}{=} v(x) - \left(v(x_j) + \frac{v(x_{j+1}) - v(x_j)}{x_{j+1} - x_j}\,(x - x_j) \right) \\ &= \int_{x_j}^{x} v' - \frac{x - x_j}{x_{j+1} - x_j} \int_{x_j}^{x_{j+1}} v' \\ &= (x - x_j)\, v'(x_j + \theta_x) - (x - x_j)\, v'(x_j + \theta_j) \\ &= (x - x_j) \int_{x_j + \theta_x}^{x_j + \theta_j} v''(t)\, dt, \end{aligned}$$

par application de la formule des accroissements finis (il existe un y tel que la fonction passe par sa moyenne) avec $0 \leqslant \theta_x \leqslant x - x_j$ et $0 \leqslant \theta_j \leqslant x_{j+1} - x_j = h$. On en déduit, en utilisant l'inégalité de Cauchy–Schwarz,

$$|v(x) - r_h v(x)|^2 \leqslant h^2 \left(\int_{x_j}^{x_{j+1}} |v''(t)|\, dt \right)^2 \leqslant h^3 \int_{x_j}^{x_{j+1}} |v''(t)|^2\, dt.$$

En intégrant par rapport à x sur l'intervalle $[x_j, x_{j+1}]$, on obtient

$$\int_{x_j}^{x_{j+1}} |v(x) - r_h v(x)|^2\, dx \leqslant h^4 \int_{x_j}^{x_{j+1}} |v''(t)|^2\, dt,$$

ce qui, par sommation en j, donne exactement (70).

La démonstration de (71) est tout à fait similaire : pour $v \in C^2([0,1])$ et $x \in]x_j, x_{j+1}[$, on écrit

$$\begin{aligned} v'(x) - (r_h v)'(x) &= v'(x) - \frac{v(x_{j+1}) - v(x_j)}{h} \\ &= \frac{1}{h}\int_{x_j}^{x_{j+1}} \big(v'(x) - v'(t)\big)\, dt = \frac{1}{h}\int_{x_j}^{x_{j+1}} \int_t^x v''(y)\, dy\, dt. \end{aligned}$$

donc en appliquant Cauchy-Schwarz deux fois

$$\begin{aligned} \big|v'(x) - (r_h v)'(x)\big|^2 &\leqslant \frac{1}{h^2}\left(\int_{x_j}^{x_{j+1}} \int_t^x v''(y)\, dy\, dt.\right)^2 \leqslant \frac{1}{h}\int_{x_j}^{x_{j+1}} \left(\int_t^x v''\right)^2 dt \\ &\leqslant \int_{x_j}^{x_{j+1}} dt \int_t^x \big|v''\big|^2 \leqslant \int_{x_j}^{x_{j+1}} dt \int_{x_j}^{x_{j+1}} \big|v''\big|^2 \\ &= h\int_{x_j}^{x_{j+1}} \big|v''\big|^2. \end{aligned}$$

□

Ceci prouve le lemme suivant

Lemma 7.4 (D'interpolation)**.** *Soit r_h l'opérateur d'interpolation $\mathbb{P}_1$. Il existe une constante C indépendante de h telle que, pour tout $v \in C^2([0,1])$ et tout $h \in]0,1]$,*

$$\|v - r_h v\|_{H^1(0,1)} \leqslant \sqrt{2} h \|v''\|_{L^2(0,1)}. \tag{73}$$

On peut maintenant prouver le théorème en utilisant Céa. On définit l'espace

$$W := \{v \in H^1([0,1], \mathbb{R}) \mid v(0) = v(1) = 0\},$$

muni du produit scalaire

$$\langle \varphi, \phi \rangle_W := \int_0^1 \varphi' \phi',$$

on note la norme associée $\|\cdot\|_{V_0}$.

Lemma 7.5. *L'application $\|\cdot\|_{V_0}$ est une norme sur $H_0^1([0,1])$.*

Proof. On a la positivité, l'homogénéité et l'inégalité triangulaire. On justifie la séparation. Soit $w \in H_0^1([0,1])$ telle que $\|w\|_{V_0} = 0$. Alors $w' = 0$ et comme $H^1([0,1]) \subset \mathcal{C}^0([0,1])$, w est constante. Puisque $w(0) = 0$, alors $w = 0$. □

Nous souhaitons appliquer le lemme de Céa. Il faut vérifier que a est bilinéaire continue symétrique coercive et que T est continue, tout cela pour l'espace V_0. On a $a(v,v) = \|v\|_{V_0}^2$ donc a est 1-coercive. De plus, $|a(v,u)| \leqslant \|v'\|_{L^2} \|u'\|_{L^2} = \|v\|_{V_0} \|u\|_{V_0}$ donc a est continue avec $M = 1$. On laisse les autres propriétés en exercice, et on peut appliquer le lemme de Céa.

On majore l'estimation en choisissant $v_h = r_h u$ qui est bien un élément de V_h^0

$$\begin{aligned}\left\|(u-u_h)'\right\|_{L^2} = \|u-u_h\|_{V_0} &\underset{\substack{\text{Céa}\\ \text{Théorème 6.4}\\ (63)}}{\leqslant} \inf_{w\in V_h^0} \|u-w\|_{V_0} \\ &\underset{r_h u\in V_h}{\leqslant} \|u-r_h u\|_{V_0} \leqslant \|u-r_h u\|_{H^1} \underset{(73)}{\leqslant} \sqrt{2}h \left\|v''\right\|_{L^2(0,1)} \\ &= \sqrt{2}h \|f\|_{L^2(0,1)}\,.\end{aligned}$$