Université Paris Diderot M1 Mathématiques – Année 2015-2016

---

# Cours de Bases des méthodes numériques

---

Matthieu Bonnivard
TD et TP : Adina Ciomaga

# Chapitre 1

# Équations différentielles et approximations numériques

## 1.1 Rappels sur les équations différentielles ordinaires (EDO)

Soit $m \geq 1$ un entier, $U \subset \mathbb{R} \times \mathbb{R}^m$ un ouvert et $f : U \to \mathbb{R}^m$ une application continue. On considère l'équation différentielle

$$(E) \quad y' = f(t, y), \quad (t, y) \in U.$$

**Définition 1.1.** Une solution de $(E)$ sur un intervalle $I \subset \mathbb{R}$ est une fonction dérivable $y : I \to \mathbb{R}^m$ telle que

(i) $\forall t \in I \quad (t, y(t)) \in U$

(ii) $\forall t \in I \quad y'(t) = f(t, y(t))$

Le qualificatif « ordinaire » associé à l'équation $(E)$ signifie que la fonction inconnue $y$ dépend d'une seule variable $t$. (Lorsqu'il y a plusieurs inconnues $t$, $x_1$, $x_2$, ..., et plusieurs dérivées partielles $\dfrac{\partial y}{\partial t}, \dfrac{\partial y}{\partial x_1}, \dfrac{\partial y}{\partial x_2}, \ldots$, on parle d'équation aux dérivées partielles (EDP)).

**Définition 1.2** (Problème de Cauchy)**.** Étant donné un point $(t_0, y_0) \in U$, le problème de Cauchy consiste à trouver une solution $y : I \to \mathbb{R}^m$ de $(E)$, définie sur un intervalle $I$ contenant $t_0$ dans son intérieur, telle que $y(t_0) = y_0$.

**Interprétation.** Dans de nombreuses situations concrètes, le paramètre $t$ représente le temps et $y = (y_1, y_2, \ldots, y_m)$ est une famille de paramètres décrivant l'état d'un système matériel donné. L'équation $(E)$ traduit la loi d'évolution du système en fonction du temps et de la valeur des paramètres. Résoudre le problème de Cauchy revient à prévoir l'évolution du système au cours du temps, sachant qu'à l'instant $t = t_0$ le système est décrit par les paramètres $(y_{01}, y_{02}, \ldots, y_{0m})$. On dit que $(t_0, y_0)$ sont les DONNÉES INITIALES du problème de Cauchy.

**Lemme 1.1** (Formulation intégrale du problème de Cauchy)**.** *Une fonction $y : I \to \mathbb{R}^m$ est solution du problème de Cauchy de données initiales $(t_0, y_0)$ si et seulement si*

*(i) $y$ est continue et $\forall t \in I \quad (t, y(t)) \in U$*

*(ii)* $(EI) \quad \forall t \in I \quad y(t) = y_0 + \displaystyle\int_{t_0}^{t} f(s, y(s))\ \mathrm{d}s$

**Preuve.** Si $y$ est continue, alors $f$ étant continue (comme fonction de plusieurs variables), l'application $t \in I \mapsto f(t, y(t)) \in \mathbb{R}^m$ est continue. Par conséquent, l'application $t \in I \mapsto \displaystyle\int_{t_0}^{t} f(s, y(s))\ \mathrm{d}s$ est de classe $\mathcal{C}^1$, et

$$\forall t \in I \quad \frac{\mathrm{d}}{\mathrm{d}t}\left(\int_{t_0}^{t} f(s, y(s))\ \mathrm{d}s\right) = f(t, y(t)).$$

Ainsi $y$ est dérivable sur $I$ et $\forall t \in I, y'(t) = f(t, y(t))$. De plus, en $t = t_0, y(t_0) = y_0$ donc $y$ est solution du problème de Cauchy.

Réciproquement, si $y$ est dérivable et satisfait $\forall t \in I, y'(t) = f(t, y(t))$, en intégrant cette relation sur l'intervalle $[t_0, t]$, on obtient l'équation intégrale $(EI)$. □

**Solutions maximales.**

**Définition 1.3** (Prolongement)**.** Soient $y : I \to \mathbb{R}^m, \tilde{y} : \tilde{I} \to \mathbb{R}^m$ deux solutions de $(E)$. On dit que $\tilde{y}$ est un PROLONGEMENT de $y$ si $I \subset \tilde{I}$ et pour tout $t \in I$, $\tilde{y}(t) = y(t)$.

**Définition 1.4** (Solution maximale)**.** On dit que $y : I \to \mathbb{R}^m$, solution de $(\mathcal{C})$, est MAXIMALE si elle n'admet pas de prolongement $\tilde{y} : \tilde{I} \to \mathbb{R}^m$ avec $I \subsetneq \tilde{I}$.

**Théorème 1.1.** *Toute solution $y$ se prolonge en une solution maximale $\tilde{y}$ (non nécessairement unique).*

**Existence de solutions.**

**Théorème 1.2** (Cauchy-Peano-Arzelà)**.** *Soit $U \subset \mathbb{R} \times \mathbb{R}^m$ un ouvert et $f : U \to \mathbb{R}^m$ une application* CONTINUE*. Alors pour tout point $(t_0, y_0) \in U$, le problème de Cauchy de données initiales $(t_0, y_0)$ possède au moins une solution $y$.*

**Corollaire 1.1.** *Si $f$ est continue, alors tout problème de Cauchy de données initiales $(t_0, y_0) \in U$ possède une solution maximale.*

**Exemple 1.1** (Non unicité de la solution du problème de Cauchy)**.** On considère le problème de Cauchy suivant :

$$(\mathcal{C}) \quad \begin{cases} y' &= 2\sqrt{|y|} \\ y(0) &= 0 \end{cases}$$

Les fonctions définies sur $\mathbb{R}$ par $y_1(t) = 0$, $y_2(t) = (\max(t, 0))^2$, $\forall t \in \mathbb{R}$, sont deux solutions maximales de $(\mathcal{C})$.

Pour assurer l'unicité de la solution, on a besoin d'une hypothèse supplémentaire, appelée condition de Lipschitz locale.

**Définition 1.5** (Condition de Lipschitz locale en $y$)**.** Soit $U \subset \mathbb{R} \times \mathbb{R}^m$ un ouvert et $f : U \to \mathbb{R}$. On dit que $f$ est LOCALEMENT LIPSCHITZIENNE en $y$ si, pour tout point $(t_0, y_0) \in U$, il existe un réel $T_0 > 0$, un rayon $r_0 > 0$ et une constante $k > 0$ tels que, en notant $\overline{B}(y_0, r_0)$ la boule fermée de $\mathbb{R}^m$ centrée en $y_0$, de rayon $r_0$, on ait :

(i) $[t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0) \subset U$

(ii) $\forall t \in [t_0 - T_0, t_0 + T_0],\ \forall y_1, y_2 \in \overline{B}(y_0, r_0) \quad \|f(t, y_1) - f(t, y_2)\| \leq k\|y_1 - y_2\|$

**Proposition 1.1** (Condition suffisante)**.** *Pour que $f : U \to \mathbb{R}^m$ soit localement lipschitzienne en $y$, il suffit que ses dérivées partielles $\dfrac{\partial f_i}{\partial y_j}, 1 \leq i, j \leq m$ soient* CONTINUES *sur $U$.*

**Preuve.** On note $C_0$ le cylindre $C_0 = [t_0 - T_0, t_0 + T_0] \times \overline{B}(y_0, r_0)$, et pour chaque composante $f_i$, on note $\nabla f_i$ le gradient de $f_i$ en les variables spatiales

$$\nabla f_i = \begin{pmatrix} \frac{\partial f_i}{\partial y_1} & \frac{\partial f_i}{\partial y_2} & \dots & \frac{\partial f_i}{\partial y_m} \end{pmatrix}^T$$

et on définit

$$M = \max_{1\leq i\leq m} \sup_{(t,y)\in C_0} \|\nabla f_i(t,y)\|.$$

$M$ est fini car $C_0$ est compact et toutes les dérivées partielles sont continues. Soit $1 \leq i \leq m$ fixé ; d'après le théorème des accroissements finis, appliqué à la restriction de $f_i(t,\cdot)$ au segment $[y_1, y_2]$, il existe $\xi_i \in ]y_1, y_2[$ tel que

$$f_i(t,y_1) - f_i(t,y_2) = \nabla f_i(t,\xi_i) \cdot (y_1 - y_2)$$

On en déduit :

$$|f_i(t,y_1) - f_i(t,y_2)| \leq \|\nabla f_i(t,\xi_i)\| \|y_1 - y_2\|.$$

En passant au max sur $i$, on obtient

$$\max_{1\leq i\leq m} |f_i(t,y_1) - f_i(t,y_2)| \leq M\|y_1 - y_2\|,$$

d'où le résultat en utilisant l'équivalence des normes sur $\mathbb{R}^m$. □

**Théorème 1.3** (Cauchy-Lipschitz)**.** *Soit $U \subset \mathbb{R}\times\mathbb{R}^m$ un ouvert, et $f : U \to \mathbb{R}^m$ une application continue, et localement lipschitzienne en $y$. Alors pour toute donnée initiale $(t_0, y_0) \in U$, le problème de Cauchy associé possède une unique solution maximale.*

## 1.2 Approximation numérique des EDO

Dans toute la suite, on se place en dimension d'espace $m = 1$. Étant donné $t_0 \in \mathbb{R}$, $T > 0$, $y_0 \in \mathbb{R}$ et une application $f : [t_0, t_0 + T] \times \mathbb{R} \to \mathbb{R}$, on cherche à calculer une approximation de la solution du problème de Cauchy

$$(\mathcal{C}) \quad \begin{cases} y' = f(t,y), \quad t \in [t_0, t_0 + T] \\ y(t_0) = y_0 \end{cases}$$

On suppose $f$ suffisamment régulière pour garantir l'existence et l'unicité de la solution de $(\mathcal{C})$. On note cette solution $y : [t_0, t_0 + T] \to \mathbb{R}$.

Le principe de l'approximation est de remplacer le problème initial, continu, par un problème discret, dans lequel on détermine des valeurs approchées de $y$ en un nombre fini de points de l'intervalle $[t_0, t_0 + T]$. Pour cela, on se donne un entier $N \in \mathbb{N}$ et on introduit une subdivision $t_0 < t_1 < \ldots < t_N = t_0 + T$ de $[t_0, t_0 + T]$. On cherche à calculer une valeurs approchées $y_n$ de la valeur exacte $y(t_n)$ prise par la solution à l'instant $t_n$.

**Vocabulaire.**

— $h_n := t_{n+1} - t_n$ est le pas du schéma,

— $h_{max} := \max_{0 \leq n \leq N-1} h_n$ est le pas maximal du schéma.

**Définition 1.6** (Méthode à un pas)**.** On appelle MÉTHODE À UN PAS une méthode numérique itérative qui, à chaque étape, permet de calculer $y_{n+1}$ en utilisant uniquement l'approximation $y_n$, ainsi que les valeurs de $t_n$, $h_n$ et la donnée $f$. On écrira une méthode à un pas sous la forme suivante :

$$\frac{y_{n+1} - y_n}{h_n} - \phi(t_n, y_n, h_n) = 0, \quad 0 \leq n < N \tag{1.1}$$

où $\phi : [t_0, t_0 + T] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ est une fonction donnée.

Dans la définition du schéma (1.1), le quotient $\dfrac{y_{n+1} - y_n}{h_n}$ est une approximation de $y'(t_n)$, et $\phi(t_n, y_n, h_n)$ est une approximation de $f(t_n, y(t_n))$.

**Exemple 1.2** (Méthode d'Euler explicite)**.** La méthode d'Euler explicite est la méthode à un pas associée à la fonction $\phi(t, y, h) = f(t, y)$ et définie par la formule de récurrence

$$y_{n+1} = y_n + h_n f(t_n, y_n).$$

**Exemple 1.3** (Méthode du point milieu)**.** C'est la méthode définie par la relation de récurrence

$$y_{n+1} = y_n + h_n f(t_n + \frac{1}{2}h_n, y_n + \frac{1}{2}h_n f(t_n, y_n)).$$

La question qui s'impose : a-t-on

$$\max_{0 \leq n \leq N} |y_n - y(t_n)| \to 0 \quad \text{quand } h_{max} \to 0 \ ?$$

Pour y répondre, on introduit deux notions : la consistance et la stabilité du schéma.

### 1.2.1 Consistance

La consistance est un moyen de quantifier si un schéma est bien adapté à l'approximation de la solution d'une équation différentielle donnée. L'idée est que les valeurs *exactes* $(y(t_n))_{0 \leq n \leq N}$ prises par la solution aux points de la subdivision $(t_n)_{0 \leq n \leq N}$, ne satisfont pas le schéma (1.1), mais que le résultat obtenu en remplaçant $y_n$ par $y(t_n)$ dans ce schéma, doit être petit si l'on veut que le schéma capture le comportement de la solution exacte. Cela conduit à la définition suivante :

**Définition 1.7** (Consistance).

1. Pour tout $0 \leq n \leq N$, l'ERREUR DE CONSISTANCE $e_n$ relative à la solution exacte $y$ pour le schéma (1.1), est définie par
$$e_n = \frac{y(t_{n+1}) - y(t_n)}{h_n} - \phi(t_n, y(t_n), h_n)$$
2. Le schéma (1.1) est CONSISTANT si
$$\max_{0 \leq n \leq N} |e_n| \to 0 \quad \text{quand } h_{max} \to 0$$
3. Soit $p \in \mathbb{N}^*$. Le schéma (1.1) est CONSISTANT D'ORDRE $p$ s'il existe une constante $C \geq 0$, dépendant de $f$, $T$, $y_0$ (mais pas de $h_{max}$) telle que
$$\max_{0 \leq n \leq N} |e_n| \leq C h_{max}^p$$

**Proposition 1.2.** *On suppose $f$ de classe $\mathcal{C}^1$ (et donc $y$ de classe $\mathcal{C}^2$). Alors le schéma d'Euler explicite est consistant d'ordre* 1.

**Preuve.** En appliquant la formule de Taylor-Lagrange à l'ordre 2 : pour tout $0 \leq n \leq N$, il existe $\theta_n \in ]0,1[$ tel que
$$y(t_{n+1}) = y(t_n) + h_n y'(t_n) + \frac{h_n^2}{2}\, y''(t_n + \theta_n h_n)$$
En utilisant $y'(t_n) = f(t_n, y(t_n))$, on obtient donc
$$\frac{y(t_{n+1}) - y(t_n)}{h_n} = f(t_n, y(t_n)) + \frac{h_n}{2}\, y''(t_n + \theta_n h_n),$$
d'où
$$\forall 0 \leq n \leq N \quad e_n = \frac{h_n}{2}\, y''(t_n + \theta_n h_n).$$
Par conséquent,
$$\max_{0 \leq n \leq N} |e_n| \leq \|y''\|_\infty \frac{h_{max}}{2}$$
Remarquons que d'après l'équation $y'(t) = f(t, y(t))$, on a
$$\begin{aligned} y''(t) &= \partial_t f(t, y(t)) + y'(t)\ \partial_y f(t, y(t)) \\ &= \partial_t f(t, y(t)) + f(t, y(t))\ \partial_y f(t, y(t)), \end{aligned}$$
donc $\|y''\|_\infty$ est contrôlé par les normes uniformes de $f$ et de ses dérivées premières. □

**Proposition 1.3.** *On suppose $f$ de classe $\mathcal{C}^2$ (et donc $y$ de classe $\mathcal{C}^3$). Alors le schéma du point milieu est consistant d'ordre* 2.

**Preuve.** En TD. □

### 1.2.2 Stabilité

Pour établir la convergence d'un schéma, on a besoin de comparer les valeurs exactes $y(t_n)$ prises par la solution aux points de la subdivision, avec les valeurs approchées $y_n$ calculées par le schéma. Bien entendu, les valeurs exactes $y(t_n)$ ne sont pas connues. Pour établir cette comparaison, on va interpréter les valeurs exactes $y(t_n)$ comme des solutions approchées du schéma (1.1), dans lequel les erreurs de consistance $e_n$ jouent le rôle d'un terme source, en écrivant :

$$\forall 0 \leq n \leq N \quad \frac{y(t_{n+1}) - y(t_n)}{h_n} - \phi(t_n, y(t_n), h_n) = e_n.$$

Pour pouvoir ensuite comparer $y(t_n)$ et $y_n$, on introduit la notion de *stabilité* du schéma (1.1).

**Définition 1.8** (Stabilité d'un schéma à un pas)**.** On dit que le schéma (1.1) est STABLE s'il existe une constante $K > 0$ telle que, étant donnée une suite $\varepsilon_n > 0$ et deux suites $y_n, \tilde{y}_n$ vérifiant les relations

$$\frac{y_{n+1} - y_n}{h_n} - \phi(t_n, y_n, h_n) = 0, \tag{1.2}$$

$$\frac{\tilde{y}_{n+1} - \tilde{y}_n}{h_n} - \phi(t_n, \tilde{y}_n, h_n) = \varepsilon_n, \tag{1.3}$$

on a

$$\max_{0 \leq n \leq N} |y_n - \tilde{y}_n| \leq K \left( |y_0 - \tilde{y}_0| + \max_{0 \leq n \leq N-1} |\varepsilon_n| \right)$$

Cette définition signifie qu'une petite erreur initiale $|\tilde{y}_0 - y_0|$ et de petites erreurs $\varepsilon_n$ dans le calcul récurrent des $\tilde{y}_n$ (par exemple, des erreurs d'arrondi dues à un calcul sur machine) provoquent une erreur finale $\max_{0 \leq n \leq N} |y_n - \tilde{y}_n|$ quantifiable.

**Proposition 1.4** (Critère de stabilité)**.** *Supposons $\phi$ lipschitzienne par rapport à $y$. Alors le schéma* (1.1) *est stable.*

**Preuve.** On note $k$ la constante de Lipschitz associée à $\phi$. En faisant la différence des relations (1.2) et (1.3), on obtient

$$y_{n+1} - \tilde{y}_{n+1} = y_n - \tilde{y}_n + h_n \left( \phi(t_n, y_n, h_n) - \phi(t_n, \tilde{y}_n, h_n) \right) - h_n \, \varepsilon_n, \quad 0 \leq n \leq N-1.$$

En utilisant la condition de Lipschitz sur $\phi$, on en déduit pour tout $0 \leq n \leq N-1$

$$\begin{aligned} |y_{n+1} - \tilde{y}_{n+1}| &\leq (1 + kh_n)\ |y_n - \tilde{y}_n| + h_n\ |\varepsilon_n| \\ &\leq e^{kh_n} |y_n - \tilde{y}_n| + h_n\ |\varepsilon_n| \end{aligned}$$

En écrivant $e^{kh_n} = e^{k(t_{n+1}-t_n)}$, on peut réécrire cette inégalité

$$e^{-kt_{n+1}} |y_{n+1} - \tilde{y}_{n+1}| \leq e^{-kt_n} |y_n - \tilde{y}_n| + h_n\ |\varepsilon_n|, \quad 0 \leq n \leq N-1.$$

On fixe un entier $1 \leq p \leq N-1$ et on somme les inégalités précédentes pour $n$ variant de 0 à $p-1$. On obtient :

$$\sum_{n=1}^{p} e^{-kt_n} |y_n - \tilde{y}_n| \leq \sum_{n=0}^{p-1} e^{-kt_n} |y_n - \tilde{y}_n| + \left( \sum_{n=0}^{p-1} h_n \right)\ \max_{0 \leq n \leq N-1} |\varepsilon_n|$$

Après simplification,

$$e^{-kt_p} |y_p - \tilde{y}_p| \leq |y_0 - \tilde{y}_0| + \left( \sum_{n=0}^{p-1} h_n \right)\ \max_{0 \leq n \leq N-1} |\varepsilon_n|.$$

Enfin, en remarquant que $e^{kt_p} \leq e^{kT}$ et que $\sum_{n=0}^{p-1} h_n \leq T$, on obtient :

$$|y_p - \tilde{y}_p| \leq e^{kT} \left( |y_0 - \tilde{y}_0| + T\ \max_{0 \leq n \leq N-1} |\varepsilon_n| \right), \quad 1 \leq p \leq N-1.$$

□

**Définition 1.9** (Convergence)**.** Un schéma est dit CONVERGENT si $\max_{0 \leq n \leq N-1} |y_n - y(t_n)| \to 0$ quand $h_{max} \to 0$.

**Théorème 1.4.** *Si un schéma est stable et consistant, il est convergent.*

**Preuve.** Par définition de l'erreur de consistance, la relation (1.3) est satisfaite avec $\tilde{y}_n = y(t_n), \tilde{y}_{n+1} = y(t_{n+1})$ et $\varepsilon_n = e_n$. Puisque $y(t_0) = y_0$, par définition de la stabilité, on peut donc écrire :

$$\max_{0 \leq n \leq N} |y_n - y(t_n)| \leq K\ \max_{0 \leq n \leq N-1} |e_n|.$$

Le schéma étant consistant, on conclut à la convergence puisque $\max_{0 \leq n \leq N-1} |e_n| \to 0$ quand $h_{max} \to 0$. □

## 1.3 Méthodes de Runge-Kutta

### 1.3.1 Principe général

On cherche à discrétiser le problème de Cauchy $(\mathcal{C})$ à l'aide d'une subdivision $t_0 < t_1 < \ldots < t_N = t_0 + T$. Pour cela, on va calculer récursivement les points $(t_n, y_n)$ en utilisant des points intermédiaires $(t_{n,i}, y_{n,i})$, où

$$t_{n,i} = t_n + c_i h_n, \quad 1 \leq i \leq q, \quad c_i \in [0,1].$$

À chacun de ces points, on associe une pente

$$p_{n,i} = f(t_{n,i}, y_{n,i}).$$

Pour calculer les approximations $y_{n,i}$, on écrit la formule intégrale

$$\begin{aligned} y(t_{n,i}) &= y(t_n) + \int_{t_n}^{t_{n,i}} f(t, y(t)) \, \mathrm{d}t \\ &= y(t_n) + h_n \int_0^{c_i} f(t_n + uh_n, y(t_n + uh_n)) \, \mathrm{d}u \end{aligned}$$

grâce au changement de variable $t = t_n + uh_n$. De même,

$$y(t_{n+1}) = y(t_n) + h_n \int_0^1 f(t_n + uh_n, y(t_n + uh_n)) \, \mathrm{d}u$$

Le principe de la méthode consiste alors à se donner, pour chaque $i = 1, 2, \ldots, q$, une méthode d'intégration approchée (ou méthode de quadrature)

$$(M_i) \qquad \int_0^{c_i} g(t) \, \mathrm{d}t \approx \sum_{1 \leq j < i} a_{ij} g(c_j),$$

ainsi qu'une méthode d'intégration approchée sur $[0,1]$ :

$$(M) \qquad \int_0^1 g(t) \, \mathrm{d}t \approx \sum_{1 \leq j \leq q} b_j g(c_j).$$

En appliquant ces méthodes à la fonction $g(u) = f(t_n + uh_n, y(t_n + uh_n))$, on obtient :

$$\begin{aligned} y(t_{n,i}) &\approx y(t_n) + h_n \sum_{1 \leq j < i} a_{ij} f(t_{n,j}, y(t_{n,j})) \\ y(t_{n+1}) &\approx y(t_n) + h_n \sum_{1 \leq j \leq q} b_j f(t_{n,j}, y(t_{n,j})) \end{aligned}$$

La méthode de Runge-Kutta correspondante est définie par l'algorithme

$$\begin{cases} \left[\begin{array}{l} t_{n,i} = t_n + c_i h_n \\ y_{n,i} = y_n + h_n \sum_{1\le j<i} a_{ij} p_{n,j} \\ p_{n,i} = f(t_{n,i}, y_{n,i}) \end{array}\right] & 1 \le i \le q \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n \sum_{1\le j\le q} b_j p_{n,j} \end{cases}$$

On la représente conventionnellement par le tableau

$$\begin{array}{cc|ccccc} (M_1) & c_1 & 0 & 0 & \dots & 0 & 0 \\ (M_2) & c_2 & a_{21} & 0 & \dots & 0 & 0 \\ & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & & 0 & 0 \\ (M_q) & c_q & a_{q1} & a_{q2} & \dots & a_{q,q-1} & 0 \\ \hline (M) & & b_1 & b_2 & \dots & b_{q-1} & b_q \end{array}$$

où les méthodes d'intégration approchées correspondent aux lignes. Par convention, $a_{ij} = 0$ pour $j \ge i$.

**Hypothèse 1.1.** On suppose que les méthodes d'intégrations $(M_i)$ et $(M)$ sont au moins d'ordre 0 (c'est-à-dire, qu'elles fournissent le résultat exact si $g$ est une constante), ce qui se traduit par

$$\sum_{1\le j<i} a_{ij} = c_i, \qquad \sum_{1\le j\le q} b_j = 1$$

(la somme des coefficients est égale à la longueur de l'intervalle d'intégration).

En particulier, on aura toujours :

$$c_1 = 0, \quad t_{n,1} = t_n, \quad y_{n,1} = y_n, \quad p_{n,1} = f(t_n, y_n).$$

### 1.3.2 Exemples

— $\mathbf{q = 1}$. Le seul choix possible est $\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}$. On a $c_1 = 0$, $a_{11} = 0$, $b_1 = 1$. L'algorithme s'écrit

$$\begin{cases} p_{n,1} = f(t_n, y_n) \\ t_{n+1} = t_n + h_n \\ y_{n+1} = y_n + h_n \; p_{n,1} \end{cases}$$

On retrouve la méthode d'Euler explicite.

— $\mathbf{q = 2}$. On considère les tableaux de la forme

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \alpha & \alpha & 0 \\ \hline & 1-\dfrac{1}{2\alpha} & \dfrac{1}{2\alpha} \end{array}, \qquad \text{où } \alpha \in ]0,1].$$

**Remarque 1.1.** Le choix du coefficient $b_2 = \dfrac{1}{2\alpha}$ permet d'obtenir une méthode d'ordre $\geq 2$.

- Pour $\alpha = \frac{1}{2}$, on retrouve la méthode du point milieu

$$y_{n+1} = y_n + h_n \ f(t_n + \frac{h_n}{2}, y_n + \frac{h_n}{2} \ f(t_n, y_n)).$$

Cette méthode est basée sur la méthode d'intégration du point milieu

$$(M) \qquad \int_0^1 g(t) \ \mathrm{d}t \approx g(\frac{1}{2}).$$

- Pour $\alpha = 1$, on obtient la *méthode de Heun* (voir le TD)

$$y_{n+1} = y_n + h_n \left( \frac{1}{2} \ f(t_n, y_n)) + \frac{1}{2} \ f(t_{n+1}, y_n + h_n f(t_n, y_n)) \right),$$

basée sur la méthode des trapèzes

$$(M) \qquad \int_0^1 g(t) \ \mathrm{d}t \approx \frac{1}{2} \ (g(0) + g(1)).$$

- $\mathbf{q = 4}$. Il s'agit de la méthode de Runge-Kutta « classique » (appelée $RK4$), définie par le tableau

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

(voir TP). C'est une méthode d'ordre 4 qui utilise plusieurs méthodes d'intégration

différentes :

$(M_2)$ $\int_0^{\frac{1}{2}} g(t)\ \mathrm{d}t \approx \frac{1}{2}\ g(0)$ : rectangles à gauche,

$(M_3)$ $\int_0^{\frac{1}{2}} g(t)\ \mathrm{d}t \approx \frac{1}{2}\ g(\frac{1}{2})$ : rectangles à droite,

$(M_4)$ $\int_0^{1} g(t)\ \mathrm{d}t \approx \frac{1}{2}\ g(\frac{1}{2})$ : point milieu,

$(M)$ $\int_0^{1} g(t)\mathrm{d}t \approx \frac{1}{6}\ g(0) + \frac{2}{6}\ g(\frac{1}{2}) + \frac{2}{6}\ g(\frac{1}{2}) + \frac{1}{6}\ g(1)$ : méthode de Simpson

# Chapitre 2

# Méthode des différences finies

La méthode des différences finies est une méthode d'approximation d'EDP, qui généralise les idées vues au chapitre précédent en approchant les opérateurs différentiels par des quotients formés à partir de valeurs discrètes.

**Rappels sur les opérateurs différentiels usuels dans $\mathbb{R}^d$.** On se donne $\Omega \subset \mathbb{R}^d$ un ouvert, $f : \Omega \to \mathbb{R}$ et $\overrightarrow{f} : \Omega \to \mathbb{R}^d$ deux applications de classe $\mathcal{C}^1$. On définit :

— le gradient $\nabla f$ par

$$\forall x \in \Omega \quad \nabla f(x) = \left( \frac{\partial f}{\partial x_1}(x) \quad \dots \quad \frac{\partial f}{\partial x_d}(x) \right)^T.$$

— la divergence div $\overrightarrow{f}$ (ou $\nabla \cdot \overrightarrow{f}$) par

$$\forall x \in \Omega \quad \text{div } \overrightarrow{f}(x) = \sum_{i=1}^{d} \frac{\partial f_i}{\partial x_i}(x).$$

C'est la trace de la matrice jacobienne de $\overrightarrow{f}$.

— le laplacien $\Delta f$ : si $f$ est de classe $\mathcal{C}^2$,

$$\forall x \in \Omega \quad (\Delta f)(x) = \sum_{i=1}^{d} \frac{\partial^2 f}{\partial x_i^2}(x).$$

On a : $\Delta f = \text{div } \nabla f = \text{tr}(Hf)$ où $Hf$ est la matrice hessienne de $f$.

**Quelques exemples d'EDP.**

— L'équation de Poisson, d'inconnue $u(x)$ :

$$\left\{\begin{array}{rcll} -\Delta u & = & f & \text{dans } \Omega \\ u & = & 0 & \text{sur } \partial\Omega \end{array}\right.$$

— L'équation de la chaleur, d'inconnue $u(t,x)$ :

$$\left\{\begin{array}{rcll} \partial_t u - \nu\Delta u & = & 0 & \text{dans } (0,T)\times\Omega \\ u(0,x) & = & u^0(x) & \text{dans } \Omega \\ u(t,x) & = & 0 & \text{sur } (0,T)\times\partial\Omega \end{array}\right.$$

C'est un problème de Cauchy dans lequel la donnée initiale $u^0$ est une fonction de $x \in \Omega$.

— L'équation des ondes, d'inconnue $u(t,x)$ :

$$\left\{\begin{array}{rcll} \partial_{tt}^2 u - c^2\Delta u & = & 0 & \text{dans } (0,T)\times\Omega \\ u(0,x) & = & u^0(x) & \text{dans } \Omega \\ \partial_t u(0,x) & = & u^1(x) & \text{dans } \Omega \\ u(t,x) & = & 0 & \text{sur } (0,T)\times\partial\Omega \end{array}\right.$$

C'est une équation d'ordre 2 en temps, qui nécessite donc de fixer une donnée initiale $u^0$ et une vitesse initiale $u^1$. On peut la voir comme un problème de Cauchy pour l'inconnue $(u(t,\cdot), \partial_t u(t,\cdot))$.

## 2.1 Principe de la méthode des différences finies

Nous allons illustrer la méthode des différences finies à l'aide de l'équation de la chaleur en dimension d'espace $d = 1$. On définit $\Omega = (0,1)$, des réels $T > 0, \nu > 0$ et on considère le problème suivant :

$$\left\{\begin{array}{rcll} \partial_t u - \nu\ \partial_{xx}^2 u & = & 0 & \text{dans } (0,T)\times(0,1) \\ u(0,x) & = & u^0(x) & \text{pour } x \in (0,1) \end{array}\right. \tag{2.1}$$

auquel il faudra ajouter des conditions aux limites $(CL)$ en $x = 0$ et $x = 1$. Ces conditions pourront être :

— des conditions de Dirichlet (homogènes) :

$$u(t,0) = 0,\ u(t,1) = 0 \quad \text{pour tout } t \in (0,T] \tag{2.2}$$

— des conditions de Neumann (homogènes) :

$$\partial_x u(t,0) = \partial_x u(t,1) \quad \text{pour tout } t \in (0,T]$$

— des conditions de périodicité :

$$u(t,0) = u(t,1) \quad \text{pour tout } t \in (0,T]$$

On supposera la régularité suivante pour les solutions du problème (2.1) :

$$u, \partial_t u, \partial_x u, \partial^2_{xx} u \in \mathcal{C}([0,T] \times [0,1]) \tag{2.3}$$

Dans toute la suite, on notera $u : [0,T] \times [0,1]$ une solution de (2.1) possédant au moins la régularité (2.3).

**Discrétisation du domaine** $[0,T] \times [0,1]$**.** On se donne un pas de temps $\Delta t$ et on définit la suite $(t_n)_{n\in\mathbb{N}}$ par

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

On fixe un entier $J \in \mathbb{N}^*$ et on subdivise l'intervalle $[0,1]$ en $J+1$ sous-intervalles de même longueur $\Delta x = \dfrac{1}{J+1}$. On note $(x_j)_{0\leq j\leq J+1}$ la subdivision définie par

$$\forall j \in [|0, J+1|] \quad x_j = j\Delta x$$

On cherche à calculer des approximations des valeurs prises par $u$ aux points $x_j$, aux différents instants $t_n$. On note $u_j^n \in \mathbb{R}$ l'approximation de $u(t_n, x_j)$. Dans toute la section, on considèrera les conditions aux limites de Dirichlet (2.2), qui impliquent que les valeurs prises par $u$ aux points $x = 0$, $x = 1$ sont fixées, et ne sont plus des inconnues du problème. On impose alors

$$u_0^n = u_{J+1}^n = 0 \quad \forall n \geq 0$$

et l'on ne conserve que $J$ inconnues à chaque instant $t_n$. On pourra les regrouper sous la forme d'un vecteur inconnu $U^n \in \mathbb{R}^J$, défini par

$$U^n = \left(u_1^n \dots u_J^n\right)^T$$

**Traitement de la condition initiale.** La donnée initiale $u^0$ étant fixée, on initialise

$$\forall 1 \leq j \leq J \quad u_j^0 = u^0(x_j).$$

**Discrétisation des opérateurs différentiels.** On écrit l'équation satisfaite par $u$ à l'instant $t_n$ et au point $x_j$ :

$$\partial_t u(t_n, x_j) - \nu \, \partial_{xx}^2 u(t_n, x_j) = 0 \quad \forall n \geq 0, \quad \forall 1 \leq j \leq J$$

On va approcher chaque quantité $\partial_t u(t_n, x_j)$, $\partial_{xx}^2 u(t_n, x_j)$ en utilisant des formules de Taylor. On supposera toujours que $u$ est suffisamment régulière pour que les formules soient valables.

— dérivée en temps : en supposant $\partial_{tt}^2 u$ bornée, on peut écrire :

$$u(t_n + \Delta t, x_j) = u(t_n, x_j) + \Delta t \, \partial_t u(t_n, x_j) + O((\Delta t)^2)$$

d'où

$$\partial_t u(t_n, x_j) = \frac{u(t_n + \Delta t, x_j) - u(t_n, x_j)}{\Delta t} + O(\Delta t)$$

ce qui conduit à l'approximation :

$$\partial_t u(t_n, x_j) \approx \frac{u(t_n + \Delta t, x_j) - u(t_n, x_j)}{\Delta t} \tag{2.4}$$

Une autre manière naturelle d'approcher $\partial_t u(t_n, x_j)$ est d'écrire la formule de Taylor au point $(t_n - \Delta t, x_j)$ :

$$u(t_n - \Delta t, x_j) = u(t_n, x_j) - \Delta t \, \partial_t u(t_n, x_j) + O((\Delta t)^2)$$

ce qui conduit à l'approximation :

$$\partial_t u(t_n, x_j) \approx \frac{u(t_n, x_j) - u(t_n - \Delta t, x_j)}{\Delta t} \tag{2.5}$$

— dérivée en espace : on écrit

$$u(t_n, x_j + \Delta x) = u(t_n, x_j) + \Delta x \, \partial_x u(t_n, x_j) + \frac{(\Delta x)^2}{2} \, \partial_{xx}^2 u(t_n, x_j) + \frac{(\Delta x)^3}{6} \, \partial_{xxx}^3 u(t_n, x_j) + O((\Delta x)^4)$$

$$u(t_n, x_j - \Delta x) = u(t_n, x_j) - \Delta x \, \partial_x u(t_n, x_j) + \frac{(\Delta x)^2}{2} \, \partial_{xx}^2 u(t_n, x_j) - \frac{(\Delta x)^3}{6} \, \partial_{xxx}^3 u(t_n, x_j) + O((\Delta x)^4)$$

En sommant, on obtient

$$\partial^2_{xx} u(t_n, x_j) = \frac{u(t_n, x_j + \Delta x) - 2u(t_n, x_j) + u(t_n, x_j - \Delta x)}{(\Delta x)^2} + O((\Delta x)^2)$$

d'où l'approximation :

$$\partial^2_{xx} u(t_n, x_j) \approx \frac{u(t_n, x_j + \Delta x) - 2u(t_n, x_j) + u(t_n, x_j - \Delta x)}{(\Delta x)^2} \tag{2.6}$$

**Schéma d'Euler explicite.** En utilisant les approximations (2.4) et (2.6), on aboutit au *schéma d'Euler explicite* :

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} - \nu \, \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} = 0, \quad n \geq 0, \quad 1 \leq j \leq J \tag{2.7}$$

En posant $c = \nu \dfrac{\Delta t}{(\Delta x)^2}$, on peut écrire cette relation sous la forme

$$u_j^{n+1} = c u_{j+1}^n + (1 - 2c) u_j^n + c u_{j-1}^n, \quad n \geq 0, \quad 1 \leq j \leq J$$

ou sous forme matricielle, en tenant compte des conditions aux limites $u_0^n = u_{J+1}^n = 0$ et en notant $U^n = \begin{pmatrix} u_1^n & \dots & u_J^n \end{pmatrix}^T$ :

$$U^{n+1} = M U^n, \quad n \geq 0 \tag{2.8}$$

où $M \in M_J(\mathbb{R})$ est définie par

$$M = \begin{pmatrix} 1-2c & c & 0 & \dots & \dots & \dots & 0 \\ c & 1-2c & c & 0 & \dots & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & \dots & 0 & c & 1-2c & c \\ 0 & \dots & \dots & \dots & 0 & c & 1-2c \end{pmatrix}$$

Le schéma (2.7) est qualifié d'explicite car il permet, à chaque étape, de calculer l'approximation à l'instant $t_{n+1}$ en utilisant les valeurs obtenues au temps $t_n$ et en appliquant une formule explicite (2.8).

**Schéma d'Euler implicite.** C'est le schéma que l'on obtient en utilisant les approximations (2.5) et (2.6) ; il s'écrit

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\nu\,\frac{u_{j+1}^{n+1}-2u_j^{n+1}+u_{j-1}^{n+1}}{(\Delta x)^2}=0,\quad n\geq 0,\quad 1\leq j\leq J \tag{2.9}$$

En notant à nouveau $c=\nu\dfrac{\Delta t}{(\Delta x)^2}$, on pourra l'écrire

$$-cu_{j+1}^{n+1}+(1+2c)u_j^{n+1}-cu_{j-1}^{n+1}=u_j^n,\quad n\geq 0,\quad 1\leq j\leq J$$

ou sous forme matricielle

$$AU^{n+1}=U^n,\quad n\geq 0 \tag{2.10}$$

où $A\in M_J(\mathbb{R})$ est définie par

$$A=\begin{pmatrix}
1+2c & -c & 0 & \dots & \dots & \dots & 0\\
-c & 1+2c & -c & 0 & \dots & \dots & \vdots\\
0 & \ddots & \ddots & \ddots & \ddots & \dots & \vdots\\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots\\
\vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0\\
\vdots & \dots & \dots & 0 & -c & 1+2c & -c\\
0 & \dots & \dots & \dots & 0 & -c & 1+2c
\end{pmatrix}$$

Le schéma (2.9) est qualifié d'implicite car, à chaque étape, le calcul de $U^{n+1}$ nécessite la résolution d'un système linéaire (2.10). Ce système possède une solution unique en vertu du lemme suivant :

**Lemme 2.1** (Hadamard)**.** *Soit $B=(b_{ij})_{1\leq i,j\leq J}$ une matrice de taille $J$, à diagonale strictement dominante, i.e.*

$$\forall i=1\dots J\quad |b_{ii}|>\sum_{j\neq i}|b_{ij}| \tag{2.11}$$

*Alors $B$ est inversible.*

**Preuve.** Par l'absurde, supposons qu'il existe $X\in\mathbb{R}^J\setminus\{0\}$ t.q. $BX=0$. On note $i$ un indice vérifiant $|x_i|=\max_{1\leq k\leq J}|x_k|$ (on a donc $|x_i|>0$). La $i$-ième ligne de la relation $BX=0$ s'écrit

$$b_{ii}x_i+\sum_{j\neq i}b_{ij}x_j=0$$

donc

$$b_{ii}x_i = -\sum_{j\neq i} b_{ij}x_j.$$

En prenant la valeur absolue puis en utilisant la définition du max, on en déduit

$$\begin{aligned} |b_{ii}||x_i| &\leq \sum_{j\neq i} |b_{ij}||x_j| \\ &\leq \sum_{j\neq i} |b_{ij}||x_i| \end{aligned}$$

Après division par $|x_i| > 0$, on trouve

$$|b_{ii}| \leq \sum_{j\neq i} |b_{ij}|.$$

Cela contredit (2.11). □

On vérifie facilement que $A$ est à diagonale strictement dominante, ce qui garantit l'inversibilité du système (2.10).

## 2.2 Erreurs de consistance et précision

On souhaite approcher la solution $u(t,x)$ d'une EDP linéaire $(\mathcal{E})$, sur le domaine $[0,T]\times[0,1]$. Nous allons considérer pour cela des schémas linéaires, de la forme

$$U^{n+1} = MU^n, \quad n \in \mathbb{N} \tag{2.12}$$

(Dans le cas du schéma d'Euler implicite, $M = A^{-1}$.) Comme dans le cas des EDO, la notion de *consistance* permet de quantifier la compatibilité entre l'EDP $(\mathcal{E})$ et le schéma aux différences finies.

**Définition 2.1** (Consistance)**.** On dit que le schéma (2.12) est CONSISTANT pour l'EDP $(\mathcal{E})$ si pour toute solution $u(t,x)$ de $(\mathcal{E})$ suffisamment régulière, en notant pour $n \geq 0$

$$\tilde{U}^n = \Big(u(t_n,x_1) \quad \dots \quad u(t_n,x_J)\Big)^T,$$

l'erreur de troncature $(\varepsilon^n)_{n\geq 0}$ définie par

$$\forall n \geq 0 \quad \varepsilon^n = \frac{\tilde{U}^{n+1} - M\tilde{U}^n}{\Delta t}$$

vérifie

$$\|\varepsilon^n\| \leq \omega(\Delta t, \Delta x) \tag{2.13}$$

où $\omega : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}$ est un module de continuité (i.e., $\lim_{\Delta t \to 0, \Delta x \to 0} \omega(\Delta t, \Delta x) = 0$). Si, de plus, $\omega(\Delta t, \Delta x) = O((\Delta t)^p + (\Delta x)^q)$, on dira que le schéma est précis d'ordre $p$ en temps et $q$ en espace.

**Remarque 2.1.** La norme $\|\cdot\|$ utilisée dans l'estimation (2.13) est une norme sur $\mathbb{R}^J$ (i.e. une norme sur l'espace de discrétisation de la variable spatiale). On montrera généralement la consistance pour la norme $L^\infty$, définie par

$$\forall U \in \mathbb{R}^J \quad \|U\|_\infty = \max_{1 \leq j \leq J} |U_j|.$$

**Proposition 2.1** (Consistance du schéma d'Euler explicite.)**.** *Le schéma d'Euler explicite* (2.7) *est consistant, précis à l'ordre* 1 *en temps et* 2 *en espace. De plus, si l'on impose la condition* $\nu \dfrac{\Delta t}{(\Delta x)^2} = \dfrac{1}{6}$*, alors ce schéma est précis à l'ordre* 2 *en temps et* 4 *en espace.*

**Preuve.** On écrit des formules de Taylor à un ordre suffisamment élevé pour atteindre la précision souhaitée (après division par $\Delta t$ et $(\Delta x)^2$, respectivement). S'il n'y a que des dérivées par rapport à $t$ ou par rapport à $x$, on notera par exemple $\partial_t^3 u$ au lieu de $\partial_{ttt}^3 u$.

$$\begin{aligned}
u(t_n + \Delta t, x_j) &= u(t_n, x_j) + \Delta t\ \partial_t u(t_n, x_j) + \frac{(\Delta t)^2}{2}\ \partial_t^2 u(t_n, x_j) + O((\Delta t)^3) \\
u(t_n, x_j + \Delta x) &= u(t_n, x_j) + \Delta x\ \partial_x u(t_n, x_j) + \frac{(\Delta x)^2}{2}\ \partial_x^2 u(t_n, x_j) \\
&\quad + \frac{(\Delta x)^3}{6}\ \partial_x^3 u(t_n, x_j) + \frac{(\Delta x)^4}{24}\ \partial_x^4 u(t_n, x_j) + \frac{(\Delta x)^5}{120}\ \partial_x^5 u(t_n, x_j) + O((\Delta x)^6) \\
u(t_n, x_j - \Delta x) &= u(t_n, x_j) - \Delta x\ \partial_x u(t_n, x_j) + \frac{(\Delta x)^2}{2}\ \partial_x^2 u(t_n, x_j) \\
&\quad - \frac{(\Delta x)^3}{6}\ \partial_x^3 u(t_n, x_j) + \frac{(\Delta x)^4}{24}\ \partial_x^4 u(t_n, x_j) - \frac{(\Delta x)^5}{120}\ \partial_x^5 u(t_n, x_j) + O((\Delta x)^6)
\end{aligned}$$

On en déduit :

$$\begin{aligned}
\partial_t u(t_n, x_j) - \frac{u(t_n + \Delta t, x_j) - u(t_n, x_j)}{\Delta t} &= -\frac{\Delta t}{2}\ \partial_t^2 u(t_n, x_j) + O((\Delta t)^2) \\
\partial_x^2 u(t_n, x_j) - \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{(\Delta x)^2} &= -\frac{(\Delta x)^2}{12}\ \partial_x^4 u(t_n, x_j) + O((\Delta x)^4).
\end{aligned}$$

Comme $u$ est solution de l'équation de la chaleur, on a la relation

$$\partial_t u(t_n, x_j) - \nu\ \partial_x^2 u(t_n, x_j) = 0$$

d'où l'erreur de troncature :

$$\begin{aligned}\varepsilon_j^n &:= \frac{u(t_n + \Delta t, x_j) - u(t_n, x_j)}{\Delta t} - \nu\ \frac{u(t_n, x_{j+1}) - 2u(t_n, x_j) + u(t_n, x_{j-1})}{(\Delta x)^2} \\ &= \frac{\Delta t}{2}\ \partial_t^2 u(t_n, x_j) - \nu\ \frac{(\Delta x)^2}{12}\ \partial_x^4 u(t_n, x_j) + O((\Delta t)^2) + O((\Delta x)^4).\end{aligned}$$

Cela montre que le schéma d'Euler explicite est consistant d'ordre 1 en temps et 2 en espace, pour n'importe quel choix de $\Delta t, \Delta x$. Pour obtenir la condition permettant d'annuler les premiers termes dans la relation précédente, qui font intervenir les dérivées d'ordre supérieur $\partial_t^2 u, \partial_x^4 u$, on dérive la relation $\partial_t u = \nu\ \partial_{xx}^2 u$ par rapport à $t$, et on permute les dérivées en $t$ et $x$ :

$$\begin{aligned}\partial_{tt}^2 u &= \nu\ \partial_{txx}^3 u \\ &= \nu\ \partial_{xxt}^3 u \\ &= \nu\ \partial_{xx}^2(\partial_t u) \\ &= \nu\ \partial_{xx}^2(\nu \partial_{xx}^2 u) \\ &= \nu^2 \partial_{xxxx}^4 u\end{aligned}$$

On peut donc écrire :

$$\varepsilon_j^n = \nu \left( \nu \frac{\Delta t}{2} - \frac{(\Delta x)^2}{12} \right) \partial_x^4 u(t_n, x_j) + O((\Delta t)^2) + O((\Delta x)^4).$$

Ainsi, sous la condition $\nu\ \dfrac{\Delta t}{(\Delta x)^2} = \dfrac{1}{6}$, le schéma est d'ordre 2 en temps et 4 en espace (ce qui est en fait équivalent, puisque dans ce cas-là, $\Delta t$ se comporte comme $(\Delta x)^2$). □

## 2.3 Propriétés qualitatives des schémas. Principes du maximum

On peut attendre des schémas numériques qu'ils préservent certaines propriétés remarquables des solutions des EDP qu'ils simulent. Dans le cas de l'équation de la chaleur, une de ces propriétés est le *principe du maximum*.

**Théorème 2.1** (Principe du maximum)**.** *Soit $\nu > 0, T > 0$ et $u^0 \in \mathcal{C}([0,1],\mathbb{R})$. On suppose qu'il existe $u \in \mathcal{C}([0,T]\times[0,1])$, solution de l'équation de la chaleur*

$$\left\{\begin{array}{rcll} \partial_t u - \nu\ \partial^2_{xx} u = & 0 & \textit{dans } (0,T)\times(0,1) \\ u(0,x) = & u^0(x) & \textit{pour } x \in [0,1] \\ u(t,0) = u(t,1) = & 0 & \textit{pour } t \in (0,T) \end{array}\right.$$

*Alors $u$ satisfait le principe du maximum :*

$$\forall (t,x) \in [0,T]\times[0,1] \quad \min(0, \min_{[0,1]} u^0) \leq u(t,x) \leq \max(0, \max_{[0,1]} u^0)$$

**Preuve.** Soit $\varepsilon > 0$ fixé. On définit la fonction $v \in \mathcal{C}([0,T]\times[0,1])$ par

$$\forall (t,x) \in [0,T]\times[0,1] \quad v(t,x) = u(t,x) - \varepsilon t.$$

$v$ étant continue sur un compact, elle possède un maximum : il existe $(t^*,x^*) \in [0,T]\times[0,1]$ t.q.

$$\forall (t,x) \in [0,T]\times[0,1] \quad v(t,x) \leq v(t^*,x^*).$$

— Supposons que $(t^*,x^*)$ est un point intérieur : $(t^*,x^*) \in (0,T)\times(0,1)$. Nous allons montrer que ce cas est impossible en utilisant l'équation.
On écrit l'équation vérifiée par $u$ au point $(t^*,x^*)$ :

$$\partial_t u(t^*,x^*) - \nu\ \partial^2_{xx} u(t^*,x^*) = 0.$$

Puisque $\partial_t v = \partial_t u - \varepsilon$, l'équation satisfaite par $v$ s'écrit

$$\varepsilon + \partial_t v(t^*,x^*) - \nu\ \partial^2_{xx} v(t^*,x^*) = 0. \tag{2.14}$$

Or, les restrictions $v(\cdot,x^*), v(t^*,\cdot)$ admettent des maxima en $t^* \in (0,T)$ et $x^* \in (0,1)$, resp. Les conditions d'optimalité donnent donc

$$\partial_t v(t^*,x^*) = 0, \quad \left\{\begin{array}{rcl} \partial_x v(t^*,x^*) & = & 0, \\ \partial^2_{xx} v(t^*,x^*) & \leqslant & 0 \end{array}\right.$$

Cela contredit (2.14).

— Si $t^* = 0$ : pour tout $(t,x)$, $v(t,x) \leq v(0,x^*)$, d'où

$$\begin{aligned} \forall (t,x) \in [0,T]\times[0,1] \quad u(t,x) &\leq u(0,x^*) + \varepsilon t \\ &\leq \max_{[0,1]} u^0 + \varepsilon T. \end{aligned}$$

— Si $x^* = 0$ : pour $(t,x)$, $v(t,x) \leq v(t,0) = 0$, donc

$$\begin{aligned}\forall (t,x) \in [0,T]\times[0,1] \quad u(t,x) &\leq \varepsilon t \\ &\leq \varepsilon T.\end{aligned}$$

De même si $x^* = 1$.

— Si $t^* = T$ : on a déjà traité le cas $x^* \in \{0,1\}$, on suppose donc que $0 < x^* < 1$. Dans ce cas, les conditions d'optimalité s'écrivent :

$$\partial_t v(T,x^*) \geq 0, \quad \left\{ \begin{array}{rcl} \partial_x v(T,x^*) & = & 0, \\ \partial^2_{xx} v(T,x^*) & \leqslant & 0 \end{array} \right.$$

On a donc

$$\partial_t u(T,x^*) - \nu \partial^2_{xx} u(T,x^*) \geq \partial_t u(T,x^*) \geq \varepsilon$$

ce qui contredit l'équation (prolongée par continuité en $(T,x^*)$) :

$$\partial_t u(T,x^*) - \nu \partial^2_{xx} u(T,x^*) = \partial_t u(T,x^*) - \nu \partial^2_{xx} u(T,x^*) = 0.$$

Conclusion : pour tout $(t,x) \in [0,T]\times[0,1]$, $u(t,x) \leq \max(\max_{[0,1]} u^0 + \varepsilon T, \varepsilon T)$. En faisant tendre $\varepsilon$ vers 0, on obtient

$$\forall (t,x) \in [0,T]\times[0,1], \quad u(t,x) \leq \max(\max_{[0,1]} u^0, 0).$$

L'autre inégalité se démontre de la même manière, en considérant le minimum de $u(t,x) + \varepsilon t$. □

Nous allons démontrer un résultat équivalent dans le cas discret pour le schéma explicite (sous condition) et pour le schéma implicite.

**Proposition 2.2** (Principe du maximum discret pour le schéma explicite)**.** *Si* $c = \nu \dfrac{\Delta t}{(\Delta x)^2} \leq \dfrac{1}{2}$, *alors le schéma explicite défini par*

$$\begin{aligned} &\frac{u_j^{n+1} - u_j^n}{\Delta t} - \nu \, \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{(\Delta x)^2} = 0, \quad n \geq 0, \quad 1 \leq j \leq J \\ &u_j^0 = u^0(x_j), \quad 1 \leq j \leq J \\ &u_0^n = u_{J+1}^n = 0, \quad n \geq 0 \end{aligned}$$

*vérifie le principe du maximum discret :*

$$\forall n \geq 0, \ \forall j \in [|0, J+1|] \quad \min(0, \min_{1\leq j\leq J} u_0^j) \leq u_j^n \leq \max(0, \max_{1\leq j\leq J} u_j^0) \tag{2.15}$$

**Preuve.** On écrit le schéma sous la forme :

$$u_j^{n+1} = cu_{j+1}^n + (1-2c)u_j^n + cu_{j-1}^n, \quad n \geq 0, \quad 1 \leq j \leq J$$

Sous la condition $c \leq \frac{1}{2}$, on voit que $1 - 2c \geq 0$, et donc $u_j^{n+1}$ est une *combinaison convexe* de $u_{j-1}^n, u_j^n, u_{j+1}^n$, ce qui lui garantit d'appartenir à l'enveloppe convexe de ces trois valeurs. Le résultat se montre alors facilement par récurrence sur $n \in \mathbb{N}$ :

— pour $n = 0$, le résultat est vrai pour tout $0 \leq j \leq J+1$ ;

— supposons le résultat vrai pour $n \in \mathbb{N}$. Si $j = 0$ ou $j = J+1$, le résultat est vrai d'après la condition de Dirichlet homogène. Sinon, on écrit pour tout $1 \leq j \leq J$,

$$u_j^{n+1} \leq \max(u_{j-1}^n, u_j^n, u_{j+1}^n) \leq \max(0, \max_{1 \leq j \leq J+1} u_j^0)$$

d'après l'hypothèse de récurrence. De même,

$$u_j^{n+1} \geq \min(u_{j-1}^n, u_j^n, u_{j+1}^n) \geq \min(0, \min_{1 \leq j \leq J+1} u_j^0)$$

□

**Remarque 2.2.** La condition $c = \nu \dfrac{\Delta t}{(\Delta x)^2} \leq \dfrac{1}{2}$ est appelée condition CFL (Courant, Friedrichs, Levy - 1928). Ce type de condition joue un rôle très important dans l'étude des schémas numériques explicites.

**Proposition 2.3** (Principe du maximum discret pour le schéma implicite)**.** *Le schéma implicite défini par*

$$\begin{aligned} &\frac{u_j^{n+1} - u_j^n}{\Delta t} - \nu \, \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{(\Delta x)^2} = 0, \quad n \geq 0, \quad 1 \leq j \leq J \\ &u_j^0 = u^0(x_j), \quad 1 \leq j \leq J \\ &u_0^n = u_{J+1}^n = 0, \quad n \geq 0 \end{aligned}$$

*vérifie le principe du maximum discret* (2.15)*.*

**Preuve.** Fixons d'abord $N \in \mathbb{N}^*$ et considérons des entiers $n \leq N$. Nous allons montrer l'inégalité

$$\forall 0 \leq n \leq N, \ \forall 0 \leq j \leq J+1 \quad u_j^n \leq \max(0, \max_{1 \leq j \leq J} u_j^0). \tag{2.16}$$

On note $n^* \in [|0, N|]$, $j^* \in [|0, J+1|]$ t.q.

$$u_{j^*}^{n^*} = \max\left\{u_j^n,\ 0 \leq n \leq N,\ 0 \leq j \leq J+1\right\}.$$

(Le maximum existe car c'est un max d'un ensemble fini de valeurs.) En choisissant l'entier $n^*$ minimal qui vérifie la propriété précédente, on peut supposer que

$$\forall 0 \leq n \leq N, \quad \forall 0 \leq j \leq J+1, \quad u_{j^*}^{n^*} = u_j^n \ \Rightarrow\ n^* \leq n. \tag{2.17}$$

Nous allons procéder d'une manière analogue à la preuve du principe du maximum continu pour l'équation de la chaleur, en distinguant plusieurs cas.

— Si $n^* = 0$ : (2.16) est vrai car le max est égal à $u_{j^*}^0$.

— Si $j^* = 0$ ou $j^* = J+1$ : (2.16) est vrai car le max est égal à 0, d'après la condition au bord de Dirichlet.

— Sinon : $n^* \geq 1$ et $1 \leq j^* \leq J$. Comme $n^* - 1 \geq 0$, on peut appliquer le schéma en $(n^*-1, j^*)$ : en notant $c = \nu\ \dfrac{\Delta t}{(\Delta x)^2}$, il s'écrit

$$u_{j^*}^{n^*-1} = -cu_{j^*-1}^{n^*} + (1+2c)u_{j^*}^{n^*} - cu_{j^*+1}^{n^*}.$$

Comme $c > 0$ et par définition du maximum,

$$-cu_{j^*-1}^{n^*} \geq -cu_{j^*}^{n^*}, \quad -cu_{j^*+1}^{n^*} \geq -cu_{j^*}^{n^*}$$

d'où

$$u_{j^*}^{n^*-1} \geq u_{j^*}^{n^*}.$$

Cela contredit (2.16).

Conclusion : dans tous les cas, la propriété (2.16) est satisfaite. Comme $N$ est arbitraire, et en raisonnant de manière analogue pour le min, on en déduit (2.15). □

## 2.4 Stabilité d'un schéma aux différences finies

Tout comme dans le cas des schémas à un pas pour l'approximation numérique des EDO (voir le chapitre 1, définition 1.8), pour assurer qu'un schéma aux différences finies est convergent, on a besoin d'introduire une notion de *stabilité* du schéma. Contrairement au cas des EDO, pour lequel la valeur approchée $y_n$ calculée à chaque itération

est un réel, ici le schéma calcule un vecteur $U^n \in \mathbb{R}^J$, dont la taille $J$ tend vers l'infini lorsque le pas d'espace $\Delta x$ tend vers 0. Nous allons voir que le choix des normes sur les espaces $\mathbb{R}^J$ va jouer un rôle important dans la stabilité des schémas ; certains schémas seront stables pour une norme, mais pas pour une autre.

**Choix des normes sur $\mathbb{R}^J$.** Nous allons considérer les normes suivantes (avec $p \geq 1$) : pour $J \in \mathbb{N}$ et $\Delta x = \frac{1}{J+1}$,

$$\forall U \in \mathbb{R}^J \quad \|U\|_p = \left( \sum_{j=1}^{J} \Delta x \ |U_j|^p \right)^{\frac{1}{p}} \tag{2.18}$$

avec la convention

$$\forall U \in \mathbb{R}^J \quad \|U\|_\infty = \max_{1 \leq j \leq J} |U_j| \tag{2.19}$$

**Remarque 2.3.**
— Les définitions (2.18), (2.19) dépendent de $\Delta x$ (et de $J$, qui lui est lié). Pour alléger l'écriture, on les note simplement $\| \cdot \|_p$.
— Dans la pratique, on utilise essentiellement $p = 2$ ou $p = +\infty$.
— Étant donné $U \in \mathbb{R}^J$, si l'on définit une fonction $v : (0,1) \to \mathbb{R}$ presque partout sur $(0,1)$, par

$$u(x) = \begin{cases} 0 & \text{si } 0 < x < \frac{\Delta x}{2} \\ U_j & \text{si } x_j - \frac{\Delta x}{2} < x < x_j + \frac{\Delta x}{2}, \quad \text{avec } 1 \leq j \leq J \\ 0 & \text{si } 1 - \frac{\Delta x}{2} < x < 1 \end{cases}$$

alors on a l'égalité

$$\|U\|_p = \|u\|_{L^p(0,1)}$$

La norme discrète $\| \cdot \|_p$ est donc appelée parfois *norme $L^p$*.

**Définition 2.2** (Stabilité d'un schéma aux différences finies.)**.** On fixe une norme $\| \cdot \|$ sur $\mathbb{R}^J$. On dit qu'un schéma aux différences finies est STABLE pour cette norme s'il existe $K > 0$ et $\varepsilon > 0$ t.q. pour toute donnée initiale $U^0$ du schéma, la suite $(U^n)_{n \in \mathbb{N}}$ satisfait

$$\forall \Delta t, \Delta x < \varepsilon, \quad \forall n \in \mathbb{N}, \qquad \|U^n\| \leq K \|U^0\| \tag{2.20}$$

Si la relation (2.20) est vraie uniquement si $\Delta t, \Delta x$ satisfont un certain critère (ex : la condition CFL), on dit que le schéma est CONDITIONNELLEMENT STABLE.

Dans le cas d'un schéma linéaire

$$\begin{cases} U^{n+1} = MU^n, \quad n \in \mathbb{N} \\ U^0 \in \mathbb{R}^J \text{ donné} \end{cases}$$

en notant $|\!|\!| M |\!|\!|$ la norme matricielle subordonnée à $\|\cdot\|$, définie par

$$|\!|\!| M |\!|\!| = \sup_{U \in \mathbb{R}^J \setminus \{0\}} \frac{\|MU\|}{\|U\|},$$

on vérifie aisément que la stabilité du schéma est équivalente à :

$$\exists K > 0,\ \exists \varepsilon > 0, \quad \forall \Delta x, \Delta t < \varepsilon,\ \forall n \in \mathbb{N} \quad |\!|\!| M^n |\!|\!| \leq K.$$

Ici, on a noté $M^n$ la puissance $n$-ième de la matrice carrée $M$.

### 2.4.1 Stabilité en norme $L^\infty$

**Proposition 2.4** (Stabilité $L^\infty$ pour les schémas d'Euler explicite et implicite.)**.** —
*Le schéma d'Euler explicite* (2.7) *est stable en norme $L^\infty$, sous la condition CFL*

$$c := \nu \, \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}.$$

— *Le schéma d'Euler implicite* (2.9) *est inconditionnellement stable en norme $L^\infty$.*

**Preuve.** C'est une conséquence directe du principe du maximum discret (2.15) établi à la section 2.3 pour ces deux schémas. En effet, on a

$$\forall n \geq 0,\ \forall 0 \leq j \leq J+1 \quad |u_j^n| \leq |u_j^0|$$

que l'on peut écrire

$$\forall n \geq 0 \quad \|U^n\|_\infty \leq \|U^0\|_\infty$$

D'où le résultat avec la constante $K = 1$ et pour tout $\Delta t, \Delta x > 0$, sous la condition CFL pour le schéma explicite et sans condition pour le schéma implicite. □

### 2.4.2 Stabilité en norme $L^2$

Certains schémas aux différences finies ne sont pas stables en norme $L^\infty$, sans pour autant être de « mauvais » schémas. La justification de leur usage pourra provenir de leur stabilité pour une autre norme, par exemple la norme $L^2$.

L'étude de la stabilité en norme $L^2$ peut se faire en utilisant l'outil puissant de l'analyse de Fourier. Pour cela, plaçons-nous dans le cas des conditions aux limites périodiques :

$$u(0,t) = u(1,t) \quad \text{pour tout } t \in (0,T).$$

Du point de vue discret, cela conduit à imposer

$$u_0^n = u_{J+1}^n \quad \text{pour tout } n \geq 0.$$

Ainsi, le vecteur inconnu $U^n$ contiendra $J+1$ valeurs, et on pourra l'écrire

$$U^n = \begin{pmatrix} u_0^n & u_1^n & \dots & u_J^n \end{pmatrix}^T.$$

En suivant la démarche expliquée à la remarque 2.3, on peut associer aux valeurs discrètes contenues dans $U^n$, une fonction $u^n \in L^2(0,1)$, constante par morceaux, définie presque partout sur $(0,1)$ par

$$u(x) = \begin{cases} u_0^n & \text{si } 0 < x < \frac{\Delta x}{2} \\ u_j^n & \text{si } x_j - \frac{\Delta x}{2} < x < x_j + \frac{\Delta x}{2}, \quad \text{avec } 1 \leq j \leq J \\ u_0^n & \text{si } 1 - \frac{\Delta x}{2} < x < 1 \end{cases}$$

On a alors l'égalité des normes :

$$\|U^n\|_2 = \|u^n\|_{L^2(0,1)}$$

où $\|\cdot\|_2$ est définie par (2.18) avec $p = 2$.

L'idée est de contrôler la norme $\|U^n\|_2$ en utilisant la décomposition en série de Fourier de la fonction $u^n$, qui s'écrit :

$$u^n(x) = \sum_{k \in \mathbb{Z}} \hat{u}^n(k)\ e^{2i\pi kx} \quad \text{p.p. } x \in (0,1)$$

où les coefficients de Fourier $\hat{u}^n(k) \in \mathbb{C}$ sont définis par

$$\hat{u}^n(k) = \int_0^1 u^n(x)\ e^{-2i\pi kx}\ \mathrm{d}x.$$

On pourra calculer les normes $L^2$ grâce à l'égalité de Parseval :

$$\int_0^1 |u^n(x)|^2 \, \mathrm{d}x = \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2$$

**Cas du schéma d'Euler explicite** En utilisant les conditions de périodicité, ce schéma s'écrit : pour tout $n \in \mathbb{N}$,

$$\left\{ \begin{array}{rcl} \dfrac{u_j^{n+1}-u_j^n}{\Delta t} - \nu \, \dfrac{u_{j-1}^n - 2u_j^n + u_{j+1}^n}{(\Delta x)^2} & = & 0, \quad 0 \le j \le J, \\ u_{J+1}^n & = & u_0^n \\ u_{-1}^n & = & u_J^n \end{array} \right.$$

Par construction de la fonction $u^n$, ce schéma est équivalent à une égalité presque partout sur $(0,1)$ :

$$\frac{u^{n+1}(x)-u^n(x)}{\Delta t} - \nu \, \frac{u^n(x-\Delta x) - 2u^n(x) + u^n(x+\Delta x)}{(\Delta x)^2} = 0 \quad \text{p.p. sur } (0,1) \quad (2.21)$$

On traduit cette relation en termes de coefficients de Fourier. Pour calculer les coefficients de Fourier des fonctions $u^n(x-\Delta x), u^n(x+\Delta x)$, on utilise la remarque suivante. En notant $v(x) = u(x+\Delta x)$ (où $u$ est une fonction périodique de période 1), les coefficients de $v$ s'écrivent :

$$\begin{aligned} \forall k \in \mathbb{Z} \quad \hat{v}(k) &= \int_0^1 v(x) \, e^{-2i\pi kx} \, \mathrm{d}x \\ &= \int_0^1 u(x+\Delta x) \, e^{-2i\pi kx} \, \mathrm{d}x \\ &= \int_{\Delta x}^{1+\Delta x} u(y) \, e^{-2i\pi k(y-\Delta x)} \, \mathrm{d}y \\ &= e^{2i\pi k\Delta x} \int_0^1 u(y) \, e^{-2i\pi ky} \, \mathrm{d}y \\ &= e^{2i\pi k\Delta x} \, \hat{u}(k) \end{aligned}$$

De même en remplaçant $\Delta x$ par $-\Delta x$. La relation (2.21) se traduit donc par :

$$\forall k \in \mathbb{Z} \quad \frac{\hat{u}^{n+1}(k)-\hat{u}^n(k)}{\Delta t} - \nu \, \frac{e^{-2i\pi k\Delta x}\hat{u}^n(k) - 2\hat{u}^n(k) + e^{2i\pi k\Delta x}\hat{u}^n(k)}{(\Delta x)^2} = 0$$

En posant $c = \nu \frac{\Delta t}{(\Delta x)^2}$, cette relation s'écrit

$$\begin{aligned} \forall k \in \mathbb{Z} \quad \hat{u}^{n+1}(k) &= [1 - 2c(1 - \cos(2\pi k \Delta x))]\, \hat{u}^n(k) \\ &= \left[1 - 4c\ \sin^2(\pi k \Delta x)\right] \hat{u}^n(k) \end{aligned}$$

où l'on a utilisé la relation $1 - \cos(2a) = 2\sin^2 a$. Pour tout $k \in \mathbb{Z}$, on note $A(k) \in \mathbb{C}$ le coefficient d'amplification défini par

$$A(k) = 1 - 4c\ \sin^2(\pi k \Delta x)$$

Supposons que pour tout mode $k \in \mathbb{Z}$, on ait $|A(k)| \leq 1$. Alors, puisque par définition, $\hat{u}^{n+1}(k) = A(k)\hat{u}^n(k)$, en appliquant l'égalité de Parseval, on obtient

$$\begin{aligned} \int_0^1 |u^{n+1}(x)|^2\ \mathrm{d}x &= \sum_{k\in\mathbb{Z}} |\hat{u}^{n+1}(k)|^2 \\ &= \sum_{k\in\mathbb{Z}} |A(k)|^2 |\hat{u}^n(k)|^2 \\ &\leq \sum_{k\in\mathbb{Z}} |\hat{u}^n(k)|^2 = \int_0^1 |u^n(x)|^2\ \mathrm{d}x \end{aligned}$$

On a ainsi $\|U^{n+1}\|_2 \leq \|U^n\|_2$, et par récurrence,

$$\forall n \in \mathbb{N} \quad \|U^n\|_2 \leq \|U^0\|_2.$$

Cela montre que le schéma est stable en norme $L^2$.

Vérifions sous quelles conditions la propriété

$$\forall k \in \mathbb{Z} \quad |A(k)| \leq 1$$

(appelée *condition de stabilité de Von Neumann*) a lieu ; on écrit

$$\begin{aligned} |A(k)| \leq 1 \quad &\Longleftrightarrow \quad -1 \leq 1 - 4c\ \sin^2(\pi k \Delta x) \leq 1 \\ &\Longleftrightarrow \quad 2c\ \sin^2(\pi k \Delta x) \leq 1 \end{aligned}$$

Cette condition est satisfaite sous la condition CFL $c \leq \dfrac{1}{2}$. On a donc montré le résultat suivant :

**Théorème 2.2** (Stabilité $L^2$ du schéma d'Euler explicite)**.** *Dans le cas de conditions aux limites périodiques, le schéma d'Euler explicite est stable en norme $L^2$ sous la condition CFL*

$$c := \nu\ \frac{\Delta t}{(\Delta x)^2} \leq \frac{1}{2}.$$

**Cas du schéma d'Euler implicite.** Le schéma implicite conduit à la relation suivante

$$\frac{u^{n+1}(x) - u^n(x)}{\Delta t} - \nu \, \frac{u^{n+1}(x-\Delta x) - 2u^{n+1}(x) + u^{n+1}(x+\Delta x)}{(\Delta x)^2} = 0 \quad \text{p.p. sur } (0,1)$$

En appliquant la même démarche, on obtient la relation suivante : pour tout mode $k \in \mathbb{Z}$, $\hat{u}^{n+1}(k) = A(k)\hat{u}^n(k)$ où $A(k)$ est défini par

$$A(k) = \frac{1}{1 + 4c \, \sin^2(\pi k \Delta x)}$$

On vérifie alors que $0 < A(k) \leq 1$ sans condition sur $c$.

**Théorème 2.3** (Stabilité $L^2$ du schéma d'Euler implicite)**.** *Dans le cas de conditions aux limites périodiques, le schéma d'Euler implicite est inconditionnellement stable en norme $L^2$.*

## 2.5 Convergence d'un schéma

Le résultat principal de convergence est le *théorème de Lax*, qui affirme qu'un schéma consistant et stable pour une certaine norme, est convergent au sens de cette norme.

**Théorème 2.4** (Lax)**.** *Soit $u(t,x)$ une solution régulière de l'équation de la chaleur* (2.1)*, avec des conditions aux limites appropriées (pour nous, Dirichlet, Neumann ou périodicité). Soit $u_j^n$ la solution numérique discrète obtenue par un schéma aux différences finies avec la donnée initiale $u_j^0 = u^0(x_j)$. On suppose que le schéma est linéaire, consistant et stable pour la norme $\|\cdot\|$. Alors le schéma est convergent au sens où*

$$\forall T > 0 \qquad \lim_{\Delta t \to 0, \Delta x \to 0} \left( \sup_{t_n \leq T} \|e^n\| = 0, \right) \tag{2.22}$$

*où $e^n$ est le vecteur d'erreur défini par ses composantes $e_j^n = u(t_n, x_j) - u_j^n$.*

*De plus, si le schéma est précis à l'ordre $p$ en espace et à l'ordre $q$ en temps, alors pour tout $T > 0$, il existe une constante $C_T > 0$ telle que*

$$\sup_{t_n \leq T} \|e^n\| \leq C_T \left( (\Delta x)^p + (\Delta t)^q \right). \tag{2.23}$$

**Preuve.** On se place dans le cas des conditions aux limites de Dirichlet. La preuve est identique dans le cas de conditions aux limites quelconques, en supposant que leur traitement n'affecte pas la précision du schéma. Pour tout $n \in \mathbb{N}$, on note $U^n = \left(u_1^n \dots u_J^n\right)^T$ le vecteur inconnu, et $\tilde{U}^n$ le vecteur

$$\tilde{U}^n = \Big(u(t_n, x_1) \dots u(t_n, x_J)\Big)^T.$$

On a alors $e^n = \tilde{U}^n - U^n$. Comme dans le cas des schémas à un pas pour les EDO (voir le chapitre 1), l'idée de la preuve est d'interpréter l'erreur $e^n$ comme la solution d'un schéma linéaire avec terme source. Pour cela, on note $\varepsilon^n \in \mathbb{R}^J$ l'erreur de consistance. Pour un schéma linéaire

$$U^{n+1} = MU^n, \quad n \in \mathbb{N},$$

l'erreur de consistance satisfait

$$\tilde{U}^{n+1} = M\tilde{U}^n + \Delta t\ \varepsilon^n,$$

d'où par différence :

$$e^{n+1} = Me^n + \Delta t\ \varepsilon^n, \quad n \in \mathbb{N}. \tag{2.24}$$

Par récurrence sur $n$, on montre que la relation (2.24) fournit la formule suivante pour $e^n$ :

$$\forall n \in \mathbb{N} \quad e^n = M^n e^0 + \Delta t \sum_{k=0}^{n-1} M^k \varepsilon^{n-1-k}$$

En prenant la norme $\|\cdot\|$ et par définition de la norme matricielle associée, on obtient alors

$$\forall n \in \mathbb{N} \quad \|e^n\| \leq |||M^n||| \, \|e^0\| + \Delta t \sum_{k=0}^{n-1} |||M^k||| \|\varepsilon^{n-1-k}\| \tag{2.25}$$

D'après les données initiales, $e^0 = 0$. De plus, la méthode étant consistante, il existe un module de continuité $\omega$ tel que, pour tous $\Delta t, \Delta x > 0$,

$$\forall n \in \mathbb{N} \quad \|\varepsilon^n\| \leq \omega(\Delta t, \Delta x).$$

D'autre part, la méthode est stable pour la norme $\|\cdot\|$, donc il existe une constante $K > 0$ t.q.

$$\forall k \in \mathbb{N} \quad |||M^k||| \leq K.$$

L'inégalité (2.25) entraîne donc :

$$\forall n \in \mathbb{N} \quad \|e^n\| \leq n\Delta t \ K \ \omega(\Delta t, \Delta x)$$

Enfin, si l'on fixe $T > 0$ et que l'on se restreint aux indices $n$ t.q. $t_n \leq T$, alors on a $n\Delta t \leq T$, d'où :

$$\sup_{t_n \leq T} \|e^n\| \leq T \ K \ \omega(\Delta t, \Delta x)$$

La propriété (2.22) s'en déduit. Pour un schéma précis à l'ordre $p$ en temps et $q$ en espace, on conclut de même en écrivant $\omega(\Delta t, \Delta x) \leq C\left((\Delta x)^p + (\Delta t)^q\right)$, d'où la propriété (2.23) avec $C_T = TKC$.

□

## 2.6 Compléments : traitement des conditions aux limites de Neumann

Considérons les conditions aux limites

$$\frac{\partial u}{\partial x}(t,0) = \frac{\partial u}{\partial x}(t,1) = 0. \tag{2.26}$$

Pour les discrétiser, on peut écrire les développements de Taylor suivant :

$$u(t,\Delta x) = u(t,0) + \Delta x \ \frac{\partial u}{\partial x}(t,0) + O((\Delta x)^2)$$
$$u(t,1-\Delta x) = u(t,1) - \Delta x \ \frac{\partial u}{\partial x}(t,1) + O((\Delta x)^2)$$

d'où

$$\frac{\partial u}{\partial x}(t,0) = \frac{u(t,\Delta x) - u(t,0)}{\Delta x} + O(\Delta x) \tag{2.27}$$
$$\frac{\partial u}{\partial x}(t,1) = \frac{u(t,1) - u(t,1-\Delta x)}{\Delta x} + O(\Delta x) \tag{2.28}$$

Les conditions (2.26) se traduisent alors au niveau discret par les relations

$$\frac{u_1^n - u_0^n}{\Delta x} = 0$$
$$\frac{u_{J+1}^n - u_J^n}{\Delta x} = 0$$

autrement dit,

$$u_0^n = u_1^n, \quad u_{J+1}^n = u_J^n. \tag{2.29}$$

Ainsi, on n'aura besoin que des $J$ inconnues « intérieures » $u_1^n, \dots, u_J^n$, les valeurs en $j = 0$ et $j = J + 1$ étant prescrites par (2.29). Si l'on considère le schéma d'Euler explicite, défini par

$$u_j^{n+1} = cu_{j-1}^n + (1 - 2c)u_j^n + cu_{j+1}^n, \quad n \geq 0,\ 1 \leq j \leq J,$$

on aura donc pour $j = 1$ et $j = J$ :

$$\begin{aligned} u_1^{n+1} &= cu_0^n + (1 - 2c)u_1^n + cu_2^n \\ &= (1 - c)u_1^n + cu_2^n \\ u_J^{n+1} &= cu_{J-1}^n + (1 - 2c)u_J^n + cu_{J+1}^n \\ &= cu_{J-1}^n + (1 - c)u_J^n \end{aligned}$$

En notant $U^n = \begin{pmatrix} u_1^n & \dots & u_J^n \end{pmatrix}^T$ le vecteur inconnu, le schéma s'écrira donc

$$U^{n+1} = MU^n$$

où la matrice $M \in M_J(\mathbb{R})$ est donnée par

$$M = \begin{pmatrix} 1-c & c & 0 & \dots & \dots & \dots & 0 \\ c & 1-2c & c & 0 & \dots & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & \dots & 0 & c & 1-2c & c \\ 0 & \dots & \dots & \dots & 0 & c & 1-c \end{pmatrix}$$

L'inconvénient de cette première approche est qu'elle utilise une approximation de $\partial_x u$ qui est seulement d'ordre 1 en espace (comme le montrent les développements (2.27) et (2.28)). Or, comme on l'a vu, les schémas d'Euler sont d'ordre 2 en espace à l'intérieur du domaine spatial $[0, 1]$. Par conséquent, ce type d'approximation génère une perte de précision du schéma près du bord. Pour y remédier, on peut construire une approximation de $\partial_x u(t, 0)$, $\partial_x u(t, 1)$ d'ordre 2 en espace.

Plaçons-nous en $x = 1$. On suppose que $u$ est prolongée au voisinage 1 (par exemple, par symétrie), en une fonction régulière, et on écrit les formules de Taylor :

$$u(t, 1 + \Delta x) = u(t, 1) + \Delta x \, \frac{\partial u}{\partial x}(t, 1) + \frac{(\Delta x)^2}{2} \, \frac{\partial^2 u}{\partial x^2}(t, 1) + O((\Delta x)^3)$$
$$u(t, 1 - \Delta x) = u(t, 1) - \Delta x \, \frac{\partial u}{\partial x}(t, 1) + \frac{(\Delta x)^2}{2} \, \frac{\partial^2 u}{\partial x^2}(t, 1) + O((\Delta x)^3)$$

Par différence, on obtient

$$\frac{\partial u}{\partial x}(t, 1) = \frac{u(t, 1 + \Delta x) - u(t, 1 - \Delta x)}{2\Delta x} + O((\Delta x)^2)$$

En introduisant un point supplémentaire « fictif » $x_{J+2} = 1 + \Delta x$, et $u^n_{J+2}$ la valeur discrète correspondante, la condition de Neumann en $x = 1$ pourra alors être discrétisée en :

$$\frac{u^n_{J+2} - u^n_J}{2\Delta x} = 0,$$

autrement dit,

$$u^n_{J+2} = u^n_J.$$

De même, on introduira le point fictif $x_{-1} = -\Delta x$ et on écrira

$$u^n_{-1} = u^n_1.$$

L'idée est ensuite d'écrire les schémas y compris aux points extrémaux (c'est-à-dire, pour $j = 0$, $j = J + 1$). Pour ces valeurs de $j$, on aura besoin de $u^n_{-1}$ et $u^n_{J+2}$ que l'on éliminera en les remplaçant par $u^n_1$ et $u^n_J$. Les inconnues du problèmes discret seront donc $u^n_0, u^n_1, \ldots, u^n_{J+1}$ (les valeurs au bord de l'intervalle sont des inconnues, contrairement au cas des conditions de Dirichlet par exemple).

Ainsi, le schéma d'Euler explicite s'écrira

$$u^{n+1}_j = cu^n_{j-1} + (1 - 2c)u^n_j + cu^n_{j+1}, \quad n \geq 0, \ 0 \leq j \leq J + 1,$$

avec pour $j = 0$ et $j = J + 1$ :

$$\begin{aligned} u^{n+1}_0 &= cu^n_{-1} + (1 - 2c)u^n_0 + cu^n_1 \\ &= (1 - 2c)u^n_0 + 2cu^n_1 \\ u^{n+1}_{J+1} &= cu^n_J + (1 - 2c)u^n_{J+1} + cu^n_{J+2} \\ &= 2cu^n_J + (1 - 2c)u^n_{J+1} \end{aligned}$$

En termes matriciels, en notant $U^n = \begin{pmatrix} u_0^n & u_1^n & \dots & u_J^n & u_{J+1}^n \end{pmatrix}^T$, le schéma s'écrira

$$U^{n+1} = MU^n$$

où la matrice $M \in M_{J+2}(\mathbb{R})$ est donnée par

$$M = \begin{pmatrix} 1-2c & 2c & 0 & \dots & \dots & \dots & 0 \\ c & 1-2c & c & 0 & \dots & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & \dots & 0 & c & 1-2c & c \\ 0 & \dots & \dots & \dots & 0 & 2c & 1-2c \end{pmatrix}$$

# Chapitre 3

# L'espace de Sobolev $H^1$

Étant donnés deux réels $a < b$ et une fonction $f \in \mathcal{C}([a,b])$, considérons le problème suivant :

$$\begin{cases} -u'' + u = f \quad \text{sur } [a,b] \\ u(a) = u(b) = 0 \end{cases} \tag{3.1}$$

Une solution *classique* de (3.1) est une fonction $u \in \mathcal{C}^2([a,b])$ t.q.

$$\forall x \in [a,b] \quad u''(x) + u(x) = f(x) \tag{3.2}$$

et $u(a) = u(b) = 0$.

Nous allons définir une nouvelle formulation du problème (3.1), basée sur le principe suivant. Multiplions l'équation (3.2) par une fonction $\varphi$ et intégrons sur $[a,b]$ ; on obtient

$$\int_a^b -u''(x)\varphi(x)\ \mathrm{d}x + \int_a^b u(x)\varphi(x)\ \mathrm{d}x = \int_a^b f(x)\varphi(x)\ \mathrm{d}x$$

puis par intégration par parties,

$$\int_a^b u'(x)\varphi'(x)\ \mathrm{d}x - u'(b)\varphi(b) + u'(a)\varphi(a) + \int_a^b u(x)\varphi(x)\ \mathrm{d}x = \int_a^b f(x)\varphi(x)\ \mathrm{d}x$$

Si l'on suppose que $\varphi$ vérifie également les conditions au bord $\varphi(a) = \varphi(b) = 0$, on obtient la relation suivante :

$$\int_a^b u'(x)\varphi'(x)\ \mathrm{d}x + \int_a^b u(x)\varphi(x)\ \mathrm{d}x = \int_a^b f(x)\varphi(x)\ \mathrm{d}x \tag{3.3}$$

On voit que cette relation intégrale a du sens sous des hypothèses moins fortes que la formulation initiale (3.1). Par exemple, en choisissant $\varphi \in \mathcal{C}^1([a,b])$, on peut écrire (3.1) en supposant simplement $u$ de classe $\mathcal{C}^1$ ; en fait, ces trois intégrales ont un sens dès lors que $u, u', f$ sont dans $L^1(a,b)$. Toutefois, afin d'utiliser les propriétés des espaces de Hilbert, nous allons plutôt considérer des fonctions $u \in L^2(a,b)$ dont la dérivée $u'$ appartient également à $L^2(a,b)$, et pour lesquelles la relation (3.3) est vérifiée pour toute fonction $\varphi \in \mathcal{C}^1_c([a,b])$. La relation (3.3) sera appelée *formulation faible* associée à (3.1), et les fonctions $\varphi$ seront appelées *fonctions test*.

Cette approche pose plusieurs difficultés :

— Si $u \in L^2(a,b)$, $u$ est définie presque partout ; par conséquent, on ne peut pas lui imposer de relation ponctuelle. Quel sens donner alors aux conditions de bord $u(a) = u(b) = 0$ ?

— Pour une fonction $u \in L^2(a,b)$, et qui n'est donc pas dérivable en général, quel sens donner à $u'$ ?

Nous allons voir qu'il est possible d'introduire une notion de *dérivée faible*, et que les fonctions $u \in L^2(a,b)$ possédant une dérivée faible dans $L^2(a,b)$ possèdent certaines propriétés remarquables, qui permettent notamment de donner du sens à $u(a), u(b)$. Ces fonctions appartiennent à l'*espace de Sobolev* $H^1(a,b)$.

## 3.1 La notion de dérivée faible

Dans toute cette section, on note $I$ un intervalle ouvert de $\mathbb{R}$ (borné ou non).

**Définition 3.1.** On note $L^1_{loc}(I)$ l'ensemble des fonctions $u : I \to \mathbb{R}$, mesurables, t.q.

$$\forall a, b \in I, \quad a < b \quad \Rightarrow \quad \int_a^b |u| < \infty.$$

Une fonction $u \in L^1_{loc}$ est dite LOCALEMENT INTÉGRABLE sur $I$.

En particulier, toute fonction $u \in L^2(I)$ appartient également à $L^1_{loc}(I)$. En effet, d'après l'inégalité de Hölder appliquée à $u$ et à la fonction constante égale à 1,

$$\int_a^b |u| \leq \sqrt{b-a} \, \left( \int_a^b u^2 \right)^{1/2} \leq \sqrt{b-a} \, \|u\|_{L^2(I)}$$

**Définition 3.2** (Dérivée faible)**.** Soit $u \in L^1_{loc}(I)$. On dit que la fonction $v \in L^1_{loc}(I)$ est la DÉRIVÉE FAIBLE de $u$ si

$$\forall \varphi \in \mathcal{C}^1_c(I) \quad \int_I u\varphi' = -\int_I v\varphi \tag{3.4}$$

On note alors $v = u'$.

**Remarque 3.1.** Dans la définition de la dérivée faible (3.2), on peut remplacer $\forall \varphi \in \mathcal{C}^1_c(I)$ par $\forall \varphi \in \mathcal{C}^\infty_c(I)$.

**Proposition 3.1.** *La dérivée faible, si elle existe, est* UNIQUE*, au sens où : si $v_1, v_2$ sont deux fonctions localement intégrables satisfaisant la propriété* (3.4)*, alors $v_1 = v_2$ presque partout sur $I$.*

**Preuve.** Ce résultat est une conséquence du lemme suivant, appliqué à la différence $w = v_1 - v_2$.

**Lemme 3.1.** *Soit $w \in L^1_{loc}(I)$ t.q.*

$$\forall \varphi \in \mathcal{C}_c(I) \quad \int_I w\psi = 0 \tag{3.5}$$

*Alors $w = 0$ p.p. sur $I$.*

**Preuve du lemme 3.1.** L'idée est de construire une fonction $\psi$, continue à support compact dans $I$, t.q. l'intégrale $\int_I w\psi$ soit arbitrairement proche de $\int_I |w|$. Pour que cette dernière intégrale soit finie, il faut que $w \in L^1(I)$ ; on commence donc par supposer que $I$ est borné.

**Étape 1.** Supposons $I$ borné, et donc $w \in L^1(I)$. Pour construire $\psi$, on commence par approcher $w$ par une fonction continue à support compact dans $L^1(I)$ (ce qui est possible par densité de $\mathcal{C}_c(I)$ dans $L^1(I)$). Soit $\varepsilon > 0$ ; il existe $w_1 \in \mathcal{C}_c(I)$ t.q. $\|w - w_1\|_{L^1(I)} \leq \varepsilon$. Alors, d'après (3.1),

$$\forall \psi \in \mathcal{C}(I) \quad \int_I (w - w_1)\psi = -\int_I w_1\psi,$$

donc

$$\left| \int_I w_1\psi \right| \leq \|w - w_1\|_{L^1(I)} \ \|\psi\|_\infty \leq \varepsilon \ \|\psi\|_\infty$$

Définissons les sous-ensembles de $I$ suivants :

$$K_1 = \{x \in I,\ w_1(x) \geq \varepsilon\}\,,\ K_2 = \{x \in I,\ w_1(x) \leq -\varepsilon\}\,.$$

$K_1$ et $K_2$ sont des compacts disjoints, donc d'après le théorème de Tietze-Urysohn, on peut construire une fonction $\psi_0 \in \mathcal{C}_c(I)$ t.q ;

$$\psi_0 = 1 \text{ sur } K_1, \quad \psi_0 = -1 \text{ sur } K_2, \quad |\psi_0(x)| \leq 1 \quad \forall x \in I.$$

On note $K = K_1 \cup K_2$. Par construction, on a donc

$$\int_K |w_1| = \int_K w_1\psi_0$$

Comme $\|\psi_0\|_\infty = 1$, on peut écrire :

$$\begin{aligned}
\left|\int_I w_1\psi_0\right| \leq \varepsilon \quad &\Rightarrow \quad \left|\int_{I\setminus K} w_1\psi_0 + \int_K w_1\psi_0\right| \leq \varepsilon \\
&\Rightarrow \quad \left|\left|\int_{I\setminus K} w_1\psi_0\right| - \left|\int_K w_1\psi_0\right|\right| \leq \varepsilon \\
&\Rightarrow \quad \int_K w_1\psi_0 \leq \varepsilon + \int_{I\setminus K} |w_1||\psi_0| \\
&\Rightarrow \quad \int_K |w_1| \leq \varepsilon + \int_{I\setminus K} |w_1|
\end{aligned}$$

D'où

$$\begin{aligned}
\int_I |w_1| = \int_K |w_1| + \int_{I\setminus K} |w_1| &\leq \varepsilon + 2\int_{I\setminus K} |w_1| \\
&\leq \varepsilon + 2\varepsilon|I|
\end{aligned}$$

où $|I|$ est la longueur de $I$, puisque $|w_1| \leq \varepsilon$ sur $I \setminus K$. Ainsi,

$$\|w\|_{L^1(I)} \leq \|w - w_1\|_{L^1(I)} + \|w_1\|_{L^1(I)} \leq 2\varepsilon + 2\varepsilon|I|.$$

Comme $\varepsilon$ est arbitraire, on en conclut que $\|w\|_{L^1(I)} = 0$ donc $w = 0$ p.p. sur $I$.
**Étape 2.** Si $I$ est non borné, on peut le recouvrir par une union dénombrable d'intervalles ouverts, bornés :

$$I = \cup_{n\in\mathbb{N}} I_n.$$

Par exemple, si $I = ]a, +\infty[$ avec $a \in \mathbb{R}$, on pose $I_n = ]a, a+n[$. On applique ce qui précédè en remplaçant $w$ par $w_n := w_{|I_n}$. Comme $w_n \in L^1(I_n)$ et satisfait (3.5) sur $I_n$ qui est borné, on obtient $w_n = 0$ p.p. sur $I_n$, pour tout $n \in \mathbb{N}$. Par conséquent, l'ensemble

$$\{x \in I,\ w(x) \neq 0\} = \cup_{n\in\mathbb{N}} \{x \in I_n,\ w_n(x) \neq 0\}$$

est de mesure nulle comme union dénombrable d'ensembles de mesure nulle.

□

**Lemme 3.2.** *Soit $u \in L^1_{loc}(I)$. On suppose que $u$ admet une dérivée faible, qui est égale à 0. Alors il existe une constante $C \in \mathbb{R}$ t.q.*

$$u = C \quad \textit{p.p. sur } I.$$

**Preuve.** Par définition de la dérivée faible, on a :

$$\forall \varphi \in \mathcal{C}^1_c(I) \quad \int_I u\varphi' = 0. \tag{3.6}$$

Nous allons montrer qu'il existe une constante $C \in \mathbb{R}$ t.q.

$$\forall \psi \in \mathcal{C}_c(I) \quad \int_I (u - C)\psi = 0. \tag{3.7}$$

On pourra alors conclure grâce au lemme 3.1.

Pour comprendre le lien entre (3.6) et (3.7), supposons que $I$ est borné. Si $u = C$ p.p. sur $I$, alors cette constante s'exprime par

$$C = \frac{1}{|I|}\int_I u.$$

L'intégrale intervenant dans (3.7) s'écrit alors

$$\begin{aligned}\int_I \left(u - \frac{1}{|I|}\int_I u\right)\psi &= \int_I u\psi - \frac{1}{|I|}\int_I u \int_I \psi \\ &= \int_I u\left(\psi - \frac{1}{|I|}\int_I \psi\right)\end{aligned}$$

On est alors tenté de choisir $\varphi \in \mathcal{C}^1_c(I)$ t.q. $\varphi' = \psi - \frac{1}{|I|}\int_I \psi$. Cependant, en prenant $\psi \in \mathcal{C}_c(I)$, la fonction $\varphi'$ correspondant n'est plus à support compact (on a retranché à $\psi$ une constante, égale à sa moyenne sur $I$). Pour y remédier, on va remplacer l'expression

$\psi - \frac{1}{|I|}\int_I \psi$ par $\psi - g\int_I \psi$, où $g$ est une fonction continue à support compact. On remarque également que pour obtenir $\varphi$ à support compact, on doit avoir $\int_I \varphi' = 0$, et donc la relation $\varphi' = \psi - g\int_I \psi$ impose $\int_I \psi - \int_I g \int_I \psi = 0$, c'est-à-dire $\int_I g = 1$.

On aboutit à la rédaction suivante. Soit $\psi \in \mathcal{C}_c(I)$, soit $g \in \mathcal{C}_c(I)$ vérifiant $\int_I g = 1$. Si l'on note $[a,b] \subset I$ un intervalle borné contenant le support de $\psi - g\int_I \psi$, on définit la fonction $\varphi$ par la formule suivante :

$$\forall x \in I \quad \varphi(x) = \int_a^x \left(\psi(t) - g(t)\int_I \psi\right) \mathrm{d}t.$$

Alors $\varphi \in \mathcal{C}_c^1(I)$. En effet, $\varphi$ est de classe $\mathcal{C}^1$ comme primitive d'une fonction continue. De plus, par définition du support, on a $\varphi(x) = 0$ si $x \leq a$, $\varphi(x) = C_1$ si $x \geq b$, où $C_1$ est une constante. Or, la condition $\int_I g = 1$ implique pour tout $x \geq b$,

$$\int_a^x \left(\psi(t) - g(t)\int_I \psi\right) \mathrm{d}t = \int_I \left(\psi(t) - g(t)\int_I \psi\right) \mathrm{d}t = \int_I \psi - \int_I g \int_I \psi = 0$$

c'est-à-dire $C_1 = 0$ et donc $\varphi$ est à support compact. On peut alors lui appliquer (3.6) :

$$\int_I u\varphi' = 0 \quad \Rightarrow \quad \int_I u\left(\psi - g\int_I \psi\right) = 0 \tag{3.8}$$

$$\Rightarrow \quad \int_I u\psi - \left(\int_I gu\right)\left(\int_I \psi\right) = 0 \tag{3.9}$$

$$\Rightarrow \quad \int_I \left(u - \left(\int_I gu\right)\right)\psi = 0 \tag{3.10}$$

Cette relation étant vraie pour toute fonction $\psi \in \mathcal{C}_c(I)$, on conclut d'après le lemme 3.1 que $u = C$ p.p. sur $I$, où la constante $C$ est donnée par $C = \int_I gu$.

□

## 3.2 L'espace de Sobolev $H^1(I)$

**Définition 3.3.** On définit l'ESPACE DE SOBOLEV $H^1(I)$ par

$$H^1(I) = \left\{u \in L^2(I), \quad u' \in L^2(I)\right\}$$

Dans cette définition, $u'$ désigne la dérivée faible de $u$ au sens de la définition 3.2.

**Remarque 3.2.** Si $I = ]a, b[$ est un intervalle borné, on écrira $H^1(a, b)$ pour signifier $H^1(]a, b[)$.

**Proposition 3.2.** *La forme bilinéaire $\langle \cdot, \cdot \rangle_{H^1} : H^1(I) \times H^1(I) \to \mathbb{R}$ définie par*

$$\forall (u, v) \in H^1(I) \times H^1(I) \quad \langle u, v \rangle_{H^1} = \int_I u'v' + uv$$

*est un produit scalaire sur $H^1(I)$, auquel on associe la norme $\| \cdot \|_{H^1}$ définie par*

$$\forall u \in H^1(I) \quad \|u\|_{H^1} = \left( \int_I (u')^2 + u^2 \right)^{1/2}$$

*Muni de ce produit scalaire, l'espace $H^1(I)$ est un espace de Hilbert séparable (c'est-à-dire, il existe des familles dénombrables denses dans $H^1$, au sens de la norme $\| \cdot \|_{H^1}$).*

**Preuve.** Par définition de la dérivée faible, on voit facilement que l'application $u \in H^1(I) \mapsto u' \in L^2(I)$ est linéaire. Par linéarité de l'intégrale, la forme $\langle \cdot, \cdot \rangle_{H^1}$ est clairement bilinéaire et symétrique. Montrons qu'elle est définie positive. Pour tout $u \in H^1(I)$,

$$\langle u, u \rangle_{H^1} = \int_I (u')^2 + u^2 \geq 0$$

et s'annule uniquement si $u = 0$ presque partout (ce qui entraîne $u' = 0$ p.p. d'après la propriété (3.4)). $\langle \cdot, \cdot \rangle_{H^1}$ est donc un produit scalaire.

Montrons que $H^1(I)$ est complet. Soit $(u_n) \in H^1(I)^{\mathbb{N}}$ une suite de Cauchy. Soit $\varepsilon > 0$. Il existe $n_0 \in \mathbb{N}$ t.q.

$$\forall (p, q) \in \mathbb{N}^2, \quad p, q \geq n_0 \quad \Rightarrow \quad \int_I (u'_p - u'_q)^2 + (u_p - u_q)^2 \leq \varepsilon$$

En particulier,

$$\forall (p, q) \in \mathbb{N}^2, \quad p, q \geq n_0 \quad \Rightarrow \quad \int_I (u_p - u_q)^2 \leq \varepsilon$$

donc $(u_n)$ est une suite de Cauchy dans $L^2(I)$. $L^2(I)$ étant un espace complet, il existe $u \in L^2(I)$ t.q. $\|u - u_n\|_{L^2(I)} \to 0$. De même, $(u'_n)$ est de Cauchy dans $L^2(I)$, donc il existe $v \in L^2(I)$ t.q. $\|v - u'_n\|_{L^2(I)} \to 0$. Montrons que $u$ est dérivable au sens faible et que $u' = v$. Soit $\varphi \in \mathcal{C}^1_c(I)$. Pour tout $n \in \mathbb{N}$, on a, par définition de la dérivée faible,

$$- \int_I u'_n \varphi = \int_I u_n \varphi'.$$

En utilisant les convergentes fortes $L^2(I)$, on obtient

$$\lim_{n\to\infty}\int_I u_n'\varphi = \int_I v\varphi, \quad \lim_{n\to\infty}\int_I u_n\varphi' = \int_I u\varphi'$$

d'où par unicité de la limite,

$$-\int_I v\varphi = \int_I u\varphi'.$$

Cela montre que $u$ est faiblement dérivable et que $u' = v$. Comme $v \in L^2(I)$, on en déduit que $u \in H^1(I)$. Enfin, en remarquant que

$$\|u - u_n\|_{H^1}^2 = \|u' - u_n'\|_{L^2}^2 + \|u - u_n\|_{L^2}^2,$$

on en déduit que

$$\lim_{n\to\infty} \|u - u_n\|_{H^1} = 0.$$

$(u_n)$ est donc convergente dans $H^1(I)$. Cela montre que $H^1(I)$ est un espace de Hilbert pour ce produit scalaire.

Enfin, pour montrer la séparabilité de $H^1(I)$, on considère l'application suivante :

$$\begin{aligned} T : H^1(I) &\to L^2(I) \times L^2(I) \\ u &\mapsto (u, u') \end{aligned}$$

Par linéarité de l'opérateur dérivée faible, $T$ est clairement linéaire. De plus, si l'on munit l'espace produit $L^2(I) \times L^2(I)$ de la norme suivante :

$$\forall (u, v) \in L^2(I) \times L^2(I) \quad \|(u, v)\|_{L^2\times L^2} = \left(\|u\|_{L^2}^2 + \|v\|_{L^2}^2\right)^{1/2}$$

on voit que

$$\forall u \in H^1(I) \quad \|T(u)\|_{L^2\times L^2} = \|u\|_{H^1}$$

Par conséquent, $T$ est une isométrie de $H^1(I)$ dans $L^2(I) \times L^2(I)$ ; en particulier, elle est injective, et donc bijective de $H^1(I)$ vers son image. Puisque $L^2(I)$ est séparable, le produit $L^2(I) \times L^2(I)$ l'est également, et par conséquent, l'image $T(H^1(I))$ est elle aussi séparable (comme sous-ensemble d'un ensemble séparable). On conclut que $H^1(I)$ est séparable, car il est isométrique à un ensemble séparable.

□

**Remarque 3.3.** On pourra retenir de la preuve précédente le fait suivant : si $(u_n)$ est une suite de $H^1$ convergeant vers une fonction $u$ dans $L^2$, et si la suite $(u_n')$ converge vers une fonction $v$ dans $L^2$, alors $u$ est faiblement dérivable, $u' = v$ et $u_n \to u$ dans $H^1$.

**Exemple 3.1.** La fonction $u : x \in ]-1,1[ \mapsto |x|$ appartient à $H^1(-1,1)$. En effet, $u \in L^2(-1,1)$ (elle est dans $L^\infty(-1,1)$). Calculons sa dérivée faible. Pour $\varphi \in \mathcal{C}_c^1(]-1;1[)$,

$$\begin{aligned}
\int_{-1}^{1} u(x)\varphi'(x)\,\mathrm{d}x &= \int_{-1}^{0} (-x)\varphi'(x)\,\mathrm{d}x + \int_{0}^{1} x\varphi'(x)\,\mathrm{d}x \\
&= \int_{-1}^{0} \varphi(x)\,\mathrm{d}x - \int_{0}^{1} \varphi(x)\,\mathrm{d}x \qquad (\text{car } \varphi(-1) = \varphi(1) = 0) \\
&= -\left( \int_{-1}^{0} (-1)\varphi(x)\,\mathrm{d}x + \int_{0}^{1} \varphi(x)\,\mathrm{d}x \right) \\
&= -\int_{-1}^{1} v(x)\varphi(x)\,\mathrm{d}x
\end{aligned}$$

où $v$ est définie presque partout sur $]-1,1[$ par

$$v(x) = \begin{cases} -1 & \text{si } x \in ]-1,0[, \\ 1 & \text{si } x \in ]0,1[. \end{cases}$$

$u$ admet donc une dérivée faible $v$, qui appartient à $L^2(-1,1)$ (elle est également dans $L^\infty(-1,1)$). $u$ appartient donc à $H^1(-1,1)$ et on peut calculer sa norme en écrivant

$$\int_{-1}^{1} \left(u'(x)^2 + u(x)^2\right) \mathrm{d}x = \int_{-1}^{1} \left(1 + x^2\right) \mathrm{d}x = \frac{8}{3}$$

d'où $\|u\|_{H^1(-1,1)} = \dfrac{2\sqrt{2}}{\sqrt{3}}$.

**Exemple 3.2.** Soit $u \in H^1(I)$ et $\eta \in \mathcal{C}^1(\overline{I})$ (c'est-à-dire, $\eta$ est de classe $\mathcal{C}^1$ et $\eta, \eta'$ sont uniformément continues sur $I$). Alors $\eta u \in H^1(I)$ et sa dérivée faible est donnée par la formule

$$(\eta u)' = \eta' u + \eta u', \tag{3.11}$$

où $\eta'$ est la dérivée classique de $\eta$, et $u'$ la dérivée faible de $u$.

En effet, soit $\varphi \in \mathcal{C}^1(I)$ ; on écrit :

$$\begin{aligned}\int_I \eta u \varphi' &= \int_I u(\eta\varphi') \\ &= \int_I u[(\eta\varphi)' - \eta'\varphi] \\ &= \int_I u(\eta\varphi)' - \int_I (u\eta')\varphi\end{aligned}$$

En remarquant que $\eta\varphi \in \mathcal{C}_c^1(I)$, on peut l'utiliser comme fonction test dans la définition de $u'$, ce qui donne :

$$\int_I u(\eta\varphi)' = -\int_I u'(\eta\varphi)$$

On obtient finalement :

$$\int_I \eta u \varphi' = -\int_I (\eta' u + \eta u')\varphi$$

d'où (3.11). Enfin, $\eta' u + \eta u' \in L^2(I)$ car $\eta, \eta'$ sont uniformément continues, donc bornées sur $I$.

**Théorème 3.1.** *Soit $u \in H^1(I)$ ; alors il existe une unique fonction $\tilde{u} \in \mathcal{C}(\overline{I})$ t.q.*

$$u = \tilde{u} \quad \textit{p.p. sur } I \tag{3.12}$$

*et*

$$\forall x, y \in I \qquad \tilde{u}(y) - \tilde{u}(x) = \int_x^y u'(t)\ \mathrm{d}t. \tag{3.13}$$

*$\tilde{u}$ s'appelle le* REPRÉSENTANT CONTINU *de $u$.*

**Remarque 3.4.** Dans la pratique, on identifie toute fonction $u \in H^1(I)$ à son représentant continu $\tilde{u} \in \mathcal{C}(\overline{I})$. Cela permet de donner un sens à des conditions aux limites du type $u(a) = u(b) = 0$, ou plus généralement à des valeurs ponctuelles $u(x)$ pour $x \in \overline{I}$, lorsque $u \in H^1(I)$.

**Preuve du théorème 3.1.** On fixe un point $y_0 \in I$. Pour tout $x \in I$, on définit $\overline{u}(x)$ par

$$\overline{u}(x) = \int_{y_0}^x u'(t)\ \mathrm{d}t.$$

Comme $u' \in L^2(I)$, en particulier $u'$ est localement intégrable sur $I$ donc $\overline{u}(x)$ est bien définie pour tout $x \in I$. De plus, pour tout $x, y \in I$, $\overline{u}(y) - \overline{u}(x) = \displaystyle\int_x^y u'(t)\ \mathrm{d}t$. En appliquant l'inégalité de Hölder, on obtient donc

$$\begin{aligned}\forall x, y \in I, \quad x < y \quad \Rightarrow \quad |\overline{u}(y) - \overline{u}(x)| &\leq \sqrt{y-x}\left(\int_x^y u'(t)^2\ \mathrm{d}t\right)^{1/2} \\ &\leq \|u'\|_{L^2(I)}\ \sqrt{y-x}\end{aligned}$$

Cela montre que la fonction $\overline{u}$ est $\frac{1}{2}$-höldérienne, par conséquent elle est uniformément continue sur $I$. Ainsi, $\overline{u} \in \mathcal{C}(\overline{I})$.

Montrons que $\overline{u}$ est faiblement dérivable et que $\overline{u}' = u'$ p.p. sur $I$. Pour cela, on considère une fonction $\varphi \in \mathcal{C}_c^1(I)$. On fixe alors un intervalle borné $[a, b] \subset I$, tel que $a < y_0 < b$ et $[a, b]$ contienne le support de $\varphi$. On a en particulier $\varphi(a) = \varphi(b) = 0$. On écrit :

$$\begin{aligned}\int_I \overline{u}(x)\varphi'(x)\ \mathrm{d}x &= \int_a^b \overline{u}(x)\varphi'(x)\ \mathrm{d}x \\ &= \int_a^b \left(\int_{y_0}^x u'(t)\ \mathrm{d}t\right)\varphi'(x)\ \mathrm{d}x \\ &= \int_a^{y_0} \left(\int_{y_0}^x u'(t)\ \mathrm{d}t\right)\varphi'(x)\ \mathrm{d}x + \int_{y_0}^b \left(\int_{y_0}^x u'(t)\ \mathrm{d}t\right)\varphi'(x)\ \mathrm{d}x \\ &= -\int_a^{y_0} \left(\int_x^{y_0} u'(t)\ \mathrm{d}t\right)\varphi'(x)\ \mathrm{d}x + \int_{y_0}^b \left(\int_{y_0}^x u'(t)\ \mathrm{d}t\right)\varphi'(x)\ \mathrm{d}x\end{aligned}$$

On applique alors le théorème de Fubini, qui permet de permuter l'ordre d'intégration. On remarque que les conditions

$$a \leq x \leq y_0, \quad x \leq t \leq y_0$$

sont équivalentes à

$$a \leq t \leq y_0, \quad a \leq x \leq t,$$

et de même, les conditions

$$y_0 \leq x \leq b, \quad y_0 \leq t \leq x$$

sont équivalentes à

$$y_0 \leq t \leq b, \quad t \leq x \leq b.$$

On obtient donc :

$$\begin{aligned}\int_I \overline{u}(x)\varphi'(x)\ \mathrm{d}x &= -\int_a^{y_0}\left(\int_a^t \varphi'(x)\ \mathrm{d}x\right)u'(t)\ \mathrm{d}t + \int_{y_0}^b\left(\int_t^b \varphi'(x)\ \mathrm{d}x\right)u'(t)\ \mathrm{d}t\\ &= -\int_a^{y_0}\varphi(t)u'(t)\ \mathrm{d}t - \int_{y_0}^b \varphi(t)u'(t)\ \mathrm{d}t\\ &= -\int_a^b \varphi(t)u'(t)\ \mathrm{d}t\\ &= -\int_I u'(t)\varphi(t)\ \mathrm{d}t.\end{aligned}$$

Ainsi, $\overline{u}$ est faiblement dérivable et $\overline{u}' = u$ p.p. sur $I$. En particulier, $u-\overline{u}$ est faiblement dérivable et $(u-\overline{u})' = 0$ p.p. sur $I$. D'après le lemme 3.2, il existe donc une constante $C \in \mathbb{R}$ t.q. $u = \overline{u} + C$ p.p. sur $I$. La fonction $\tilde{u}$ définie par

$$\forall x \in I \quad \tilde{u}(x) = \overline{u}(x) + C$$

est uniformément continue sur $I$, et vérifie les conditions (3.12), (3.13).

Enfin, l'unicité provient du fait que si $u$ possède deux représentants continus $\tilde{u}_1, \tilde{u}_2$, alors d'après (3.12), ces deux fonctions sont égales presque partout sur $I$. Puisqu'elles sont continues, elles sont en fait égales en tout point de $I$.

□

**Remarque 3.5.** On pourra retenir de la preuve précédente une autre manière de calculer une dérivée faible, et donc de montrer qu'une fonction est dans $H^1(I)$. En effet, si $v \in L^2(I)$ et si, pour un $y_0 \in I$ fixé, une fonction $u \in L^2(I) \cap \mathcal{C}(\overline{I})$ s'écrit sous la forme

$$\forall x \in I \quad u(x) = \int_{y_0}^x v(t)\ \mathrm{d}t,$$

alors $u \in H^1(I)$ et $u' = v$.

**Théorème 3.2** (Opérateur de prolongement)**.** *Il existe un* OPÉRATEUR DE PROLONGEMENT $P : H^1(I) \to H^1(\mathbb{R})$ *linéaire et continu, t.q.*

*(i)* $(Pu)_{|I} = u \quad \forall u \in H^1(\mathbb{R})$

*(ii)* $\|Pu\|_{L^2(\mathbb{R})} \leq C\|Pu\|_{L^2(I)} \quad \forall u \in H^1(\mathbb{R})$

*(iii)* $\|Pu\|_{H^1(\mathbb{R})} \leq C\|Pu\|_{H^1(I)} \quad \forall u \in H^1(\mathbb{R})$

*(où $C > 0$ est une constante dépendant uniquement de $I$).*

**Théorème 3.3** (Densité)**.** *Soit $u \in H^1(I)$. Alors il existe une suite $(u_n)_{n\in\mathbb{N}} \in \mathcal{C}_c^\infty(\mathbb{R})$ t.q. $u_{n|I} \to u$ dans $H^1(I)$.*

Ces deux théorèmes sont admis. Le lecteur intéressé pourra se référer à [2], chapitre 8. Les théorèmes 3.2 et 3.3 permettent d'obtenir le résultat suivant :

**Théorème 3.4** (Injection de $H^1$ dans $L^\infty$)**.** *Il existe une constante $C > 0$ (dépendant uniquement de $I$) t.q.*

$$\|u\|_{L^\infty(I)} \leq C\|u\|_{H^1(I)} \quad \forall u \in H^1(I), \tag{3.14}$$

*c'est-à-dire $H^1(I) \subset L^\infty(I)$ avec injection* CONTINUE*.*

*De plus, lorsque $I$ est* BORNÉ*, l'injection $H^1(I) \subset \mathcal{C}(\overline{I})$ est* COMPACTE*, c'est-à-dire : de toute suite $(u_n)_{n\in\mathbb{N}} \in H^1(I)^{\mathbb{N}}$ bornée dans $H^1(I)$, on peut extraire une sous-suite $(u_{n_k})$ t.q. la suite des représentants continus $\tilde{u}_{n_k}$ converge uniformément vers une fonction $u \in \mathcal{C}(\overline{I})$.*

**Preuve.** Commençons par établir (3.14) dans le cas $I = \mathbb{R}$ ; le cas général s'en déduira par le théorème de prolongement 3.2. Pour $I = \mathbb{R}$, l'espace $\mathcal{C}_c^1(\mathbb{R})$ est dense dans $H^1(\mathbb{R})$ (d'après le théorème 3.3) ; on va donc raisonner par densité et considérer $v \in \mathcal{C}_c^1(\mathbb{R})$. Puisque $v$ est à support compact, en remarquant que $(v^2)' = 2v'v$, on peut écrire :

$$\begin{aligned}
\forall x \in \mathbb{R} \quad v(x)^2 &= 2\int_{-\infty}^{x} v'(t)v(t)\ \mathrm{d}t \\
&\leq 2\left(\int_{-\infty}^{x} v'(t)^2\ \mathrm{d}t\right)^{\frac{1}{2}}\left(\int_{-\infty}^{x} v(t)^2\ \mathrm{d}t\right)^{\frac{1}{2}} \\
&\leq 2\|v'\|_{L^2(\mathbb{R})}\|v\|_{L^2(\mathbb{R})} \\
&\leq \left(\|v\|_{L^2(\mathbb{R})}^2 + \|v'\|_{L^2(\mathbb{R})}^2\right) \\
&\leq \|v\|_{H^1(\mathbb{R})}^2
\end{aligned}$$

où l'on a utilisé successivement l'inégalité de Hölder, puis l'inégalité de Young ($ab \leq \frac{1}{2}(a^2 + b^2)$ pour $a, b \geq 0$). On en déduit :

$$\|v\|_{L^\infty(\mathbb{R})} \leq \|v\|_{H^1(\mathbb{R})}. \tag{3.15}$$

À présent, considérons $u \in H^1(\mathbb{R})$ ; il existe une suite $(u_n)_{n\in\mathbb{N}} \in \mathcal{C}_c^1(\mathbb{R})$ t.q.

$$\lim_{n\to\infty} \|u - u_n\|_{H^1(\mathbb{R})} = 0.$$

$(u_n)$ étant convergente dans $H^1(\mathbb{R})$, elle est de Cauchy. En appliquant l'inégalité (3.15) à des différences de la forme $v = u_p - u_q$, on voit que $(u_n)$ est également de Cauchy dans $L^\infty(\mathbb{R})$. Ce dernier étant un espace complet, $(u_n)$ converge dans $L^\infty(\mathbb{R})$. Comme $(u_n)$ converge vers $u$ dans $H^1(\mathbb{R})$, elle converge vers $u$ dans $L^2(\mathbb{R})$, et donc presque partout dans $\mathbb{R}$ ; cela montre que sa limite dans $L^\infty(\mathbb{R})$ est nécessairement égale à $u$. Ainsi, $u \in L^\infty(\mathbb{R})$ et en passant à la limite dans l'inégalité (3.15), on obtient

$$\|u\|_{L^\infty(\mathbb{R})} \leq \|u\|_{H^1(\mathbb{R})}.$$

Dans le cas général, on considère $I \subset \mathbb{R}$ un intervalle ouvert et $u \in H^1(I)$. On applique l'inégalité précedente au prolongement $Pu$ de $u$ dans $H^1(\mathbb{R})$ ; on obtient

$$\|Pu\|_{L^\infty(\mathbb{R})} \leq \|Pu\|_{H^1(\mathbb{R})}.$$

Or, par continuité du prolongement $P$, il existe une constante $C > 0$ t.q. $\|Pu\|_{H^1(\mathbb{R})} \leq C\|u\|_{H^1(I)}$ ; de plus, puisque $Pu = u$ presque partout sur $I$, on a $\|u\|_{L^\infty(I)} \leq \|Pu\|_{L^\infty(\mathbb{R})}$. Par conséquent, on obtient

$$\|u\|_{L^\infty(I)} \leq C\|u\|_{H^1(I)}$$

Dans le cas où $I$ est borné, montrons que l'injection $H^1(I) \subset \mathcal{C}(\overline{I})$ est compacte. On note $I = ]a, b[$ et on considère une suite $(u_n)_{n\in\mathbb{N}} \in H^1(a,b)^{\mathbb{N}}$, bornée dans $H^1(a,b)$. Pour tout $n \in \mathbb{N}$, on note $\tilde{u}_n$ le représentant continu de $u_n$ ; les fonctions $\tilde{u}_n$ sont continues sur $\overline{I}$, qui est un sous-ensemble compact de $\mathbb{R}$. Nous allons montrer que la suite $(\tilde{u}_n)$ est uniformément équicontinue, et conclure en appliquant le théorème d'Ascoli. Pour cela, on applique la relation (3.13) puis l'inégalité de Hölder :

$$\begin{aligned}
\forall n \in \mathbb{N},\ \forall x, y \in I \quad |\tilde{u}_n(x) - \tilde{u}_n(y)| &= \left| \int_{\min(x,y)}^{\max(x,y)} u_n'(t)\ \mathrm{d}t \right| \\
&\leq \sqrt{|x-y|}\ \left( \int_a^b (u_n'(t))^2\ \mathrm{d}t \right)^{\frac{1}{2}} \\
&\leq \sqrt{|x-y|}\ \|u_n\|_{H^1(I)} \\
&\leq C\sqrt{|x-y|}
\end{aligned}$$

où $C = \sup_{n\in\mathbb{N}} \|u_n\|_{H^1(I)} < \infty$. On en déduit que $(u_n)$ est uniformément équicontinue. Par conséquent, d'après le théorème d'Ascoli, il existe une fonction $u \in \mathcal{C}(\overline{I})$ et une

suite extraite $(\tilde{u}_{n_k})_{k\in\mathbb{N}}$ t.q.

$$\|u - \tilde{u}_{n_k}\|_\infty \to 0 \quad \text{quand } k \to \infty.$$

□

**Remarque 3.6.** L'inégalité (3.14) entraîne que si $(u_n)_{n\in\mathbb{N}} \in H^1(I)^{\mathbb{N}}$ converge vers une fonction $u$ dans $H^1(I)$, la convergence a également lieu dans $L^\infty(I)$. En effet, on a l'estimation :

$$\|u - u_n\|_{L^\infty(I)} \leq C\|u - u_n\|_{H^1(I)}.$$

**Corollaire 3.1.** *On suppose que $I$ n'est pas borné et soit $u \in H^1(I)$. Alors*

$$\lim_{x\in I,\ |x|\to\infty} u(x) = 0.$$

**Preuve.** D'après le théorème 3.3, il existe une suite $(u_n)$ de $\mathcal{C}^1_c(\mathbb{R})$ t.q. $u_{n|I} \to u$ dans $H^1(I)$. On déduit de (3.14) que $\|u_n - u\|_{L^\infty} \to 0$. Cela montre le résultat puisque, pour $\varepsilon > 0$ fixé, on peut choisir $n$ assez grand pour que $\|u_n - u\|_{L^\infty} < \varepsilon$ ; on choisit alors $|x|$ assez grand pour que $u_n(x) = 0$ ($u_n$ est à support compact), et donc $|u(x)| < \varepsilon$. □

**Corollaire 3.2** (Dérivation d'un produit)**.** *Soit $u, v \in H^1(I)$, alors $uv \in H^1(I)$ et*

$$(uv)' = u'v + uv'. \tag{3.16}$$

*De plus, on a la formule d'intégration par parties :*

$$\forall x, y \in \overline{I} \quad \int_x^y u'v = u(y)v(y) - u(x)v(x) - \int_x^y uv' \tag{3.17}$$

**Preuve.** Soit $u, v \in H^1(I)$ ; en particulier, $u \in L^\infty(I)$ (théorème 3.4), $v \in L^2(I)$ donc $uv \in L^2(I)$. De plus, $u_n \to u$ et $v_n \to v$ dans $L^2$ et $L^\infty$, donc $u_n v_n \to uv$ dans $L^2$. (On pourra le vérifier, à titre d'exercice, en écrivant $u_n v_n - uv = \frac{1}{2}\left((u_n - u)(v_n + v) + (u_n + u)(v_n - v)\right)$.)

Soit $(u_n), (v_n)$ deux suites de fonctions de classe $\mathcal{C}^1$, telles que les restrictions $u_{n|I} \to u$, $v_{n|I} \to v$ fortement dans $H^1(I)$. Au sens classique, on a la formule $(u_n v_n)' = u_n' v_n + u_n v_n'$. Nous allons montrer que $u_n' v_n + u_n v_n' \to u'v + uv'$ dans $L^2(I)$, ce qui conclura la preuve en vertu de la remarque 3.3. Comme $u_n' \to u'$ dans $L^2$ et que $v_n \to v$ dans $L^\infty$, on a $u_n' v_n \to u'v$ dans $L^2$ ; de même, $u_n v_n' \to uv'$ dans $L^2$, d'où le résultat.

**Théorème 3.5** (Dérivation d'un produit de composition)**.** *Soit $G \in \mathcal{C}^1(\mathbb{R})$ t.q. $G(0) = 0$, et soit $u \in H^1(I)$. Alors*

$$G \circ u \in H^1(I) \quad et \quad (G \circ u)' = (G' \circ u)u'.$$

**Remarque 3.7.** L'hypothèse $G(0) = 0$ est inutile si $I$ est borné.

**Preuve.** Montrons que $G \circ u \in L^2(I)$. Pour cela, on note $M = \|u\|_{L^\infty(I)}$ (qui est finie d'après le théorème 3.4). Comme $G(0) = 0$ et que $G'$ est continue, $G'$ est borné sur le compact $[-M, M] \subset \mathbb{R}$, donc d'après l'inégalité des accroissements finis, il existe une constante $C \geq 0$ t.q.

$$\forall s \in [-M, M] \quad |G(s)| \leq C|s|.$$

Par définition de la norme $L^\infty$, on a donc

$$|G(u(x))| \leq C|u(x)| \quad \text{p.p. } x \in I.$$

Puisque $u \in L^2(I)$, on en déduit que $G \circ u \in L^2(I)$. De même, on a

$$|G'(u(x))| \leq C \quad \text{p.p. } x \in I,$$

ce qui montre que $|(G' \circ u)u'| \leq C|u'|$ p.p. sur $I$, d'où $(G' \circ u)u' \in L^2(I)$ puisque $u' \in L^2(I)$.

Nous allons montrer la propriété suivante :

$$\forall \varphi \in \mathcal{C}_c^1(I) \quad \int_I (G \circ u)\varphi' = -\int_I (G' \circ u)u'\varphi \tag{3.18}$$

Pour cela, on utilise le théorème de densité 3.3 : il existe une suite $(u_n)$ de fonctions de $\mathcal{C}^\infty(\mathbb{R})$ t.q. $u_{n|I} \to u$ dans $H^1(I)$. On a donc, au sens des dérivées classiques,

$$\forall \varphi \in \mathcal{C}_c^1(I) \quad \int_I (G \circ u_n)\varphi' = -\int_I (G' \circ u_n)u_n'\varphi \tag{3.19}$$

Montrons que $G \circ u_n \to G \circ u$ dans $L^\infty(I)$. D'après le théorème 3.4, $u_{n|I} \to u$ également dans $L^\infty(I)$ ; par conséquent, on peut supposer que

$$\forall n \in \mathbb{N},\ \forall x \in I \quad |u_n(x)| \leq 2M.$$

$G$ étant continue, elle est uniformément continue sur $[-2M, 2M]$. Soit $\varepsilon > 0$ fixé ; il existe donc $\delta > 0$ t.q.

$$\forall s_1, s_2 \in [-2M, 2M] \quad |s_1 - s_2| \leq \delta \Rightarrow |G(s_1) - G(s_2)| \leq \varepsilon.$$

De plus, il existe $n_0 \in \mathbb{N}$ t.q.

$$\forall n \geq n_0 \quad |u(x) - u_n(x)| \leq \delta \quad \text{p.p. } x \in I.$$

On en déduit :

$$\forall n \geq n_0 \quad |G(u(x)) - G(u_n(x)) \leq \varepsilon \quad \text{p.p. } x \in I.$$

Par le même raisonnement, $G' \circ u_n \to G' \circ u$ dans $L^\infty(I)$. Comme $u_n' \to u'$ dans $L^2(I)$, on peut passer à la limite dans chaque intégrale de (3.19) pour obtenir (3.18).

□

## 3.3 L'espace $H_0^1(I)$

**Définition 3.4.** On note $H_0^1(I)$ la fermeture de $\mathcal{C}_c^1(I)$ pour la norme $H^1(I)$. Muni du produit scalaire induit par $H^1$, c'est un espace de Hilbert séparable.

**Remarque 3.8.** Lorsque $I = \mathbb{R}$, on a vu que $\mathcal{C}_c^1(I)$ est dense dans $H^1(\mathbb{R})$ (théorème (3.3), donc $H_0^1(\mathbb{R}) = H^1(\mathbb{R})$.

**Remarque 3.9.** En utilisant une suite régularisante $(\rho_n)$, on peut vérifier facilement que :

— $\mathcal{C}_c^\infty(I)$ est dense dans $H_0^1(I)$,
— si $u \in H^1(I) \cap \mathcal{C}_c(I)$ alors $u \in H_0^1(I)$.

**Théorème 3.6.** *Soit $u \in H^1(I)$, alors $u \in H_0^1(I)$ si et seulement si $u = 0$ sur $\partial I$ (au sens du représentant continu).*

**Preuve.** Si $u \in H_0^1(I)$ (identifiée à son représentant continu), alors il existe une suite $(u_n)_{n\in\mathbb{N}}$ de fonctions de $\mathcal{C}_c^1(I)$ t.q. $u_n \to u$ dans $H^1(I)$. Alors d'après la remarque (3.6), $u_n \to u$ uniformément, ce qui entraîne que $u = 0$ sur $\partial u$.

Réciproquement, soit $u \in H^1(I)$ t.q. $u = 0$ sur $\partial I$. On fixe une fonction $G \in C^1(\mathbb{R})$ t.q.

$$G(t) = \begin{cases} 0 & \text{si } |t| \leq 1 \\ t & \text{si } |t| \leq 2 \end{cases}$$

et $|G(t)| \leq |t|$ pour tout $t \in \mathbb{R}$. On définit la suite $u_n = \frac{1}{n} G(nu)$ de sorte que $u_n \in H^1(I)$ (voir théorème 3.5). D'autre part,

$$\text{Supp} u_n \subset \left\{ x \in I,\ |u(x)| \geq \frac{1}{n} \right\}$$

et donc $\text{Supp} u_n$ est un compact inclus dans $I$ (utiliser le fait que $u = 0$ sur $\partial I$ et $u(x) \to 0$ quand $|x| \to \infty$, $x \in I$). Par conséquent, $u_n \in H_0^1(I)$ (voir la remarque 3.9). Enfin on vérifie aisément à l'aide du théorème de convergence dominée que $u_n \to u$ dans $H^1(I)$. □

**Proposition 3.3** (Inégalité de Poincaré)**.** *On suppose que $I = (a, b)$ est* BORNÉ. *Alors il existe une constante $C > 0$ (dépendant de $|I| = b - a$) t.q.*

$$\|u\|_{H^1(I)} \leq C\ \|u'\|_{L^2(I)} \quad \forall u \in H_0^1(I)$$

**Preuve.** Soit $u \in H_0^1(I)$ (identifié à son représentant continu). En utilisant la propriété (3.13) et l'inégalité de Hölder, on peut écrire pour tout $x \in \overline{I}$ :

$$\begin{aligned} |u(x)| = |u(x) - u(a)| = \left| \int_a^x u'(t)\ \mathrm{d}t \right| &\leq \int_a^x |u'(t)|\ \mathrm{d}t \\ &\leq (b-a)^{1/2}\ \left( \int_a^b |u'(t)|^2\ \mathrm{d}t \right)^{1/2} \end{aligned}$$

En élevant au carré et en intégrant sur $(a, b)$, on en déduit

$$\int_a^b |u(x)|^2\ \mathrm{d}x \leq (b-a)^2\ \int_a^b |u'(t)|^2\ \mathrm{d}t,$$

d'où le résultat avec $C = b - a$.

□

**Remarque 3.10.** Si $I$ est borné, la forme bilinéaire

$$(u, v) \in H_0^1(I) \times H_0^1(I) \mapsto \langle u', v' \rangle_{L^2(I)}$$

définit donc un produit scalaire sur $H_0^1(I)$, et la norme associée (c'est-à-dire $\|u'\|_{L^2(I)}$) est une norme sur $H_0^1(I)$, qui est équivalente à la norme $\|u\|_{H^1(I)}$.

# Chapitre 4

# Approche variationnelle. Méthode des éléments finis

## 4.1 Approche variationnelle

### 4.1.1 Théorème de Lax-Milgram

Soit $V$ un espace de Hilbert. On note $\langle \cdot, \cdot \rangle$ le produit scalaire et $|\cdot|$ la norme associée.

**Définition 4.1.** On dit qu'une forme bilinéaire $a : V \times V \to \mathbb{R}$ est

(i) CONTINUE s'il existe une constante $M \geq 0$ telle que

$$|a(u,v)| \leq M|u||v| \quad \forall u, v \in V, \tag{4.1}$$

(ii) COERCIVE s'il existe une constante $\alpha > 0$ telle que

$$a(v,v) \geq \alpha |v|^2 \quad \forall v \in V, \tag{4.2}$$

(iii) SYMÉTRIQUE si

$$a(u,v) = a(v,u) \quad \forall u, v \in V.$$

**Théorème 4.1** (Lax-Milgram)**.** *Soit $a : V \times V \to \mathbb{R}$ une forme bilinéaire continue et coercive. Soit $b : V \to \mathbb{R}$ une forme linéaire continue. Alors il existe un unique $u \in V$ tel que*

$$a(u,v) = b(v) \quad \forall v \in V. \tag{4.3}$$

*Si, de plus, $a$ est symétrique, alors en notant*

$$E(v) = \frac{1}{2}a(v,v) - b(v) \quad \forall v \in V$$

*l'énergie associée au problème* (4.3)*, $u$ est caractérisé par la propriété suivante :*

$$E(u) = \min\{E(v),\ v \in V\}. \tag{4.4}$$

**Preuve. Cas général.** La preuve repose sur le théorème de représentation de Riesz. $b$ étant une forme linéaire continue sur $V$, il existe un unique $f \in V$ tel que

$$\forall v \in V \quad b(v) = \langle f, v\rangle.$$

De même, à $u \in V$ fixé, l'application $v \in V \mapsto a(u,v)$ est une forme linéaire continue, donc il existe un unique $\ell_u \in V$ tel que

$$\forall v \in V \quad a(u,v) = \langle \ell_u, v\rangle.$$

En utilisant la bilinéarité de $a$, on peut montrer facilement que l'application $u \in V \mapsto \ell_u \in V$ est linéaire ; on note $\ell_u = Au$. Ainsi, l'équation (4.3) se réécrit

$$\langle Au, v\rangle = \langle f, v\rangle \quad \forall v \in V$$

qui est équivalente à l'égalité

$$Au = f. \tag{4.5}$$

Pour montrer l'existence et l'unicité de $u \in V$ satisfaisant (4.5), nous allons montrer que $A$ est bijectif de $V$ dans lui-même.

D'après (4.1), on a :

$$\forall v \in V \quad |\langle Au, v\rangle| = |a(u,v)| \leq M|u||v|$$

ce qui montre que $|Au| \leq M|u|$, donc $A$ est continue. En utilisant la coercivité (4.2) et l'inégalité de Cauchy-Schwarz, on écrit :

$$\forall v \in V \quad |Av||v| \geq \langle Av, v\rangle \geq \alpha|v|^2$$

d'où

$$\forall v \in V \quad |Av| \geq \alpha|v| \tag{4.6}$$

ce qui entraîne que $A$ est injectif ($Av = 0$ entraîne $v = 0$).

Montrons que $A$ est surjectif. On note $R(A) \subset V$ son image (*range* en anglais). Tout d'abord, la propriété (4.6) entraîne que $R(A)$ est fermé. En effet, si $(Av_n)_{n\in\mathbb{N}}$ est une suite de $R(A)$, convergeant vers $w^* \in V$, alors (4.6) implique que $(v_n)$ est une suite bornée dans $V$ ; par conséquent, il existe une sous-suite $(v_{n_k})$ et un élément $v^* \in V$ tels que $v_{n_k} \rightharpoonup v^*$ dans $V$. Mais par continuité de $A$, cela entraîne $Av_{n_k} \rightharpoonup Av^*$. $(Av_{n_k})$ étant une suite extraite de $(Av_n)$, elle converge nécessairement vers la même limite, d'où $w^* = Av^*$ donc $w^* \in R(A)$.

$R(A)$ étant un sous-espace fermé de $V$, on peut écrire la décomposition suivante :

$$V = R(A) \oplus [R(A)]^{\perp}.$$

Il reste à montrer que $[R(A)]^{\perp}$ est réduit à $\{0\}$. Pour cela, on considère $w \in [R(A)]^{\perp}$ ; puisque $Aw \in R(A)$, on a par définition $\langle Aw, w\rangle = 0$. Mais alors, la coercivité de $a$ (4.2) entraîne que $w = 0$. D'où $V = R(A)$ et $A$ est surjectif.

Ainsi, il existe un unique $u \in V$ satisfaisant (4.5), ou de manière équivalente, (4.3).
**Cas où $A$ est symétrique.** On peut alors démontrer la propriété suivante :

$$\forall v \in V \quad E(v) - E(u) \geq \frac{\alpha}{2}|v - u|^2$$

ce qui montre que $u$ est l'unique élément de $V$ satisfaisant (4.4). Pour le voir, on remplace $v$ par $(v - u)$ dans (4.3), pour écrire

$$\forall v \in \mathbb{N} \quad a(u, v - u) = b(v - u).$$

On utilise la symétrie de $a$ pour écrire l'identité suivante :

$$\forall v \in \mathbb{N} \quad a(v - u, v - u) = a(v, v) + a(u, u) - 2a(u, v).$$

On effectue alors le calcul suivant :

$$\begin{aligned}
\forall v \in \mathbb{N} \quad E(v) - E(u) &= \frac{1}{2}a(v, v) - b(v) - \frac{1}{2}a(u, u) + b(u) \\
&= \frac{1}{2}\left(a(v, v) - a(u, u) - 2b(v - u)\right) \\
&= \frac{1}{2}\left(a(v, v) - a(u, u) - 2a(u, v - u)\right) \\
&= \frac{1}{2}\left(a(v, v) + a(u, u) - 2a(u, v)\right) \\
&= \frac{1}{2}a(v - u, v - u) \geq \frac{\alpha}{2}|v - u|^2
\end{aligned}$$

### 4.1.2 Application à la résolution d'une EDP linéaire du second ordre

Reprenons l'exemple introductif du chapitre précédent. En notant $I = ]0,1[$, on considère le problème suivant :

$$\begin{cases} -u'' + u = f \quad \text{sur } I \\ u(0) = u(1) = 0 \end{cases} \tag{4.7}$$

où $f$ est une fonction donnée (par exemple, $f \in \mathcal{C}(\overline{I})$ ou $f \in L^2(I)$).

**Définition 4.2.** — On appelle SOLUTION CLASSIQUE de (4.7) une fonction $u \in \mathcal{C}^2(\overline{I})$ qui vérifie (4.7) au sens usuel (ce qui suppose $f \in \mathcal{C}(\overline{I})$).

— On appelle SOLUTION FAIBLE une fonction $u \in H_0^1(I)$ vérifiant

$$\forall \varphi \in H_0^1(I) \qquad \int_I u'\varphi' + \int_I u\varphi = \int_I f\varphi. \tag{4.8}$$

La propriété (4.8) est appelée FORMULATION FAIBLE (ou FORMULATION VARIATIONNELLE) du problème (4.7).

**Remarque 4.1.** Dans la relation (4.8), $u'$ désigne la dérivée faible de $u$, définie au chapitre 3.

Le lien entre solution classique et solution faible est établi par la proposition suivante.

**Proposition 4.1.** *On suppose $f \in \mathcal{C}(\overline{I})$. Alors :*

*(i) toute solution classique $u \in \mathcal{C}^2(\overline{I})$ de (4.7), est une solution faible ;*

*(ii) toute solution faible $u \in H_0^1(I)$, de classe $\mathcal{C}^2(\overline{I})$, est une solution classique.*

**Preuve.** $(i)$ On raisonne par densité de $\mathcal{C}_c^1(I)$ dans $H_0^1(I)$. Soit $\varphi \in \mathcal{C}_c^1(I)$. On multiplie l'équation $-u'' + u = f$ par $\varphi$ et on intègre sur $I$ ; on obtient

$$\int_I -u''\varphi + \int_I u\varphi = \int_I f\varphi$$

En intégrant par parties dans la première intégrale, et en utilisant $\varphi(0) = \varphi(1) = 0$, on en déduit

$$\forall \varphi \in \mathcal{C}_c^1(I) \qquad \int_I u'\varphi' + \int_I u\varphi = \int_I f\varphi.$$

À présent, on se donne $\varphi \in H_0^1(I)$. D'après les résultats du chapitre 3, il existe une suite $(\varphi_n)_{n\in\mathbb{N}}$ de fonctions de $\mathcal{C}_c^1(I)$ telle que $\varphi_n \to \varphi$ en norme $H^1(I)$. En particulier, $\varphi_n' \to \varphi'$ et $\varphi_n \to \varphi$ dans $L^2(I)$. Puisque $\varphi_n \in \mathcal{C}_c^1(I)$, on peut écrire

$$\int_I u'\varphi_n' + \int_I u\varphi_n = \int_I f\varphi_n$$

et passer à la limite dans chaque intégrale en utilisant les convergences fortes dans $L^2(I)$, pour obtenir la propriété (4.8).

(*ii*) Si $u \in H_0^1(I)$ est une solution faible, en choisissant des fonctions tests régulières à support compact et en effectuant une intégration par parties, on obtient :

$$\forall \varphi \in \mathcal{C}_c^1(I) \qquad \int_0^1 (-u'' + u - f)\varphi = 0.$$

Si $u \in \mathcal{C}^2(\overline{I})$, et comme $f \in \mathcal{C}(\overline{I})$, la fonction $-u'' + u - f \in \mathcal{C}(\overline{I})$ ; en particulier, elle est dans $L^2(I)$. Par densité de $\mathcal{C}_c^1(I)$ dans $L^2(I)$, la relation précédente est vérifiée pour tout $\varphi \in L^2(I)$ ; en prenant $\varphi = -u'' + u - f$, on obtient donc

$$-u'' + u - f = 0 \qquad \text{presque partout sur } I.$$

Mais la fonction $-u'' + u - f$ étant continue sur $\overline{I}$, on a en fait la relation ponctuelle

$$-u''(x) + u(x) - f(x) = 0 \qquad \forall x \in [0,1].$$

$u$ est donc une solution classique.

Pour démontrer l'existence et l'unicité de la solution classique du problème (4.7) (dans le cas où $f \in \mathcal{C}(\overline{I})$), on pourra donc procéder en deux étapes.

**Étape 1.** Montrer l'existence et l'unicité de la solution faible $u \in H_0^1(I)$, définie par (4.8).

**Étape 2.** Montrer que le représentant continu $\tilde{u}$ de $u$, possède la régularité $\mathcal{C}^2(\overline{I})$. $\tilde{u}$ sera alors l'unique solution classique de (4.7).

L'étape 1 repose sur le résultat suivant.

**Proposition 4.2.** *Pour tout $f \in L^2(I)$, il existe un unique $u \in H_0^1(I)$ satisfaisant* (4.8). *De plus, $u$ est la solution du problème de minimisation :*

$$\min_{v\in H_0^1} \left\{ \frac{1}{2}\int_I (v'^2 + v^2) - \int_I fv \right\};$$

*c'est ce qu'on appelle le* PRINCIPE DE DIRICHLET.

**Preuve.** On applique le théorème de Lax-Milgram (théorème 4.1) dans l'espace de Hilbert $V = H_0^1(I)$, avec la forme bilinéaire

$$a(u,v) = \int_I u'v' + \int_I uv = \langle u, v\rangle_{H^1}$$

et la forme linéaire $b : v \mapsto \int_I fv$.

□

Pour l'étape 2, on commence par appliquer la formulation variationnelle en testant contre des fonctions $\varphi \in \mathcal{C}_c^1(I)$ ; en écrivant

$$\forall \varphi \in \mathcal{C}_c^1(I) \qquad \int_I u'\varphi' = -\int_I (u-f)\varphi,$$

on voit que $u'$ est faiblement dérivable et que

$$u''(x) = u(x) - f(x) \qquad \text{p.p. } x \in I.$$

En supposant simplement $f \in L^2(I)$, on en déduit que $u'' \in L^2(I)$, et donc $u' \in H^1(I)$ ; on peut donc l'identifier à son représentant continu, et écrire :

$$\forall x \in \overline{I} \qquad u'(x) - u'(0) = \int_0^x u''(t)\ \mathrm{d}t = \int_0^x (u(t) - f(t))\ \mathrm{d}t$$

On identifie également $u$ à son représentant continu ; si $f$ est également continue, l'identité précédente montre que $u'$ est en fait une fonction de classe $\mathcal{C}^1$ (en tant que primitive de la fonction continue $u - f$). En écrivant de même

$$\forall x \in \overline{I} \qquad u(x) = \int_0^x u'(t)\ \mathrm{d}t,$$

on en déduit finalement que $u \in \mathcal{C}^2(\overline{I})$. En appliquant la proposition 4.1, $(ii)$, on conclut que $u$ est une solution classique de (4.7).

## 4.2 Méthode des éléments finis en dimension 1

### 4.2.1 Principes généraux : méthode de Galerkin

On se place dans le cadre du théorème de Lax-Milgram : on se donne un espace de Hilbert $V$, une forme bilinéaire continue et coercive $a : V \times V \to \mathbb{R}$ et une forme linéaire

continue $b : V \to \mathbb{R}$. On note $|\cdot|$ la norme sur $V$, $M$ la constante de continuité de $a$, et $\alpha$ sa constante de coercivité (*i.e.* les meilleures constantes satisfaisant les propriétés (4.1) et (4.2), resp.).

On cherche à approcher dans $V$, l'unique solution $u \in V$ du problème variationnel suivant :

$$\forall v \in V \qquad a(u, v) = b(v) \tag{4.9}$$

La méthode de Galerkin consiste à se donner une famille de sous-espaces vectoriels $V_h \subset V$, de dimension finie, et à définir l'approximation $u_h \in V_h$ comme l'unique solution du problème suivant :

$$\forall v_h \in V_h \qquad a(u_h, v_h) = b(v_h) \tag{4.10}$$

Le problème (4.10) est appelé problème d'approximation interne, car la fonction $u_h$ appartient au même espace $V$ que la solution exacte $u$ (contrairement aux méthodes de différences finies, qui construisent des solutions approchées qui sortent du domaine de régularité des solutions exactes). Étant donné que pour tout $h > 0$, $V_h$ est un sous-espace de dimension finie de $V$, c'est en particulier un sous-espace fermé ; par conséquent, c'est un espace de Hilbert pour la norme induite par celle de $V$. En appliquant le théorème de Lax-Milgram, on obtient donc l'existence et l'unicité de $u_h \in V_h$, solution de (4.10).

Dans la pratique, les sous-espaces $V_h$ seront obtenus à partir d'une discrétisation du domaine spatial, associée à un pas d'espace $h$ ; par exemple, en dimension 1, $h$ sera le pas de la subdivision d'un intervalle borné $I \subset \mathbb{R}$. L'idée générale est que lorsque $h \to 0$, la dimension de l'espace $V_h$ tend vers l'infini ($V$ étant lui-même de dimension infinie) ; lorsque le nombre de dégrés de liberté augmente, les fonctions $u_h$ doivent décrire de plus en plus fidèlement le comportement de $u$, au sens de la norme de $V$.

L'écart entre $u_h$ et $u$ est quantifié par le lemme suivant.

**Lemme 4.1** (Lemme de Céa)**.** *Sous les hypothèses précédentes, on a l'inégalité :*

$$\forall h > 0 \qquad |u - u_h| \leq \frac{M}{\alpha} \inf_{v_h \in V_h} |u - v_h| = \frac{M}{\alpha}\, d(u, V_h) \tag{4.11}$$

*(où $d(u, V_h)$ désigne la distance de $u$ au sous-espace $V_h$).*

**Remarque 4.2.** Le terme $d(u, V_h)$ peut s'interpréter comme une erreur d'interpolation : c'est l'erreur commise en projetant $u$ sur le sous-espace $V_h$.

**Preuve.** Soit $w_h \in V_h$ ; puisque $w_h \in V$, on peut appliquer (4.9) et (4.8) pour écrire

$$a(u, w_h) = b(w_h), \quad a(u_h, w_h) = b(w_h)$$

d'où par bilinéarité de $a$ :

$$a(u - u_h, w_h) = 0 \qquad \forall w_h \in V_h. \tag{4.12}$$

À présent, soit $v_h \in V_h$ ; la coercivité, la bilinéarité et la continuité de $a$ permettent d'estimer $|u - u_h|$ en procédant ainsi :

$$\begin{aligned} \alpha|u - u_h|^2 \leq a(u - u_h, u - u_h) &= a(u - u_h, u - v_h) + a(u - u_h, v_h - u_h) \\ &= a(u - u_h, u - v_h) \\ &\leq M|u - u_h||u - v_h| \end{aligned}$$

(remarquer que $w_h := v_h - u_h \in V_h$). Par conséquent : soit $u_h = u$ et (4.11) est vraie, soit on peut diviser par $\alpha|u - u_h| > 0$ dans l'inégalité précédente, pour obtenir le résultat.

□

**Remarque 4.3.** Dans le cas où $a$ est symétrique, l'application $(u, v) \in V \times V \mapsto a(u, v) \in \mathbb{R}$ définit un produit scalaire sur $V$. La propriété (4.12) signifie alors que $u_h$ est la projection orthogonale de $u$ sur $V_h$, pour ce produit scalaire.

**Définition 4.3.** On dira que l'approximation de l'espace $V$ par la famille de sous-espaces $V_h$ est

— CONSISTANTE si

$$\lim_{h \to 0} d(v, V_h) = 0 \qquad \text{pour tout } v \in V;$$

— d'ordre $k \in \mathbb{N}$ s'il existe un constante $C > 0$ telle que

$$d(v, V_h) \leq Ch^k \qquad \text{pour tout } h > 0, \quad \text{pour tout } v \in V_h.$$

Contrairement au cas des différences finies, une méthode consistante est donc nécessairement convergente, d'après le lemme 4.1.

**Écriture d'un système linéaire à partir du problème** (4.10)**.** Pour résoudre numériquement le problème d'approximation interne (4.10), il faut se ramener à un système linéaire. Pour cela, considérons un sous-espace $V_h$ de dimension $N$ ; $V_h$ est décrit à l'aide d'une base $\phi_1^h, \ldots, \phi_N^h$, et $u_h$ se décompose sur cette base sous la forme d'une combinaison linéaire

$$u_h = \sum_{j=1}^{N} U_j^h \phi_j^h \tag{4.13}$$

Une application linéaire étant entièrement déterminée par ses valeurs sur une base, le problème (4.10) s'écrit de manière équivalente :

$$\forall i = 1 \ldots N \qquad a(u_h, \phi_i^h) = b(\phi_i^h) \tag{4.14}$$

En utilisant la bilinéarité de $a$ et la décomposition (4.13), on en déduit la relation :

$$\forall i = 1 \ldots N \qquad \sum_{j=1}^{N} a(\phi_j^h, \phi_i^h)\ U_j^h = b(\phi_i^h)$$

Ainsi, en définissant la matrice $A$ et les vecteurs $U, B$ par

$$A = \Big(a(\phi_j^h, \phi_i^h)\Big)_{1 \le i,j \le N} \quad U = (U_i^h)_{1 \le i \le N} \quad B = (b(\phi_i^h))_{1 \le i \le N} \tag{4.15}$$

le problème (4.10) est finalement équivalent à la résolution du système linéaire $AU = B$.

### 4.2.2 Éléments finis de Lagrange en dimension 1

Pour simplifier la présentation, nous allons considérer le problème modèle suivant :

$$\begin{cases} -u'' = f \quad \text{sur } ]0,1[ \\ u(0) = u(1) = 0 \end{cases} \tag{4.16}$$

où $f$ est une fonction de $L^2(0,1)$. On note $V$ l'espace $H_0^1(0,1)$ muni de la norme $H^1$, et on définit la forme bilinéaire $a : V \times V \to \mathbb{R}$ et la forme linéaire $b : V \to \mathbb{R}$ par

$$\forall u, v \in H_0^1(0,1) \qquad a(u,v) = \int_0^1 u'(x)v'(x)\ \mathrm{d}x, \qquad b(v) = \int_0^1 f(x)v(x)\ \mathrm{d}x.$$

$a$ est une forme bilinéaire symétrique ; elle est continue puisque d'après l'inégalité de Hölder,

$$\forall u, v \in H_0^1(0,1) \qquad \int_0^1 u'(x)v'(x)\ \mathrm{d}x \le \|u'\|_{L^2(0,1)} \|v'\|_{L^2(0,1)} \le \|u\|_{H^1(0,1)} \|v\|_{H^1(0,1)}$$

et elle est coercive d'après l'inégalité de Poincaré (voir chapitre 3, proposition 3.3). En effet, il existe une constante $C_P > 0$ telle que

$$\forall u \in H_0^1(0,1) \quad \|u'\|_{L^2(0,1)} \geq \frac{1}{C_P}\|u\|_{H^1(0,1)},$$

autrement dit,

$$\forall u \in H_0^1(0,1) \quad a(u,u) \geq \alpha \|u\|^2_{H^1(0,1)}$$

avec $\alpha = 1/C_P^2$. De plus, $b$ est continue puisque

$$\forall v \in H_0^1(0,1) \quad \int_0^1 f(x)v(x)\,\mathrm{d}x \leq \|f\|_{L^2(0,1)}\|v\|_{L^2(0,1)} \leq \|f\|_{L^2(0,1)}\|v\|_{H^1(0,1)}.$$

D'après le théorème de Lax-Milgram, il existe donc une unique solution faible du problème (4.16), c'est-à-dire une unique fonction $u \in H_0^1(0,1)$ satisfaisant la formulation variationnelle :

$$\forall v \in H_0^1(0,1) \qquad a(u,v) = b(v).$$

**Éléments finis de Lagrange en dimension 1.** Dans cette méthode, chaque espace $V_h$ est associé à une subdivision de l'intervalle d'étude, ici $I = [0,1]$. Étant donné un entier $J \in \mathbb{N}$, on définit le pas d'espace $h = 1/(J+1)$ et la subdivision $(x_j)_{0 \leq j \leq J+1}$ associée, où $x_j = jh$ pour tout $j$. On définit alors $V_h$ par :

$$V_h = \Big\{ v_h \in \mathcal{C}([0,1]), \quad v_h(0) = v_h(1) = 0 \quad \text{et} \quad \forall j = 0, \ldots, J, \quad v_{h|[x_j, x_{j+1}]} \quad \text{est affine} \Big\}$$

**Proposition 4.3.** *Les ensembles d'approximation $V_h$ définis ci-dessus, sont des sous-espaces vectoriels de $V = H_0^1(I)$.*

**Preuve.** La preuve est laissée en exercice. (On pourra vérifier que si $v_h \in V_h$, alors sa dérivée faible existe et est donnée par $v_h' = \sum_{j=0}^{J} \lambda_i \chi_{]x_j, x_{j+1}[}$ où $\chi_{]x_j,x_{j+1}[}$ est l'indicatrice de l'intervalle $]x_j, x_{j+1}[$ et où $\lambda_i$ est la pente de la restriction de $v_h$ au même intervalle.) □

**Remarque 4.4.** L'hypothèse de continuité signifie que les limites à gauche et à droite des points de subdivision internes (les $x_j$ pour $1 \leq j \leq J$), sont les mêmes, et égales à $v_h(x_j)$. Cette hypothèse est essentielle pour que les fonctions $v_h$ soient bien des fonctions de $H^1(0,1)$.

**Proposition 4.4.** *Si $J \in \mathbb{N}$ et $h = 1/(J+1)$, alors $V_h$ est un sous-espace de $V$ de dimension $J$.*

**Preuve.** Il suffit de remarquer qu'une fonction $v_h \in V_h$ est entièrement déterminée par la donnée des valeurs prises aux points $x_j$, pour $j = 1, \ldots, J$ (faire un dessin). Ainsi, $V_h$ est en bijection avec $\mathbb{R}^J$ ; de plus, on voit facilement que cette bijection est linéaire.

□

**Choix d'une base de $V_h$.** Nous allons utiliser une base $(\phi_i^h)_{1 \leq i \leq J}$ de $V_h$, pour laquelle le système $AU = B$ (voir (4.15)) sera facile à résoudre numériquement. Pour $i = 1, \ldots, J$, on définit la fonction $\phi_i^h : [0,1] \to \mathbb{R}$ par

$$\phi_i^h(x) = \begin{cases} \dfrac{x - x_{i-1}}{h} & \text{si } x \in [x_{i-1}, x_i], \\ \dfrac{x_{i+1} - x}{h} & \text{si } x \in [x_i, x_{i+1}], \\ 0 & \text{sinon.} \end{cases} \tag{4.17}$$

Pour tout $i = 1, \ldots, J$, $\phi_i^h \in V_h$ et sa dérivée faible $(\phi_i^h)'$ est définie presque partout sur $[0,1]$ par

$$(\phi_i^h)'(x) = \begin{cases} \dfrac{1}{h} & \text{si } x \in ]x_{i-1}, x_i[, \\ \dfrac{-1}{h} & \text{si } x \in ]x_i, x_{i+1}[, \\ 0 & \text{sinon.} \end{cases}$$

Pour tout couple $(i,j) \in [|1, \ldots, J|]^2$, $a(\phi_j^h, \phi_i^h) = \int_0^1 (\phi_i^h)'(\phi_j^h)'$. Or, $(\phi_i^h)'$ et $(\phi_j^h)'$ sont à supports disjoints si $|i-j| \geq 2$, donc dans ce cas, le coefficient matriciel associé $a(\phi_j^h, \phi_i^h)$

est nul. Il suffit donc de traiter les cas $j = i$, $j = i - 1$, $j = i + 1$.

$$\begin{aligned} a(\phi_i^h, \phi_i^h) = \int_0^1 [(\phi_i^h)'(x)]^2 \, \mathrm{d}x &= \int_{x_{i-1}}^{x_i} \left(\frac{1}{h}\right)^2 \, \mathrm{d}x + \int_{x_i}^{x_{i+1}} \left(\frac{1}{h}\right)^2 \, \mathrm{d}x \\ &= \frac{2}{h} \\ a(\phi_{i+1}^h, \phi_i^h) = \int_0^1 (\phi_{i+1}^h)'(x) \; (\phi_i^h)'(x) \, \mathrm{d}x &= \int_{x_i}^{x_{i+1}} \frac{-1}{h}\frac{1}{h} \, \mathrm{d}x \\ &= -\frac{1}{h} \\ a(\phi_{i-1}^h, \phi_i^h) = \int_0^1 (\phi_{i-1}^h)'(x) \; (\phi_i^h)'(x) \, \mathrm{d}x &= \int_{x_{i-1}}^{x_i} \frac{1}{h}\frac{-1}{h} \, \mathrm{d}x \\ &= -\frac{1}{h} \end{aligned}$$

D'où la matrice $A$ :

$$A = \begin{pmatrix} \frac{2}{h} & -\frac{1}{h} & 0 & \dots & \dots & \dots & 0 \\ -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} & 0 & \dots & \dots & \vdots \\ 0 & \ddots & \ddots & \ddots & \ddots & \dots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & 0 \\ \vdots & \dots & \dots & 0 & -\frac{1}{h} & \frac{2}{h} & -\frac{1}{h} \\ 0 & \dots & \dots & \dots & 0 & -\frac{1}{h} & \frac{2}{h} \end{pmatrix}$$

**Remarque 4.5.** Pour n'importe quelle base de $V_h$, la matrice $A$ que l'on obtient est nécessairement inversible, puisque d'après le théorème de Lax-Migram, le système $AU = B$ possède une solution unique (qui est le vecteur des coordonnées de $u_h$, solution de (4.10), dans la base choisie). En choisissant la base définie par (4.17), on voit que l'on obtient une matrice tri-diagonale, facile à inverser. Étant donné $V \in \mathbb{R}^J$, en notant $v_h = \sum_{j=1}^J V_j \phi_j^h$, on peut vérifier que

$$\alpha \|v_h\|_{H^1}^2 \leq a(v_h, v_h) = \sum_{i,j} V_j V_i \; a(\phi_j^h, \phi_i^h) = V^T A V,$$

ce qui montre que $A$ est définie positive. On pourra donc utiliser la décomposition de Cholesky (qui est implémentée en Scilab dans le cadre des matrices creuses) pour

résoudre le système $AU = B$ de manière efficace, même pour un grand nombre de variables $J$.

Il reste à calculer le vecteur $B$, défini par $B = (b(\phi_i^h))_{1 \leq i \leq J}$. Cependant,

$$b(\phi_i^h) = \int_0^1 f(x)\ \phi_i^h(x)\ \mathrm{d}x,$$

et cette quantité n'est pas calculable de manière exacte, pour des $f$ très générales. En pratique, on utilisera des formules de quadrature pour calculer un vecteur $\tilde{B}$, approximation du vecteur $B$ ; pour cela, on décompose

$$\int_0^1 f(x)\ \phi_i^h(x)\ \mathrm{d}x = \int_{x_{i-1}}^{x_i} f(x)\frac{x - x_{i-1}}{h}\ \mathrm{d}x + \int_{x_i}^{x_{i+1}} f(x)\frac{x_{i+1} - x}{h}\ \mathrm{d}x$$

et on approche chaque intégrale par une méthode de quadrature. On pourra utiliser, par exemple :

— une méthode de point milieu, basée sur l'approximation :

$$\int_a^b g(x)\ \mathrm{d}x \approx (b-a)\ g\left(\frac{a+b}{2}\right),$$

ce qui donne ici :

$$\begin{aligned}\int_0^1 f(x)\ \phi_i^h(x)\ \mathrm{d}x &\approx hf\left((i-\frac{1}{2})h\right)\frac{1}{2} + hf\left((i+\frac{1}{2})h\right)\frac{1}{2} \\ &\approx \frac{h}{2}\left[f\left((i-\frac{1}{2})h\right) + f\left((i+\frac{1}{2})h\right)\right]\end{aligned}$$

— une méthode de Simpson, basée sur l'approximation :

$$\int_a^b g(x)\ \mathrm{d}x \approx \frac{b-a}{6}\left[g(a) + 4g\left(\frac{a+b}{2}\right) + g(b)\right],$$

d'où

$$\begin{aligned}\int_0^1 f(x)\ \phi_i^h(x)\ \mathrm{d}x &\approx \frac{h}{6}\left[0 + 4f\left((i-\frac{1}{2})h\right)\frac{1}{2} + f(ih)\right] + \frac{h}{6}\left[f(ih) + 4f\left((i+\frac{1}{2})h\right)\frac{1}{2} + 0\right] \\ &\approx \frac{h}{3}\left[f\left((i-\frac{1}{2})h\right) + f\left((i+\frac{1}{2})h\right) + f(ih)\right]\end{aligned}$$

## 4.3 Vers la dimension 2

### 4.3.1 La formule de Green, une généralisation de l'intégration par parties

Pour travailler avec des formulations faibles de problèmes définis sur un domaine $\Omega \subset \mathbb{R}^2$, on a besoin de généraliser l'intégration par parties (qui nous a permis d'établir des formulations faibles de problèmes posés sur un intervalle $I \subset \mathbb{R}$), à de tels domaines du plan. Ces formules d'intégration par parties généralisées sont appelées formules de Green ; elles font intervenir des intégrales sur $\Omega$ (intégrales multiples) et des intégrales sur le bord $\partial\Omega$ (intégrales curvilignes). Commençons par rappeler quelques définitions et résultats utiles.

**Proposition 4.5** (Intégration par tranches)**.** *Soit $a < b$ des réels, et $\varphi_1, \varphi_2 : [a,b] \subset \mathbb{R} \to \mathbb{R}$ deux fonctions continues, telles que $\varphi_1(x) \leq \varphi_2(x)$ pour tout $x \in [a,b]$. On définit la partie $D$ de $\mathbb{R}^2$ par*

$$D = \left\{(x,y) \in \mathbb{R}^2,\ a \leq x \leq b \text{ et } \varphi_1(x) \leq y \leq \varphi_2(x)\right\}.$$

*$D$ est fermé donc mesurable (pour la mesure de Lebesgue $\mathcal{L}^2$). De plus, si $f : D \subset \mathbb{R}^2 \to \mathbb{R}$ est une fonction intégrable sur $D$, alors pour tout $x \in [a,b]$, la fonction $y \mapsto f(x,y)$ est intégrable sur $[\varphi_1(x), \varphi_2(x)]$ ; enfin, la fonction $x \mapsto \int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\ \mathrm{d}y$ est elle-même intégrable sur $[a,b]$, et on a la formule d'intégration par tranches :*

$$\int_D f = \int_a^b \left( \int_{\varphi_1(x)}^{\varphi_2(x)} f(x,y)\ \mathrm{d}y \right) \mathrm{d}x.$$

**Définition 4.4** (Courbe paramétrée)**.** On appelle COURBE PARAMÉTRÉE une application continue $\gamma$, définie sur un intervalle $I = [0, L]$, à valeurs dans $\mathbb{R}^2$. Le SUPPORT de la courbe $\gamma$ est l'ensemble

$$\Gamma := \{\gamma(t),\ t \in [0,L]\}.$$

On dit aussi que $\gamma$ est une PARAMÉTRISATION de $\Gamma$.

**Définition 4.5** (Courbe régulière)**.** Soit $\gamma : [0,L] \to \mathbb{R}^2, t \mapsto \gamma(t) = (x(t), y(t))$ une courbe paramétrée. Si $\gamma$ possède une extension de classe $\mathcal{C}^1$ à un intervalle ouvert contenant $[0,L]$, on dit que $\gamma$ est une courbe paramétrée de classe $\mathcal{C}^1$. Si de plus sa dérivée

$\gamma'(t) = (x'(t), y'(t))$ ne s'annule pas sur $[0, L]$, on dit que $\gamma$ est une courbe RÉGULIÈRE. Le vecteur $\gamma'(t)$ est un VECTEUR TANGENT à la courbe au point $\gamma(t)$ ; sa direction définit la tangente à la courbe en $\gamma(t)$ .

**Exemple 4.1** (Paramétrisation d'un segment)**.** Étant donnés deux points $A, B \in \mathbb{R}^2$, le segment $[A, B]$ est le support de la courbe $\gamma : [0, 1] \to \mathbb{R}^2, t \mapsto A + t(B - A)$.

**Exemple 4.2** (Paramétrisation d'un cercle)**.** Le cercle de centre $(x_0, y_0) \in \mathbb{R}^2$ et de rayon $r > 0$ est le support de la courbe $\gamma : [0, 1] \to \mathbb{R}^2, t \mapsto (x_0 + r\cos t, y_0 + r\sin t)$.

**Exemple 4.3** (Graphe d'une fonction)**.** Soit $F : [0, L] \to \mathbb{R}$ une fonction de classe $\mathcal{C}^1$. L'application $\gamma : [0, L] \to \mathbb{R}^2, t \mapsto (t, F(t))$ est une courbe paramétrée régulière, dont le support est le graphe de $F$. La tangente en tout point $(t, F(t))$ est dirigée par le vecteur $(1, F'(t))$.

**Définition 4.6** (Intégrale curviligne)**.** Soit $U \subset \mathbb{R}^2$ un ouvert, $\gamma : [0, L] \to U, t \mapsto (x(t), y(t))$ une courbe régulière et $f : U \to \mathbb{R}$ une application continue. L'intégrale de $f$ le long de $\gamma$ est définie par

$$\int_\gamma f := \int_0^L f(x(t), y(t))\sqrt{x'(t)^2 + y'(t)^2}\ \mathrm{d}t.$$

**Remarque 4.6.** L'intégrale curviligne est invariante par changement de paramétrisation : si $\gamma : [0, L] \to \mathbb{R}^2$ est une courbe paramétrée de classe $\mathcal{C}^1$ et si l'application $\phi : [0, L'] \to [0, L]$ est un $\mathcal{C}^1$-difféomorphisme, alors

$$\int_\gamma f = \int_{\gamma\circ\phi} f.$$

C'est une conséquence directe de la formule de changement de variables dans une intégrale simple.

**Remarque 4.7.** Si $\Gamma$ est le support d'une courbe paramétrée régulière $\gamma : [0, L] \to \mathbb{R}^2$, telle que $\gamma_{|(0,L)}$ soit injective (avec éventuellement $\gamma(0) = \gamma(L)$), on notera généralement $\int_\Gamma$ l'intégrale curviligne le long de $\gamma$.

**Définition 4.7** (Normale extérieure)**.** Soit $D \subset \mathbb{R}^2$ un domaine borné, dont le bord $\partial D$ est le support d'une courbe paramétrée $\gamma : [0, L] \to \mathbb{R}^2$, régulière. On suppose que $\gamma_{|(0,L)}$ est injective et que $\gamma$ est orientée dans le sens trigonométrique, c'est-à-dire qu'un

observateur imaginaire se marchant sur le plan le long de $\partial D$, dans le sens de $\gamma$, a le domaine $D$ sur sa gauche.

Le VECTEUR NORMAL EXTÉRIEUR en tout point $\gamma(t)$ de $\partial D$, est le vecteur obtenu par une rotation d'angle $-\pi/2$ du vecteur tangent $\gamma'(t)$, et division par sa norme. Si $x$ est un point du bord, on note $\mathbf{n}(x)$ le vecteur normal extérieur.

**Proposition 4.6** (Formule de Green en dimension 2)**.** *Soit $D \subset \mathbb{R}^2$ un domaine borné, dont le bord $\partial D$ est le support d'une courbe paramétrée régulière. Soit $u \in \mathcal{C}^1(\overline{D})$. La formule de Green s'écrit :*

$$\forall i = 1, 2, \qquad \int_D \frac{\partial u}{\partial x_i} \, \mathrm{d}x\mathrm{d}y = \int_{\partial D} u \; n_i,$$

*$n_i$ étant la $i$-ème composante du vecteur normal extérieur à $\partial D$.*

**Proposition 4.7** (Intégration par parties généralisée)**.** *Si $D$ est un domaine borné de $\mathbb{R}^2$, à bord régulier, alors on a la formule suivante pour $u, v \in \mathcal{C}^1(\overline{D})$ :*

$$\forall i = 1, 2, \qquad \int_D \frac{\partial u}{\partial x_i} v = -\int_D u \frac{\partial v}{\partial x_i} + \int_{\partial D} uv \; n_i.$$

**Preuve :** C'est une conséquence immédiate de la formule de Green.

**Preuve de la proposition 4.6.** Nous allons faire la preuve dans le cas d'un domaine élémentaire du plan.

**Définition 4.8** (Domaine élémentaire du plan)**.** Un DOMAINE ÉLÉMENTAIRE $D$ du plan est la donnée de fonctions de classe $\mathcal{C}^1$, $\phi_1, \phi_2 : [a, b] \subset \mathbb{R} \to \mathbb{R}$, $\psi_1, \psi_2 : [c, d] \subset \mathbb{R} \to \mathbb{R}$ telles que

$$D = \{(x, y), \; a \leqslant x \leqslant b \text{ et } \phi_1(x) \leqslant y \leqslant \phi_2(x)\} = \{(x, y), \; c \leqslant y \leqslant d \text{ et } \psi_1(y) \leqslant x \leqslant \psi_2(y)\}.$$

Soit donc $D$ un tel domaine élémentaire. On applique le théorème de Fubini pour calculer les intégrales sur $D$ :

$$\begin{aligned} \int_D \frac{\partial u}{\partial y} &= \int_a^b \left( \int_{\phi_1(x)}^{\phi_2(x)} \frac{\partial u}{\partial y}(x, y) \mathrm{d}y \right) \mathrm{d}x \\ &= \int_a^b \left( u(x, \phi_2(x)) - u(x, \phi_1(x)) \right) \mathrm{d}x \end{aligned}$$

On note $\Gamma_2$ la partie du bord décrite par le graphe de $\phi_2$ et $\Gamma_1$ la partie décrite par le graphe de $\phi_1$. Les normales extérieures se calculent par les formules :

$$\mathbf{n}(x,\phi_2(x)) = \frac{(-\phi_2'(x),1)}{\sqrt{1+\phi_2'(x)^2}}, \qquad \mathbf{n}(x,\phi_1(x)) = \frac{(\phi_1'(x),-1)}{\sqrt{1+\phi_1'(x)^2}}$$

On peut donc écrire :

$$\begin{aligned}\int_{\Gamma_2} u\ n_y &= \int_a^b \frac{u(x,\phi_2(x))}{\sqrt{1+\phi_2'(x)^2}}\sqrt{1+\phi_2'(x)^2}\mathrm{d}x \\ &= \int_a^b u(x,\phi_2(x))\mathrm{d}x\end{aligned}$$

et de même

$$\int_{\Gamma_1} u\ n_y = -\int_a^b u(x,\phi_1(x))\mathrm{d}x.$$

Ainsi

$$\begin{aligned}\int_a^b \left(u(x,\phi_2(x)) - u(x,\phi_1(x))\right)\mathrm{d}x &= \int_a^b u(x,\phi_2(x))\mathrm{d}x - \int_a^b u(x,\phi_1(x))\mathrm{d}x \\ &= \int_{\Gamma_2} u\ n_y + \int_{\Gamma_1} u\ n_y \\ &= \int_{\partial D} u\ n_y.\end{aligned}$$

On procède de manière analogue pour l'intégrale de $\frac{\partial u}{\partial x}$, en écrivant

$$\int_D \frac{\partial u}{\partial x} = \int_c^d \left(\int_{\psi_1(y)}^{\psi_2(y)} \frac{\partial u}{\partial x}(x,y)\mathrm{d}x\right)\mathrm{d}y.$$

□

**Exemple 4.4** (aire d'un disque)**.** On considère le disque $D$ de centre $(x_0,y_0)\in\mathbb{R}^2$, de rayon $r>0$. On introduit la paramétrisation suivante du cercle :

$$\begin{aligned}\gamma : [0,2\pi] &\to \mathbb{R}^2 \\ t &\mapsto (x_0 + r\cos t, y_0 + r\sin t).\end{aligned}$$

La normale extérieure est définie par $\mathbf{n}(\gamma(t)) = (\cos t, \sin t)$. En appliquant la formule de Green, on obtient :

$$\begin{aligned}\int_D 1 &= \int_{\partial D} x\ n_x \\ &= \int_0^{2\pi} (x_0 + r\cos t)(\cos t) r\ \mathrm{d}t \\ &= \int_0^{2\pi} r^2 \cos^2 t\ \mathrm{d}t \\ &= r^2 \int_0^{2\pi} \frac{1+\cos(2t)}{2}\ \mathrm{d}t \\ &= \pi r^2.\end{aligned}$$

### 4.3.2 Formulation faible d'EDP en dimension 2

Soit $D \subset \mathbb{R}^2$ un domaine borné, à bord régulier et $\mathbf{u} \in \mathcal{C}(\overline{D}), \mathbb{R}^N)$ un champ de vecteurs. On note $u_i$ la $i-$ème composante de $\mathbf{u}$. On rappelle que div $\mathbf{u}$ est le champ scalaire défini par

$$\forall x \in D \qquad (\text{div } \mathbf{u})(x) = \sum_{i=1}^{2} \frac{\partial u_i}{\partial x_i}(x).$$

En appliquant la formule de Green à chaque composante de $\mathbf{u}$ et en faisant la somme, on obtient la formule suivante (formule de Stokes) :

$$\int_D \text{div } \mathbf{u} = \int_{\partial D} \mathbf{u} \cdot \mathbf{n} \tag{4.18}$$

Si $\phi \in \mathcal{C}^1(\overline{D})$, en utilisant la formule

$$\text{div } (\phi \mathbf{u}) = \phi \text{ div } \mathbf{u} + \mathbf{u} \cdot \nabla \phi,$$

on obtient la formule d'intégration par parties :

$$\int_D (\text{div } \mathbf{u})\phi = -\int_D \mathbf{u} \cdot \nabla \phi + \int_{\partial D} \phi \mathbf{u} \cdot \mathbf{n}. \tag{4.19}$$

Rappelons que si $u \in \mathcal{C}^2(\overline{D})$, on a la relation suivante :

$$\Delta u = \text{div } (\nabla u).$$

En appliquant la formule (4.19) avec $\mathbf{u} = \nabla u$, on en déduit la formule suivante :

$$-\int_D (\Delta u)\ \phi = \int_D \nabla u \cdot \nabla \phi - \int_{\partial D} \phi \frac{\partial u}{\partial \mathbf{n}} \tag{4.20}$$

où on a noté $\dfrac{\partial u}{\partial \mathbf{n}} = \nabla u \cdot \mathbf{n}$.

La formule (4.20) généralise la formule

$$-\int_a^b u''\phi = \int_a^b u'\phi' - \left[u'\phi\right]_a^b$$

qui permet de définir la formulation faible des problèmes de type $-u'' = f$ sur $[a, b]$, avec conditions aux limites de Dirichlet ou de Neumann en $a$ et $b$. De même, en utilisant la relation intégrale (4.20), on définit une solution faible du problème

$$\begin{aligned} -\Delta u &= f \quad \text{dans } D, \\ u &= 0 \quad \text{sur } \partial D \end{aligned}$$

(avec $f \in L^2(D)$) comme une fonction $u \in H_0^1(D)$ telle que

$$\forall \phi \in H_0^1(D) \quad \int_D \nabla u \cdot \nabla \phi = \int_D f\phi.$$

L'espace $H_0^1(D)$ est défini comme l'adhérence pour la norme $H^1$ de l'ensemble des fonctions de classe $\mathcal{C}^\infty$ à support compact dans $D$.

De même, on appellera solution faible du problème

$$\begin{aligned} -\Delta u &= f \quad \text{dans } D, \\ \frac{\partial u}{\partial n} &= 0 \quad \text{sur } \partial D \end{aligned}$$

fonction $u \in H^1(D)$ telle que

$$\forall \phi \in H^1(D) \quad \int_D \nabla u \cdot \nabla \phi = \int_D f\phi.$$

# Bibliographie

[1] Grégoire ALLAIRE, Analyse numérique et optimisation

[2] Haïm BREZIS, Analyse fonctionnelle

[3] Jean-Pierre DEMAILLY, Analyse numérique et équations différentielles