MAO Calcul scientifique

3 juin 2020

Table des matières

Introduction

L'objectif de ce module MAO Calcul scientifique est de proposer une introduction à la discrétisation et à la résolution numérique des équations aux dérivées partielles. Comme il s'agit d'une introduction, nous avons choisi de brosser un panorama des méthodes numériques pour les équations aux dérivées partielles :

- La première partie du cours portera sur la discrétisation d'EDP d'évolution par la méthode des différences finies. Cette première partie peut constituer une bonne préparation à l'option "Calcul scientifique" de l'épreuve de Modélisation de l'agrégation. Les outils utilisés sont très élémentaires.
- La seconde partie du cours portera sur la méthode des éléments finis pour des équations aux dérivées partielles elliptiques (équation de Poisson, fonctions propres du Laplacien, problème d'obstacle). Cette partie empruntera (et rappellera) quelques outils de la théorie des distributions, et plus précisément des espaces de Sobolev.

Nous chercherons à répondre à plusieurs questions de nature très différentes, allant de questions théoriques (discrétisation des EDP et étude de convergence du discret vers le continu) à des questions pratiques (algorithmes de résolution de systèmes discrets, mise en œuvre informatique en Python) :

- Comment discrétiser les équations aux dérivées partielles ? Les équations aux dérivées partielles linéaires (équation de la chaleur, des ondes, de transport, équation de Poisson, etc.) peuvent être vues comme des systèmes linéaires dans des espaces de fonctions (e.g. espaces de Sobolev). Comment passer d'un système linéaire en dimension infinie à un système linéaire en dimension finie ?
- Les systèmes discrets construits constituent-ils une bonne approximation des EDP qu'on considère ? Si oui, en quel sens ? Peut-on estimer l'erreur commise ? Ces question relèvent de l'analyse numérique ; les démonstrations de convergence du discret vers le continu reposent souvent sur des versions discrètes de principes utilisés en analyse des EDP (par exemple, principe du maximum, étude de stabilité, méthodes hilbertiennes).
- Comment résoudre les systèmes discrets ? Il s'agit d'introduire et d'étudier des algorithmes pour la résolution de systèmes linéaires (Ax=b), la recherche de valeurs propres, l'optimisation avec contraintes, etc.

Chapitre 1

Méthodes numériques pour les EDO

Avant de s'attaquer à la discrétisation des équations aux dérivées partielles d'évolution, on fait quelques rappels sur celle des équations différentielles ordinaires. On insistera en particulier sur la ressemblance entre l'étude de stabilité des solutions à une équation différentielle ordinaire (EDO) et les démonstration de convergence des méthodes numériques pour les EDO : les deux reposent sur le lemme de Gronwall.

Dans la suite, on considère un système d'équations différentielles de la forme

$$\begin{cases} y'(t) = F(t, y(t)) \quad t \in [0, T] \\ y(0) = y_0, \end{cases} \tag{1.1}$$

où $F : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$. Pour simplifier l'exposition, on suppose que F est continue et globalement Lipschitzienne en sa seconde variable,

$$\forall (t, y_1, y_2) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d, \quad \|F(t, y_1) - F(t, y_2)\| \leq L \|y_1 - y_2\| . \tag{1.2}$$

Le théorème de Cauchy-Lipschitz (version globale) garantit l'existence d'une solution définie sur $\mathbb{R}$ du système (1.1) pour toute donnée initiale $y_0 \in \mathbb{R}^d$.

1.1 Stabilité des solutions à une EDO

La proposition suivante montre que les solutions d'une EDO dépendent continûment de la fonction dans le second membre. La convergence des schémas numériques pour l'approximation des équations différentielles peut être vue comme une variante (ou un raffinement) de ce résultat de stabilité.

Proposition 1.1 (Stabilité des solutions)**.** *Soient $y \in \mathcal{C}^1([0, T], \mathbb{R}^d)$ une solution du problème de Cauchy* (1.1) *et $y_\tau \in \mathcal{C}^1([0, T], \mathbb{R}^d)$ une solution du problème de Cauchy*

$$\begin{cases} y_\tau'(t) = F_\tau(t, y_\tau(t)) \quad t \in [0, T] \\ y_\tau(0) = y_0, \end{cases}$$

où $F_\tau : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ vérifie les deux hypothèses suivantes :

(i) ("stabilité") *il existe une constante* $K \geq 0$ *telle que*

$$\forall (t, y_1, y_2) \in [0,T] \times \mathbb{R}^d \times \mathbb{R}^d, \quad \|F_\tau(t, y_1) - F_\tau(t, y_2)\| \leq K \|y_1 - y_2\|.$$

(ii) ("consistance à l'ordre $k \geq 1$") : *il existe une constante* $C \geq 0$ *telle que*

$$\|F - F_\tau\|_\infty \leq C\tau^k.$$

Alors

$$\forall t \in [0,T], \quad \|y(t) - y_\tau(t)\| \leq CT \exp(KT)\tau^k.$$

Démonstration. On pose $E(t) = y(t) - y_\tau(t)$ et on cherche à majorer $\|E(t)\|$. Pour contrôler $\|E(t)\|$, on commence par majorer $\|E'(t)\|$ en fonction de $\|E(t)\|$:

$$\begin{aligned}
\|E'(t)\| &= \|y'(t) - y'_\tau(t)\| \\
&= \|F(t, y(t)) - F_\tau(t, y_\tau(t))\| \\
&\leq \|F(t, y(t)) - F_\tau(t, y(t))\| + \|F_\tau(t, y(t)) - F_\tau(t, y_\tau(t))\| \\
&\leq C\tau^k + K \|y(t) - y_\tau(t)\|
\end{aligned}$$

où l'on a utilisé l'inégalité triangulaire pour la première inégalité et la stabilité et la consistance à l'ordre k pour la seconde inégalité. Ainsi,

$$\|E'(t)\| \leq K \|E(t)\| + C\tau^k.$$

Si l'on pose $e(t) = \|E(t)\|$, on a en utilisant l'inégalité précédente

$$\begin{aligned}
e(t) - e(0) \leq \|E(t) - E(0)\| &= \left\| \int_0^t E'(s) \mathrm{d}s \right\| \\
&\leq \int_0^t \|E'(s)\| \, \mathrm{d}s \\
&\leq \int_0^t K \|E(s)\| + C\tau^k \mathrm{d}s \\
&= C\tau^k t + K \int_0^t e(s) \mathrm{d}s.
\end{aligned}$$

En posant $\alpha(t) = C\tau^k t$, et en remarquant que $e(0) = 0$, cette inégalité se réécrit

$$e(t) \leq \alpha(t) + K \int_0^t e(s) \mathrm{d}s.$$

Comme α est croissante, le lemme de Gronwall (lemme 1.2) permet donc de conclure

$$e(t) \leq \exp(Kt)\alpha(t) \leq CT \exp(KT)\tau^k. \qquad \square$$

Lemme 1.2 (Gronwall, version intégrale)**.** *Soit* e, α *deux fonctions continues sur* $[0,T]$*,* α *croissante, vérifiant*

$$e(t) \leq \alpha(t) + K \int_0^t e(s) \mathrm{d}s.$$

Alors,

$$e(t) \leq \exp(Kt)\alpha(t).$$

Démonstration. Posons $f(t) = \exp(-Kt)\int_0^t e(s)\mathrm{d}s$, de sorte que

$$f'(t) = -K\exp(-Kt)\int_0^t e(s)\mathrm{d}s + \exp(-Kt)e(t) = -Kf(t) + \exp(-Kt)e(t)$$

L'hypothèse donne

$$\exp(-Kt)e(t) \leq \exp(-Kt)\alpha(t) + Kf(t),$$

Les deux équations précédentes combinées nous donnent

$$f'(t) \leq \exp(-Kt)\alpha(t),$$

soit encore, en utilisant $f(0) = 0$ et la croissance de la fonction α,

$$\begin{aligned} f(t) &= f(0) + \int_0^t f'(s)\mathrm{d}s \\ &\leq \int_0^t \exp(-Ks)\alpha(s)\mathrm{d}s \\ &\leq \alpha(t)\left[\frac{\exp(-Ks)}{-K}\right]_{s=0}^{s=t} \\ &= \frac{\alpha(t)}{K}(1-\exp(-Kt)) \end{aligned}$$

On en déduit finalement, en réutilisant à nouveau l'hypothèse, que

$$\begin{aligned} e(t) &\leq \alpha(t) + K\int_0^t e(s)\mathrm{d}s \\ &= \alpha(t) + K\exp(Kt)f(t) \\ &\leq \alpha(t) + K\exp(Kt)\frac{\alpha(t)}{K}(1-\exp(-Kt)) \\ &\leq \exp(Kt)\alpha(t). \end{aligned}$$

□

1.2 Approximation numérique

On considère maintenant un *schéma à un pas* pour approcher l'équation différentielle. Pour un certain pas de temps $\tau > 0$, on se donne une application $F_\tau : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ et on définit par récurrence une suite $(y_\tau^n)_{n\geq 0}$

$$\begin{cases} y_\tau^0 = y_0 \\ y_\tau^{n+1} = y_\tau^n + \tau F_\tau(t^n, y_\tau^n). \end{cases} \tag{1.3}$$

La proposition suivante permet de quantifier la distance entre la "solution approchée" y_τ^n et la "solution exacte" $y(t^n)$ du système différentiel (1.1), où $t^n = n\tau$.

Proposition 1.3 (Convergence du schéma)**.** *Soit y la solution du système différentiel* (1.1) *et soit $(y^n)_{n\geq 0}$ la suite définie par* (1.3)*, où la fonction F_τ vérifie*

(i) ("stabilité") : *il existe une constante* $K \geq 0$ *telle que*

$$\forall (t, y_1, y_2) \in [0,T] \times (\mathbb{R}^d)^2, \|F_\tau(t,y_1) - F_\tau(t,y_2)\| \leq K \|y_1 - y_2\| \tag{1.4}$$

(ii) ("consistance à l'ordre $k \geq 1$") : *il existe* $C \geq 0$ *telle que*

$$\forall t \in [0, T-\tau], \ \left\| \frac{y(t+\tau) - y(t)}{\tau} - F_\tau(t, y(t)) \right\| \leq C\tau^k. \tag{1.5}$$

Alors,

$$\max_{0 \leq n \leq \frac{T}{\tau}} \|y^n - y(t^n)\| \leq CT \exp(KT) \tau^k. \tag{1.6}$$

Remarque 1.1. Noter que la notion d'ordre dépend de la solution y : pour que le schéma puisse être d'ordre k, il faut en général que y soit de classe $\mathcal{C}^k$.

Remarque 1.2. En pratique, un schéma est défini pour une famille de pas de temps tendant vers zéro, par exemple $\tau \in \mathcal{T} := \{T/k \mid k \in \mathbb{N}^*\}$. Dans ce cas, pour avoir convergence du schéma, il faut que le second membre de la majoration (1.6) tende vers zéro lorsque $\tau \to 0$: pour cela, il suffit que les constantes K et C dans les hypothèses de stabilité et de consistance soient indépendantes de $\tau \in \mathcal{T}$. En revanche, ces constantes peuvent dépendre de la solution y à l'EDO (cf remarque précédente).

Démonstration. En s'inspirant de la preuve de stabilité, on pose $t^n = n\tau$ et on note $E^n = y(t^n) - y_\tau^n$, la différence entre la solution de l'EDO au temps t^n et la solution approchée y^n, et on pose $e^n = \|E^n\|$. Alors, en utilisant $\|A\| - \|B\| \leq \|A - B\|$,

$$\begin{aligned} e^{n+1} - e^n &= \left\| y(t^{n+1}) - y_\tau^{n+1} \right\| - \|y(t^n) - y_\tau^n\| \\ &\leq \left\| (y(t^{n+1}) - y_\tau^{n+1}) - y(t^n) - y_\tau^n \right\| \\ &= \left\| y(t^{n+1}) - (y(t^n) + \tau F_\tau(t^n, y_\tau^n)) \right\|, \end{aligned}$$

où l'on a utilisé l'équation (1.3) vérifiée par la suite (y_τ^n) pour obtenir la deuxième égalité. Ainsi,

$$e^{n+1} - e^n \leq \left\| y(t^{n+1}) - (y(t^n) + \tau F_\tau(t^n, y(t^n))) \right\| + \|\tau F_\tau(t^n, y(t^n)) - \tau F_\tau(t^n, y_\tau^n)\|.$$

L'hypothèse de consistance d'ordre k permet de majorer le premier terme par $C\tau^{k+1}$, tandis que la stabilité permet de majorer le second terme par $\tau K \|y(t^n) - y_\tau^n\| = \tau K e^n$. On obtient donc

$$\frac{e^{n+1} - e^n}{\tau} \leq K e^n + C\tau^k.$$

En utilisant le lemme de Gronwall discret (Lemme 1.4), avec $B = C\tau^k$, on a pour tout n tel que $t_n \leq T$ (soit $n\tau \leq T$),

$$E^n \leq (E^0 + n\tau B) \exp(n\tau K) \leq CT \exp(KT) \tau^k. \qquad \square$$

Lemme 1.4 (Gronwall discret)**.** *Soit* $(e^n)_{n \geq 0}$ *une suite de réels positifs vérifiant*

$$\frac{e^{n+1} - e^n}{\tau} \leq K e^n + B$$

Alors,

$$e^n \leq (e^0 + n\tau B) \exp(n\tau K).$$

Démonstration. L'inégalité peut être mise sous la forme

$$e^{n+1} \leq (1+\tau K)e^n + \tau B.$$

En divisant cette inégalité par $(1+\tau K)^{n+1}$, on a

$$\begin{aligned}(1+\tau K)^{-(n+1)}e^{n+1} &\leq (1+\tau K)^{-n}e^n + \tau B(1+\tau K)^{-(n+1)} \\ &\leq (1+\tau K)^{-n}e^n + \tau B.\end{aligned}$$

En sommant cette inégalité de 0 à $n-1$, et par téléscopage, on a

$$(1+\tau K)^{-n}e^n \leq e^0 + n\tau B$$

soit en utilisant l'inégalité de convexité $1+x \leq \exp(x)$,

$$e^n \leq (e^0 + n\tau B)(1+\tau K)^n \leq (e^0 + n\tau B)\exp(n\tau K). \qquad \square$$

Exemple 1.1 (Schéma d'Euler explicite). Soit $F : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ une fonction continue et vérifiant (1.2). Le schéma d'Euler implicite pour résoudre le système (1.1) est défini de la manière suivante : pour $\tau > 0$, on construit une suite (y_τ^n) par

$$\begin{cases} y_\tau^0 = y_0 \\ \frac{y_\tau^{n+1} - y_\tau^n}{\tau} = F(t^n, y_\tau^n), \end{cases} \tag{1.7}$$

où $t^n = n\tau$. En d'autres termes,

$$y_\tau^{n+1} = y_\tau^n + \tau F(t^n, y_\tau^n). \tag{1.8}$$

Le schéma est donc de la forme (1.3), avec $F_\tau = F_\tau^{\exp} := F$. On peut noter que la première expression (1.7) peut être vue comme l'approximation de $y'(t)$ par

$$\frac{y(t^{n+1}) - y(t^n)}{t^{n+1} - t^n} \simeq y'(t^n) = F(t^n, y(t^n)),$$

tandis que la deuxième expression (1.8) peut être comprise en se rappelant que la solution y du système (1.1) vérifie

$$\begin{aligned} y(t^{n+1}) &= y(t^n) + \int_{t^n}^{t^{n+1}} y'(t)\mathrm{d}t \\ &= y(t^n) + \int_{t^n}^{t^{n+1}} F(t, y(t))\mathrm{d}t \\ &\simeq y(t^n) + (t^{n+1} - t^n)F(t^n, y(t^n)), \end{aligned}$$

où l'intégrale est approchée en utilisant la méthode des rectangles à gauche. Ce deuxième point de vue est en général plus fécond, permettant d'interpréter la plupart des schémas classiques pour les équations différentiels.

Supposons maintenant que la fonction F soit Lipschitzienne en espace et Hölderienne en temps :

$$\|F(t,y) - F(s,z)\| \leq L(|s-t|^\alpha + \|y-z\|), \tag{1.9}$$

avec $\alpha \in (0,1]$. Alors, le schéma est stable, i.e. $F_\tau^{\exp}$ vérifie l'hypothèse (1.4) avec $K = L$. Montrons la consistance du schéma :

$$\begin{aligned}\|y(t+\tau) - (y(t) + \tau F_\tau^{\exp}(t, y(t)))\| &= \left\| \int_t^{t+\tau} F(s, y(s))\mathrm{d}s - \tau F_\tau^{\exp}(t, y(t)) \right\| \\ &\leq \int_t^{t+\tau} \|F(s, y(s)) - F_\tau^{\exp}(t, y(t))\| \, \mathrm{d}s \\ &\leq L \int_t^{t+\tau} (|s-t|^\alpha + \|y(s) - y(t)\|)\mathrm{d}s,\end{aligned}$$

où la deuxième inégalité vient de l'hypothèse que F est lipschitzienne en ses deux variables. On note $M = \max_{t\in[0,T]} F(t, y(t))$, de sorte que

$$\forall 0 \leq s \leq t \leq T, \quad \|y(t) - y(s)\| \leq \int_s^t \left\|y'(u)\right\| \mathrm{d}u \leq M \left|t - s\right|.$$

Ainsi,

$$\|y(t+\tau) - (y(t) + \tau F_\tau^{\exp}(t, y(t)))\| \leq L(\tau^{1+\alpha} + M\tau^2),$$

et le schéma d'Euler explicite est donc d'ordre $\min(\alpha, 1)$. Noter que la constante dans l'hypothèse de consistance (1.5) dépend de la solution, via la constante M.

1.3 Exercices

Exercice 1.1. *Lemme de Gronwall, version différentielle.* Soit $e \in \mathcal{C}^1([0,T])$ et $\beta \in \mathrm{L}^1([0,T])$ vérifiant l'inégalité $e'(t) \leq Ke(t) + \beta(t)$.

1. Démontrer que $e(t) \leq e^{Kt}\left(e(0) + \int_0^t e^{-Ks}\beta(s)\mathrm{d}s\right)$ pour tout $t \in [0,T]$.
   (Indication : poser $f(t) = e^{-Kt}e(t)$ et majorer f'.)
2. En déduire que si $\beta \geq 0$, alors $e(t) \leq e^{Kt}\left(e(0) + \int_0^t \beta(s)\mathrm{d}s\right)$.
3. Redémontrer le résultat de la deuxième question en utilisant le Lemme 1.2.

Exercice 1.2. *Schéma d'Euler implicite.* Soit $F : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ une fonction continue et vérifiant (1.2). Pour $\tau > 0$, une suite (y_τ^n) est construite par schéma d'Euler implicite de la manière suivante :

$$\begin{cases} y_\tau^0 = y_0 \\ \frac{y_\tau^{n+1} - y_\tau^n}{\tau} = F(t^{n+1}, y_\tau^{n+1}), \end{cases}$$

où $t^n = n\tau$. La seconde équation s'écrit aussi $y_\tau^{n+1} = y_\tau^n + \tau F(t^{n+1}, y_\tau^{n+1})$. Notons que l'existence (et le calcul) d'une suite vérifiant cette récurrence n'est pas évidente a priori, et nécessite la résolution d'un problème de point fixe.

1. [Existence] Soit $z \in \mathbb{R}^d$ et $t, \tau \in \mathbb{R}_+$ et soit $S_\tau : x \in \mathbb{R}^d \mapsto z + \tau F(t+\tau, x)$.
   (a) Montrer que la fonction S_τ est contractante si $\tau L < 1$.
   (b) En déduire que, si $\tau L < 1$, pour tout $(t, z) \in \mathbb{R}\times\mathbb{R}^d$, il existe un unique point noté $X_\tau^{\mathrm{imp}}(t, z)$ vérifiant la relation

$$X_\tau^{\mathrm{imp}}(t, z) = z + \tau F(t+\tau, X_\tau^{\mathrm{imp}}(t, z)).$$

(c) Montrer que $\left\|X_\tau^{\text{imp}}(t,z_1) - X_\tau^{\text{imp}}(t,z_2)\right\| \leq 1/(1-\tau L)\,\|z_1 - z_2\|$

On suppose dorénavant que $\tau < 1/L$. Le schéma d'Euler implicite est alors défini par $y^{n+1} = X_\tau^{\text{imp}}(t^n, y_\tau^n)$, ou, en posant $F_\tau^{\text{imp}}(t,x) = F(t, X_\tau^{\text{imp}}(t,x))$,

$$y_\tau^{n+1} = y_\tau^n + \tau F_\tau^{\text{imp}}(t^n, y_\tau^n),$$

qui est de la forme attendue (1.3).

2. [Stabilité du schéma] Déduire de la question précédente que le schéma est stable, i.e. que la fonction F_τ^{imp} vérifie (1.4).
3. [Consistance] Montrons la consistance du schéma d'Euler implicite. Pour cela, considérons $y \in \mathcal{C}^1([0,T])$ la solution du système (1.1) et posons

$$N := \sup_{t\in[0,T-\tau]} \left\|F(t+\tau, X_\tau^{\text{imp}}(t, y(t)))\right\|.$$

(a) Montrer que $N < +\infty$ et que $\forall t \in [0, T-\tau]$, $\left\|X_\tau^{\text{imp}}(t,y(t)) - y(t)\right\| \leq N\tau$.

(b) On suppose maintenant que F vérifie l'hypothèse (1.9). Déduire des questions précédentes que

$$\begin{aligned}\left\|y(t) + \tau F_\tau^{\text{imp}}(t,y(t)) - y(t+\tau)\right\| \leq \|(y(t) + \tau F(t,y(t))) - y(t+\tau)\| \\ + L\tau(\tau^\alpha + N\tau)\end{aligned}$$

puis que le schéma d'Euler implicite est consistant d'ordre $\min(1,\alpha)$.

Exercice 1.3. *Schéma du point du milieu.* Soit $F : \mathbb{R}\times\mathbb{R}^d \to \mathbb{R}^d$ continue, vérifiant (1.2) et $\|F\|_\infty \leq M$. Le schéma du point du milieu pour un problème de Cauchy de la forme (1.1) est défini par $y_\tau^{n+1} = y_\tau^n + \tau F_\tau^{\text{mp}}(t^n, y_\tau^n)$

$$F_\tau^{\text{mp}}(t,y) = F\left(t + \frac{\tau}{2}, y + \frac{\tau}{2}F(t,y)\right)$$

1. Montrer que le schéma est stable et majorer la constante K dans (1.4).
2. Montrer que le schéma est consistant d'ordre 2 si F est de classe $\mathcal{C}^2$.

1.4 Correction des exercices

Correction de l'exercice 1.*. 1. Si $f(t) = e^{-Kt}e(t)$, alors*

$$f'(t) = -Ke^{-Kt}e(t) + e^{-Kt}e'(t) \leq e^{-Kt}\beta(t),$$

où l'inégalité vient de l'hypothèse. Ainsi,

$$f(t) = f(0) + \int_0^t f'(s)\mathrm{d}s \leq f(0) + \int_0^t e^{-Ks}\beta(s)\mathrm{d}s.$$

En multipliant cette inégalité par e^{Kt} on trouve

$$e(t) \leq e^{Kt}e(0) + e^{Kt}\int_0^t e^{-Ks}\beta(s)\mathrm{d}s.$$

2. Si $\beta \geq 0$ et $s \geq 0$, alors $e^{-Ks}\beta(s) \leq \beta(s)$, soit

$$\int_0^t e^{-Ks}\beta(s)\mathrm{d}s \leq \int_0^t \beta(s)\mathrm{d}.s$$

3. On commence par intégrer l'inégalité dans l'hypothèse :

$$e(t) - e(0) = \int_0^t e'(s)\mathrm{d}s \leq \int_0^t Ke(s) + \beta(s)\mathrm{d}s,$$

ce qui donne

$$e(t) \leq e(0) + \int_0^t K\beta(s)\mathrm{d}s + K\int_0^t e(s)\mathrm{d}s = \alpha(t) + K\int_0^t e(s)\mathrm{d}s,$$

où $\alpha(t) = e(0) + \int_0^s \beta(s)\mathrm{d}s$ est croissante. Le lemme de Gronwall intégral (lemme 1.2) nous donne

$$e(t) \leq e^{Kt}\alpha(t) = e^{Kt}(e(0) + \int_0^s \beta(s)\mathrm{d}s).$$

***Correction de l'exercice 2.**. 1.a. On majore la constante de Lipschitz de S_τ en utilisant le fait que F est L-Lipschitzienne en sa variable d'espace :*

$$\begin{aligned}\left\|S_\tau(x) - S_\tau(x')\right\| &= \left\|z + \tau F(t+\tau, x) - (z + \tau F(t+\tau, x'))\right\| \\ &= \tau\left\|F(t+\tau, x) - F(t+\tau, x')\right\| \\ &\leq \tau L\left\|x - x'\right\|\end{aligned}$$

Ainsi, si $\tau L < 1$, la fonction S_τ est contractante et par théorème du point fixe, elle admet un unique point fixe.
1.b. On note $X_\tau^{\mathrm{imp}}(t,z)$ l'unique point fixe de S_τ (si $\tau L < 1$). Il est caractérisé par

$$X_\tau^{\mathrm{imp}}(t,z) = S_\tau(X_\tau^{\mathrm{imp}}(t,z)) = z + \tau F(t+\tau, X_\tau^{\mathrm{imp}}(t,z)).$$

1.c. Par définition de X_τ^{imp}, on a

$$X_\tau^{\mathrm{imp}}(t,z_i) = z_i + \tau F(t+\tau, X_\tau^{\mathrm{imp}}(t,z_i))$$

Ainsi, en soustrayant l'égalité pour $i = 2$ à celle pour $i = 1$, et en utilisant l'inégalité triangulaire,

$$\begin{aligned}&\left\|X_\tau^{\mathrm{imp}}(t,z_1) - X_\tau^{\mathrm{imp}}(t,z_2)\right\| \\ &\quad\leq \|z_1 - z_2\| + \tau\left\|F(t+\tau, X_\tau^{\mathrm{imp}}(t,z_1)) - F(t+\tau, X_\tau^{\mathrm{imp}}(t,z_2))\right\| \\ &\quad\leq \|z_1 - z_2\| + \tau L\left\|X_\tau^{\mathrm{imp}}(t,z_1) - X_\tau^{\mathrm{imp}}(t,z_2)\right\|\end{aligned}$$

Ainsi,

$$(1 - \tau L)\left\|X_\tau^{\mathrm{imp}}(t,z_1) - X_\tau^{\mathrm{imp}}(t,z_2)\right\| \leq \|z_1 - z_2\|.$$

On en déduit l'inégalité souhaitée en divisant par $1 - \tau L$, lorsque $0 < \tau L < 1$.

2. Pour montrer la stabilité du schéma, il faut montrer que la fonction F_τ^{imp} est Lipschitzienne en sa variable d'espace :

$$\begin{aligned}\left\|F_\tau^{\text{imp}}(t,x_1) - F_\tau^{\text{imp}}(t,x_2)\right\| &= \left\|F(t, X_\tau^{\text{imp}}(t,x_1)) - F(t, X_\tau^{\text{imp}}(t,x_2))\right\| \\ &\leq L\left\|X_\tau^{\text{imp}}(t,x_1) - X_\tau^{\text{imp}}(t,x_2)\right\| \\ &\leq \frac{L}{1-\tau L}\left\|x_1 - x_2\right\|,\end{aligned}$$

où l'on a utilisé la lipschitzité de F (hypothèse) pour la première inégalité, et celle de X_τ^{imp} (Q1.c) pour la seconde inégalité. Ainsi, le schéma est stable (i.e. vérifie l'inégalité (1.4)) pour $K = L/(1-\tau L)$.
3.a La constante $N := \sup_{t\in[0,T-\tau]} \left\|F(t+\tau, X_\tau^{\text{imp}}(t,y(t)))\right\|$ est bornée comme suprémum de la fonction continue

$$t \mapsto F(t+\tau, X_\tau^{\text{imp}}(y(t))),$$

composée des fonctions continues F, X_τ^{imp} et y, sur le segment (compact) $[0, T-\tau]$. Par définition de X_τ^{imp},

$$X_\tau^{\text{imp}}(t,y(t)) = y(t) + \tau F(t+\tau, X_\tau^{\text{imp}}(t,y(t))),$$

de sorte que pour $t \in [0, T-\tau]$,

$$\left\|X_\tau^{\text{imp}}(t,y(t)) - y(t)\right\| = \tau\left\|F(t+\tau, X_\tau^{\text{imp}}(t,y(t)))\right\| \leq N\tau,$$

par définition de N.
3.b. Montrons que le schéma a au moins le même ordre de consistance que le schéma d'Euler explicite $F_\tau^{\text{exp}} = F$:

$$\begin{aligned}&\left\|y(t) + \tau F_\tau^{\text{imp}}(t,y(t)) - y(t+\tau)\right\| \\ &\leq \left\|y(t) + \tau F_\tau^{\text{exp}}(t,y(t)) - y(t+\tau)\right\| + \tau\left\|F_\tau^{\text{imp}}(t,y(t)) - F_\tau^{\text{exp}}(t,y(t))\right\|.\end{aligned}$$

De plus,

$$\begin{aligned}\left\|F_\tau^{\text{imp}}(t,y(t)) - F_\tau^{\text{exp}}(t,y(t))\right\| &= \left\|F(t+\tau, X_\tau^{\text{imp}}(t,y(t)) - F(t,y(t))\right\| \\ &\leq L(|t+\tau-t|^\alpha + \left\|X_\tau^{\text{imp}}(t,y(t)) - y(t)\right\| \\ &\leq L(\tau^\alpha + N\tau),\end{aligned}$$

où l'on a utilisé, dans l'ordre, la définition des schémas, l'hypothèse sur F (1.9), et la question Q3.a pour la dernière inégalité. D'autre part, l'estimation de consistance d'ordre 1 du schéma d'Euler explicite (Exemple 1.1) nous donne

$$\left\|y(t+\tau) - (y(t) + \tau F_\tau^{\text{exp}}(t,y(t)))\right\| \leq L(\tau^{1+\alpha} + M\tau^2)$$

En combinant les inégalités précédentes, on trouve

$$\left\|y(t+\tau) - (y(t) + \tau F_\tau^{\text{imp}}(t,y(t)))\right\| \leq L(2\tau^{1+\alpha} + (M+N)\tau^2),$$

le schéma d'Euler implicite est donc également d'ordre $\min(1,\alpha)$.

Chapitre 2

Méthode des différences finies pour l'équation de la chaleur

L'objectif de ce chapitre et des chapitres suivants est d'introduire la méthode des différences finies pour les équations d'évolution : équation de la chaleur et équation de transport. On commencera par l'équation de la chaleur, principalement en dimension 1 d'espace. Cette équation modélise l'évolution de la température, notée u, dans un domaine $\Omega \subseteq \mathbb{R}^d$ en fonction de la position d'espace $x \in \Omega$ et du temps $t \in [0, T]$.

Notations Le domaine spatial est un ouvert borné Ω de $\mathbb{R}^d$. Étant donnée une fonction u sur $[0, T] \times \overline{\Omega}$, on note $\partial_t u$ dénote la dérivée partielle par rapport au temps et $\partial_i u$ la dérivée partielle par rapport à la $i$ème coordonnée d'espace, $\partial_{ij} u$ la dérivée partielle seconde, etc. On considèrera l'espace des fonctions admettant une dérivée partielle continue en temps et des dérivées spatiales secondes continues :

$$\mathcal{C}_1^2(]0, T[\times\Omega) = \{u \in \mathcal{C}^0(]0, T[\times\Omega) \mid \partial_i u, \partial_{ij} u \in \mathcal{C}^0(]0, T[\times\Omega) \text{ pour tout } i, j \in [\![1, d]\!]\},$$

On utilisera les notations usuelles suivantes pour le gradient et la matrice hessienne

$$\nabla u(t, x) = (\partial_i u(t, x))_{1 \leq i \leq d} \in \mathbb{R}^d,$$

$$D^2 u(t, x) = (\partial_{ij} u(t, x))_{1 \leq i, j \leq d} \in \mathcal{M}_d(\mathbb{R}).$$

On rappelle enfin que le *laplacien* de u est défini par

$$\Delta u(t, x) = \mathrm{Tr}(\mathrm{D}^2 u(t, x)) = \sum_{1 \leq i \leq d} \partial_{ii} u(t, x).$$

Définition 2.1 (Équation de la chaleur)**.** Soit $f \in \mathcal{C}^2([0, T] \times \overline{\Omega})$ représentant une source de chaleur, et $u_0 \in \mathcal{C}^2(\overline{\Omega})$ représentant la distribution de température au temps initial $t = 0$. Une *solution (forte) à l'équation de la chaleur* est une fonction $u \in \mathcal{C}^0([0, T] \times \overline{\Omega}) \cap \mathcal{C}_1^2(]0, T[\times\Omega)$ vérifiant le système d'équations aux dérivées partielles

$$\begin{cases} \partial_t u - \Delta u = f & \text{dans }]0, T[\times\Omega \\ u = 0 & \text{sur } [0, T] \times \partial\Omega \\ u = u_0 & \text{sur } \{0\} \times \overline{\Omega} \end{cases} \tag{2.1}$$

Remarque 2.1 (Modélisation). L'équation (2.2) peut modéliser (entre autres) l'évolution de la chaleur dans un domaine Ω. Décrivons brièvement chacun des termes de l'équation :

- Lorsque $f = 0$, la première équation se réécrit

$$\partial_t u = \Delta u$$

ce qu'on appelle une *équation de diffusion*. Afin de comprendre le comportement de cette équation considérons l'évolution de la quantité totale de chaleur contenue dans un sous-ensemble compact $A \subseteq \Omega$, dont le bord est régulier. Une suite de calcul classique (qu'il n'est pas nécessaire de bien comprendre pour suivre le reste du chapitre) donne

$$\begin{aligned}\frac{d}{dt}\int_A u(t,x)\mathrm{d}x &= \int_A \partial_t u(t,x)\mathrm{d}x \\ &= \int_A \Delta u(t,x)\mathrm{d}x \\ &= \int_A \mathrm{div}(\nabla u(t,x))\mathrm{d}x \\ &= \int_{\partial A} \langle \nabla u(t,x) | n_A(x)\rangle \mathrm{d}x,\end{aligned}$$

où l'on a utilisé la formule $\Delta u = \mathrm{div}(\nabla u)$ pour l'avant dernière inégalité et la formule de la divergence pour obtenir la dernière égalité. Ainsi, la chaleur s'échappe de A par les points $x \in \partial A$ tels que $\langle \nabla u(t,x) | n_A(x)\rangle < 0$, ce qui correspond intuitivemnt au fait qu'au voisinage de x les points à l'intérieur de A sont "plus chauds" que ceux à l'extérieur de A. Autrement dit, l'équation de la chaleur tend à homogéniser la température.
- Le terme f dans le second membre de la première équation correspond à un apport externe de chaleur (par exemple un radiateur).
- La condition initiale $u(0,\cdot) = u_0$ décrit simplement la distribution de température au temps initial.
- La seconde équation du système est la plus inhabituelle : elle modélise une "condition au bord", prescrivant la distribution de la température au bord du domaine $\partial\Omega$. La condition $u \equiv 0$ sur $\partial\Omega$ est appelée condition de Dirichlet homogène. Mathématiquement, cette condition garantit l'unicité des solutions à l'équation de la chaleur. En effet, soit u est solution du système auquel on a enlevé la seconde équation,

$$\begin{cases}\partial_t u - \Delta u = f & \text{dans }]0,T[\times\Omega \\ u = f & \text{sur } \{0\}\times\Omega\end{cases} \tag{2.2}$$

Alors pour toute fonction harmonique v sur Ω (c'est-à-dire vérifiant $\Delta v = 0$), la fonction $u+v$ est également solution du système. Les conditions de Dirichlet suppriment cette invariance.

2.1 Principe du maximum et stabilité des solutions régulières

On commence par montrer le principe (faible) du maximum pour des solutions suffisamment régulières de l'équation (2.2). On en déduira un résultat de stabilité, montrant que la solution de (2.2), si elle existe, dépend continûment de la donnée initiale u_0 et de la source f. On verra dans §2.3 une version "discrète" de ces résultats.

Proposition 2.1 (Principe du maximum). *Soit* $u \in \mathcal{C}^0([0,T]\times\overline{\Omega}) \cap \mathcal{C}^2_1(]0,T[\times\Omega)$ *vérifiant*

$$\partial_t u - \Delta u \geq 0 \text{ sur }]0,T[\times\Omega.$$

Alors, en notant $\Gamma = ([0,T]\times\partial\Omega)\cup\{0\}\times\overline{\Omega}$,

$$\min_{[0,T]\times\overline{\Omega}} u \geq \min_{\Gamma} u.$$

Démonstration. On notera $\Omega^T =]0,T[\times\Omega$ et $\Gamma^T = ([0,T]\times\partial\Omega)\cup\{0\}\times\overline{\Omega}$.
Étape 1. On suppose dans un premier temps que $\partial_t u - \Delta u > 0$ sur Ω^T, et on choisit $T' \in [0,T[$. On considère (t_0,x_0) un minimiseur de u sur l'ensemble compact $\overline{\Omega}^{T'}$.

Montrons par l'absurde que le minimiseur (t_0,x_0) de u ne peut pas appartenir à l'intérieur $\Omega^{T'}$ de $\overline{\Omega^{T'}}$. Supposons donc que $(t_0,x_0)\in\Omega^{T'}$ est un minimiseur de u sur l'ouvert $\Omega^{T'}$, ce qui implique que toutes ses dérivées partielles s'annulent :

$$\begin{cases}\partial_t u(t_0,x_0) = 0\\ \nabla u(t_0,x_0) = 0.\end{cases}$$

L'hypothèse $\partial_t u(t_0,x_0) - \Delta u(t_0,x_0) > 0$, nous permet d'en déduire que

$$\Delta u(t_0,x_0) = \mathrm{Tr}(D^2u(t_0,x_0)) < 0.$$

La matrice symétrique $D^2u(t_0,x_0)$ est diagonalisable en base orthonormée, et l'inégalité précédente montre que la somme des valeurs propres de strictement négative. Il existe donc un vecteur $v\in\mathbb{R}^d$ tel que

$$\langle \mathrm{D}^2u(t_0,x_0)v|v\rangle < 0.$$

Par développement de Taylor de u en (t_0,x_0) dans la direction v, on en déduit

$$u(t_0,x_0+\varepsilon v) = u(t_0,x_0) + \varepsilon\underbrace{\langle\nabla u(t_0,x_0)|v\rangle}_{=0} + \frac{\varepsilon^2}{2}\underbrace{\langle D^2u(t_0,x_0)v|v\rangle}_{<0} + \mathrm{o}(\varepsilon^2).$$

Ceci contredit le fait que (t_0,x_0) soit un minimiseur de u sur $[0,T']\times\overline{\Omega}$. Ainsi, tout minimiseur (t_0,x_0) de u doit apparentir au bord $\partial\Omega^{T'}$ du domaine spatio-temporel.

Supposons maintenant que le minimum de u soit atteint en $(t_0,x_0)\in\partial\Omega^{T'}\setminus\Gamma^{T'}$, ce qui signifie simplement que $t_0 = T'$ et $x_0\in\overline{\Omega}$ (faire un dessin du domaine). Pour

tout $\varepsilon \in [0, T']$, le point $(T' - \varepsilon, x_0)$ appartient à $\Omega^{T'}$. La minimalité de (x_0, T) nous donne $u(T - \varepsilon, x_0) \geq u(T, x_0)$. Ainsi,

$$\forall \varepsilon \in [0, T], u(t_0, x_0) \leq u(t_0 - \varepsilon, x_0) = u(t_0, x_0) - \varepsilon \partial_t u(t_0, x_0) + \mathrm{o}(\varepsilon).$$

On en déduit que $\partial_t u(t_0, x_0) \leq 0$. On peut alors suivre ligne à ligne le raisonnement du paragraphe précédent en remplaçant $\partial_t u(t_0, x_0) = 0$ par $\partial_t u(t_0, x_0) \leq 0$ pour en déduire une absurdité.

On vient donc de démontrer que le minimiseur (t_0, x_0) de u sur $\overline{\Omega}_{T'}$ appartient à l'ensemble $\Gamma^{T'}$, d'où l'on déduit

$$\min_{\overline{\Omega}'_T} u \geq \min_{\Gamma^{T'}} u,$$

et on obtient la conclusion souhaitée en faisant tendre T' vers T.

Étape 2. On suppose maintenant seulement $\partial_t u - \Delta u \geq 0$. On introduit la fonction $u_\varepsilon(t, x) = u(t, x) + \varepsilon t$. Alors,

$$\partial_t u_\varepsilon(t, x) - \Delta u_\varepsilon(t, x) = \varepsilon + \partial_t u(t, x) - \Delta u(t, x) \geq \varepsilon > 0.$$

Ainsi,

$$\min_{\overline{\Omega}_T} u_\varepsilon \geq \min_{\Gamma_T} u_\varepsilon,$$

et on obtient l'inégalité souhaitée en passant à la limite lorsque $\varepsilon \to 0$. □

Corollaire 2.2 (Stabilité en $\|.\|_\infty$). *Soit $u \in \mathcal{C}^0([0, T] \times \overline{\Omega}) \cap \mathcal{C}^2_1(]0, T[\times\Omega)$ vérifiant*

$$\begin{cases} \partial_t u - \Delta u = f & \text{dans }]0, T[\times\Omega \\ u = 0 & \text{sur } [0, T] \times \partial\Omega \\ u = u_0 & \text{sur } \{0\} \times \Omega \end{cases}$$

Alors,

$$\|u\|_\infty \leq \|u_0\|_\infty + T \|f\|_\infty \tag{2.3}$$

Démonstration. L'idée est d'ajouter une fonction à u qui lui fait vérifier les hypothèses de la proposition précédente. Pour cela, on considère $v(t, x) = u(t, x) + \|f\|_\infty t$. Cette fonction vérifie

$$\partial_t v - \Delta v = \|f\|_\infty + \partial_t u - \Delta u = \|f\|_\infty + f \geq 0.$$

La proposition précédente permet d'en conclude que $v \geq \min_{\Gamma^T} v$. Notons que $\Gamma^T = A \cup B$ où $A = [0, T] \times \partial\Omega$ et $B = \{0\} \times \Omega$. Comme $u \equiv 0$ sur A, $v \equiv t \|f\|_\infty$ sur A, et

$$\min_A v = 0.$$

D'autre part, en utilisant $v(0, \cdot) = u(0, \cdot) = u_0$,

$$\min_B v = \min_B u_0 \geq - \|u_0\|_\infty .$$

Ainsi, $v \geq -\|u_0\|_\infty$, et comme $v(t,x) = u(t,x) + t\|f\|_\infty$, on en déduit que

$$\min_{\overline{\Omega}_T} u \geq -\|u_0\|_\infty - T\|f\|_\infty .$$

En faisant le même raisonnement avec $v(t,x) = \|f\|_\infty t - u(t,x)$, qui vérifie

$$\partial_t v - \Delta v = \|f\|_\infty - f \geq 0,$$

on obtient

$$-\max_{\overline{\Omega}_T} u = \min_{\overline{\Omega}_T} -u \geq -\|u_0\|_\infty - T\|f\|_\infty ,$$

ce qui permet de conclure. □

Corollaire 2.3 (Stabilité en $\|.\|_\infty$, bis)**.** *Soient* $u^1, u^2 \in \mathcal{C}^0([0,T]\times\overline{\Omega}) \cap \mathcal{C}^2_1(]0,T[\times\Omega)$ *vérifiant*

$$\begin{cases} \partial_t u^i - \Delta u^i = f^i & \text{dans }]0,T[\times\Omega \\ u^i = 0 & \text{sur } [0,T]\times\partial\Omega \\ u^i = u_0 & \text{sur } \{0\}\times\Omega \end{cases}$$

Alors,

$$\|u^1 - u^2\|_\infty \leq \|u_0^1 - u_0^2\|_\infty + T\|f^1 - f^2\|_\infty \tag{2.4}$$

Démonstration. Considérer $u = u^1 - u^2$ dans le corollaire précédent. □

2.2 Discrétisation par différences finies en dimension 1

Pour transformer l'équation aux dérivées partielles de la chaleur (2.2) en un problème de dimension finie, il faut discrétiser le domaine spatio-temporel $]0,T]\times\Omega$. Dans la suite, on suppose pour simplifier que $d = 1$ et $\Omega =]a,b[$. On utilisera les notations suivantes :

- $N \in \mathbb{N}^*$ le nombre de pas de temps, $\tau = T/N$ le pas de temps, $t^n = n\tau$ et

$$\overline{I}_\tau = \{t^n \mid n \in [\![0, N]\!]\}.$$

- $M \in \mathbb{N}^*$ le nombre de pas d'espace, $h = (b-a)/(M+1)$ le pas d'espace, $x_j = a + jh$ et

$$\begin{aligned} \overline{\Omega}_h &= \{x_j \mid j \in [\![0, M+1]\!]\} \\ \Omega_h &= \{x_j \mid j \in [\![1, M]\!]\}, \\ \partial\Omega_h &= \{a, b\} \end{aligned}$$

  Les ensembles $\partial\Omega^h$ et Ω^h doivent être compris comme le "bord" et "l'intérieur" du domaine spatial discret $\overline{\Omega}_h$.
- $F(X)$ désignera l'ensemble des fonctions à valeurs réelles définies sur l'ensemble X. En particulier, $F(\overline{I}_\tau \times \overline{\Omega}_h)$ désigne l'espace de dimension fini dans lequel on construira les solutions discrétisées de l'équation de la chaleur.

Remarque 2.2. L'espace de fonctions $F(\overline{I}_\tau \times \overline{\Omega}_h)$ est de dimension $(N+1)\times(M+2)$.

Définition 2.2 (Laplacien discret)**.** Le laplacien discret d'une fonction $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ en un point $(t,x) \in \overline{I}_\tau \times \Omega_h$ est défini par

$$\Delta_h v(t,x) = \frac{v(t,x+h) - 2v(t,x) + v(t,x-h)}{h^2}.$$

Remarque 2.3. Nous montrerons plus tard que si v est suffisamment régulière,

$$\frac{v(t,x+h) - 2v(t,x) + v(t,x-h)}{h^2} = \partial_{xx} v(t,x) + \mathrm{O}(h^2),$$

ce qui justifiera cette définition du laplacien discret.

Schéma d'Euler explicite Nous sommes maintenant en mesure de décrire deux schémas pour l'équation de la chaleur. Nous commençons par le schéma d'Euler *explicite*, qui consiste à approcher le système d'équations (2.2) en approchant le laplacien Δv par $\Delta_h v$ et la dérivée temporelle par

$$\partial_t v(t,x) = \frac{v(t+\tau,x) - v(t,x)}{\tau} + \mathrm{O}(\tau).$$

Une fonction $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ est solution du schéma d'Euler exlicite si

$$\begin{cases} \dfrac{v(t+\tau,x) - v(t,x)}{\tau} - (\Delta_h v)(t,x) = f(t,x) & \text{pour } (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ v = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \\ v = u_0 & \text{sur } \{0\} \times \overline{\Omega}_h \end{cases} \tag{2.5}$$

Noter que ce système d'équations permet de construire par récurrence sur le pas de temps la solution v : on commence par $v(t^0,\cdot) = u_0$, puis on définit par récurrence sur $n \in [\![0, N]\!]$, $v(t^n,\cdot)$ via

$$v(t^{n+1},x) = v(t^n,x) + \tau \frac{v(t^n,x+h) - 2v(t^n,x) + v(t^n,x-h)}{h^2} + \tau f(t^n,x)$$

L'existence et l'unicité des solutions au système (2.5) en découle directement.

Schéma d'Euler implicite Nous décrivons maintenant le schéma d'Euler *implicite*. Ce schéma provient d'une autre discrétisation de la dérivée temporelle :

$$\partial_t v(t,x) = \frac{v(t,x) - v(t-\tau,x)}{\tau} + \mathrm{O}(\tau).$$

Une fonction $v \in F(\overline{I}_\tau \times \Omega_h)$ est solution du schéma d'Euler implicite si

$$\begin{cases} \dfrac{v(t+\tau,x) - v(t,x)}{\tau} - (\Delta_h v)(t+h,x) = f(t+h,x) & \text{pour } (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ v = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \\ v = u_0 & \text{sur } \{0\} \times \overline{\Omega}_h \end{cases} \tag{2.6}$$

À cause du caractère implicite, l'existence de solution à ce schéma n'est pas évident.

2.3 Stabilité des schémas d'Euler explicites et implicites

Dans cette section, nous montrons qu'il existe un analogue discret du corollaire 2.2 pour les schémas d'Euler implicite et explicite. En plus de son intérêt intrinsèque, ce résultat de stabilité est un des ingrédients cruciaux de la démonstration de convergence de ces schémas numériques.

Proposition 2.4 (Stabilité pour Euler explicite)**.** *Soit $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ une fonction vérifiant*

$$\begin{cases} \dfrac{v(t+\tau,x)-v(t,x)}{\tau} - (\Delta_h v)(t,x) = f(t,x) & \text{pour } (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ v = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \end{cases} \tag{2.7}$$

où $f \in F((\overline{I}_\tau \setminus \{T\}) \times \Omega_h)$. On suppose que les pas de temps et d'espace vérifient

$$\frac{\tau}{h^2} \leq \frac{1}{2}. \tag{2.8}$$

Alors,

$$\|v\|_\infty \leq \|v(0,\cdot)\|_\infty + T\,\|f\|_\infty .$$

Remarque 2.4. L'inégalité (2.8) est appelé condition CFL, pour Courant-Friedrichs, Lewy. Sans cette condition, le schéma d'Euler explicite pour l'équation de la chaleur est instable, donc en pratique inutilisable. En pratique, la condition (2.8) est très contraignante : elle nécessite que le nombre de pas de temps soit le carré du nombre de pas d'espace !

Démonstration. Par définition du schéma (2.5), pour tout $(t,x) \in \overline{I}_\tau \times \Omega_h$,

$$\begin{aligned} v(t+\tau,x) &= v(t,x) + \tau\Delta_h v(t,x) + \tau f(t,x) \\ &= v(t,x) + \frac{\tau}{h^2}(v(t,x-h) - 2v(t,x) + v(t,x+h)) + \tau f(t,x) \\ &= \left(1 - 2\frac{\tau}{h^2}\right) v(t,x) + \frac{\tau}{h^2} v(t,x-h) + \frac{\tau}{h^2} v(t,x+h) + \tau f(t,x) \end{aligned}$$

L'hypothèse $\frac{\tau}{h^2} \leq \frac{1}{2}$ garantit que les coefficients devant $v(t,x), v(t,x-h)$ et $v(t,x+h)$ sont tous positifs, de sorte que

$$\begin{aligned} |v(t+\tau,x)| &\leq \left[(1 - 2\frac{\tau}{h^2}) + \frac{\tau}{h^2} + \frac{\tau}{h^2}\right] \|v(t,\cdot)\|_\infty + \tau\,|f(t,x)| \\ &= \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty \end{aligned}$$

Ainsi, $\|v(t+\tau,\cdot)\|_\infty \leq \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty$, et donc par récurrence, pour $n \leq N$,

$$\|v(n\tau,\cdot)\|_\infty \leq \|v(0,\cdot)\|_\infty + n\tau\,\|f\|_\infty \leq \|v(0,\cdot)\|_\infty + T\,\|f\|_\infty .$$

Le résultat s'en déduit directement. □

Proposition 2.5 (Stabilité pour Euler implicite). *Soit $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ une fonction vérifiant*

$$\begin{cases} \dfrac{v(t+\tau,x)-v(t,x)}{\tau} - (\Delta_h v)(t+h,x) = f(t+h,x) & \forall (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ v = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \end{cases} \tag{2.9}$$

où $f \in F((\overline{I}_\tau \setminus \{0\}) \times \Omega_h)$. Alors,

$$\|v\|_\infty \leq \|v(0,\cdot)\|_\infty + T\,\|f\|_\infty .$$

Remarque 2.5. Pour le schéma d'Euler implicite, il n'y a pas de condition CFL : on dit qu'il est *inconditionellement stable*. C'est un gros avantage sur le schéma d'Euler explicite : il permet de choisir des pas de temps plus grands ; le prix à payer est de devoir résoudre un système linéaire à chaque pas de temps (cf exercice 2.2).

Démonstration. Comme dans le cas d'Euler explicite, nous allons démontrer que

$$\|v(t+\tau,\cdot)\|_\infty \leq \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty .$$

Soit $t \in \overline{I}_\tau \setminus \{T\}$ et $x^* \in \overline{\Omega}_h$ un maximiseur de $v(t+\tau,\cdot)$, i.e.

$$v(t+\tau,x^*) = \max_{x\in\Omega_h} v(t+\tau,x).$$

Si $x^* \in \partial\Omega_h$, alors en utilisant la deuxième équation de (2.9),

$$v(t+\tau,\cdot) = 0 \leq \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty .$$

Sinon, $x^* \in \Omega_h$ et on peut donc utiliser la première équation de (2.9) :

$$v(t+\tau,x^*) = v(t,x^*) + \tau\Delta_h v(t+\tau,x^*) + \tau f(t+\tau,x^*).$$

La maximalité de x^* implique que $v(t+\tau,x^* \pm h) \leq v(t+\tau,x^*)$, soit

$$\Delta_h v(t+\tau,x^*) = \frac{v(t+\tau,x^*+h) - 2v(t+\tau,x^*) + v(t+\tau,x^*-h)}{h^2} \leq 0.$$

Combiné à l'équation précédente, cette inégalité donne

$$v(t+\tau,x^*) \leq v(t,x^*) + \tau f(t+\tau,x^*) \leq \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty .$$

On a donc montré dans les deux cas possibles que

$$\max_{\Omega_h} v(t+\tau,\cdot) = v(t+\tau,x^*) \leq \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty ,$$

et on montre par la même méthode que

$$\min_{\Omega_h} v(t+\tau,\cdot) \geq -(\|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty),$$

d'où l'on déduit que $\|v(t+\tau,\cdot)\|_\infty \leq \|v(t,\cdot)\|_\infty + \tau\,\|f\|_\infty$. On peut conclure comme dans la preuve précédente. □

Corollaire 2.6. *Le schéma d'Euler implicite admet une unique solution.*

Démonstration. On considère l'application linéaire L qui envoie une fonction $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ vers la fonction $Lv \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ définie par

$$(Lv)(t,x) = \begin{cases} v(t,x) & \text{si } (t,x) \in (\{0\} \times \Omega_h) \cup (\overline{I}_\tau \times \partial\Omega_h) \\ \dfrac{v(t,x) - v(t-\tau,x)}{\tau} - (\Delta_h v)(t,x) & \text{sinon.} \end{cases}$$

La proposition précédente montre directement que si $Lv = 0$, alors $v = 0$. L'application linéaire L est donc injective, et comme son espace de départ a la même dimension que son espace d'arrivée, elle est bijective. On en déduit directement l'existence et l'unicité d'une solution au schéma (2.6). □

2.4 Consistance et convergence des schémas d'Euler

La notion de consistance [1] est la même que pour les EDO (cf équation (1.5)) : il s'agit de vérifier que la solution de l'EDP qu'on cherche à approcher est "presque" solution du schéma.

Définition 2.3 (Consistance)**.** Soit $u \in \mathcal{C}^0([0,T] \times \overline{\Omega}) \cap \mathcal{C}^2_1(]0,T[\times\Omega)$ une solution classique de l'équation de la chaleur (2.2). Les *erreurs locale de consistance* des schémas explicites (2.5) et implicites (2.6) en la solution u, et en un point $(t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h$ sont définies par

$$\varepsilon^{\text{ee}}_{u,\tau,h}(t,x) = \frac{u(t+\tau,x) - u(t,x)}{\tau} - (\Delta_h u)(t,x) - f(t,x), \tag{2.10}$$

$$\varepsilon^{\text{ei}}_{u,\tau,h}(t,x) = \frac{u(t+\tau,x) - u(t,x)}{\tau} - (\Delta_h u)(t+\tau,x) - f(t+\tau,x). \tag{2.11}$$

Le schéma s ($s \in \{\text{ee}, \text{ei}\}$) est dit *consistant en norme infinie* si

$$\lim_{(\tau,h)\to 0} \left\|\varepsilon^s_{u,\tau,h}\right\|_\infty = 0.$$

Le schéma est *d'ordre $k \in \mathbb{N}$ en temps et $\ell \in \mathbb{N}$ en espace en norme infinie* si

$$\left\|\varepsilon^s_{u,\tau,h}\right\|_\infty = \text{const}(u) \cdot (\tau^k + h^\ell).$$

Proposition 2.7 (Consistance)**.** *Les schémas d'Euler explicite et implicite sont :*

- *d'ordre 1 en temps et en espace si $u \in \mathcal{C}^3_2([0,T] \times \overline{\Omega})$*
- *d'ordre 1 en temps et 2 en espace si $u \in \mathcal{C}^4_2([0,T] \times \overline{\Omega})$.*

Démonstration. La démonstration de la consistance d'un schéma est un exercice (parfois difficile) d'application de la formule de Taylor. Nous démontrons le résultat dans le cas Euler explicite, le cas implicite se traite de la même manière. Pour commencer,

1. Le mot consistance est une mauvaise traduction du mot anglais *consistency*, qui devrait être traduit en français par *cohérence*

nous supposons seulement $u \in \mathcal{C}_2^3([0,T]\times\overline{\Omega})$. Par la formule de Taylor-Lagrange, pour tout $(t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h$, il existe $\xi^\pm \in [0,h]$ tel que

$$u(t,x\pm h) = u(t,x) \pm \partial_x u(t,x)h + \partial_{xx}u(t,x)\frac{h^2}{2} \pm \partial_{xxx}u(t,x+\xi^\pm)\frac{h^3}{6},$$

de sorte que

$$u(t,x+h)-2u(t,x)+u(t,x-h) = \partial_{xx}u(t,x)h^2+(\partial_{xxx}u(t,x+\xi^+)-\partial_{xxx}u(t,x+\xi^-))\frac{h^3}{6},$$

soit

$$|\Delta_h u(t,x) - \partial_{xx}u(t,x)| \leq \frac{1}{3}\left\|\partial_{xxx}u\right\|_\infty h.$$

De même, par une formule de Taylor-Lagrange à l'ordre 2 en temps, on montre qu'il existe $\zeta \in [0,\tau]$ tel que

$$u(t+\tau,x) = u(t,x) + \partial_t u(t,x)\tau + \partial_{tt}u(t+\zeta,x)\frac{\tau^2}{2},$$

soit

$$\left\|\frac{u(t+\tau,x)-u(t,x)}{\tau}\right\| \leq \frac{1}{2}\left\|\partial_{tt}u\right\|_\infty \tau$$

En utilisant la première ligne de l'équation (2.2), $f = \partial_t u - \Delta u$, l'erreur locale de consistance $\varepsilon := \varepsilon^{\text{ee}}_{u,\tau,h}$ peut se réécrire

$$\begin{aligned}\varepsilon(t,x) &= \frac{u(t+\tau,x)-u(t,x)}{\tau} - (\Delta_h u)(t,x) - f(t,x)\\ &= \frac{u(t+\tau,x)-u(t,x)}{\tau} - \partial_t u - ((\Delta_h u)(t,x) - \Delta u(t,x))\end{aligned}$$

soit

$$\begin{aligned}|\varepsilon(t,x)| &\leq \left|\frac{u(t+\tau,x)-u(t,x)}{\tau} - \partial_t u\right| + |(\Delta_h u)(t,x) - \Delta u(t,x)|\\ &\leq \frac{1}{2}\left\|\partial_{tt}u\right\|_\infty \tau + \frac{1}{3}\left\|\partial_{xxx}u\right\|_\infty h\end{aligned}$$

Ainsi, $\|\varepsilon\|_\infty \leq \text{const}(u)\cdot(\tau+h)$ et le schéma d'Euler explicite est bien d'ordre 1 en espace et en temps.

Lorsque u est plus régulière en espace, $u \in \mathcal{C}_2^4([0,T]\times\overline{\Omega})$, on peut pousser le développement de Taylor à un ordre de plus en espace :

$$u(t,x\pm h) = u(t,x)\pm\partial_x u(t,x)h+\partial_{xx}u(t,x)\frac{h^2}{2}\pm\partial_{xxx}u(t,x)\frac{h^3}{6}+\partial_{xxxx}u(t,x+\xi^\pm)\frac{h^4}{24}.$$

Lorsqu'on fait la somme de ces deux développements (en $+h$ et $-h$), les termes d'ordre impair s'annulent, et on obtient donc

$$\Delta_h u(t,x) = \partial_{xx}u(t,x) + (\partial_{xxxx}u(t,x+\xi^-) + \partial_{xxxx}u(t,x+\xi^+))\frac{h^2}{24},$$

soit

$$|\Delta_h u(t,x) - \partial_{xx}u(t,x)| \leq \frac{1}{24}\left\|\partial_{xxxx}u\right\|_\infty h^2.$$

Ainsi, dans ce cas le schéma est d'ordre 2 en espace et toujours 1 en temps. □

Théorème 2.8 (Convergence). *Soit $u \in \mathcal{C}^0([0,T]\times\overline{\Omega}) \cap \mathcal{C}_1^2(]0,T[\times\Omega)$ une solution classique de l'équation de la chaleur. Soit v^{ei} une solution du schéma d'Euler implicite* (2.6) *et soit $\hat{u} = u|_{\overline{I}_\tau \times \Omega_h}$ la restriction de la solution exacte à la grille discrète. Alors,*

$$\left\|\hat{u} - v^{\mathrm{ei}}\right\|_\infty \leq \begin{cases} \mathrm{const}(u)\cdot(\tau+h) & \text{si } u \in \mathcal{C}_2^3([0,T]\times\overline{\Omega}) \\ \mathrm{const}(u)\cdot(\tau+h^2) & \text{si } u \in \mathcal{C}_2^4([0,T]\times\overline{\Omega}) \end{cases}$$

Soit v^{ee} une solution du schéma d'Euler explicite (2.5). *Si la condition CFL* (2.8) *est vérifiée, alors*

$$\|\hat{u} - v^{\mathrm{ee}}\|_\infty \leq \begin{cases} \mathrm{const}(u)\cdot(\tau+h) & \text{si } u \in \mathcal{C}_2^3([0,T]\times\overline{\Omega}) \\ \mathrm{const}(u)\cdot(\tau+h^2) & \text{si } u \in \mathcal{C}_2^4([0,T]\times\overline{\Omega}) \end{cases}$$

Remarque 2.6. Ce théorème est un cas particulier du théorème de Lax-Richtmyers, qui dit que si un schéma numérique pour une EDP linéaire est stable et consistant, alors il est convergent.

Démonstration. La démonstration des deux énoncés est similaire, nous ne ferons donc que la démonstration dans le cas Euler implicite. On fixe $\tau, h > 0$ et on note $\varepsilon(t,x) = \varepsilon_{u,\tau,h}(x,t)$ l'erreur de consistance locale définie dans l'équation (2.11). Alors, par définition, pour tout $(t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h$,

$$\frac{u(t+\tau,x) - u(t,x)}{\tau} - (\Delta_h u)(t+h,x) = f(t+h,x) + \varepsilon(t,x)$$

De plus, par définition du schéma, on sait que

$$\frac{v^{\mathrm{ei}}(t+\tau,x) - v^{\mathrm{ei}}(t,x)}{\tau} - (\Delta_h v^{\mathrm{ei}})(t+h,x) = f(t+h,x)$$

En posant $w = u - v^{\mathrm{ei}}$ et en soustrayant la deuxième équation à la première, on voit que w vérifie

$$\begin{cases} \dfrac{w(t+\tau,x) - w(t,x)}{\tau} - (\Delta_h w)(t+h,x) = \varepsilon(t,x) & \text{pour } (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ w = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \end{cases}$$

Comme le schéma d'Euler implicite est stable (proposition 2.5), et que $w(0,\cdot) = u_0 - u_0 = 0$, on en déduit

$$\|w\|_\infty \leq \|w(0,\cdot)\|_\infty + T\,\|\varepsilon\|_\infty = T\,\|\varepsilon\|_\infty .$$

Enfin, en utilisant la consistance du schéma d'Euler implicite (proposition 2.7), on a

$$\|\varepsilon\|_\infty \leq \mathrm{const}(u)\cdot(\tau^k + h^\ell).$$

avec $(k,\ell) = (1,1)$ si $u \in \mathcal{C}_2^3([0,T]\times\overline{\Omega})$ et $(k,\ell) = (1,2)$ si $u \in \mathcal{C}_2^3([0,T]\times\overline{\Omega})$. Ainsi,

$$\left\|u - v^{\mathrm{ei}}\right\|_\infty \leq T\mathrm{const}(\tau^k + h^\ell). \qquad \square$$

2.5 Stabilité et convergence en norme L^2

La stabilité en norme infinie est une condition très forte, qui n'est satisfaite que par un petit nombre de schémas. Dans cette section, nous montrons comment obtenir de la stabilité, et en déduire la convergence de schémas, pour une norme de type L^2. Cette partie sera un peu moins formalisée (i.e. on n'introduira pas de nouvelle définition).

Comme dans la partie précédente, on montre d'abord la stabilité dans le cas continu pour montrer la similarité entre ce cas et le cas discret.

Proposition 2.9 (Stabilité L^2, cas continu)**.** *Soit $u \in \mathcal{C}_1^2([0,T]\times\overline{\Omega})$ une solution classique de l'équation de la chaleur* (2.2)*. Alors,*

$$\|u(t,\cdot)\|_{\mathrm{L}^2(\Omega)} \leq \|u(0,\cdot)\|_{\mathrm{L}^2(\Omega)} + \int_0^t \|f(t,\cdot)\|_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s.$$

Démonstration. Posons $E(t) = \int_\Omega u(t,x)^2\mathrm{d}t$. Alors,

$$\begin{aligned} E'(t) &= \frac{\mathrm{d}}{\mathrm{d}t}\int_\Omega u(t,x)^2\mathrm{d}x \\ &= 2\int_\Omega u(t,x)\partial_t u(t,x)\mathrm{d}x \\ &= 2\int_\Omega u(t,x)(f(t,x)+\Delta u(t,x))\mathrm{d}x \end{aligned}$$

On utilise la formule de Stokes et $u=0$ sur $\partial\Omega$ pour obtenir

$$\int_\Omega u\Delta u = \int_\Omega u\mathrm{div}(\nabla u) = -\int_\Omega \langle\nabla u|\nabla u\rangle + \underbrace{\int_{\partial\Omega}\langle\nabla u|un_\Omega\rangle}_{=0} \leq 0$$

soit

$$E'(t) \leq 2\langle u(t,\cdot)|f(t,\cdot)\rangle_{\mathrm{L}^2(\Omega)} \leq 2\,\|u(t,\cdot)\|_{\mathrm{L}^2(\Omega)}\,\|f(t,\cdot)\|_{\mathrm{L}^2(\Omega)}\,.$$

Ainsi, si on pose pour $\varepsilon > 0$,

$$E_\varepsilon(t) = \sqrt{\varepsilon + \|u(t,\cdot)\|^2_{\mathrm{L}^2(\Omega)}} = \sqrt{\varepsilon + E(t)},$$

alors

$$E'_\varepsilon(t) = \frac{E'(t)}{2E_\varepsilon(t)} \leq \frac{2\,\|u(t,\cdot)\|_{\mathrm{L}^2(\Omega)}}{2\sqrt{\varepsilon+\|u(t,\cdot)\|^2_{\mathrm{L}^2(\Omega)}}}\,\|f(t,\cdot)\|_{\mathrm{L}^2(\Omega)} \leq \|f(t,\cdot)\|_{\mathrm{L}^2(\Omega)}\,.$$

Ainsi,

$$E_\varepsilon(t) \leq E_\varepsilon(0) + \int_0^t \|f(t,\cdot)\|_{\mathrm{L}^2(\Omega)}\,\mathrm{d}s.$$

On obtient la proposition en prenant la limite lorsque $\varepsilon \to 0$. □

Pour étudier la stabilité L^2 dans le cas discret, nous utiliserons le théorème de Gershgorin-Hadamard. Ce théorème montre que le spectre d'une matriceq A est inclus dans une union de boules centrées en les éléments diagonaux de A. C'est un résultat très utile pour localiser grossièrement et très rapidement le spectre d'une matrice (il permet par exemple souvent de voir si la matrice est positive ou négative).

Théorème 2.10 (Gerschgorin-Hadamard)**.** *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice carrée. Alors, pour toute valeur propre λ (réelle ou complexe) de A,*

$$\exists i_0 \in [\![1,n]\!],\ \ tq\ \ |\lambda - a_{i_0,i_0}| \leq \sum_{i \in [\![1,n]\!] \setminus \{i_0\}} |a_{i_0,i}|\,.$$

Démonstration. Soit $\lambda \in \mathbb{R}$ une valeur propre de A et $v \in \mathbb{R}^n \setminus \{0\}$ le vecteur propre associé. Soit $i_0 \in \{1,\dots,n\}$ l'indice de la coordonnée de A la plus grande en module (i.e. $|v_{i_0}| = \max_{i\in[\![1,n]\!]} |v_i|$). Alors, comme $Av = \lambda v$,

$$\lambda v_{i_0} = |Av|_{i_0} = a_{i_0,i_0} v_{i_0} + \sum_{i \neq i_0} a_{i_0,i} v_i,$$

soit

$$(\lambda - a_{i_0,i_0}) v_{i_0} = \sum_{i \leq i_0} a_{i_0,i} v_i.$$

En passant au module, on obtient

$$|\lambda - a_{i_0,i_0}|\,|v_{i_0}| = \left|\sum_{i\neq i_0} a_{i_0,i} v_i\right| \leq \sum_{i\neq i_0} |a_{i_0,i}|\,|v_i| \leq \left(\sum_{i\neq i_0} |a_{i_0,i}|\right) |v_{i_0}|\,.$$

Comme $|v_{i_0}| \neq 0$ (sinon v serait nul), on en déduit que λ est contenu dans la boule centrée en a_{i_0,i_0} et de rayon $\sum_{i\neq i_0} |a_{i_0,i}|$. □

Corollaire 2.11. *Soit $A_h \in \mathcal{M}_M(\mathbb{R})$ la matrice tridiagonale carrée*

$$A_h = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & & \\ 1 & -2 & 1 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & \ddots & 1 & -2 & 1 \\ & & 0 & 1 & -2 \end{pmatrix} \tag{2.12}$$

Alors, A_h est diagonalisable et ses valeurs propres sont dans l'intervalle $[-\frac{4}{h^2}, 0[$.

Démonstration. La matrice A_h est symétrique, donc diagonalisable en base orthonormée. Par le théorème de Gerschgorin-Hadamard, les valeurs propres de A sont toutes incluses dans la boule $\mathrm{B}(-2,2) = [-4,0]$. Ainsi, A est symétrique négative ; il reste à montrer que 0 n'est pas valeur propre, c'est-à-dire que $\mathrm{Ker} A = \{0\}$ (exercice). □

Nous énonçons maintenant un résultat de stabilité L^2 pour le schéma d'Euler explicite, dans le cas 1-dimensionnel. On réutilise toutes les notations des sections précédente. On considère de plus la norme suivante sur $F(\overline{\Omega}_h)$:

$$\|v\|_{2,h} := \left(h \sum_{x \in \overline{\Omega}_h} v(x)^2 \right)^{1/2}$$

Proposition 2.12 (Stabilité L^2 pour Euler explicite)**.** *Soit $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ une fonction vérifiant*

$$\begin{cases} \dfrac{v(t+\tau,x) - v(t,x)}{\tau} - (\Delta_h v)(t,x) = f(t,x) & \text{pour } (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ v = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \end{cases} \tag{2.13}$$

où $f \in F((\overline{I}_\tau \setminus \{T\}) \times \Omega_h)$. On suppose la condition CFL (2.8) satisfaite. Alors,

$$\|v(t,\cdot)\|_{2,h} \leq \|v(0,\cdot)\|_{2,h} + \tau \sum_{s \in [0,t[\cap \overline{I}_\tau} \|f(t,\cdot)\|_{2,h} \,.$$

Démonstration. Comme dans l'exercice 2.2, on définit $V^n = (v(t^n, x_j))_{1 \leq j \leq M}, F^n = f(t^n, x_j)_{1 \leq j \leq M}$ et $U_0 = (u_0(x^j))_{1 \leq j \leq M}$. Alors, $V^0 = U_0$ et

$$V^{n+1} = (\mathrm{Id} + \tau A_h) V^n + \tau F^n,$$

où A_h est la matrice du laplacien définie dans (2.12). Par le corollaire du théorème de Gerschgorin-Hadamard, les valeurs propres de A_h sont dans $[-4, 0[$, donc celles de la matrices $B_h = \mathrm{Id}_M + \tau A_h$ sont contenues dans l'intervalle $[1 - 4\frac{\tau}{h^2}, 1[$. Ainsi, si $4\frac{\tau}{h^2} \leq 2$, les valeurs propres de B_h appartiennent à l'intervalle $[-1, 1]$. Comme B_h est symétrique, on en déduit que la norme d'opérateur de B_h, induite par la norme Euclidienne, est plus petite que 1. Ainsi,

$$\left\|V^{n+1}\right\| = \|(\mathrm{Id} + \tau A_h) V^n + \tau F^n\| \leq \|V^n\| + \tau \left\|F^n\right\|,$$

et par récurrence on obtient

$$\|V^n\| \leq \left\|V^0\right\| + \tau \sum_{0 \leq m \leq n-1} \|F^m\| \,,$$

qui donne exactement l'inégalité voulue en multipliant tout par $\sqrt{h}$. □

La démonstration de convergence pour la norme $\|.\|_{2,h}$ fonctionne exactement comme pour la démonstration de convergence en norme infinie (Théorème 2.8), nous ne la faisons donc pas. L'exercice 2.7 montre la stabilité, consistance et convergence en norme $\|.\|_{2,h}$ du schéma de Crank-Nicolson.

2.6 Exercices

Exercice 2.1. *Modélisation.* Faire les calculs de la remarque 2.1 dans le cas $d = 1$, en prenant $A = [a, b]$.

Exercice 2.2. *Formulation matricielle des schémas d'Euler.* Soit $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$, et définissons $V^n = (v(t^n, x_j))_{1\leq j\leq M}, F^n = f(t^n, x_j)_{1\leq j\leq M}$. et $U_0 = (u_0(x^j))_{1\leq j\leq M}$. Noter que V^n est de dimension M et pas $M+2$ (on a oublié les valeurs de v sur $\partial\Omega_h$ car elles sont nulles par hypothèse). Soit enfin A_h la matrice tridiagonale carrée

$$A_h = \frac{1}{h^2}\begin{pmatrix} -2 & 1 & 0 & & \\ 1 & -2 & 1 & \ddots & \\ 0 & \ddots & \ddots & \ddots & 0 \\ & \ddots & 1 & -2 & 1 \\ & & 0 & 1 & -2 \end{pmatrix}$$

1. Montrer que le schéma d'Euler explicite (2.5) peut être reformulé en

$$\begin{cases} V^0 = U_0 \\ V^{n+1} = (\mathrm{Id}_M + \tau A_h)\, V^n + \tau F^n. \end{cases}$$

2. De même, montrer que le schéma d'Euler implicite (2.6) s'écrit récurrence

$$\begin{cases} V^0 = U_0 \\ (\mathrm{Id}_M - \tau A_h)\, V^{n+1} = (V^n + \tau F^{n+1}). \end{cases}$$

3. Démontrer que

$$h^2\langle V|A_hV\rangle = -V_1^2 - V_M^2 - \sum_{i=1}^{M-1}(V_{i+1} - V_i)^2.$$

En déduire que la matrice A_h est définie négative, puis que $\mathrm{Id}_M + \tau A_h$ est définie positive donc inversible. Conclure le schéma d'Euler implicite admet une unique solution.

Exercice 2.3. *Principe du maximum pour Euler explicite.* Soit (τ, h) des pas de temps et d'espace vérifiant l'inégalité (2.8). On considère une fonction $v \in F(\overline{I}_\tau \times \overline{\Omega}_h)$ vérifiant

$$\forall (t,x) \in \overline{I}_\tau \setminus \{T\} \times \overline{\Omega}_h, \qquad \frac{v(t+\tau, x) - v(t,x)}{\tau} - \Delta_h v(t,x) \leq 0. \tag{2.14}$$

1. Dans cette question, on suppose que le maximum de v est atteint en un point

$$(t^*, x^*) \in (\overline{I}_\tau \setminus \{0\}) \times \Omega_h.$$

a) Montrer que

$$v(t^*,x^*) \leq (1-2\frac{\tau}{h^2})v(t^*-\tau,x^*)+\frac{\tau}{h^2}v(t^*-\tau,x^*-h)+\frac{\tau}{h^2}v(t^*-\tau,x^*-h),$$

puis que $v(t^*-\tau,x^*)=v(t^*,x^*)$.

b) En déduire que $v(t^*,x^*)=v(0,x^*)$, puis que v atteint son maximum en un point de $\{0\}\times\overline{\Omega}_h$.

2. Montrer que pour toute fonction v vérifiant l'hypothèse (2.15) (mais pas nécessairement celle de la question précédente),

$$\max_{\overline{I}_\tau\times\overline{\Omega}_h} v = \max_{(\{0\}\times\overline{\Omega}_h)\cup((\overline{I}_\tau\setminus\{T\})\times\partial\Omega_h)} v$$

Le but du prochain exercice est d'établir un analogue discret du principe du maximum fort, qui affirme que si $\partial_t v-\Delta v\leq 0$ dans $]0,T[\times\Omega$ et si v atteint son maximum en un point de $]0,T]\times\Omega$, alors v est constante.

Exercice 2.4. *Principe du maximum pour Euler implicite.* Soit $v\in F(\overline{I}_\tau\times\overline{\Omega}_h)$ une fonction vérifiant

$$\forall(t,x)\in(\overline{I}_\tau\setminus\{T\}\times\overline{\Omega}_h),\qquad \frac{v(t+\tau,x)-v(t,x)}{\tau}-\Delta_h v(t+\tau,x)\leq 0. \tag{2.15}$$

On suppose que le maximum de v est atteint en un point

$$(t^*,x^*)\in(\overline{I}_\tau\setminus\{0\})\times\Omega_h.$$

1. Montrer que $\Delta_h v(t^*,x^*)\leq 0$ et que $\frac{v(t^*,x)-v(t^*-\tau,x)}{\tau}\geq 0$.

2. En utilisant le schéma, en déduire que $v(t^*,x^*\pm h)=v(t^*-\tau,x^*)=v(t^*,x^*)$.

3. Conclure que v est constante sur $\{0,\tau,2\tau,\ldots,t^*\}\times\overline{\Omega}_h$, égale à $v(t^*,x^*)$.

Exercice 2.5. *Exemple d'instabilité du schéma d'Euler explicite.* Dans cet exercice, on suppose que M est impair. On définit

$$u_0(x_j)=\sin(\frac{\pi}{2}j),$$

de sorte que les valeurs de u sont $0,1,0,-1,\ldots,0$. Soit v la solution de (2.5).

1. Calculer explicitement $v(t^n,\cdot)$ en fonction de $v(t^0,\cdot)=u_0$.

2. En déduire que si $\frac{\tau}{h^2}>1$, alors $\lim_{n\to+\infty}\|v(t^n,\cdot)\|_\infty=+\infty$.

Exercice 2.6. *Discrétisation d'ordre 4 du laplacien.* Étant donnée $u\in\mathcal{C}^6(I)$, I un segment et $x\in\mathbb{R}$ tel que $[x-2h,x+2h]\subseteq I$, on pose

$$\Delta_{5,h}u(x)=\frac{-u(x+2h)+16u(x+h)-30u(x)+16u(x-h)-u(x-2h)}{12h^2}$$

1. Montrer que

$$\left|u''(x)-\Delta_{5,h}u(x)\right|\leq C\left\|u^{(6)}\right\|_\infty h^4, \tag{2.16}$$

où C est une constante universelle.

2. Montrer que l'unique laplacien discret $\overline{\Delta}$ défini par une formule de la forme

$$\overline{\Delta}_h u(x) = \frac{cu(x+2h) + bu(x+h) + au(x) + bu(x-h) + cu(x-2h)}{h^2},$$

où $a, b, c \in \mathbb{R}$ et vérifiant (2.16) pour toute fonction $u \in \mathcal{C}^6(I)$ doit être égal à $\Delta_{5,h}$.

Exercice 2.7. *Schéma de Crank-Nicolson.* On s'intéresse à la résolution de l'équation de la chaleur (2.2) dans le cas $f \equiv 0$. Une fonction $v \in F(\overline{I}_\tau \times \Omega_h)$ est solution du schéma de Crank-Nicholson si

$$\begin{cases} \dfrac{v(t+\tau,x) - v(t,x)}{\tau} - \frac{1}{2}\left[(\Delta_h v)(t,x) + (\Delta_h v)(t+\tau,x)\right] = 0 & \text{pour } (t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \Omega_h \\ v = 0 & \text{sur } \overline{I}_\tau \times \partial\Omega_h \\ v = u_0 & \text{sur } \{0\} \times \overline{\Omega}_h \end{cases} \tag{2.17}$$

1. On note V^n le vecteur $(v(t_n, x_j))_{j \in [\![1,M]\!]} \in \mathbb{R}^M$. Montrer que le schéma de Crank-Nicolson peut être écrit sous forme matricielle (où le laplacien discret A_h est défini dans (2.12))

$$\begin{cases} V^0 = (u_0(x_1), \dots, u_0(x_M)), \\ (\mathrm{Id}_M - \frac{\tau}{2}\Delta_h)V^{n+1} = (\mathrm{Id}_M + \frac{\tau}{2}\Delta_h)V^n \end{cases}$$

2. Montrer que $(\mathrm{Id}_M - \frac{\tau}{2}A_h)$ est inversible, puis calculer les valeurs propres de la matrice $B_h = (\mathrm{Id}_M - \frac{\tau}{2}A_h)^{-1}(\mathrm{Id}_M + \frac{\tau}{2}A_h)$ en fonction de celles de A_h.

3. Montrer que la norme de B_h est majorée par 1
(Indication : passer par le rayon spectral.)

4. *Stabilité* Soit $(W^n)_{n \in [\![0,N]\!]}$ et $(\varepsilon^n)_{n \in [\![0,N-1]\!]}$ des vecteurs de $\mathbb{R}^M$ vérifiant

$$\begin{cases} W^0 = 0 \\ (\mathrm{Id}_M - \frac{\tau}{2}A_h)W^{n+1} = (\mathrm{Id}_M + \frac{\tau}{2}A_h)W^n + \varepsilon^n \quad \text{pour } n \in [\![0, N-1]\!] \end{cases}$$

Montrer l'inégalité de stabilité

$$\forall n \in [\![1, n]\!], \quad \|W^n\| \leq \sum_{k \in [\![0,n-1]\!]} \left\|\varepsilon^k\right\|. \tag{2.18}$$

5. *Consistance* Soit $u \in \mathcal{C}_2^3([0,T] \times \overline{\Omega})$ une solution de l'équation de la chaleur (2.2) avec $f \equiv 0$. L'erreur de consistance locale de Crank-Nicolson est

$$\varepsilon_{u,\tau}^{\mathrm{cn}}(t,x) = \frac{u(t+\tau,x) - u(t,x)}{\tau} - \frac{1}{2}\left[(\Delta_h u)(t,x) + (\Delta_h u)(t+\tau,x)\right].$$

a) Montrer que $\left|\frac{u(t+\tau,x)-u(t,x)}{\tau} - \partial_t u(t+\frac{\tau}{2}, x)\right| = O(\tau^2)$.
(Indication : les développements limités doivent être au point $(t+\frac{\tau}{2}, x)$.)

b) En utilisant l'inégalité $\|\Delta u(x,t) - \Delta_h u(x,t)\| \leq \text{const}(u)\cdot h$ vue en cours, montrer que

$$\left\| \frac{1}{2}\left[(\Delta_h u)(t,x) + (\Delta_h u)(t+\tau,x)\right] - \Delta u(t+\frac{\tau}{2},x)\right\| = O(h+\tau^2).$$

c) En déduire la majoration suivante de l'erreur de consistance $\varepsilon = \varepsilon^{\text{cn}}_{u,\tau}$:

$$\forall t \in \overline{I}_\tau, \quad \|\varepsilon(t,\cdot)\|_{2,h} \leq \text{const}(u)\cdot(\tau^2+h). \tag{2.19}$$

6. *Convergence* On garde les hypothèses de la question précédente et on note

- $\hat{U}^n = (u(t^n,x_j))_{j\in[\![1,M]\!]} \in \mathbb{R}^M$ où u est solution $\mathcal{C}^4$ de (2.2) avec $f\equiv 0$,
- $V^n = (v(t^n,x_j))_{j\in[\![1,M]\!]} \in \mathbb{R}^M$, où v est solution du schéma (2.17),
- $W^n = \hat{U}^n - V^n$ l'erreur de convergence,
- $\varepsilon^n = (\varepsilon(t^n,x_j))_{j\in[\![1,M]\!]}$ l'erreur de consistance.

a) Montrer que $(\text{Id}_M - \frac{\tau}{2}A_h)W^{n+1} = (\text{Id}_M + \frac{\tau}{2}A_h)W^n + \varepsilon^n$ et $W^0 = 0$.

b) Déduire de (2.18) et (2.19) que

$$\forall t \in \overline{I}_\tau, \quad \|u(t,\cdot) - v(t,\cdot)\|_{2,h} \leq \text{const}(u)\cdot(\tau^2+h).$$

Remarque : on peut en fait montrer que l'erreur de consistance est en $O(\tau^2+h^2)$; c'est un avantage du schéma de Crank-Nicolson sur Euler implicite.

2.7 Correction des exercices

Correction de l'exercice 2.3.. *1. a. Il suffit de réécrire l'inégalité (2.15) en $(t,x) = (\tau^*-\tau, x^*)$. La condition CFL garantit que les coefficients devant $v(t^*-\tau,x^*)$ et $v(t^*-\tau,x^*\pm h)$ sont positifs. Comme (t^*,x^*) est le maximum de v, chacun de ces termes est majoré par $v(t^*,x^*)$. Autrement dit,*

$$\begin{aligned} v(t^*,x^*) &\leq (1-2\frac{\tau}{h^2})v(t^*-\tau,x^*) + \frac{\tau}{h^2}v(t^*-\tau,x^*-h) + \frac{\tau}{h^2}v(t^*-\tau,x^*-h) \\ &= \left((1-2\frac{\tau}{h^2}) + \frac{\tau}{h^2} + \frac{\tau}{h^2}\right) v(t^*,x^*) \\ &= v(t^*,x^*) \end{aligned}$$

Ainsi, les trois inégalités $v(t^*-\tau,x^*) \leq v(t^*,x^*)$ et $v(t^*\pm\tau,x^*) \leq v(t^*,x^*)$ sont en fait des égalités, et en particulier $v(t^*-\tau,x^*) = v(t^*,x^*)$. 1.b. Par récurrence, on en déduit que $v(t^*,x^*) = v(t^*-n\tau)$ pour tout $n \leq t^*/\tau$, donc $v(t^*,x^*) = v(t^0,x^*)$. Ainsi,

$$\max_{\overline{I}_\tau\times\overline{\Omega}_h} v = v(t^*,x^*) = v(t^0,x^*),$$

et v atteint donc son maximum en $(t^0,x^*) \in \{0\}\times\overline{\Omega}_h$.
2. Soit v vérifiant (2.15) et (t^*,x^*) un maximiseur de v. Si (t^*,x^*) vérifie l'hypothèse de la question 1, le maximum de v est atteint en $\{0\}\times\overline{\Omega}_h$ par 1.b., et l'affirmation

de la question 2. est donc vrai. Si (t^*, x^*) ne vérifie pas l'hypothèse de la question 1, alors

$$\begin{aligned}(t^*, x^*) &\in (\overline{I}_\tau \times \overline{\Omega}_h) \setminus ((\overline{I}_\tau \setminus \{0\}) \times \Omega_h) \\ &= (\{0\} \times \overline{\Omega}_h) \cup ((\overline{I}_\tau \setminus \{T\}) \times \partial\Omega_h),\end{aligned}$$

et v atteint aussi son maximum sur $\{0\} \times \overline{\Omega}_h) \cup ((\overline{I}_\tau \setminus \{T\}) \times \partial\Omega_h)$.

***Correction de l'exercice 2.5**. 1. Un calcul direct montre que pour tout $x \in \overline{\Omega_h}$,*

$$v(t^1, x) = \left(1 - 2\frac{\tau}{h^2}\right) u^0(x),$$

de sorte que

$$v(t^n, x) = \left(1 - 2\frac{\tau}{h^2}\right)^n u^0(x).$$

2. Si $\frac{\tau}{h^2} > 1$, alors $(1 - 2\frac{\tau}{h^2}) < -1$, ce qui implique

$$\|v(t^n, \cdot)\|_\infty = \left|1 - 2\frac{\tau}{h^2}\right|^n \left\|u^0\right\|^\infty \xrightarrow{n\to\infty} +\infty.$$

***Correction de l'exercice 2.6**. Il suffit d'écrire le développement à l'ordre 6 de $u(x \pm h)$ et $u(x \pm 2h)$: il existe $\xi^\pm \in [0, h]$ et $\zeta \in [0, 2h]$ tels que*

$$u(x\pm h) = u(x) \pm u'(x)h + u''(x)\frac{h^2}{2} \pm u^{(3)}(x)\frac{h^3}{6} + u^{(4)}(x)\frac{h^4}{24} + \pm u^{(5)}(x)\frac{h^5}{120} + u^{(6)}(x \pm \xi^\pm)\frac{h^6}{720}$$

$$u(x\pm 2h) = u(x) \pm u'(x)h + u''(x)\frac{4h^2}{2} \pm u^{(3)}(x)\frac{8h^3}{6} + u^{(4)}(x)\frac{16h^4}{24} + \pm u^{(5)}(x)\frac{64h^5}{120} + u^{(6)}(x \pm \zeta^\pm)\frac{256h^6}{720}$$

On considère des coefficients $a, b, c \in \mathbb{R}$ et une discrétisation de la forme $\overline{\Delta}_h$ défini dans la deuxième question. Ainsi, en remarquant que les termes en h, h^3, h^5 s'annulent, on obtient

$$h^2\overline{\Delta}_h u(x) = (2a + 2b + c)u(x) + \frac{h^2}{2}(2b + 8c)u''(x) + \frac{h^4}{24}(2b + 32c)u^{(4)} + \mathrm{O}(h^6)$$

Pour que la discrétisation soit d'ordre 4, il faut et il suffit que

$$\begin{cases} 2c + 2b + a = 0 \\ b + 4c = 1 \\ b + 16c = 0 \end{cases}$$

soit $-12c = 1$, $c = -\frac{1}{12}$, $b = -16c = \frac{16}{12}$ et $a = -2c + 2b = \frac{30}{12}$.

***Correction de l'exercice 2.7**. 2. L'inversibilité de la matrice $\mathrm{Id}_M - \frac{\tau}{2}A_h$ a déjà été montrée dans l'exercice 2.2 (ses valeurs propres sont de la forme $1 - \lambda$, où λ est une valeur propre de A_h, donc négative). En raisonnant dans une base diagonalisant A_h, on voit facilement que toute valeur propre de B_h est de la forme*

$$\frac{1 + \frac{\tau}{2}\lambda}{1 - \frac{\tau}{2}\lambda},$$

où λ est une valeur propre de A_h.
3. On se sert de la question précédente pour calculer la norme matricielle de B_h via le rayon spectral :

$$\begin{aligned}\|B_h\| &= \max\{|\mu| \mid \mu \in \mathrm{Spec}(B_h)\} \\ &= \max\{f(\lambda) \mid \lambda \in \mathrm{Spec}(A_h)\}\end{aligned}$$

où l'on a posé $f(\lambda) = \frac{1+\frac{\tau}{2}\lambda}{1-\frac{\tau}{2}\lambda}$. Comme f est croissante, on en déduit $f(t) \leq f(0) = 1$ pour $t \leq 0$. D'autre part, $\lim_{t\to-\infty} f(t) = -1$. Ainsi, $f(\lambda) \in [-1,1]$ pour $\lambda \in \mathrm{Spec}(A_h) \subseteq [-4,0]$, on conclut que

$$\|B_h\| = \max_{\mathrm{Spec}(A_h)} |f| \leq 1.$$

4. Calculons la norme de W^{n+1} :

$$\begin{aligned}\left\|W^{n+1}\right\| &= \left\|B_h W^n + (\mathrm{Id}_M - \frac{\tau}{2}A_h)^{-1}\varepsilon^n\right\| \\ &\leq \|B_h\|\,\|W^n\| + \left\|(\mathrm{Id}_M - \frac{\tau}{2}A_h)^{-1}\right\|\|\varepsilon^n\| \\ &\leq \|W^n\| + \|\varepsilon^n\|\,.\end{aligned}$$

Le résultat s'en déduit directement en prenant sommant ces inégalités.

Chapitre 3

Méthode des différences finies pour l'équation de transport

Dans ce court chapitre, on s'intéresse à l'équation de transport, aussi appelée équation de convection. Cette équation modélise un phénomène physique de transport d'une quantité (particules, énergie) par un champ de vecteurs. Pour éviter les difficultés lié au bord du domaine nous travaillerons dans un domaine périodique.

Définition 3.1 (Fonctions périodiques)**.** Une fonction f sur $\mathbb{R}^d$ est $\mathbb{Z}^d$-*périodique* si

$$\forall x \in \mathbb{R}^d, \forall z \in \mathbb{Z}^d, f(x+z) = f(x).$$

Remarque 3.1. De manière équivalente, une fonction $\mathbb{Z}^d$ périodique peut être vue comme une fonction sur le tore d-dimensionel $\mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$. Une fonction $f : \mathbb{R}^d \to \mathbb{R}$ $\mathbb{Z}^d$-périodique est uniquement définie par sa restriction à $[0,1[^d$.

On notera

$$\mathcal{C}^k(\mathbb{T}^d, \mathbb{R}^\ell) = \{f \in \mathcal{C}^k(\mathbb{R}^d, \mathbb{R}^\ell) \mid f \text{ est } \mathbb{Z} - \text{périodique}\}.$$

Définition 3.2 (Équation de transport)**.** Une fonction $u \in \mathcal{C}^1([0,T]\times\mathbb{T}^d)$ est solution classique de l'équation de transport si

$$\begin{cases} \partial_t u(t,x) + \langle b(x) | \nabla u(t,x) \rangle = f(t,x) & \text{pour } (t,x) \in [0,T] \times \mathbb{T}^d \\ u(0,\cdot) = u_0 \end{cases} \tag{3.1}$$

où $u_0 \in \mathcal{C}^1(\mathbb{T}^d)$ est la condition initiale et $b \in \mathcal{C}^1(\mathbb{T}^d, \mathbb{R}^d)$ est un champ de vecteurs.

3.1 Existence et unicité des solutions régulières

Pour construire une solution régulière à l'équation de transport, nous utiliserons le flot du champ de vecteurs b.

Proposition 3.1 (Flot associé à un champ de vecteurs)**.** *Soit $b \in \mathcal{C}^1(\mathbb{T}^d, \mathbb{R}^d)$. Alors, il existe une unique application $X : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^d$ vérifiant :*

(i) $\partial_t X(t,x) = b(X(t,x))$ *pour tout* $(t,x) \in \mathbb{R}\times\mathbb{R}^d$;
(ii) $X \in \mathcal{C}^1(\mathbb{R}\times\mathbb{R}^d,\mathbb{R}^d)$;
(iii) $X(t,X(s,x)) = X(t+s,x)$ *pour tout* $(t,s,x)\in\mathbb{R}\times\mathbb{R}\times\mathbb{R}^d$;
(iv) pour tout $t\in\mathbb{R}$, $X_t : x\in\mathbb{R}^d \mapsto X(t,x)\in\mathbb{R}^d$ *est un difféomorphisme ;*
(v) pour tout $(x,z)\in\mathbb{R}^d\times\mathbb{Z}$ *et* $t\in\mathbb{R}$, $X(t,x+z) = X(t,x)+z$.
En particulier, X *passe au quotient et définit une application* $X\in\mathcal{C}^1(\mathbb{R}\times\mathbb{T}^d,\mathbb{T}^d)$.

Démonstration. Avant toute chose, notons que comme la différentielle de b est $\mathbb{Z}^d$-périodique, et comme $b\in\mathcal{C}^1(\mathbb{R}^d)$, $\sup_{\mathbb{R}^d}\|\mathrm{D}_x b\| \leq \max_{[0,1]^d}\|\mathrm{D}_x b\| := M < +\infty$. Ainsi, la fonction b est Lipschitzienne. Par théorème de Cauchy-Lipschitz (global), on en déduit que pour tout $x\in\mathbb{R}^d$ il existe une unique courbe $\gamma_x\in\mathcal{C}^1(\mathbb{R})$ solution du problème de Cauchy

$$\begin{cases}\gamma_x(0) = x\\ \gamma_t'(x) = b(\gamma_x(t)).\end{cases}$$

On définit alors $X(t,x) = \gamma_x(t)$. La fonction X admet une dérivée partielle temporelle et vérifie bien (i) par construction.

(iii) Soit $x\in\mathbb{R}^d$, et $s,t\in\mathbb{R}$. On pose $y = \gamma_x(s)$. Soit $\gamma : t\in\mathbb{R} = \gamma_x(s+t)$. Alors,

$$\begin{cases}\gamma(0) = y\\ \gamma'(t) = \gamma_x'(s+t) = b(\gamma_x(s+t)) = b(\gamma(t))\end{cases},$$

par unicité dans le théorème de Cauchy-Lipschitz, on en déduit que $\gamma\equiv\gamma_y$. Ainsi,

$$X(t,y) = \gamma_y(t) = \gamma(t) = \gamma_x(s+t) = X(t+s,x).$$

Le point (v) se démontrer de la même manière (prendre $y = x+z$ et utiliser la périodicité du champ de vecteur).

(ii) Pour montrer que X est différentiable en la variable d'espace, nous allons dans un premier temps essayer de "deviner" la formule pour la matrice jacobienne

$$Z(t,x) := \mathrm{D}_x X(t,x) = (\partial_{x_j} X_i(t,x))_{1\leq i,j\leq d}\in\mathcal{M}_d(\mathbb{R}).$$

En supposant que les dérivées partielles commutent, on trouve

$$\partial_t Z(t,x) = \partial_t \mathrm{D}_x X(t,x) = \mathrm{D}_x \partial_t X(t,x).$$

Or, $\partial_t X(t,x) = b(X(t,x))$, et par dérivation de fonctions composées on trouve

$$\partial_t Z(t,x) = \mathrm{D}_x b(X(t,x))\mathrm{D}_x X(t,x) = \mathrm{D}_x b(X(t,x))Z(t,x).$$

Ainsi, pour tout $x\in\mathbb{R}^d$, la courbe $z_x : t\mapsto Z(\cdot,x)\in\mathcal{M}_d(\mathbb{R})$ vérifie le système d'équations différentielles ordinaires

$$\begin{cases}z_x'(t) = \mathrm{D}_x b(X(t,x))z_x(t) =: F_x(t,z_x(t))\\ z_x(0) = \mathrm{Id}_d.\end{cases}$$

où $F_x(t,z) = \mathrm{D}_x b(X(t,x))z$ est *linéaire* en z et lipschitzienne car $\|\mathrm{D}_x b\|\leq M$. On en déduit par Cauchy-Lipschitz l'existence d'une courbe $z_x\in\mathcal{C}^1(\mathbb{R})$ vérifiant le système

précédent, puis d'une fonction $Z : \mathbb{R}\times\mathbb{R}^d \to \mathcal{M}_d(\mathbb{R})$ admettant des dérivées partielles temporelles, et vérifiant

$$\begin{cases} \partial_t Z(t,x) = \mathrm{D}_x b(X(t,x))Z(t,x) & \forall (t,x) \in \mathbb{R}\times\mathbb{R}^d \\ Z(0,x) = \mathrm{Id}_d. & \forall x \in \mathbb{R}^d \end{cases}$$

Cette fonction Z étant maintenant construite, nous allons maintenant démontrer que $X(t,\cdot)$ est différentiable en et que $\mathrm{D}_x X(t,\cdot) = Z(t,\cdot)$. Pour cela, il faut montrer que $\|X(t,x+h) - (X(t,x) + Z(t,x)h)\| = \mathrm{o}(h)$ pour $h \in \mathbb{R}^d$. Posons

$$g(t) = \|X(t,x+h) - (X(t,x) + Z(t,x)h)\|.$$

Nous allons contrôler g en utilisant le lemme de Gronwall. Comme

$$\begin{cases} X(t,x) = X(0,x) + \int_0^t \partial_t X(s,x)\mathrm{d}s = x + \int_0^t b(X(s,x))\mathrm{d}s, \\ X(t,x+h) = x + h + \int_0^t b(X(s,x+h))\mathrm{d}s, \\ Z(t,x) = Z(0,x) + \int_0^t \partial_t Z(s,x)\mathrm{d}s = \mathrm{Id}_d + \int_0^t \mathrm{D}_x b(X(s,x))Z(s,x)\mathrm{d}s. \end{cases},$$

on en déduit que

$$\begin{aligned} g(t) &= \|X(t,x+h) - (X(t,x) + Z(t,x)h)\| \\ &= \left\| x + h - (x + \mathrm{Id}_d h) + \int_0^t (b(X(s,x+h)) - (b(X(s,x)) + \mathrm{D}_x b(X(s,x))Z(s,x)h)\mathrm{d}s \right\| \\ &\leq \int_0^t \|b(X(s,x+h)) - (b(X(s,x)) + \mathrm{D}_x b(X(s,x))Z(s,x)h)\| \,\mathrm{d}s \end{aligned}$$

On remarque de plus que

$$\begin{aligned} b(X(s,x+h)) &= b(X(s,x) + (X(s,x+h) - X(s,x))) \\ &= b(X(s,x)) + \mathrm{D}_x b(X(s,x))(X(s,x+h) - X(s,x)) + \mathrm{o}(\|X(s,x+h) - X(s,x)\|). \end{aligned}$$

Par l'inégalité de stabilité [1] $\|X(s,x+h) - X(s,x)\| \leq e^{M|s|}\|h\|$, le terme de reste dans l'égalité précédente est $\mathrm{o}(\|X(s,x+h) - X(s,x)\|) = e^{M|s|}\mathrm{o}(h)$. Ainsi,

$$\begin{aligned} g(t) &\leq \int_0^t \|\mathrm{D}_x b(X(s,x))(X(s,x+h) - X(s,x) - Z(s,x)h)\| \,\mathrm{d}s + e^{M|t|}t\mathrm{o(h)} \\ &\leq \|\mathrm{D}_x b\|_\infty \int_0^t g(s)\mathrm{d}s + (e^{M|t|}t)\cdot \mathrm{o(h)} \end{aligned}$$

1. Démonstration de cette inégalité : poser $e(s) = \frac{1}{2}\|X(s,x+h) - X(s,x)\|^2$. Alors

$$\begin{aligned} e'(s) &= \langle \partial_t X(s,x+h) - \partial_t X(s,x) | X(s,x+h) - X(s,x)\rangle \\ &= \langle b(X(s,x+h)) - b(X(s,x+h)) | X(s,x+h) - X(s,x)\rangle \\ &\leq M\|X(s,x+h) - X(s,x)\|^2 = 2Me(s) \end{aligned}$$

où l'inégalité vient du fait que b est M-Lipschitzienne combinée à Cauchy-Schwarz. Par le lemme de Gronwall différentiel (exercice 1.1 du premier chapitre), on en déduit $e(t) \leq e^{2M|t|}e(0)$, soit exactement $\|X(s,x+h) - X(s,x)\| \leq e^{M|s|}\|h\|$.

Ainsi, par lemme de Gronwall (Lemme 1.2),

$$g(t) \leq \exp(\|\mathrm{D}_x b\|_\infty |t|)\, |t| \cdot \mathrm{o}(h) = \mathrm{o}(h),$$

soit exactement $\|X(t, x+h) - (X(t,x) + Z(t,x)h)\| = g(t) = \mathrm{o}(h)$. Ceci montre que $X(t,\cdot)$ est différentiable et que $\mathrm{D}_x X(t,\cdot) = Z(t,x)$. Enfin, la fonction Z est continue [2], de sorte que que X admet des dérivées partielles spatiales continues. Comme $\partial_t X(t,x) = b(X(t,x))$ existe et est aussi continue, $X \in \mathcal{C}^1(\mathbb{R} \times \mathbb{R}^d, \mathbb{R}^d)$.

(iv) On sait que $X_t : x \mapsto X(t,x) \in \mathcal{C}^1(\mathbb{R}^d)$ par (iii). Comme de plus par (ii), $X_t \circ X_{-t} = X_0 = \mathrm{id}$, on en déduit que X_t est une bijection, dont l'inverse X_{-t} est de classe $\mathcal{C}^1$. Ainsi, X_t est un difféomorphisme. □

Théorème 3.2 (Existence et stabilité). *Si $u_0 \in \mathcal{C}^1(\mathbb{T}^d)$ et si $b \in \mathcal{C}^1(\mathbb{T}^d, \mathbb{R}^d)$, l'équation de transport* (3.1) *admet une unique solution régulière $u \in \mathcal{C}^1(\mathbb{R} \times \mathbb{T}^d)$, définie par la formule*

$$u(t,x) = u_0(X(-t,x)) + \int_0^t f(s, X(s-t,x))\mathrm{d}s.$$

En particulier, sur $[0,T] \times \mathbb{T}^d$,

$$\|u\|_\infty \leq \|u_0\|_\infty + T \|f\|_\infty .$$

Remarque 3.2 (Interprétation des solutions). La formule (3.2) permet de bien comprendre la solution de l'équation de transport :

- la valeur $u_0(x)$ est transportée le long du chemin $t \mapsto X(t,x)$.
- le long de chemin, la masse créee par la source f est accumulée, donnant le terme $\int f(s, X(s,t))\mathrm{d}s$.

Remarque 3.3 (Champs de vecteurs plus généraux). On s'est placé dans le cas le plus simple, à savoir celui où b est suffisamment régulier (globalement Lipschitz) pour avoir existence et unicité des trajectoires de l'EDO $y' = b(y)$. Il existe de nombreuses situations intéressantes où ça n'est pas le cas, notamment lorsque le champ de vecteurs dépend de la solution.

Démonstration. On commence par l'unicité : soit $u \in \mathcal{C}^1(\mathbb{R} \times \mathbb{T}^d)$ une solution classique de l'équation de transport (3.1). Alors,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t} u(t, X(t,x)) &= \partial_t u(t, X(t,x)) + \langle \nabla_x u(t, X(t,x)) | \partial_t X(t,x) \rangle \\ &= \partial_t u(t, X(t,x)) + \langle \nabla_x u(t, X(t,x)) | b(X(t,x)) \rangle = f(t, X(t,x)).\end{aligned}$$

On en déduit que

$$\begin{aligned}u(t, X(t,x)) &= u(0, X(0,x)) + \int_0^t f(s, X(s,x))\mathrm{d}s \\ &= u_0(x) + \int_0^t f(s, X(s,x))\mathrm{d}s. \qquad (3.2)\end{aligned}$$

2. Exercice : pour le démontrer, utiliser la Proposition 1.1

En utilisant la formule $X(s, X(-t,x)) = X(s-t,x)$, qui provient de la proposition précédente, on obtient

$$u(t,x) = u(t, X(t, X(-t,x))) = u_0(X(-t,x)) + \int_0^t f(s, \underbrace{X(s, X(-t,x))}_{=X(s-t,x)})\mathrm{d}s.$$

Réciproquement, on vérifie que la fonction u définie de cette manière est de classe $\mathcal{C}^1$ comme composée de fonction $\mathcal{C}^1$. De plus,

$$\partial_t u(t,x) = \langle \nabla u_0(X(-t,x)) | -\partial_t X(-t,x)\rangle + f(t,x) - \int_0^t \langle \nabla_x f(s, X(s-t,x)) | \partial_t X(s-t,x)\rangle \mathrm{d}s$$

$$\partial_{x_j} u(t,x) = \langle \nabla u_0(X(-t,x)) | \partial_{x_j} X(-t,x)\rangle + \int_0^t \langle \nabla_x f(s, X(s-t,x)) | \partial_{x_j} X(s-t,x)\rangle,$$

où $\partial_{x_j} X = (\partial_{x_j} X_i)_{1\le i\le d} \in \mathbb{R}^d$. Ainsi,

$$\begin{aligned}
&\partial_t u(t,x) + \langle b(x) | \nabla u(t,x)\rangle \\
&\quad = f(t,x) + \langle \nabla u_0(X(-t,x)) | - \partial_t X(-t,x) + \sum_j b_j(x) \partial_{x_j} X(-t,x)\rangle \\
&\qquad + \int_0^t \langle \nabla_x f(s, X(s-t,x)) | - \partial_t X(s-t,x) + \sum_j b_j(x) \partial_{x_j} X(s-t,x)\rangle \mathrm{d}s
\end{aligned}$$

Pour montrer que le second membre de cette égalité est égal à $f(t,x)$, il suffit donc de démontrer que pour tout $u \in \mathbb{R}$,

$$-\partial_t X(u,x) + \sum_j b_j(x) \partial_{x_j} X(u,x) = 0$$

En dérivant la relation $X(u,y) = X(u-t, X(u+t,y))$ par rapport à t, on a

$$\begin{aligned}
0 &= -\partial_t X(u-t, X(t,y)) + \sum_j \partial_{x_j} X(u-t, X(u+t,y)) \cdot \partial_t X_j(u-t, X(u+t,y)) \\
&= -\partial_t X(u-t, X(u+t,y)) + \sum_j b_j(X(u+t,y)) \partial_{x_j} X(u-t, X(u+t,y)),
\end{aligned}$$

en prenant $t=0$, et en posant $y = X(-u,x)$ ou de manière équivalente $x = X(u,y)$,

$$0 = -\partial_t X_j(u,x) + \sum_j \langle \nabla X_j(u,x) | b_j(x)\rangle.$$

Ainsi, $\partial_t u + \langle b | \nabla u\rangle = 0$ et u est bien solution de (3.1). □

3.2 Schéma décentré amont pour l'équation de transport

Dans cette section, on s'intéresse à l'équation de transport en dimension 1, sur $[0,T] \times \mathbb{T}$. On supposera toujours que le champ de vecteur b appartient à $\mathcal{C}^1(\mathbb{T}, \mathbb{R})$. Pour un nombre de pas de temps $N \in \mathbb{N}^*$ et un nombre de pas d'espace $M \in \mathbb{N}^*$, on définit

- $\tau = T/N$ le pas de temps, $t^n = n\tau$ et $\overline{I}_\tau = \{t^n \mid n \in [\![0, N]\!]\}$.
- $h = 1/M$ le pas d'espace, $x_j = jh$ et $\mathbb{T}_h = h\mathbb{Z}/\mathbb{Z} \simeq_{\text{bij}} \{x_j \mid j \in [\![1, M]\!]\}$.

Comme précédemment, $F(\overline{I}_\tau \times \mathbb{T}_h) \simeq \mathbb{R}^{(N+1)\times M}$ désigne l'espace de dimension finie dans lequel on construira les solutions de l'équation de transport discrétisées. On considèrera dans le cours un unique schéma, appelé *schéma décentré amont*. Le nom de ce schéma est dû à la manière dont la dérivée spatiale $\frac{\partial u}{\partial x}(t,x)$ est discrétisée :

- Si $b(x) \geq 0$, on utilise une différence finie à gauche

$$\frac{\partial u}{\partial x}(t,x) \simeq \frac{u(t,x) - u(t,x-h)}{h}$$

- Si $b(x) \leq 0$, on utilise une différence finie à droite

$$\frac{\partial u}{\partial x}(t,x) \simeq \frac{u(t,x+h) - u(t,x)}{h}$$

Définition 3.3 (Schéma décentré amont)**.** Une fonction $v \in F(\overline{I}_\tau \times \mathbb{T}_h)$ est solution du schéma décentré amont si

$$\begin{cases} \frac{v(t+\tau,x)-v(t,x)}{\tau} + b(x)\dfrac{v(t,x) - v(t,x-h)}{h} = f(t,x) & \forall (t,x) \in \overline{I}_\tau \setminus \{T\} \times \mathbb{T}_h \text{ tq } b(x) \leq 0 \\ \frac{v(t+\tau,x)-v(t,x)}{\tau} + b(x)\dfrac{v(t,x+h) - v(t,x)}{h} = f(t,x) & \forall (t,x) \in \overline{I}_\tau \setminus \{T\} \times \mathbb{T}_h \text{ tq } b(x) \geq 0 \\ v(0,\cdot) = u_0 \end{cases} \tag{3.3}$$

Proposition 3.3 (Stabilité)**.** *Le schéma décentré amont* (3.3) *est stable en norme infinie sous la condition CFL* $\|b\|_\infty \frac{\tau}{h} \leq 1$ *: si* v *vérifie* (3.3)*, alors*

$$\|v\|_\infty \leq \|u_0\|_\infty + T \|f\|_\infty$$

Démonstration. Soit x tel que $b(x) \leq 0$, et posons $\lambda(x) = -\frac{\tau}{h} b(x) \geq 0$. Notons que la condition CFL montre que $0 \leq \lambda(x) \leq 1$. Alors,

$$\begin{aligned} v(t+\tau,x) &= v(t,x) + \frac{\tau}{h} b(x)(v(t,x) - v(t,x-h)) + \tau f(t,x) \\ &= (1 - \lambda(x))\, v(t,x) + \lambda(x) v(t,x-h) + \tau f(t,x) \\ &\leq \|v(t,\cdot)\|_\infty + \tau \|f(t,\cdot)\|_\infty \end{aligned}$$

De même, on peut montrer que

$$v(t+\tau,x) \geq -(\|v(t,\cdot)\|_\infty + \tau \|f(t,\cdot)\|_\infty).$$

On obtient les mêmes majorations en x tel que $b(x) \geq 0$ en posant $\lambda(x) = \frac{\tau}{h} b(x) \in [0,1]$. Ainsi,

$$\|v(t+\tau,\cdot)\|_\infty \leq \|v(t,\cdot)\|_\infty + \tau \|f(t,\cdot)\|_\infty .$$

L'estimation de stabilité se déduit alors d'une simple récurrence. □

Proposition 3.4 (Consistance). *Le schéma décentré amont* (3.3) *est consistant et d'ordre* 1 *en temps et en espace : si u est solution de l'équation de transport* (3.1) *et si $u \in \mathcal{C}^2(\mathbb{R}\times\mathbb{T})$, alors la fonction $\varepsilon_{u,\tau,h} : \overline{I}_\tau \setminus \{T\} \times \mathbb{T}_h \to \mathbb{R}$ définie par*

$$\varepsilon_{u,\tau,h}(x,t) := \begin{cases} \frac{u(t+\tau,x)-u(t,x)}{\tau} + b(x)\dfrac{u(t,x)-u(t,x-h)}{h} - f(t,x) & \text{si } b(x) \geq 0 \\ \frac{u(t+\tau,x)-u(t,x)}{\tau} + b(x)\dfrac{u(t,x+h)-u(t,x)}{h} - f(t,x) & \text{si } b(x) \leq 0 \end{cases},$$

vérifie $\|\varepsilon_{u,\tau,h}\|_\infty \leq \text{const}(u)\cdot(\tau+h)$.

Démonstration. Laissée en exercice. □

Corollaire 3.5 (Convergence). *Supposons que l'équation de transport* (3.1) *admette une solution $u \in \mathcal{C}^2(\mathbb{R}\times\mathbb{T})$. Alors, si v est solution du schéma décentré amont* (3.3),

$$\left\| u|_{\overline{I}_\tau\times\mathbb{T}_h} - v \right\| \leq \text{const}(u)\cdot(\tau+h).$$

Démonstration. Par définition de l'erreur de consistance $\varepsilon_{u,\tau,h}$, la fonction $\hat{u} = u|_{\overline{I}_\tau\times\mathbb{T}_h}$ est solution du schéma décentré amont (3.3) avec comme second membre $f + \varepsilon_{u,\tau,h}$ et condition initiale $\hat{u}(0,\cdot) = u_0$. Par hypothèse, v est solution de (3.3) avec comme second membre f et condition initial $v(0,\cdot) = u_0$. Ainsi, $w = \hat{u} - v$ est solution de (3.3) avec comme second membre $\varepsilon_{u,\tau,v}$ et condition initiale $w(0,\cdot) = 0$. Par stabilité (Proposition 3.3),

$$\|w\|_\infty \leq \underbrace{\|w(0,\cdot)\|_\infty}_{=0} + T\,\|\varepsilon_{u,\tau,h}\|_\infty .$$

Or, par consistance (Proposition 3.4), $\|\varepsilon_{u,\tau,h}\|_\infty \leq \text{const}(u)\cdot(\tau+h)$. Ainsi,

$$\|\hat{u} - v\|_\infty = \|w\|_\infty \leq T\text{const}(u)\cdot(\tau+h).$$ □

3.3 Exercices

Exercice 3.1. *[Transport à vitesse constante]* Soit $c \in \mathbb{R}$ une constante, $u_0 \in \mathcal{C}^1(\mathbb{T})$ et $u \in \mathcal{C}^1([0,T]\times\mathbb{T})$ la solution de l'équation

$$\begin{cases} \partial_t u(t,x) + c\partial_x u(t,x) & \forall (t,x) \in [0,T]\times\mathbb{T}, \\ u(0,\cdot) = u_0 \end{cases} \tag{3.4}$$

1. Expliciter la solution de l'équation (3.4).
2. *[Schémas classiques]* On suppose que M est pair, $h = 1/M$ et on pose que $u_0(jh) = (-1)^j$. Calculer la solution des schémas décentrés à gauche, à droite et du centré pour cette condition initiale : $v^{\mathrm{g}}(0,\cdot) = v^{\mathrm{d}}(0,\cdot) = v^{\mathrm{c}}(0,\cdot) = u_0$ et

$$\frac{v^{\mathrm{g}}(t+\tau,x)-v^{\mathrm{g}}(t,x)}{\tau} + c\frac{v^{\mathrm{g}}(t,x)-v^{\mathrm{g}}(t,x-h)}{h} = 0$$
$$\frac{v^{\mathrm{d}}(t+\tau,x)-v^{\mathrm{d}}(t,x)}{\tau} + c\frac{v^{\mathrm{d}}(t,x+h)-v^{\mathrm{d}}(t,x)}{h} = 0$$
$$\frac{v^{\mathrm{c}}(t+\tau,x)-v^{\mathrm{c}}(t,x)}{\tau} + c\frac{v^{\mathrm{c}}(t,x+h)-v^{\mathrm{c}}(t,x-h)}{2h} = 0$$

En déduire que le schéma aval (i.e. à droite si $c \geq 0$ et à gauche si $c \leq 0$) et le schéma décentré sont instables.

Exercice 3.2. *[Schémas monotones]* Dans cet exercice, qui réutilise les notations et hypothèses du précédent, on s'intéresse à un schéma général linéaire, de la forme $v(t+\tau,\cdot) = Av(t,\cdot)$ où $A : F(\mathbb{T}_h) \to F(\mathbb{T}_h)$ est défini par

$$Au(x) = \sum_{\ell \in [\![-L,L]\!]} \gamma(\ell) u(x+\ell h), \qquad \gamma \in F([\![-L,L]\!]).$$

On suppose que $M > 2L$, de sorte que $2L \not\equiv 0 \mathrm{mod} M$.

1. *[Monotonie]* Montrer que le schéma est monotone (i.e. si $u, v \in F(\mathbb{T}_h)$ vérifient $u \geq v$, alors $Au \geq Av$) si et seulement si $\gamma \geq 0$.
2. *[Ordre]* On suppose que pour $i \in [\![0,k]\!]$,

$$\sum_{\ell \in [\![-L,L]\!]} \gamma(\ell)\ell^i = \left(-c\frac{\tau}{h}\right)^i. \tag{3.5}$$

Montrer alors que pour toute fonction fonction $u \in \mathcal{C}^{k+1}([0,T]\times\mathbb{T})$ vérifiant (3.4) et pour tout $(t,x) \in (\overline{I}_\tau \setminus \{T\}) \times \mathbb{T}_h$,

$$|Au(t,x) - u(t+h,x)| \leq \mathrm{const}(k,L) \cdot \left\|\mathrm{D}^{k+1}u\right\|_\infty \cdot (\tau^{k+1} + h^{k+1}).$$

(Indication : écrire les développements de Taylor et utiliser $\partial_t u = -c\partial_x u$.)

3. *[Stabilité en norme infinie]* Montrer que si le schéma vérifie (3.5) pour $i = 0$, alors il préserve les constante (i.e. si $u \equiv C \in \mathbb{R}$, $Au \equiv C$). En déduire la stabilité en norme infinie : $\forall u \in F(\mathbb{T}_h)$, $\|Au\|_\infty \leq \|u\| \infty$.
4. *[Théorème de Godunov]* Montrer que s'il existe schéma monotone vérifiant la relation (3.5) pour $0 \leq i \leq 2$, alors $-c\frac{\tau}{h} \in [\![-L,L]\!]$.
(Indication : poser $\alpha = -c\frac{\tau}{h}$ et calculer $\sum_{\ell\in[\![-L,L]\!]}(\ell-\alpha)^2\gamma(\ell)$).

Exercice 3.3. *[Schéma de Lax-Wendroff]* On garde les notations des exercices précédents et on considère le schéma de Lax-Wendroff :

$$\frac{v^{\mathrm{lw}}(t+\tau,x) - v^{\mathrm{lw}}(t,x)}{\tau} + c\frac{v^{\mathrm{lw}}(t,x+h) - v^{\mathrm{lw}}(t,x-h)}{2h} - \frac{c^2\tau}{2}\Delta_h v^{\mathrm{lw}}(t,x) = 0$$

1. *[Ordre]* Montrer que le schéma est d'ordre 2 en espace et en temps.
2. *[Monotonie]* Montrer que le schéma n'est monotone que si $\frac{\tau c}{h} = 1$.
3. *[Stabilité L^2]* Nous montrons maintenant la stabilité L^2 du schéma, en introduisant la norme et le produit scalaire suivants sur $F(\mathbb{T}_h)$:

$$\langle u|v\rangle_h = h\sum_{x\in\mathbb{T}_h} u(x)v(x), \quad \|u\|_h = \sqrt{\langle u|v\rangle_h}.$$

   a) Mettre le schéma de Lax-Wendroff sous la forme $v^{\mathrm{lw}}(t+\tau,\cdot) = Av^{\mathrm{lw}}(t,\cdot)$ où $A : F(\mathbb{T}_h) \to F(\mathbb{T}_h)$ est à déterminer, puis montrer que si l'on pose $u_k(x) = e^{2i\pi kx}$, alors $Au_k = \lambda_k u_k$. Calculer les coefficients λ_k en fonction de $\alpha = c\frac{\tau}{h}$.

b) Montrer que si $|\alpha| \leq 1$, alors $|\lambda_k| \leq 1$.
c) Conclure que $\left\|v^{\mathrm{lw}}(t+\tau,\cdot)\right\|_h \leq \left\|v^{\mathrm{lw}}(t,\cdot)\right\|_h$.

Exercice 3.4. *Équation de Hamilton-Jacobi.* Dans cet exercice nous donnons quelques éléments d'étude théorique et numérique d'une équation modélisant une propagation de front. Étant donnée $u_0 : \mathbb{T} \to \mathbb{R}$, il s'agit de trouver $u : [0,T] \times \mathbb{T} \to \mathbb{R}$ vérifiant

$$\begin{cases} \partial_t u(t,x) = |\partial_x u(t,x)| \quad \forall (t,x) \in [0,T] \times \mathbb{T}, \\ u(0,\cdot) = u_0. \end{cases} \tag{3.6}$$

1. *[Solutions régulières]* Nous supposons dans cette question que l'équation (3.6) possède une solution $u \in \mathcal{C}_1^2([0,T] \times \mathbb{T})$.
   a) Soit $\gamma \in \mathcal{C}^1([0,t],\mathbb{R})$ $(t \leq T)$ une courbe vérifiant $|\gamma'|_\infty \leq 1$. Montrer que $\frac{\mathrm{d}}{\mathrm{d}t} u(t,\gamma(t)) \geq 0$, et en déduire que si $\|x-y\| \leq t$, alors $u(t,x) \geq u_0(y)$.
   b) En considérant le chemin $\gamma_\varepsilon : [0,t] \to \mathbb{R}$ défini par

$$\begin{cases} \gamma_\varepsilon(t) = x \\ \gamma_\varepsilon'(t) = -\frac{\nabla u(t,\gamma_\varepsilon(t))}{\varepsilon + \|\nabla u(t,\gamma_\varepsilon(t))\|}, \end{cases}$$

   et $y_\varepsilon = \gamma_\varepsilon(0)$, montrer qu'il existe y tel que $\|x-y\| \leq t$ et $u(t,x) = u_0(y)$.
   c) Conclure que pour tout $(t,x) \in [0,T] \times \mathbb{R}^d$, $u(t,x) = \max_{y \in B(x,t)} u_0(y)$.

Étant donné un couple pas de temps/d'espace $(\tau, h) = (\frac{T}{N}, \frac{1}{M+1})$, nous considérons le schéma explicite suivant, dont $v \in F(\overline{I}_\tau \times \mathbb{T}_h)$ est solution si $v(0,\cdot) = u_0$ et si

$$\begin{aligned} &\forall t \in (\overline{I}_\tau \setminus \{T\}), \quad \frac{v(t+\tau,\cdot) - v(t,\cdot)}{\tau} = H_h v(t,\cdot) \\ &\text{où } H_h v(t,x) = \max\left(0, \frac{v(t,x+h) - v(t,x)}{h}, \frac{v(t,x-h) - v(t,x)}{h}\right) \end{aligned} \tag{3.7}$$

2. *([Monotonie et stabilité.]* Mettre le schéma sous la forme $v(t+h,\cdot) = A_{\tau,h}(v(t,\cdot))$, où $A_{\tau,h} : F(\mathbb{T}_h) \to F(\mathbb{T}_h)$. Montrer que l'opérateur $A_{\tau,h}$ est monotone et préserve les constantes (au sens de l'exercice 3.2), et en déduire que

$$\left\|A_{\tau,h} u\right\|_\infty \leq \|u\|_\infty .$$

3. *[Consistance]* Soit $u \in \mathcal{C}^2([0,T] \times \mathbb{T})$ une solution de (3.6), montrer que pour tout $(t,x) \in [0, T-\tau] \times \mathbb{T}$ on a

$$\left| \frac{u(t+\tau,x) - u(t,x)}{\tau} - H_h u(t,x) \right| \leq \mathrm{O}(\tau + h)$$

Noter que comme l'équation (3.6) *n'est pas linéaire (et de plus, n'admet en général pas de solutions régulières), on ne peut* pas *en déduire directement la convergence des solutions du schéma vers celle de l'équation. L'étude de convergence repose sur la notion de solution de viscosité pour l'équation 3.6, introduite par Crandall et Lions.*

3.4 Correction des exercices

Correction de l'exercice 3.2.
1. Si $\gamma \geq 0$ et si $u \geq v$, alors

$$Au(x) = \sum_{\ell \in [\![-L,L]\!]} \underbrace{\gamma(\ell)}_{\geq 0} \underbrace{u(x+\ell h)}_{\geq v(x+\ell h)} \geq Av(x).$$

Réciproquement, supposons que $u \geq v \Longrightarrow Au \geq Av$. Fixons $\ell \in [\![-L, L]\!]$ et définissons

$$u(x) = \begin{cases} 1 & si\ x = \ell h + z, z \in \mathbb{Z} \\ 0 & sinon \end{cases}$$

et $v = \frac{1}{2}u$. Alors

$$Au(0) = \gamma(\ell) \geq Av(0) = \frac{1}{2}\gamma(\ell).$$

On en déduit que $\gamma(\ell) \geq 0$.
2. On écrit les développement de Taylor à l'ordre $k+1$: il existe $\xi_t \in [0, \tau]$ et $\xi_\ell \in [-\ell h, \ell h]$ pour tout $\ell \in [\![-L, L]\!]$ tels que

$$u(t, x+\ell h) = \sum_{i=0}^{k} \partial_x^{(i)} u(t,x) \frac{(\ell h)^i}{i!} + \frac{(\ell h)^{k+1}}{(k+1)!} \partial_x^{(k+1)} u(t, x+\xi_\ell)$$

$$u(t+\tau, x) = \sum_{i=0}^{k} \partial_t^{(i)} u(t,x) \frac{\tau^i}{i!} + \frac{\tau^{k+1}}{(k+1)!} \partial_t^{(k+1)} u(t+\xi_t, x)$$

Comme de plus pour tout $i \in \{0, \dots, k\}$, $\partial_t^{(i)} u = (-c)^i \partial_x^{(i)} u$, on en déduit que

$$\begin{aligned} Au(t,x) - u(t+\tau,x) &= \sum_{\ell \in [\![-L,L]\!]} \gamma(\ell) u(t, x+\ell h) - u(t+\tau, x) \\ &= \sum_{i=0}^{k} \left[\left(\sum_{\ell \in [\![-L,L]\!]} \gamma(\ell)(\ell h)^i \right) - (-c)^i (\tau)^i \right] \frac{\partial_x^{(i)} u(t,x)}{i!} + \mathrm{O}(\tau^{k+1} + h^{k+1}) \end{aligned}$$

Par hypothèse, le terme devant $\frac{\partial_x^{(i)} u(t,x)}{i!}$ s'annule, ce qui démontre que $|Au(t,x) - u(t+\tau,x)| \leq O(\tau^{k+1} + h^{(k+1)})$. Le terme en O est de la forme

$$\frac{1}{k!} \left[\left(\sum_{\ell \in [\![-L,L]\!]} |\gamma(\ell)| \right) \left\| \partial_x^{(k+1)} u \right\|_\infty h^{k+1} + \left\| \partial_t^{(k+1)} u \right\|_\infty \tau^{k+1} \right].$$

3. Il suffit de remarquer que sous l'hypothèse $\sum_\ell \gamma = 1$, si $u \equiv C$, alors

$$Au(x) = \sum_{\ell \in [\![-L,L]\!]} \gamma(\ell) C = C,$$

Pour en déduire la stabilité en norme infinie, on utilise $-\|u\|_\infty \le u \le \|u\|_\infty$, *de sorte que par monotonie,* $-A\|u\|_\infty \le Au \le \|u\|_\infty$ *et par préservation des constantes* $-\|u\|_\infty \le Au \le \|u\|_\infty$, *soit* $\|Au\|_\infty \le \|u\|_\infty$.
4. En posant $\alpha = -c\frac{\tau}{h}$, *et en utilisant le fait que le schéma est d'ordre* ≥ 2,

$$\sum_{\ell\in[\![-L,L]\!]} (\ell-\alpha)^2\gamma(\ell) = \underbrace{\sum_{\ell\in[\![-L,L]\!]} \ell^2\gamma(\ell)}_{=\alpha^2} + \alpha^2 \underbrace{\sum_{\ell\in[\![-L,L]\!]} \gamma(\ell)}_{=1} - 2\alpha \underbrace{\sum_{\ell} \ell\gamma(\ell)}_{=\alpha} = 0.$$

Comme chacun des termes de la somme du membre de gauche est positif, on en déduit que $\forall \ell \in [\![-L,L]\!](\ell-\alpha)^2\gamma(\ell) = 0$. *Ainsi, si* $\gamma(\ell) > 0$, *alors* $\alpha = \ell \in [\![-L,L]\!]$.

Correction de l'exercice 3.3.
1. En posant $v = v^{\mathrm{lw}}$,

$$\begin{aligned} v(t+\tau,x) &= v(t,x) - \frac{\tau c}{2h}(v(t,x+h)-v(t,x-h)) + \frac{c^2\tau^2}{2h^2}(v(t,x+h)-2v(t,x)+v(t,x-h)) \\ &= \underbrace{\left(\frac{\tau c}{2h} + \frac{c^2\tau^2}{2h^2}\right)}_{\gamma(-1)} v(t,x-h) + \underbrace{\left(1 - c^2\frac{\tau^2}{h^2}\right)}_{\gamma(0)} v(t,x) + \underbrace{\left(-\frac{\tau c}{2h} + \frac{c^2\tau^2}{2h^2}\right)}_{\gamma(1)} v(t,x+h) \end{aligned}$$

Ainsi,

$$\sum_\ell \gamma(\ell) = \gamma(-1)+\gamma(0)+\gamma(1) = 1$$
$$\sum_\ell \ell\gamma(\ell) = \gamma(1)-\gamma(-1) = -\frac{\tau c}{h}$$
$$\sum_\ell \ell^2\gamma(\ell) = \gamma(1)+\gamma(-1) = \frac{c^2\tau^2}{h^2}$$

Par le résultat de l'exercice précédent, le schéma est donc d'ordre 2.
2. En utilisant à nouveau l'exercice précédent, le schéma de Lax-Wendroff est monotone si $\gamma \ge 0$ *soit*

$$\begin{cases} 1 - c^2\frac{\tau^2}{h^2} \ge 0 \\ -\frac{\tau c}{2h} + \frac{c^2\tau^2}{2h^2} \ge 0 \\ \frac{\tau c}{2h} + \frac{c^2\tau^2}{2h^2} \ge 0 \end{cases}$$

Les deuxième et troisièmes lignes donnent

$$c^2\frac{\tau^2}{h^2} \ge \left|\frac{\tau c}{h}\right|,$$

soit $\frac{\tau|c|}{h} \ge 1$.
3.a. b. Calculons l'effet du schéma sur un mode de Fourier $u_k(x) = e^{2i\pi kx}$. *Remarquons d'abord que* $u_k(x\pm h) = e^{\pm 2i\pi kh}u_k(x)$. *Ainsi, en posant* $\alpha = \frac{\tau c}{h}$,

$$\begin{aligned} Au_k(x) &= u_k(x) - \frac{\alpha}{2}(u_k(x+h)-u_k(x-h)) + \frac{\alpha^2}{2}(u_k(x+h)-2u_k(x)+u_k(x-h)) \\ &= \left(1 - \frac{\alpha}{2}(e^{2i\pi kh}-e^{-2i\pi kh}) + \frac{\alpha^2}{2}(e^{2i\pi kh}+e^{-2i\pi kh}-2)\right)u_k(x), \end{aligned}$$

soit $Au_k = \lambda_k u_k$ *avec*

$$\begin{aligned}\lambda_k &= 1 - \frac{\alpha}{2}(e^{2i\pi kh} - e^{-2i\pi kh}) + \frac{\alpha^2}{2}(e^{2i\pi kh} + e^{-2i\pi kh} - 2) \\ &= 1 - 4i\frac{\alpha}{2}\sin(\pi kh)\cos(\pi kh) - 4\frac{\alpha^2}{2}\sin(\pi kh)^2\end{aligned}$$

car

$$\begin{aligned}e^{2i\pi kh} + e^{-2i\pi kh} - 2 &= 2(\cos(2\pi kh) - 1) = -4\sin^2(\pi kh) \\ e^{2i\pi kh} - e^{-2i\pi kh} &= 2i\sin(2\pi kh) = 4i\sin(\pi kh)\cos(\pi kh)\end{aligned}$$

Ainsi,

$$\begin{aligned}|\lambda_k|^2 &= (1 - 2\alpha^2\sin^2(\pi kh))^2 + 4\alpha^2\sin(\pi kh)^2\cos(\pi kh)^2 \\ &= 1 + 4\alpha^4\sin(\pi kh)^4 - 4\alpha^2\sin^2(\pi kh) + 4\alpha^2\sin^2(\pi kh)(1 - \sin^2(\pi kh)) \\ &= 1 + 4(\alpha^4 - \alpha^2)\sin^4(\pi kh)\end{aligned}$$

Ainsi, si $|\alpha| \leq 1$, $\alpha^4 \leq \alpha^2$ *et* $|\lambda|_k \leq 1$.
c. Remarque : comme l'opérateur A *n'est pas auto-adjoint, on sait a priori seulement que* $\|A\| \leq \rho(A)$ *ou* $\rho(A)$ *est le rayon inverse. Pour avoir l'inégalité inverse, nous utilisons que* A *est diagonalisable dans la base des* u_k*, qui est orthogonale. Plus précisément, on considère le produit scalaire hermitien sur* $F(\mathbb{T}_h, \mathbb{C})$

$$\langle u|v\rangle_h = h\sum_{x\in\mathbb{T}_h}\overline{u}(x)v(x) = h\sum_{0\leq i\leq M-1}\overline{u}(ih)v(ih),$$

et $\|\cdot\|_h$ *la norme associée. et on remarque que si* $k \neq k'$*, alors*

$$\begin{aligned}\sum_{x\in\mathbb{T}_h} u_k(x)\overline{u_{k'}(x)} &= \sum_{0\leq j\leq M-1} e^{2i\pi(k-k')jh} \\ &= \sum_{0\leq j\leq M-1}(e^{2i\pi(k-k')h})^j \\ &= \frac{1 - e^{2i\pi(k-k')hM}}{1 - e^{2i\pi hM}} = 0.\end{aligned}$$

Ainsi, la famille $(u_k)_{0\leq k\leq M-1}$ *est orthogonale. On en déduit que pour tout coefficients* $\alpha_1, \ldots, \alpha_{M-1} \in \mathbb{C}$*, et tout* $u = \sum_k \alpha_k u_k$*,*

$$\begin{aligned}\|Au\|_h^2 &= \left\|\sum_k \lambda_k\alpha_k u_k\right\|_h^2 \\ &= \sum_k |\lambda_k|^2 |\alpha_k|^2 \|u_k\|_h^2 \\ &\leq \sum_k |\alpha_k|^2 \|u_k\|_h^2 = \|u\|_h^2.\end{aligned}$$

En particulier, l'inégalité précédente est vraie si $u \in F(\mathbb{T}_h, \mathbb{R})$.

Chapitre 4

Méthode des éléments finis pour les problèmes variationnels

4.1 Problèmes variationnels et leurs approximations

Définition 4.1. Soit V un espace de Hilbert et $a : V \times V \to \mathbb{R}$ une forme bilinéaire continue sur V. On dit que a est coercive si et seulement si

$$\exists \alpha > 0, \quad \forall v \in V, \quad |a(v,v)| \geq \alpha \|v\|^2 .$$

et qu'elle est continue (ou bornée) si

$$\exists M > 0, \quad \forall u, v \in V, \quad |a(u,v)| \leq M \|u\| \|v\| .$$

Théorème 4.1 (Lax-Milgram)**.** *Soit a une forme bilinéaire continue et coercive sur un espace de Hilbert V. Alors, pour toute forme linéaire continue ℓ, il existe un unique vecteur $u \in V$ tel que*

$$\forall v \in V, a(u,v) = \ell(v). \tag{4.1}$$

De plus, on a l'estimation de stabilité suivante

$$\|u\| \leq \frac{1}{\alpha} \|\ell\|_{V'} .$$

Démonstration. Pour tout $u \in V$, $v \in V \mapsto a(u,v)$ est une forme linéaire continue, de sorte que par théorème de Riesz, il existe un unique $A(u) \in V$ tel que $a(u,v) = \langle A(u)|v\rangle$ pour tout $v \in V$. De plus, l'application $u \mapsto A(u)$ est linéaire continue (exercice). Soit $f \in V$ tel que $\ell = \langle f|\cdot\rangle$. Alors, u est solution de (4.1) si et seulement si $A(u) = f$. Il suffit donc de montrer que l'opérateur A est bijectif.

Soit $T(w) := w - \mu(A(w) - f)$. On va montrer que pour μ bien choisi, l'opérateur T est (strictement) contractant, et possède donc un unique point fixe, ce qui établit l'existence et l'unicité du u tel que $A(u) = f$. On note α la constante de coercivité

et M la norme d'opérateur de a, de sorte que

$$\begin{aligned}\|T(v) - T(w)\|^2 &= \|v - w - \mu A(v-w)\|^2 \\ &= \|v-w\|^2 - 2\mu\langle v-w|A(v-w)\rangle + \mu^2 \|A(v-w)\|^2 \\ &\geq \|v-w\|^2 - 2\mu\alpha \|v-w\|^2 + \mu^2 \|A\|_{\text{op}} \|v-w\|^2 \\ &= (1 - 2\mu\alpha + \mu^2 M) \|v-w\|^2,\end{aligned}$$

et en effet T est contractant dès que $\mu > 0$ est choisi suffisamment petit, et on conclut par théorème du point fixe contractant.

Soit u la solution de $A(u) = f$. Alors,

$$\alpha \|u\|^2 \leq a(u,u) = \langle f|u\rangle \leq \|f\| \|u\| = \|\ell\|_{V'} \|u\| \qquad \square$$

Proposition 4.2. *Si a est symétrique, alors u est solution du problème variationnel (4.1) si et seulement si u est le minimum global de $J(v) = \frac{1}{2}a(v,v) - \ell(v)$.*

Démonstration. En effet,

$$\begin{aligned}J(u+tv) - J(u) &= \frac{1}{2}t^2 a(v,v) + \frac{1}{2}ta(u,v) + \frac{1}{2}ta(v,u) - t\ell(v) \\ &= t(a(u,v) - \ell(v)) + \frac{t^2}{2}a(v,v),\end{aligned}$$

de sorte que si u est minimum global de J, alors u est solution du problème variationnel. La réciproque est aussi vraie, car si u est solution du problème variationnel, alors $J(u+tv) - J(u) = t^2/2a(v,v) \geq 0$ et u est minimum global de J. $\square$

4.1.1 Approximation dans un sous-espace

Pour approcher le problème variationnel (4.1) on considère un sous-espace vectoriel $V_h \subseteq V$ de dimension finie indexés par un paramètre $h \in \mathbb{R}$ qui a vocation à tendre vers zéro. Le problème variationnel discret consiste à trouver une fonction $u_h \in V_h$ tel que

$$\forall v_h \in V_h, a(u_h, v_h) = \ell(v_h). \tag{4.2}$$

Comme la forme bilinéaire a reste coercive sur V_h, il existe une unique solution u_h à ce problème. On a de plus la caractérisation suivante :

Proposition 4.3 (Lemme d'orthogonalité)**.** *Si u est la solution de (4.1) et u_h la solution de (4.2), alors*

$$\forall v_h \in V_h, a(u - u_h, v_h) = 0.$$

En particulier, si a est symétrique, alors u_h est la projection orthogonale de u sur V_h, au sens du produit scalaire $\langle\cdot|\cdot\rangle_a := a(\cdot,\cdot)$.

Démonstration. Pour tout $v_h \in V_h$, $a(u_h, v_h) = \ell(v_h) = a(u, v_h)$. $\square$

Notation (Distance à un ensemble). Si $V' \subseteq V$ est un sous-espace fermé de l'espace de Hilbert V et si $v \in V$, on note $\mathrm{d}(v, V') = \min_{v' \in V'} \|v - v'\|$.

Le lemme de Céa permet d'estimer la distance entre la solution u_h du problème variationnel discret (4.2) et la solution u du problème variationnel (4.1) en fonction de la distance entre la solution u et le sous-espace V_h.

Théorème 4.4 (Lemme de Céa)**.** *Si a est α-coercive et bornée par M, que u est une solution de* (4.1) *et u_h est une solution de* (4.2)*, alors*

$$\|u - u_h\| \leq \frac{M}{\alpha} \mathrm{d}(u, V_h).$$

Si de plus a est symétrique, alors

$$\|u - u_h\| \leq \sqrt{\frac{M}{\alpha}} \mathrm{d}(u, V_h)$$

Démonstration. Soit $v_h \in V_h$. Alors, comme $v_h - u_h \in V_h$, en utilisant la relation d'orthogonalité, on a

$$\begin{aligned} & 0 = a(u_h - u, v_h - u_h) = a(u_h - u, v_h - u + u - u_h) = 0 \\ \text{soit} \quad & a(u_h - u, u_h - u) = a(u_h - u, v_h - u) \end{aligned}$$

Comme a est coercive et bornée,

$$\alpha \|u - u_h\|^2 \leq a(u_h - u, u_h - u) = a(u_h - u, v_h - u) \leq M \|u_h - u\| \|v_h - u\|.$$

Ainsi,

$$\|u - u_h\|^2 \leq \frac{M}{\alpha} \min_{v_h \in V_h} \|v_h - u\|$$

Dans le cas symétrique, u_h est la projection de u sur V_h au sens de $\|u\|_a := \sqrt{a(u,u)}$. On obtient alors

$$\frac{1}{\alpha} \|u - u_h\|^2 \leq \|u - u_h\|_a^2 = \inf_{v_h \in V_h} \|u - v_h\|^2 \leq M \inf_{v_h \in V_h} \|u - v_h\|^2. \qquad \square$$

Exemple 4.1 (Méthode de Galerkin). Soit V un espace de Hilbert séparable et $(e_i)_{i \in \mathbb{N}}$ une base hilbertienne. Pour $h_n = 1/n$, soit V_{h_n} le sous-espace vectoriel engendré par $e_1, \ldots, e_n$. Soit u la solution de (4.1). Alors,

$$\mathrm{d}(u, V_{h_n})^2 \leq \left\| u - \sum_{1 \leq i \leq n} \langle u | e_i \rangle e_i \right\|^2 \xrightarrow{n \to \infty} 0,$$

de sorte que par Lemme de Céa, on obtient la convergence de la solution u_{h_n} vers u_h lorsque $n \to \infty$.

Intéressons nous maintenant au problème (4.2). On cherche la solution u_{h_n} sous la forme $u_{h_n} = \sum_{1 \leq i \leq n} x_i e_i$. Alors, (4.2) est équivalent au système linéaire

$$\forall 1 \leq j \leq n, \quad \sum_{1 \leq i \leq n} a(e_j, e_i) x_i = \langle f | e_j \rangle,$$

que l'on peut écrire sous forme matricielle $Kx = b$ où $K_{ij} = a(e_j, e_i)$. Une difficulté pratique, qui limite fortement l'applicabilité de la méthode de Galerkin est que la matrice A est généralement une matrice *pleine*, au sens où les entrées K_{ij} sont (en général) non nulles.

On conclut cette section par un critère permettant de montrer la convergence de u_h vers u même quand on ne sait pas estimer la distance entre u et V_h.

Proposition 4.5 (Critère de convergence)**.** *S upposons qu'il existe un sous-ensemble dense $W \subseteq V$ et pour tout $h \in \mathcal{H} \subseteq \mathbb{R}$ un sous-espace vectoriel de dimension finie $V_h \subseteq V$ et application $r_h : W \to V_h$ telle que*

$$\forall w \in W,\ \lim_{h\to 0} \|w - r_h w\| = 0.$$

Alors, si a est α-coercive et bornée par M, que u est la solution de (4.1) et u_h celle de (4.2), alors

$$\lim_{h\to 0, h\in\mathcal{H}} \|u - u_h\| = 0.$$

Démonstration. Soit $\varepsilon > 0$ et $v \in W$ tel que $\|w - u\| \leq \frac{1}{2}\varepsilon$. Comme par hypothèse $\lim_{h\to 0} \|w - r_h w\| = 0$, pour tout h suffisamment petit, on a la majoration $\|w - r_h w\| \leq \frac{1}{2}\varepsilon$. Ainsi,

$$\mathrm{d}(u, V_h) \leq \|r_h w - u\| \leq \varepsilon.$$

On conclut en appliquant le lemme 4.4 (Céa). □

4.2 Problème de Poisson avec conditions de Dirichlet

4.2.1 Espace de Sobolev $H^1(\Omega)$

Définition 4.2 (Dérivée partielle faible)**.** Soit $\Omega \subseteq \mathbb{R}^d$ un ouvert et $u \in \mathrm{L}^2(\Omega)$. On dira que u admet une $i$ème dérivée partielles faible au sens L^2 s'il existe $w_i \in \mathrm{L}^2(\Omega)$ telles que

$$\forall \phi \in \mathcal{C}_c^\infty(\Omega),\quad \int (\partial_i \phi) u = -\int \phi w_i.$$

Remarque 4.1 (Intégrabilité). Pour toute fonction $w_i \in \mathrm{L}^2(\Omega)$ et tout ensemble compact K, le théorème de Cauchy-Schwarz donne, $\|w_i \mathbf{1}_K\|_{\mathrm{L}^1(\Omega)} \leq \sqrt{|K|}\, \|w_i\|_{\mathrm{L}^2(K)}$ $w_i|_K \in \mathrm{L}^1(K)$. En particulier, toute fonction $\phi \in \mathcal{C}_c^\infty(K)$ est bornée et supportée sur un compact K, de sorte que la fonction ϕw_i est intégrable.

Remarque 4.2 (Unicité de la dérivée partielle). Si $w_i, w_i' \in \mathrm{L}^2(\Omega)$ sont deux dérivées partielles faibles de $u \in \mathrm{L}^2(\Omega)$, alors $\int \phi(w_i - w_i') = 0$, pour tout $\phi \in \mathcal{C}_c^\infty(\Omega)$ ce qui implique que $w_i = w_i'$ par densité de $\mathcal{C}_c^\infty(\Omega)$ dans $\mathrm{L}^2(\Omega)$. On notera ainsi sans ambiguité $\partial_i u$ la dérivée partielle faible, si elle existe.

Lemme 4.6. *S'il existe une constante C telle que*

$$\forall \phi \in \mathcal{C}_c^\infty(\Omega),\quad \langle \partial_i \phi | u \rangle_{\mathrm{L}^2(\Omega)} \leq C \|\phi\|_{\mathrm{L}^2(\Omega)},$$

alors u admet une $i$ème dérivée partielles faible au sens L^2.

Démonstration. Soit $\ell(\phi) = \langle \partial_i \phi | u \rangle_{\mathrm{L}^2(\Omega)}$. L'estimation montre que cette forme linéaire est (uniformément) continue sur $\mathcal{C}_c^\infty(\Omega) \subseteq \mathrm{L}^2(\Omega)$ et par densité peut être étendue en une unique forme linéaire sur $\mathrm{L}^2(\Omega)$. Par théorème de Riesz, il existe $w_i \in \mathrm{L}^2(\Omega)$ tel que

$$\forall \phi \in \mathrm{L}^2(\Omega), \ell(\phi) = \langle \phi | w_i \rangle,$$

ce qui montre en particulier que w_i est la dérivée partielle faible. □

Définition 4.3 (Espace de Sobolev)**.** L'espace de Sobolev $H^1(\Omega)$ sur un ouvert $\Omega \subseteq \mathbb{R}^d$ est défini par

$$H^1(\Omega) = \{u \in \mathrm{L}^2(\Omega) \mid \forall i, \partial_i u \in \mathrm{L}^2(\Omega)\},$$

et est normé par

$$\|u\|_{H^1(\Omega)}^2 := \|u\|_{\mathrm{L}^2(\Omega)}^2 + \sum_{1 \leq i \leq d} \|\partial_i u\|_{\mathrm{L}^2(\Omega)}^2 .$$

Nous supposerons une certaine familiarité avec les espaces de Sobolev. Nous admettrons en particulier les énoncés suivants.

- L'espace de Sobolev $H^1(\Omega)$ muni de la norme $\|\cdot\|_{H^1(\Omega)}$ est un espace de Hilbert ;
- L'application linéaire qui à une fonction associe son gradient, $u \in H^1(\Omega) \mapsto \nabla u \in \mathrm{L}^2(\Omega)^d$, est continue.

4.2.2 Équation de Poisson

Soit $\Omega \subseteq \mathbb{R}^d$ un ouvert. On s'intéresse à l'équation aux dérivées partielles suivante, appelée *équation de Poisson* :

$$\begin{cases} -\Delta u = f & \text{sur } \Omega \\ u = 0 & \text{sur } \partial\Omega \end{cases} \tag{4.3}$$

Supposons que cette équation admette une solution $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$ et que le bord de l'ouvert soit de classe $\mathcal{C}^1$. En multipliant la première ligne de (4.4) par $\phi \in \mathcal{C}_c^\infty(\Omega)$ et en appliquant la formule de Green [1], on obtient

$$\begin{aligned} \int_\Omega f\phi &= -\int_\Omega (\Delta u)\phi \\ &= -\int_\Omega \sum_i (\partial_{ii} u)\phi \\ &= \int_\Omega \sum_i (\partial_i u)(\partial_i \phi) - \underbrace{\int_{\partial\Omega} (\partial_i u)\phi n_i \mathrm{d}\sigma}_{=0} = \int_\Omega \langle \nabla u | \nabla \phi \rangle \end{aligned}$$

Ainsi pour toute fonction test $\phi \in \mathcal{C}_c^1(\Omega)$, on a l'égalité

$$\int_\Omega \langle \nabla u | \nabla \phi \rangle = \int_\Omega f\phi.$$

1. $\int_\Omega u\partial_i v = -\int_\Omega v\partial_i u + \int_{\partial\Omega} uvn_i(x)\mathrm{d}\sigma$, où $n_i : \partial\Omega \to \mathbb{R}^d$ est la normale unitaire sortante et σ est la mesure de surface sur $\partial\Omega$

Définition 4.4 (Solution faible à l'équation de Poisson)**.** On note $H_0^1(\Omega)$ l'adhérence de $\mathcal{C}_c^\infty(\Omega)$ vu comme sous ensemble de $H^1(\Omega)$. On appelle solution faible à l'équation de Poisson une fonction $u \in H^1(\Omega)$ vérifiant

$$\forall \phi \in H_0^1(\Omega), \quad \int_\Omega \langle \nabla u | \nabla \phi \rangle = \int_\Omega f\phi \tag{4.4}$$

Cette formulation est un problème variationnel de la forme (4.1), où $V = H_0^1(\Omega)$ et où la forme bilinéaire est donnée par $a(u,v) = \int_\Omega \langle \nabla u | \nabla v \rangle$ et la forme linéaire $\ell(v) = \int fv$.

Théorème 4.7. *Soit Ω est un ouvert borné. Alors,*

(i) Si Ω est $\mathcal{C}^1$, toute solution $u \in \mathcal{C}^2(\Omega) \cap \mathcal{C}^0(\overline{\Omega})$ de l'équation (4.4) est une solution faible.

(ii) Si $f \in \mathrm{L}^2(\Omega)$, il existe exactement une solution faible.

Le point (i) est déjà établi, et pour avoir (ii) il suffit de pouvoir appliquer le théorème de Lax-Milgram. Il suffit donc de montrer que a est coercive, ce qui est une conséquence directe de l'inégalité de Poincaré.

Proposition 4.8 (Inégalité de Poincaré)**.** *Si le domaine Ω est inclus dans une boule de rayon r, alors*

$$\forall v \in H_0^1(\Omega), \quad \int_\Omega v^2 \leq 4r^2 \int_\Omega \|\nabla v\|^2 .$$

Ainsi, $(4r^2+1)a(v,v) \geq \|v\|^2_{H^1(\Omega)}$ et a est coercive.

Démonstration. Par densité et par continuité des deux membres de l'inégalité, il suffit de savoir établir cette inégalité pour toute fonction $v \in \mathcal{C}_c^\infty(\Omega)$. Quitte à appliquer une rotation et une translation, on suppose que $\Omega \subseteq [-r,r] \times \mathbb{R}^{d-1}$. On applique le théorème de Fubini, qui nous donne

$$\int_\Omega v^2 = \int_{\mathbb{R}^{d-1}} \int_{[-r,r]} v(x_1, \overline{x})^2 \mathrm{d}x_1 \mathrm{d}\overline{x}.$$

Or, comme $v(-r, \overline{x}) = 0$,

$$\begin{aligned} v(x_1,\overline{x})^2 &\leq \left(\int_{-r}^{x_1} |\partial_1 v(-r+s, \overline{x})| \, \mathrm{d}s \right)^2 \\ &\leq 2r \int_{-r}^{r} |\partial_1 v(-r+s,\overline{x})|^2 \, \mathrm{d}s \end{aligned}$$

En utilisant $|\partial_1 v| \leq \|\nabla v\|$ il vient

$$\int_\Omega v^2 \leq 4r^2 \int_{\mathbb{R}^{d-1}} \int_{[-r,r]} \|\nabla v(x_1,\overline{x})\|^2 \, \mathrm{d}x_1 \mathrm{d}\overline{x}. \qquad \square$$

4.3 Éléments finis $\mathbb{P}_1$ dans $\mathbb{R}^1$

Pour simplifier, on considère $\Omega = (0,1)$ et $x_i = ih$ pour $1 \leq i \leq M+1$ où $h = 1/(M+1)$. On note $\mathbb{P}_1 = \mathbb{R}_1[X]$ l'espace des fonctions affines sur $\mathbb{R}$ et on note

$$V_h = \{v \in \mathcal{C}^0([0,1]) \mid \forall i \in [\![0, M]\!],\, v|_{[x_i, x_{i+1}]} \in \mathbb{P}_1\}$$
$$V_{0h} = \{v \in V_h \mid v(0) = v(1) = 0\}.$$

Soit $\phi(x) = \max(1 - |x|, 0)$ la fonction "chapeau" et pour $0 \leq i \leq M+1$,

$$\phi_i(x) = \phi\left(\frac{x - x_i}{h}\right).$$

On vérifie sans peine que $\phi_i \in V_h$.

Lemme 4.9. *L'espace V_h (resp. V_{0h}) est de dimension $M+2$ (resp. M) et admet pour base $e_0, \ldots, e_{M+1}$ (resp $e_1, \ldots, e_M$). De plus,*

$$V_h \subseteq H^1(\Omega), \qquad V_{0h} \subseteq H^1_0(\Omega).$$

Démonstration. Montrons d'abord que la famille est libre : si $v_h = \sum_i \alpha_i \phi_i \equiv 0$ alors, en utilisant la relation $\phi_i(x_j) = \delta_{ij}$ on en déduit que $0 = v_h(x_i) = \alpha_i$. Montrons ensuite qu'elle est génératrice : soit $v_h \in V_h$ et $w_h = \sum_i v_h(x_i)\phi_i$. Alors $(v_h - w_h)(x_i) = 0$ pour tout $i \in [\![0, M+1]\!]$, et comme $v_h - w_h|_{[x_i, x_{i+1}]}$ est affine pour tout $i \in [\![0, M]\!]$, on en déduit $v_h - w_h \equiv 0$ soit $v_h = w_h$.

Soit maintenant $v_h \in V_h$ et $\phi \in \mathcal{C}^0_c(\Omega)$. Alors, par intégration par partie et télescopage,

$$\begin{aligned}
\int_\Omega v_h \phi' &= \sum_{0 \leq i \leq M} \int_{x_i}^{x_{i+1}} v_h \phi' \\
&= \sum_{0 \leq i \leq M} - \int_{x_i}^{x_{i+1}} v_h' \phi + [v_h \phi]_{x_i}^{x_{i+1}} \\
&= - \int_\Omega v_h' \phi + (v_h(x_{M+1}) \underbrace{\phi(x_{M+1})}_{=0} - v_h(x_0) \underbrace{\phi(x_0)}_{=0}.
\end{aligned}$$

Comm $v_h' \in \mathrm{L}^2(\Omega)$, on en déduit que $v_h \in H^1(\Omega)$. □

Définition 4.5 (Opérateur de restriction)**.** On définit $r_h : \mathcal{C}^0(\overline{\Omega}) \to V_h$ par

$$r_h v = \sum_{0 \leq i \leq N+1} v(x_i)\phi_i.$$

On vérifie facilement que si $v \in \mathcal{C}^\infty_c(\Omega)$, alors $r_h v \in V_{0h}$.

Proposition 4.10. *Soit $v \in \mathcal{C}^2(\overline{\Omega})$. Alors,*

$$\|v - r_h v\|_{\mathrm{L}^2(\Omega)} \leq ch^2 \left\|v''\right\|_{\mathrm{L}^2(\Omega)}$$
$$\left\|v' - (r_h v)'\right\|_{\mathrm{L}^2(\Omega)} \leq ch \left\|v''\right\|_{\mathrm{L}^2(\Omega)}$$

En particulier,

$$\|v - r_h v\|_{H^1(\Omega)} \leq ch \left\|v''\right\|_{\mathrm{L}^2(\Omega)}$$

Lemme 4.11. *Soit $\tilde{v} \in \mathcal{C}^2([0,1])$, et $\tilde{e}(t) = \tilde{v}(t) - ((1-t)\tilde{v}(0) + t\tilde{v}(1))$. Alors,*

$$\int_{[0,1]} \tilde{e}(t)^2 \mathrm{d}t \leq c \int_{[0,1]} \tilde{v}''(t)^2 \mathrm{d}t.$$

$$\int_{[0,1]} \tilde{e'}(t)^2 \mathrm{d}t \leq c \int_{[0,1]} \tilde{v}''(t)^2 \mathrm{d}t.$$

Démonstration. □

Démonstration de la proposition 4.10. Comme $\Omega = \bigcup_{i\in\{1,M}[x_i, x_{i+1}]$, il suffit d'obtenir l'estimation sur chacun des segments $\sigma_i = [x_i, x_{i+1}]$. Par définition, $r_h v$ est affine sur $[x_i, x_{i+1}]$ et vérifie $r_h v(x_i) = v(x_i)$ et $r_h v(x_{i+1}) = v(x_{i+1})$, i.e.

$$r_h v(x) = \frac{x_{i+1} - x}{h} v(x_i) + \frac{x - x_i}{h} v(x_{i+1}).$$

Ainsi,

$$e(x) := v(x) - r_h v(x) = v(x) - \left(\frac{x_{i+1} - x}{h} v(x_i) + \frac{x - x_i}{h} v(x_{i+1}) \right).$$

On pose $\tilde{e}(t) = e(x_i + th)$ et $\tilde{v}(t) = v(x_i + th)$, de sorte que $\tilde{e}(x) = \tilde{v}(x) - ((1-t)\tilde{v}(0) + t\tilde{v}(1))$. Le lemme précédent nous donne alors

$$\int_{[0,1]} \tilde{e}(t)^2 \mathrm{d}t \leq c \int_{[0,1]} \tilde{v}''(t)^2 \mathrm{d}t.$$

Avec le changement de variable $x = x_i + th$, on obtient

$$\int_{[x_i, x_{i+1}]} \tilde{e} \left(\frac{x - x_i}{h} \right)^2 \mathrm{d}x \leq c \int_{[x_i, x_{i+1}]} \tilde{v}'' \left(\frac{x - x_i}{h} \right)^2 \mathrm{d}x.$$

Comme $\tilde{v}''(t) = h^2 v''(x_i + th)$, on en déduit que

$$\int_{[x_i, x_{i+1}]} e(x)^2 \mathrm{d}x \leq ch^4 \int_{[x_i, x_{i+1}]} v''(x)^2 \mathrm{d}x.$$

Ainsi, comme souhaité, on en déduit que

$$\int_{\Omega} e^2 \mathrm{d}x \leq ch^4 \int_{\Omega} v''(x)^2 \mathrm{d}x.$$

□

Théorème 4.12. *Soit $\Omega =]0,1[$ et $u \in V = H_0^1(\Omega)$. On considère une forme bilinéaire $a : V \times V \to \mathbb{R}$, qu'on suppose α-coercive et bornée (par M) et $\ell \in V'$. Soit $u \in V$ et $u_h \in V_{0h}$ solution des problèmes variationnels*

$$\forall v \in V, \qquad a(u, v) = \ell(v),$$
$$\forall v_h \in V_{0h}, \qquad a(u_h, v_h) = \ell(v_h).$$

Alors,

- *Si $u \in \mathcal{C}^2([0,1])$, alors $\|u - u_h\| \leq c\frac{M}{\alpha} h$*
- *Si $u \in H_0^1(\Omega)$, alors $\lim_{h\to 0} \|u - u_h\| = 0$*

4.4 Éléments finis $\mathbb{P}_1$ dans $\mathbb{R}^d$

4.4.1 Coordonnées barycentriques

Afin de pouvoir définir l'espace des éléments finis en dimension $d > 1$, nous aurons besoin de quelques notions de géométrie affine :

Définition 4.6. Soit $x_0, \dots, x_k \in \mathbb{R}^d$. On appelle

- *espace affine engendré par* $x_0, \dots, x_k$ l'ensemble

$$\mathrm{aff}(\{x_1, \dots, x_k\}) = \left\{ \sum_{1 \le i \le k} \lambda_i x_i \mid \lambda_i \in \mathbb{R}^k, \text{ t.q. } \sum_{1 \le i \le k} \lambda_i = 1 \right\}$$

- *enveloppe convexe de* $x_1, \dots, x_k$ l'ensemble

$$\mathrm{conv}(\{x_1, \dots, x_k\}) = \left\{ \sum_{1 \le i \le k} \lambda_i x_i \mid \lambda_i \in \mathbb{R}_+^k, \text{ t.q. } \sum_{1 \le i \le k} \lambda_i = 1 \right\}$$

On dit que les points sont *affinement indépendants* si l'espace affine qu'ils engendrent est de dimension k, c-à-d

$$\dim(\mathrm{aff}(\{x_0, \dots, x_k\} - x_0) = \dim(\mathrm{vect}(x_1 - x_0, \dots, x_k - x_0)) = k.$$

Exercice 4.1. Montrer que pour tout point $z \in A := \mathrm{aff}(\{x_0, \dots, x_k\}$, l'ensemble $A - z = \{x - z \mid x \in A\}$ est un sous-espace vectoriel de $\mathbb{R}^d$, qui ne dépend pas de z. (Cela montre que la notion d'indépendance affine ne dépend pas du choix de x_0)

On appelle *fonction affine* sur A toute fonction qui peut s'écrire comme la somme d'une fonction linéaire et d'une constante (ou de manière équivalente, comme un polynôme de degré 1).

Proposition 4.13. *Soient $x_0, \dots, x_k \in \mathbb{R}^d$ affinement indépendants. Pour tout point x dans $A = \mathrm{aff}(\{x_0, \dots, x_k\})$, il existe un unique $(\lambda_0, \dots, \lambda_k) \in \mathbb{R}^{k+1}$ tel que*

$$\begin{cases} \sum_{0 \le i \le k} \lambda_i x_i = x \\ \sum_{0 \le i \le k} \lambda_i = 1 \end{cases}$$

On appelle ce $\lambda_0, \dots, \lambda_k \in \mathbb{R}$ les coordonnées barycentriques *du point x. Chacune des applications $\lambda_i : A \to \mathbb{R}$ est affine et vérifie $\lambda_i(x_j) = \delta_{ij}$. De plus, si $\phi : A \to \mathbb{R}$ est une fonction affine, alors*

$$\forall x \in A, \phi(x) = \sum_{0 \le i \le k} \phi(x_i)\lambda_i(x).$$

Enfin, $\mathrm{conv}(\{x_0, \dots, x_k\}) = \{x \in A \mid \forall i, \lambda_i(x) \ge 0\}$.

Démonstration. Si $\lambda \in \mathbb{R}^k$ vérifie l'équation, alors $\lambda_0 = 1 - \lambda_1 - \ldots - \lambda_k$, de sorte que la première égalité s'écrit

$$\sum_{1 \leq i \leq k} \lambda_i (x_i - x_0) = x - x_0.$$

Elle adment exactement une solution car $x - x_0 \in \text{vect}(\{x_i - x_0\}_{1 \leq i \leq k})$ et que la famille $\{x_i - x_0\}_{1 \leq i \leq k}$ est libre. On peut donc définir les applications "coordonnées barycentriques" $\lambda_i : A \to \mathbb{R}$. Le fait que λ_i est affine est laissé en exercice.

Soit maintenant ϕ une fonction affine, et

$$\hat{\phi}(x) = \sum_{0 \leq i \leq k} \phi(x_i) \lambda_i(x),$$

alors $(\phi - \hat{\phi})(x_j) = 0$ pour $0 \leq j \leq k$. Soit $V = \text{vect}(x_1 - x_0, \ldots, x_k - x_0)$ (qui est de dimension k) et $\psi : V \to \mathbb{R}$ définie par $\psi(z) = (\phi - \hat{\psi})(z + x_0)$. Alors, $\psi(0) = 0$, donc ψ est une forme affine, qui s'annule en chaque point $x_i - x_0$ où $i \geq 1$. Ainsi $\psi \equiv 0$ sur V, soit $\phi = \hat{\phi}$ sur A. □

4.4.2 Définitions

Définition 4.7 (Simplexe)**.** On appelle *k-simplexe* de $\mathbb{R}^d$ un ensemble $\sigma \subseteq \mathbb{R}^d$ obtenu en prenant l'enveloppe convexe de $k+1$ points $x_0, \ldots, x_k$ affinement indépendants,

$$\sigma = \left\{ \sum_{0 \leq i \leq k} \lambda_i x_i \mid \lambda_i \geq 0, \sum_j \lambda_j = 1 \right\}$$

Les points $x_0, \ldots, x_k$ sont appelés les *sommets* de σ. On note aussi $\sigma = [x_0, \ldots, x_k]$.

Définition 4.8 (Triangulation)**.** On appelle *triangulation* d'un ouvert Ω de $\mathbb{R}^d$ une famille finie $\mathcal{T}$ de d-simplexes vérifiant :

(i) $\overline{\Omega} = \cup_{\sigma \in \mathcal{T}} \sigma$ (en particulier, Ω est polygonal)

(ii) Pour tout $\sigma, \tau \in \mathcal{T}$, l'intersection $\sigma \cap \tau$ est un simplexe de dimension $k \leq d$ dont les sommets appartiennent à l'intersection des sommets de σ et τ.

Lorsque qu'une triangulation de Ω est notée $\mathcal{T}_h$, avec $h \in \mathbb{R}$, on suppose implicitement que

(iii) Pour tout $\sigma \in \mathcal{T}_h$, le diamètre de $\text{diam}(\sigma) \leq h$.

Définition 4.9 (Espace $\mathbb{P}_1$)**.** Soit $\mathcal{T}_h$ une triangulation d'un ouvert Ω. On pose

- $V_h := \{\phi \in \mathcal{C}^0(\Omega) \mid \forall \sigma \in \mathcal{T}_h, \quad \phi|_\sigma \in \mathbb{P}_1\}$, où $\mathbb{P}_1$ est l'espace des fonctions affines sur $\mathbb{R}^d$.
- $V_{0h} := \{\phi \in V_h \mid \phi|_{\partial\Omega} = 0\}$.
- $S_h = \{x_1, \ldots, x_{N_h}\}$ l'ensemble des sommets des simplexes composant $\mathcal{T}_h$.

Proposition 4.14. *Soit Ω un ouvert (borné) et $\mathcal{T}_h$ une triangulation de Ω.*

(i) L'application linéaire $L : v \in V_h \mapsto (v(x_i))_{1\leq i\leq N_h} \in \mathbb{R}^{N_h}$ est bijective, i.e. une fonction de V_h est uniquement définie par ses valeurs au sommets S_h de la triangulation. En particulier,

- *V_h admet pour base les "fonctions chapeau" $(\phi_i)_{i\in[\![1,N_h]\!]}$ définies par $\phi_i = L^{-1}(e_i)$ ou plus explicitement par*

$$\phi_i \in V_h \text{ et } \phi_i(x_j) = \delta_{ij}.$$

- *V_{0h} admet pour base l'ensemble des ϕ_i tels que $x_i \notin \partial\Omega$.*

(ii) V_h (resp V_{0h}) est un sous-espace de $H^1(\Omega)$ (resp. $H^1_0(\Omega)$). Le gradient d'une fonction $v \in V_h$ est la fonction constante par morceaux égale à $\nabla v|_\sigma$ sur chaque simplexe $\sigma \in \mathcal{T}_h$.

Démonstration. (i) L'injectivité de l'application linéaire L est directe : si $v \in V_h$ vérifie $v(x_i) = 0$ pour tout $i \in [\![1, N_h]\!]$, alors $v \equiv 0$ sur tout simplexe, donc $v \equiv 0$ sur Ω. Montrons la surjectivité. Étant donnée $w \in \mathbb{R}^{N_h}$, on peut définir pour tout simplexe $\sigma = [x_{i_0}, \dots, x_{i_d}]$ de la triangulation $\mathcal{T}_h$ une fonction affine $w_\sigma : \sigma \to \mathbb{R}$ vérifiant $w_\sigma(x_{i_j}) = w_{i_j}$, via

$$w_\sigma(x) = \sum_{0\leq i\leq d} \lambda_j(x) w_{i_j},$$

où λ_i sont les coordonnées barycentriques dans le simplexe (cf. proposition 4.13). De plus pour tout $\sigma, \tau \in \mathcal{T}_h$, l'intersection $\sigma \cap \tau$ est un k-simplexe engendré par des sommets communs à τ et σ. Ainsi $w_\sigma|_{\sigma\cap\tau} = w_\tau|_{\sigma\cap\tau}$. Ceci permet de définir définir globalement une fonction continue v sur $\overline{\Omega} = \bigcup_{\sigma\in\mathcal{T}_h} \sigma$ telle que $v|_\sigma = w_\sigma$ pour tout simplexe $\sigma \in \mathcal{T}_h$. En particulier, $v(x_i) = w_i$ pour tout $i \in [\![1, N_h]\!]$, et l'application linéaire L est donc bien surjective.

(ii) On donne une preuve élémentaire, reposant uniquement sur la définition des dérivées partielles faibles (qui peut être sautée en première lecture). Notons d'abord que u est différentiable à l'intérieur de chaque simplexe, i.e. sur l'ensemble

$$\bigcup_{\sigma\in\mathcal{T}_h} \mathring{\sigma},$$

qui est de mesure pleine dans Ω. On notera $\partial_i u$ la dérivée partielle définie sur l'intérieur des simplexes, qu'on étendra (de manière arbitraire) en une fonction $\partial_i u \in \mathrm{L}^\infty(\Omega) \subseteq \mathrm{L}^2(\Omega)$.

Soit maintenant $\phi \in \mathcal{C}^\infty_c(\Omega)$ et $i \in \{1, \dots, d\}$. Pour tout $t > 0$, on note

$$B_t = \bigcup_{\sigma\in\mathcal{T}_h} \bigcup_{x\in\partial\sigma} B(x, t).$$

L'ensemble B_t est inclus dans une union finie de "tranches" d'espace de la forme $\{x \in \Omega \mid |\langle x|v\rangle - a| \leq h\}$, une par facette de chaque simplexe. Sa mesure de Lebesgue est donc bornée par $\mathrm{O}(t)$. Ainsi,

$$\int_\Omega u\partial_i\phi = \lim_{t\to 0} \int_\Omega u(x) \frac{\phi(x + te_i) - \phi(x)}{t} \mathrm{d}x$$

et

$$\int_\Omega u(x)\frac{\phi(x+te_i)-\phi(x)}{t}\mathrm{d}x = \int_\Omega \frac{u(x-te_i)-u(x)}{t}\phi(x)\mathrm{d}x$$

Soit $x \in \Omega \setminus B_t$. Comme Ω est l'union des simplexes de $\mathcal{T}_h$, il existe un simplexe σ tel que $x \in \sigma$. Alors, $x - te_i \in \sigma$ car sinon, x serait à distance plus petite que t de $\partial\sigma$, ce qui contredirait l'hypothèse $x \notin B_t$. Comme u est affine sur σ, $u(x-te_i) = u(x) - t\partial_i u(x)$. On en déduit que

$$\int_\Omega \frac{u(x-te_i)-u(x)}{t}\phi(x)\mathrm{d}x = \int_\Omega \partial_i u(x)\phi(x) + \int_{B_t}\left(\frac{u(x-te_i)-u(x)}{t} - \partial_i u(x)\right)\phi(x).$$

De plus, comme u est Lipschitzienne, $|u(x-te_i)-u(x)| \leq Lt$ pour une certaine constant L. En utilisant $|B_t| = \mathrm{O}(t)$ on en conclut

$$\int_\Omega u(x)\frac{\phi(x+te_i)-\phi(x)}{t}\mathrm{d}x = \int_\Omega \partial_i u(x)\phi(x) + \mathrm{O}(t),$$

ce qui démontre que $\partial_i u$ est la dérivée faible de u.
(NB. La démonstration de l'inclusion $V_{0h} \subseteq H_0^1(\Omega)$ est assez technique et sera admise.) □

4.4.3 Convergence

Définition 4.10 (Opérateur de restriction)**.** Soit $\mathcal{T}_h$ une triangulation de Ω. On définit $r_h : \mathcal{C}^0(\overline{\Omega}) \to V_h$ par

$$r_h v(x) = \sum_{i\in[\![1,N_h]\!]} v(x_i)\phi_i(x),$$

où $x_1, \ldots, x_{N_h}$ sont les sommets de la triangulation $\mathcal{T}_h$ et où ϕ_i sont les "fonctions chapeau" définies dans la proposition 4.14.

En d'autre termes, la fonction $r_h v \in V_h$ est l'unique fonction continue, affine par morceau sur la triangulation $\mathcal{T}_h$ et vérifiant par $r_h v(x_i) = v(x_i)$. En dimension $d \geq 2$, le contrôle de l'erreur $\|r_h v - v\|$ nécessite une hypothèse sur la triangulation.

Définition 4.11 (Triangulation ρ-régulière)**.** Une triangulation $\mathcal{T}_h$ d'un ouvert $\Omega \subseteq \mathbb{R}^d$ est appelée ρ-régulière (où $\rho \in]0,1[$) si tout simplexe $\sigma \in \mathcal{T}_h$ contient une boule de diamètre plus grand que ρh.

Cette définition impose en particulier que les simplexes de $\mathcal{T}_h$ ne soient pas trop "aplatis". La construction en pratique de triangulations ρ-régulières est un problème difficile, en particulier en dimension $d \geq 3$ et pour des domaines admettant des coins.

On admet temporairement la proposition suivante :

Proposition 4.15. *Soit $\mathcal{T}_h$ une triangulation ρ-régulière d'un ouvert Ω et V_{0h} l'espace d'éléments finis $\mathbb{P}_1$ défini par $\mathcal{T}_h$. Alors,*

$$\forall v \in \mathcal{C}^2(\overline{\Omega}), \quad \|v - r_h v\|_{H^1(\Omega)} \leq c \left\|\mathrm{D}^2 v\right\|_\infty \frac{h}{\rho},$$

où la constante c ne dépend que de Ω.

Théorème 4.16 (Convergence de la méthode des éléments finis)**.** *Soit*

- *Ω est un ouvert borné de $\mathbb{R}^d$,*
- *$V = H^1_0(\Omega)$ (resp. $V = H^1(\Omega)$),*
- *$a : V \times V \to \mathbb{R}$ une forme bilinéaire continue et α-coercive,*
- *$\ell : V \to \mathbb{R}$ une forme linéaire continue.*

On se donne également

- *$(\mathcal{T}_h)_{h\in\mathcal{H}}$ une famille de triangulations ρ-régulière de Ω,*
- *l'espace l'espace d'éléments finis V_{0h} (resp. V_h) $\mathbb{P}_1$ défini par $\mathcal{T}_h$.*

Soit $u \in V$ la solution du problème variationnel continu (4.1) et $u_h \in V_{0h}$ (resp. V_h) la solution du problème variationnel discret (4.2). Alors

(i) Si $u \in \mathcal{C}^2(\Omega)$, alors $\|u - u_h\|_{H^1(\Omega)} \leq \frac{c\|a\|}{\rho\alpha} \left\|\mathrm{D}^2 u\right\|_\infty h$.

(ii) Dans le cas général $u \in H^1(\Omega)$ on a $\lim_{h\to 0, h\in\mathcal{H}} \|u - u_h\|_{H^1(\Omega)} = 0$.

Démonstration. On traite le cas $V = H^1_0(\Omega)$, le cas $V = H^1(\Omega)$ étant analogue.
(i) Par le lemme de Céa, on sait que

$$\|u - u_h\|_V \leq \frac{\|a\|}{\alpha} \mathrm{d}(u, V_{0h}) \leq \frac{\|a\|}{\alpha} \|u - r_h u\|_V ,$$

que l'on majore en utilisant la proposition 4.15.
(ii) Il suffit d'appliquer le critère de convergence de la proposition 4.5. □

4.5 Interpolation avec des éléments $\mathbb{P}_1$

L'objectif de cette partie est de démontrer la proposition suivante.

Proposition 4.17. *Soit $\sigma \subseteq \mathbb{R}^d$ un d-simplexe, $v \in \mathcal{C}^2(\sigma)$ et $\hat{v} \in \mathbb{P}_1$ telle que $\hat{v} = v$ aux sommets de σ. Alors,*

$$\|v - \hat{v}\|^2_{\mathrm{L}^2(\sigma)} \leq \left\|\mathrm{D}^2 v\right\|^4_\infty \mathrm{Leb}(\sigma) h^4 \tag{4.5}$$

$$\|\nabla v - \nabla \hat{v}\|^2_{\mathrm{L}^2(\sigma)} \leq c \left\|\mathrm{D}^2 v\right\|^2_{\mathrm{L}^2(\Omega)} \mathrm{Leb}(\sigma) \frac{h^4}{r^2} \tag{4.6}$$

où h est le diamètre de σ, r le diamètre de la plus grande boule contenue dans σ, $c = c(d)$ est une constante et Leb *est la mesure de Lebesgue.*

Montrons comment cette proposition implique la Proposition 4.15.

Démonstration de la Proposition 4.15. Soit $v \in \mathcal{C}^2(\Omega)$. Alors, par la proposition précédente

$$\begin{aligned}\|v - r_h v\|^2_{H^1(\Omega)} &\leq \sum_{\sigma\in\mathcal{T}_h} \|v - r_h v\|^2_{H^1(\sigma)} \\ &\leq \left(h^4 + c\frac{h^4}{r^2}\right) \sum_{\sigma\in\mathcal{T}_h} \mathrm{Leb}(\sigma) \left\|\mathrm{D}^2 v\right\|^2_\infty \\ &\leq c(\Omega)\frac{h^2}{\rho^2} \left\|\mathrm{D}^2 v\right\|^2_\infty .\end{aligned}$$

où l'on a utilisé $r \geq \rho h$. □

Soit $\sigma = [A_0, \ldots, A_d]$ un d-simplexe de $\mathbb{R}^d$, et soit $\lambda_i : \mathbb{R}^d \to \mathbb{R}$ les coordonnées barycentriques, i.e. l'unique fonction affine telle que $\lambda_i(A_j) = \delta_{ij}$.

Lemme 4.18. *Les λ_i vérifient les propriétés suivantes pour tout $x \in \mathbb{R}^2$:*
(i) $\sum_{0\leq i\leq d} \lambda_i(x) = 1$;
(ii) $\sum_{0\leq i\leq d} \lambda_i(x) A_i = x$;
(iii) si σ contient une boule de diamètre r, alors $\|\nabla\lambda_i(x)\| \leq \frac{1}{r}$.

Démonstration. Nous avons déjà vu (i) et (ii). (iii) Comme λ_i est affine, son gradient est un vecteur constant $p_i \in \mathbb{R}^2$. Comme σ contient une boule de rayon r, il existe deux points $x, y \in \sigma$ tels que

$$\frac{y - x}{r} = \frac{p_i}{\|p_i\|}$$

de sorte qu'en utilisant $\lambda_i(y) = \lambda_i(x) + \langle p_i | y - x\rangle$ et $0 \leq \lambda_i(y), \lambda_i(x) \leq 1$,

$$\|p_i\| = \frac{1}{r}\langle p_i | y - x\rangle = \frac{1}{r}(\lambda_i(y) - \lambda_i(x)) \leq \frac{1}{r}.$$ □

Démonstration de la proposition 4.17. Soit $v \in \mathcal{C}^2(\sigma)$ et soit $\hat{v}$ l'interpolation linéaire de v sur le simplexe :

$$\hat{v}(x) = \sum_i \lambda_i(x) v(A_i)$$

On pose $T_i(x,t) = x + t(A_i - x) = (1-t)x + tA_i$. La formule de Taylor avec reste intégral donne

$$\begin{aligned} v(A_i) &= v(x + A_i - x) \\ &= v(x) + \langle \nabla v(x) | A_i - x\rangle + \int_0^1 \langle A_i - x | \mathrm{D}^2 v(T_i(x,t))(A_i - x)\rangle (1-t)\mathrm{d}t \end{aligned}$$

En utilisant cette formule dans $\hat{v}(x) = \sum_{0\leq i\leq d} \lambda_i(x) v(A_i)$ et en la combinant avec les égalités $\sum_i \lambda_i(x) = 1$ et $\sum_i \lambda_i(x) A_i = x$, on obtient

$$\begin{aligned}
\hat{v}(x) &= \sum_{i\in[\![0,d]\!]} \lambda_i(x) v(A_i) \\
&= \sum_{i\in[\![0,d]\!]} \lambda_i(x) \left(v(x) + \langle \nabla v(x) | A_i - x \rangle + \int_0^1 \langle A_i - x | \mathrm{D}^2 v(T_i(x,t))(A_i - x) \rangle (1-t) \mathrm{d}t \right) \\
&= \underbrace{(\sum_i \lambda_i(x))}_{=1} v(x) + \langle \nabla v(x) | \underbrace{\sum_i \lambda_i(x)(A_i - x)}_{=0} \rangle + \dots \\
&= v(x) + \sum_i \int_0^1 \lambda_i(x) \langle A_i - x | \mathrm{D}^2 v(T_i(x,t))(A_i - x) \rangle (1-t) \mathrm{d}t
\end{aligned}$$

Ceci donne

$$\begin{aligned}
\|\hat{v}(x) - v(x)\|^2 &= \left(\int_0^1 \sum_{0\leq i\leq d} \lambda_i(x) \langle A_i - x | \mathrm{D}^2 v(T_i(x,t))(A_i - x) \rangle (1-t) \mathrm{d}t \right)^2 \\
&\leq \int_0^1 \left(\sum_{0\leq i\leq d} \lambda_i(x) \langle A_i - x | \mathrm{D}^2 v(T_i(x,t))(A_i - x) \rangle (1-t) \right)^2 \mathrm{d}t \\
&\leq \int_0^1 \sum_{0\leq i\leq d} \lambda_i(x) \|A_i - x\|^4 \left\| \mathrm{D}^2 v(T_i(x,t)) \right\|^2 (1-t)^2 \mathrm{d}t
\end{aligned}$$

où on a utilisé deux fois l'inégalité de Jensen (la deuxième application utilise que $\sum_{0\leq i\leq d} \lambda_i(x) = 1$ et $\lambda_i(x) \geq 0$), l'inégalité de Cauchy-Schwartz et la définition de la norme d'opérateur. De plus, en se rappelant que le diamètre de σ est majoré par h (soir $\|A_i - x\| \leq h$) on obtient en intégrant sur σ que

$$\int_\sigma \|\hat{v}(x) - v(x)\|^2 \, \mathrm{d}x \leq \mathrm{Leb}(\sigma) \left\| D^2 v \right\|_\infty^2 h^4$$

Majorons maintenant la norme $\|\nabla v - \nabla \hat{v}\|_{\mathrm{L}^2(\sigma)}$. Soit $p_i = \nabla \lambda_i$. On a

$$\begin{aligned}
\nabla \hat{v}(x) &= \sum_{0\leq i\leq d} v(A_i) p_i \\
&= \sum_{0\leq i\leq d} \left(v(x) + \langle \nabla v(x) | A_i - x \rangle + \int_0^1 \langle A_i - x | \mathrm{D}^2 v(T_i(x,t))(A_i - x) \rangle t \mathrm{d}t \right) p_i
\end{aligned}$$

Posons $g(x) := \sum_{1\leq i\leq d} \langle \nabla v(x) | A_i - x \rangle p_i$. En utilisant les propriétés des fonctions λ_i

on obtient que pour tout $j \in [\![0,d]\!]$,

$$\begin{aligned}\langle g(x)|A_j - x\rangle &= \sum_{0\le i\le d} \langle \nabla v(x)|A_i - x\rangle\langle p_i|A_j - x\rangle \\ &= \sum_{0\le i\le d} \langle \nabla v(x)|A_i - x\rangle(\underbrace{\lambda_i(A_j)}_{=\delta_{ij}} - \lambda_i(x)) \\ &= \langle \nabla v(x)|A_j - x\rangle - \langle \nabla v(x)|\underbrace{\sum_{0\le i\le d} \lambda_i(x)(A_i - x)}_{=0}\rangle \\ &= \langle \nabla v(x)|A_j - x\rangle,\end{aligned}$$

et comme les vecteurs $A_j - x$ engendrent $\mathbb{R}^d$, on en déduit que $g(x) = \nabla v(x)$. Ainsi,

$$\nabla \hat{v}(x) - \nabla v(x) = \sum_{1\le i\le d} \left(\int_0^1 \langle A_i - x|\mathrm{D}^2 v(T_i(x,t))(A_i - x)\rangle (1-t)\mathrm{d}t \right) p_i.$$

On majore la norme de la différence comme précédemment, en utilisant de plus l'inégalité $\|p_i\| \le 1/r$, démontrée dans le lemme 4.18. □

4.6 Exercices

Exercice 4.2. *Triangle plat et approximation du gradient.* Soit $\sigma \subseteq \mathbb{R}^2$ le triangle de sommets $A = (h^2, 0)$, $B_\pm = (0, \pm h)$, qui est de diamètre $\mathrm{O}(h)$, et $v(x) = x_1^2 + x_2^2$.

1. Calculer $\hat{v} = r_h v$.

2. Montrer que $\|\nabla \hat{v} - \nabla v\|_{\mathrm{L}^2(\sigma)} \ge \mathrm{Leb}(\sigma)$.

Conclusion : sur un tel triangle, l'approximation du gradient de $v : x \mapsto \|x\|^2$ par celui de $\hat{v}$ est très mauvaise !

Exercice 4.3. *Discrétisation sur un ouvert non-polygonal.* Soit Ω un ouvert borné, a une forme bilinéaire, continue et coercive sur $V = H_0^1(\Omega)$, $\ell \in V^*$. On se donne $(\mathcal{T}_h)_{h\in\mathcal{H}}$ une famille de triangulations ρ-régulière d'ouverts $\Omega_h \subseteq \Omega$ et on note V_{0h} les espaces d'éléments finis associé. On fait l'hypothèse que Ω_h converge vers Ω au sens suivant : pour tout compact $K \subseteq \Omega$, $\exists h_0 > 0$ tel que $\forall h \in \mathcal{H} \cap [0, h_0]$, $K \subseteq \Omega_h$.

Soit $u_h \in V_{0h}$ la solution du problème variationnel discret

$$\forall v_h \in V_{0h},\ a(u_h, v_h) = \ell(v_h),$$

et $u \in V$ la solution du problème variationnel continu

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

On admet que $H_0^1(\Omega_h) \subseteq H_0^1(\Omega)$.

1. Soit $v \in H_0^1(\Omega)$. En approchant v par des fonctions $\mathcal{C}_c^\infty(\Omega)$, démontrer que utilisant la définition de $H_0^1(\Omega)$, montrer que

$$\lim_{h\to 0} \mathrm{d}(v, V_h) = \lim_{h\to 0} \min_{v_h \in V_h} \|v - v_h\|_V = 0.$$

2. Conclure à la convergence du schéma : $\lim_{h\to 0} u_h = u$.

Chapitre 5

Problème d'obstacle

Soit Ω un ouvert de $\mathbb{R}^d$ et $V := H_0^1(\Omega)$, et $J(u) := \|\nabla u\|^2_{\mathrm{L}^2(\Omega)}$. On se donne une fonction $\Psi \in \mathcal{C}^\infty(\overline{\Omega}) \cap H_0^1(\Omega)$, et on s'intéresse au problème suivant :

$$\min_{u \in K} J(u) \tag{5.1}$$

où

$$K = \{v \in V \mid v \geq \Psi \text{ p.p.}\}.$$

Remarque 5.1. L'énergie élastique d'une membrane attachée aux bords du domaine est donnée par

$$\tilde{J}(u) = \int_\Omega \sqrt{1 + \|\nabla u\|^2}.$$

L'énergie qu'on considère peut être vue comme une linéarisation de cette énergie, valable lorsque ∇u est "petit".

5.1 Rappels sur la convergence faible

Définition 5.1. Soit V un espace de Hilbert. Une suite $(u_n) \in V$ *converge faiblement* vers $u \in V$ si

$$\forall v \in V, \quad \lim_{n \to +\infty} \langle u_n | v \rangle = \langle u | v \rangle.$$

On appelle valeur d'adhérence faible d'une suite (u_n) tout vecteur $u \in V$ qui est obtenu comme limite faible d'une suite extraite de (u_n).

Théorème 5.1 (Banach-Alaoglu)**.** *Si (u_n) est une suite bornée, alors il existe une suite extraite $(u_{\sigma(n)})$ faiblement convergente.*

Corollaire 5.2. *Si (u_n) est une suite bornée n'admettant qu'une valeur d'adhérence faible u, alors elle converge faiblement vers u.*

Remarque 5.2.
- La fonction $\|\cdot\|_V^2$ n'est pas faiblement continue. Par exemple, considérer $(e_n)_{n \geq 1}$ une base hilbertienne de l'espace V. Alors, $\lim_{n \to +\infty} e_n = 0$ mais $\lim_{n \to +\infty} \|e_n\|^2 = 1$.

- Si (u_n) converge faiblement vers u et si $\lim_{n\to+\infty} \|u_n\| = \|u\|$, alors u_n converge fortement vers u.

Définition 5.2. Une fonction $J : V \to \mathbb{R} \cup \{+\infty\}$ est dite (séquentiellement) faiblement continue (resp. semi-continue inférieurement) si pour toute suite u_n convergeant faiblement vers u, on a $J(u) = \lim_{n\to+\infty} J(u_n)$ (resp. $J(u) \leq \liminf_{n\to+\infty} J(u_n)$).

Exemple 5.1.
- Soit $v \in V$. Alors $J : u \in V \mapsto \langle u|v\rangle$ est faiblement continue.
- Si $(J_i)_{i\in I}$ est une famille quelconque de fonctions faiblement sci, alors $J(x) = \sup_{i\in I} J_i(x)$ est aussi faiblement sci.
- Si $f : \mathbb{R} \to \mathbb{R}$ est continue et $J : V \to \mathbb{R}$ est faiblement sci, alors $f \circ J$ est également aussi sci.
- Ainsi, $\|\cdot\|_V^2$ est faiblement sci, comme composée de la fonction carré et de $\|\cdot\| : u \in V \mapsto \sup_{\|v\|=1} \langle u|v\rangle$.

5.2 Existence et caractérisation de la solution

Proposition 5.3. *Le problème* (5.1) *admet une unique solution.*

Lemme 5.4. *On a* $K = \{v \in V \mid \forall \phi \in \mathcal{C}_c^\infty(\Omega, \mathbb{R}^+),\ \int \phi v \geq \int \phi\Psi\}$. *Ainsi, l'ensemble* K *est faiblement fermé.*

Démonstration. Montrons d'abord l'égalité entre les deux ensembles. Si $v \geq \Psi$, alors on a $\int \phi v \geq \int \phi\Psi$ pour toute fonction $\phi \in \mathcal{C}_c^\infty(\Omega), \phi \geq 0$.

Montrons l'inclusion réciproque. Soit $v \in V$ tel que $v \not\geq \Psi$ p.p. Alors, il existe $\varepsilon > 0$ et un ensemble A de mesure > 0 tel que $v < \Psi - \varepsilon$ sur A. Il existe une suite de fonction $\phi_n \in \mathcal{C}_c^\infty(\Omega)$, $\phi_n \geq 0$ approchant la fonction caractéristique χ_A de A (i.e. $\lim_{n\to+\infty} \|\phi_n - \chi_A\|_{\mathrm{L}^2(\Omega)} = 0$), de sorte que

$$0 < \varepsilon\,|A| \leq \int_A \Psi - v \leq \lim\inf_{n\to+\infty} \int \phi_n(\Psi - v).$$

Ainsi, il existe $\phi \in \mathcal{C}_c^\infty(\Omega, \mathbb{R}^+)$ tel que $\int \phi v < \int \phi\Psi$ soit $v \notin K$. □

Démonstration de la proposition 5.3. Soit v_n une suite minimisante, c'est-à-dire que $v_n \in K$ et $J(v_n)$ converge vers $m := \inf_{u\in K} J$. Sans perte de généralité, on suppose que la suite $(J(v_n))_{n\in\mathbb{N}}$ est décroissante, de sorte que $J(v_n) \leq J(v_0)$. Ainsi, la suite $(v_n)_{n\in\mathbb{N}}$ est bornée dans V et par théorème de Banach-Alaoglu, elle admet une sous-suite $(v_{\sigma(n)})$ qui converge faiblement vers v^*. Or, par le lemme précédent, l'ensemble K est faiblement fermé, donc $v^* \in K$. De plus, J étant semi-continue inférieurement,

$$J(v^*) \leq \lim_{n\to\infty} J(v_{\sigma(n)}) = \lim_{n\to\infty} J(v_n) = m = \inf_K J$$

Ainsi, v^* est minimiseur du problème.

Pour montrer l'unicité de la solution, raisonnons par l'absurde : soient $u_0, u_1 \in K$ deux minimiseurs de J sur K, que l'on suppose distincts. Alors, $u = \frac{1}{2}(u_0 + u_1) \in K$

— car K est convexe comme intersection de demi-espaces — et u est donc admissible. De plus, comme J est strictement convexe,

$$J(u) < \frac{1}{2}(J(u_0) + J(u_1)) = J(u_0),$$

ce qui contredit la minimalité de u_0. On en déduit l'unicité. □

Proposition 5.5. *Une fonction $u \in K$ est solution de (5.1) si et seulement si*

$$\forall v \in K, \quad \int_\Omega \langle \nabla u | \nabla(v-u) \rangle \geq 0.$$

Démonstration. Notons d'abord que pour tout $u, v \in K$ et $t \in [0,1]$,

$$J((1-t)u + tv) = J(u) + 2t \int \nabla u(\nabla v - \nabla u) + t^2 \left\| \nabla u - \nabla v \right\|_{\mathrm{L}^2}^2 .$$

Si u est minimum de J sur K et si $v \in K$, alors

$$J(u) \leq J((1-t)u + tv) = J(u) + 2t \int \nabla u(\nabla v - \nabla u) + t^2 \left\| \nabla u - \nabla v \right\|_{\mathrm{L}^2}^2$$

d'où l'on déduit que $\int \nabla u(\nabla v - \nabla u) \geq 0$.

Réciproquement, si $v \in K$ et si $\int \nabla u(\nabla v - \nabla u) \geq 0$ alors $\forall t \in [0,1], J((1-t)u + tv) \geq J(u)$, d'où en particulier $J(v) \geq J(u)$. □

Proposition 5.6. *Si la solution u de (5.1) appartient à $\mathcal{C}^2(\overline{\Omega})$, alors*

$$\begin{cases} u \geq \Psi \text{ et } -\Delta u \geq 0 & \text{sur } \Omega \\ -\Delta u = 0 & \text{sur } \Omega \cap \{u > \Psi\} \end{cases}$$

Remarque 5.3. En dimension $d = 1$, la solution est donc concave.

Démonstration. Soit $v \in \mathcal{C}_c^\infty(\Omega)$, $v \geq 0$. Si $u \in K$, alors pour tout $t > 0$, $u + tv$ appartient aussi à K. Par la proposition précédente, on obtient

$$\int_\Omega \langle \nabla u | \nabla(u + tv - u) \rangle \geq 0,$$

En intégrant cette inégalité par partie, on obtient

$$\forall v \in \mathcal{C}_c^\infty(\Omega, \mathbb{R}_+), \quad -\int_\Omega (\Delta u) v \leq 0.$$

Comme Δu est continue par hypothèse, on en déduit que $-\Delta u \geq 0$.

Comme u est continue, l'ensemble $O = \{u > \Psi\}$ est ouvert. Soit $v \in \mathcal{C}_c^\infty(\Omega)$. Alors, comme u et Ψ sont continues et $u > \Psi$ sur Ω, il existe $\varepsilon > 0$ tel que

$$\forall t \in [-\varepsilon, \varepsilon], \quad u + tv > \Psi,$$

c'est-à-dire $u + tv \in K$ pour tout $t \in [-\varepsilon, \varepsilon]$. Alors, toujours par la proposition précédente et par intégration par partie, on obtient $-\Delta u = 0$ sur O. □

5.3 Discrétisation du problème d'obstacle

Pour simplifier, on suppose que le domaine Ω est polygonal, et on se donne une famille uniformément régulières de triangulations $(\mathcal{T}_h)_{h\in\mathcal{H}}$ de Ω. On note toujours

$$V_h = \{\phi \in \mathcal{C}^0(\Omega) \mid \forall \sigma \in \mathcal{T}_h,\, \phi|_\sigma \in \mathbb{P}_1\}$$

$$V_{0h} = \{\phi \in V_h \mid \psi|_{\partial\Omega} = 0\}$$

Étant donné $\phi \in \mathcal{C}^2(\overline{\Omega})$, on définit $r_h\phi$ l'unique fonction appartenant à V_h et vérifiant $r_h\phi(x) = \phi(x)$ pour tout sommet de la triangulation. On sait alors par la proposition 4.15 que

$$\forall \phi \in \mathcal{C}^2(\overline{\Omega}),\quad \lim_{h\to 0} \|r_h\phi - \phi\|_{H^1_0(\Omega)} = 0.$$

On pose enfin

$$K_h := \{\phi \in V_{0h} \mid \phi \geq r_h\Psi\}.$$

On remarquera deux choses :

(i) *a priori*, K_h n'est pas contenu dans K (car $r_h\Psi$ n'a aucune raison, a priori, de majorer de Ψ)

(ii) pour tout $v \in K \cap \mathcal{C}^\infty(\overline{\Omega})$, on a $r_h v \in K_h$.

Finalement, on considère le problème discret

$$\min_{v\in K_h} J(v). \tag{5.2}$$

En utilisant les mêmes arguments que dans la démonstration de (5.3) (il suffit de montrer que K_h est fermé), on obtient l'existence d'une unique solution à (5.2).

Théorème 5.7. *Soit u la solution continu* (5.1) *et u_h celle du problème discret* (5.2). *Si l'une des trois hypothèses suivantes est vérifiée :*

(i) $u \in \mathcal{C}^2(\overline{\Omega})$ *;*

(ii) $K \cap \mathcal{C}^\infty_c(\overline{\Omega})$ *est dense dans* K *;*

(iii) l'obstacle vérifie $\Psi \leq 0$ *dans un voisinage de zéro,*

alors $\lim_{h\to 0} \|u - u_h\|_V = 0$

On ne traite que le second cas, le premier cas étant plus simple (le rédiger constitue un bon exercice !). Pour traiter le cas (iii), il suffira d'utiliser le lemme suivant :

Lemme 5.8. *On suppose que $\Psi \leq 0$ dans un voisinage de $\partial\Omega$. Alors, $K \cap \mathcal{C}^\infty_c(\overline{\Omega})$ est dense dans K.*

Démonstration du théorème 5.7. Soit $u_h \in K_h$ la solution du problème discret. Comme $\Psi \in \mathcal{C}^\infty(\overline{\Omega})$ et $\Psi|_{\partial\Omega} = 0$, on sait que $r_h\Psi$ est bien défini et $r_h\Psi \in K_h$. Ainsi,

$$J(u_h) \leq J(r_h\Psi)$$

Comme de plus, $\lim_{h\to 0} \|r_h\Psi - \Psi\|_{H^1(\Omega)} = 0$, on sait que la suite $r_h\Psi$ est bornée. A fortiori, la suite u_h est également bornée, et on peut supposer à sous-suite près

que u_h admet une limite faible $u^* \in V$ lorsque $h \to 0$. Comme $u_h \in K_h$, on sait que $u_h \geq r_h\Psi$ ponctuellement, de sorte que

$$\forall \phi \in \mathcal{C}_c^\infty(\Omega, \mathbb{R}^+), \int u_h\phi \geq \int (r_h\Psi)\phi \geq \int \Psi\phi - \|r_h\Psi - \Psi\|_{\mathrm{L}^2(\Omega)} \|\phi\|_{\mathrm{L}^2(\Omega)}.$$

En passant à la limite $h \to 0$, on obtient $\int u^*\phi \geq \int \Psi\phi$ pour toute fonction $\phi \in \mathcal{C}_c^\infty(\Omega)$, soit encore $u^* \in K$. Ainsi, $J(u^*) \geq \min_K J$.

Passons à l'autre inégalité. Soit u la solution du problème continu (5.1). Alors, par le lemme 5.8, pour tout $\eta > 0$ il existe $u^\eta \in K \cap \mathcal{C}_c^\infty(\Omega)$ tel que $\|u - u^\eta\|_{H^1(\Omega)} \leq \eta$. Alors, $r_h u^\eta$ appartient à K_h et donc $J(u_h) \leq J(r_h u^\eta)$. Or, comme $r_h u^\eta$ converge vers u^η fortement, et comme J est sci, on obtient

$$J(u^*) \leq \liminf_{h\to 0} J(u_h) \leq \lim_{h\to 0} J(r_h u^\eta) = J(u^\eta),$$

où l'on a utilisé la semi-continuité faible de J pour obtenir l'inégalité de gauche. En prenant la limite (forte) du membre de gauche lorsque η converge vers zéro, on en déduit que $J(u^*) \leq J(u)$. Ainsi $J(u^*) = J(u)$ et $u^* \in K$. Ainsi, par unicité de la solution de (5.1) (Proposition 5.3), on sait que $u^* = u$. Ainsi, la suite bornée (u_h) possède une unique valeur d'adhérence faible et converge donc faiblement vers celle-ci, i.e. $\lim_{h\to 0} u_h = u$ faiblement. Comme de plus $J(u) = \|u\|_V^2 = \lim_{h\to 0} J(u_h)$, on en déduit que u_h converge en fait fortement vers u. □

La démonstration du lemme 5.8 est délicate, et repose sur le résultat suivant, que nous admettons :

Lemme 5.9 (Stampacchia)**.** *Soit Ω un ouvert borné. Alors l'application*

$$(v, w) \in H^1(\Omega) \times H^1(\Omega) \mapsto \max(v, w) \in H^1(\Omega)$$

est (fortement) continue.

Démonstration du lemme 5.8. **Étape 1.** On commence par démontrer que $K \cap \mathcal{C}_c^0(\Omega)$ est dense dans K. Pour cela, on considère $v \in K$, et on considère une suite $w_n \in \mathcal{C}_c^\infty(\Omega)$ convergeant fortement vers v, on pose

$$v_n = \max(w_n, \Psi),$$

qui est continue et à support compact (car $\Psi \leq 0$ dans un voisinage de $\partial\Omega$) et contenue dans K. Par le lemme de Stampacchia,

$$\lim_{n\to\infty} v_n = \max\left(\lim_{n\to\infty} w_n, \Psi\right) = \max(v, \Psi) = v.$$

Ceci montre que v peut être approché par des fonctions de $K \cap \mathcal{C}_c^0(\Omega)$.

Étape 2. Nous montrons maintenant que toute fonction $v \in \mathcal{C}_c^0(\Omega) \cap K$ peut être approché (en norme H^1) par des fonctions de $\mathcal{C}_c^\infty(\Omega) \cap K$. Pour cela, on considère

une suite $\rho_n \in \mathcal{C}_c^\infty(\mathbb{R}^d, \mathbb{R}^+)$ telle que $\int \rho_n = 1$, et $\mathrm{spt}(\rho_n) \subseteq B(0, r_n)$ où la suite (r_n) est décroissante et tend vers 0. Quitte à renuméroter la suite, on peut supposer que

$$2r_1 \leq \gamma := \min\left(\min_{v(x)\neq 0} \mathrm{d}(x, \partial\Omega\}), \min_{\Psi(x)>0} \mathrm{d}(x, \partial\Omega)\right)$$

de sorte que $w_n := \rho_n * v$ appartient à $\mathcal{C}_c^\infty(\Omega)$ et que de plus, pour tout point x à distance $\leq r_1$ de $\partial\Omega$, on a $w_n(x) = 0 \geq \Psi(x)$. De plus, w_n converge uniformément et pour $\|\cdot\|_{H^1}$ vers v. Il se peut que cependant $w_n \not\geq \Psi$, et l'on pose donc

$$\varepsilon_n := \max_{\overline{\Omega}} \max(\Psi - w_n, 0) \geq 0,$$

de sorte que $w_n + \varepsilon_n \geq \Psi$ et comme $w_n \xrightarrow{\|\cdot\|_\infty} v \geq \Psi$ on sait que ε_n converge vers zéro lorsque $n \to \infty$. Soit enfin χ une fonction $\mathcal{C}_c^\infty(\Omega)$ telle que $\chi \geq 1$ sur l'ensemble

$$\Omega' = \{x \in \Omega \mid \mathrm{d}(x, \partial\Omega) > r_1\}.$$

On pose $v_n = w_n + \varepsilon_n \chi$, de sorte que

$$\begin{aligned} &\forall x \notin \Omega', \quad w_n(x) + \chi(x)\varepsilon_n \geq v_n(x) = 0 \geq \Psi(x) \\ &\forall x \in \Omega', \quad w_n(x) = v_n(x) + \varepsilon_n \geq \Psi(x), \end{aligned}$$

ce qui montre que $v_n \in K$. Enfin, on vérifie facilement que v_n converge vers v dans $\mathcal{C}^\infty(\overline{\Omega})$ et donc à fortiori dans $H^1(\Omega)$. □

Remarque 5.4. La démonstration du théorème se généralise *mutatis mutandis* aux hypothèses suivantes :

(a) $J : V \to \mathbb{R}$ est uniformément convexe et continue
(b) K est un convexe fermé
(c) Pour toute suite $(v_h)_h$ bornée telle que $v_h \in K_h$, tous les points d'adhérence faible de (v_h) sont dans K
(d) Il existe $\mathcal{K} \subseteq V$ et tel que $\overline{\mathcal{K}} = K$ et $r_h : \mathcal{K} \to K_h$ tel que

$$\forall v \in \mathcal{K}, \quad \lim_{h\to 0} r_h v = v \text{ fortement.}$$

5.4 Exercices

Rappel : projection sur un convexe fermé Dans les deux exercices ci-dessous, on utilisera le théorème suivant : si V est un espace de Hilbert, et si $K \subseteq V$ est un convexe fermé, alors pour tout $u \in V$, le problème de minimisation suivant

$$\inf_{v\in K} \|u - v\|^2$$

admet un unique minimiseur, noté $\Pi_K u$, appelé *projection orthogonale* de u sur K,. $\Pi_K u$ est l'unique élément de K vérifiant $\|u - \Pi_K u\| = \min_{v\in K} \|u - v\|$.

Exercice 5.1. *Problème d'obstacle sur $\Omega =]0,1[$ (examen 2019).*
Soit $V = H_0^1(\Omega)$ muni de la norme $\|v\|_V^2 = \int_0^1 v'(x)^2 \mathrm{d}x$. On rappelle que comme $d = 1$, les fonctions de V admettent un unique représentant continu, permettant de considérer V comme un sous-espace vectoriel de $\mathcal{C}^0(\overline{\Omega})$. Enfin, on rappelle que si (v_n) converge vers $v \in V$ pour $\|\cdot\|_V$, alors (v_n) converge uniformément vers v.

Problème d'obstacle Soit $\Psi \in V$ et $K = \{v \in V \mid v \geq \Psi\}$ Le problème d'obstacle s'écrit

$$E := \min_{v \in K} \|v\|_V^2 \tag{5.3}$$

Problème d'obstacle discret Étant donné $N \geq 1$ on considère $(x_i)_{0 \leq i \leq N+1}$ dans Ω, où $x_i = hi$ et $h = 1/(N+1)$. On note $\mathbb{P}^1 \subseteq \mathcal{C}^0(\mathbb{R})$ les fonctions affines et

$$V_N = \{v \in \mathcal{C}^0(\overline{\Omega}) \mid \forall 0 \leq i \leq N,\, v|_{[x_i, x_{i+1}]} \in \mathbb{P}^1 \text{ et } v(0) = v(1) = 0\}$$

$$K_N = \{v \in V_N \mid \forall 1 \leq i \leq N, v(x_i) \geq \Psi(x_i)\}.$$

Le problème d'obstacle discret s'écrit alors

$$E_N := \min_{v_N \in K_N} \|v\|_V^2 \tag{5.4}$$

1. Démontrer que K et K_N sont convexes et fermés, et que les problèmes (5.3) et (5.4) admettent chacun un unique minimiseur, noté $u \in V$ et $u_N \in V_N$.
2. Pour $v \in V$, donner l'expression de l'unique fonction $\tilde{v} \in V_N$ telle que $\forall i \in [\![0, N+1]\!]$, $\tilde{v}(x_i) = v(x_i)$. Montrer que $r_N : v \in V \mapsto \tilde{v} \in V_N$ est linéaire.
3. Démontrer que pour tout $v \in \mathcal{C}^1(\overline{\Omega}) \cap V$ et $\tilde{v}$ comme dans la question précédente,

$$\int_{x_i}^{x_{i+1}} \left|\tilde{v}'(x)\right|^2 \mathrm{d}x \leq \int_{x_i}^{x_{i+1}} \left|v'(x)\right|^2 \mathrm{d}x.$$

   En déduire que $\|r_N v\|_V^2 \leq \|v\|_V^2$ puis que r_N est continue.
4. Démontrer que $\forall v \in V$, $\lim_{N \to +\infty} \|v - r_N v\|_V = 0$.
5. Dans cette question, on démontre que $(u_N)_{N \geq 1}$ converge faiblement vers u.
   a) Démontrer que si $v \in K$, alors $r_N v \in K_N$, en déduire $E_N \leq E$.
   b) Montrer que si $(u_{\sigma(N)})_{N \geq 1}$ est une sous-suite de $(u_N)_{N \geq 1}$ qui converge faiblement vers $\overline{u} \in V$, alors $\|\overline{u}\|^2 \leq \|u\|^2$.
   c) En admettant que $\overline{u} \in K$, en déduire que $\overline{u} = u$.
   d) Montrer que la suite $(u_N)_{N \geq 1}$ est bornée, conclure.

Exercice 5.2. *Théorème de Stampacchia.* Soit V un espace de Hilbert, $a : V \times V \to \mathbb{R}$ une forme bilinéaire continue et α-coercive et $K \subseteq V$ un convexe fermé. L'objectif est de démontre l'existence et l'unicité de $u \in V$ tel que

$$\forall v \in K, \quad a(u, v - u) \geq 0. \tag{5.5}$$

(Remarquer la similarité avec la caractérisation de la solution au problème d'obstacle, dans la proposition 5.5.)

Pour cela, on considèrera (comme dans la démonstration du théorème de Lax-Milgram) l'opérateur $A : V \to V$ défini par

$$\forall v \in V, \langle A(u)|v\rangle = a(u,v).$$

On admet que la projection orthogonale sur K, définie par (**??**) vérifie

$$\forall v \in K, \quad \langle u - \Pi_K u | u - v\rangle \geq 0.$$

1. Démontrer que u est solution de (5.5) si et seulement si u est point fixe de l'application $T_\tau : v \in V \mapsto \Pi_K(v - \tau A(v))$, pour $\tau > 0$.
2. Montrer que l'application Π_K vérifie $\forall u, v \in V, \langle \Pi_K u - \Pi_K v | u - v\rangle \geq 0$, puis que Π_K est 1-Lipschitzienne.
3. En suivant la démonstration du théorème de Lax-Milgram (théorème 4.1), montrer que si $\tau > 0$ est suffisamment petit, la fonction T_τ est contractante.
4. Conclure.