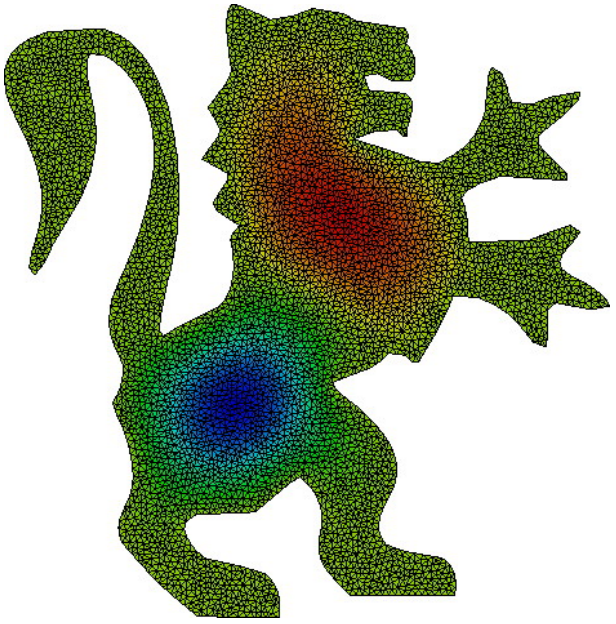
ÉCOLE CENTRALE LYON

Méthodes numériques pour les EDP

Différences finies, éléments finis

Grégory Vial



Module Ouvert Disciplinaire 3A

28 janvier 2019

Table des matières

Chapitre 1

Rappels sur les méthodes de différences finies

1.1 Introduction

Nous allons revenir, à l'aide d'un exemple, sur les principes et difficultés de l'approximation par différences finies d'une équations aux dérivées partielles (EDP) de type convection-réaction-diffusion. Considérons donc le problème, posé pour $x \in \Omega \subset \mathbb{R}$ et $t > 0$:

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} - \nu\dfrac{\partial^2 u}{\partial x^2} = \gamma u\left(1 - \dfrac{u}{K}\right), \\ u|_{t=0} = \varphi_0, \\ \Omega = \mathbb{R} \quad \text{ou} \quad u|_{\partial\Omega} = g \quad \text{ou} \quad \dfrac{\partial u}{\partial n}|_{\partial\Omega} = h. \end{cases} \tag{1.1}$$

Le « ou » est ici exclusif. Ainsi, des *conditions aux limites* (ou *conditions aux bord*) sont imposées uniquement dans le cas où le domaine (ouvert) Ω n'est pas $\mathbb{R}$ tout-entier. La condition $u|_{\partial\Omega} = g$ est appelée condition de *Dirichlet*, tandis que $\frac{\partial u}{\partial n}|_{\partial\Omega} = h$ est une condition dite de *Neumann*. La condition $u|_{t=0} = \varphi_0$, quant à elle, est appelée *condition initiale*.

Les paramètres sont choisis tels que

$$c \in \mathrm{Lip}(\mathbb{R} \times]0, +\infty[, \mathbb{R}), \quad \alpha, \nu, \gamma, K > 0.$$

On peut interpréter le problème comme suit :

- $u(x,t)$ est une *densité* (ou concentration, de population bactérienne, par exemple) au point x et à l'instant t,
- $c(x,t)$ est la *vitesse d'advection* (ou convection) de la population, par exemple sous l'effet d'un écoulement. Elle peut dépendre des variables d'espace-temps, mais pourra supposée constante au besoin.
- ν est un coefficient de *diffusion*, rendant compte de la propension de la population à se déplacer des fortes concentrations vers les plus faibles, et cela proportionnellement au gradient de densité (cf. lois de Fick ou Fourier),
- γ est un *taux de croissance*,
- K prend en compte une limitation de la croissance de la population due à des ressources limitées.

1.2 Cas homogène en espace : EDO

Si on cherche des solution au problème (1.1) indépendantes de la variable d'espace, on est ramené à l'étude d'une équation différentielle ordinaire (EDO) :

$$\begin{cases} u'(t) = \gamma u(t) \left(1 - \frac{u(t)}{K}\right), \\ u(0) = \varphi_0, \end{cases} \tag{1.2}$$

où $\varphi_0 \in \mathbb{R}$ est une constante donnée. Il est possible de résoudre explicitement ce problème de Cauchy ou d'en faire une étude qualitative pour montrer qu'il y a existence globale et que, si $\varphi_0 \in]0, K[$, alors pour tout temps $t > 0$, on a $u(t) \in]0, K[$.

La résolution numérique d'une équation de ce type [1] peut être effectuée à l'aide d'une méthode de type Runge-Kutta. Ayant fixé un pas de temps $\Delta t > 0$, on construit une suite $(u_n)_{n\geq 0}$ (on choisit, bien sûr, pour u_0 la condition initiale donnée φ_0) dont on vérifiera qu'elle fournit une approximation de $(u(n\Delta t)$. Voici trois schémas simples :

◇ *La méthode d'Euler explicite.*

$$u_{n+1} = u_n + \gamma \Delta t\, u_n \left(1 - \frac{u_n}{K}\right).$$

◇ *La méthode d'Euler implicite.*

$$u_{n+1} = u_n + \gamma \Delta t\, u_{n+1} \left(1 - \frac{u_{n+1}}{K}\right).$$

◇ *Une méthode semi-implicite.*

$$u_{n+1} = u_n + \gamma \Delta t\, u_n \left(1 - \frac{u_{n+1}}{K}\right).$$

La méthode explicite est très simple à mettre en œuvre, puisqu'un simple calcul permet de déterminer u_{n+1} à partir de u_n. En revanche, la méthode implicite nécessite la résolution d'une équation à chaque étape. L'équation est ici quadratique, donc de résolution peu coûteuse ; toutefois pour des EDO plus complexes, la mise en place d'une méthode de type Newton peut s'avérer nécessaire. La méthode semi-implicite, quant à elle, est permise par la forme particulière de l'équation, et permet d'exprimer u_{n+1} facilement en fonction de u_n.

Exercice 1

Montrer que pour le schéma semi-implicite, si $\varphi_0 \in]0, K[$, alors $u_n \in]0, K[$ pour tout $n \geq 0$. Qu'en est-il pour les deux autres schémas ?

Les trois méthodes convergent lorsque le pas de temps Δt tend vers 0, dans le sens suivant : pour chaque N fixé,

$$\max_{n=0}^{N} |u_n - u(n\Delta t)| \longrightarrow 0 \quad \text{lorsque} \quad \Delta t \to 0.$$

Pour les trois méthodes proposées, la vitesse de convergence est d'ordre $\mathcal{O}(\Delta t)$. On renvoie à [CM84] pour l'analyse des méthodes d'approximation des EDO, et la description de schémas plus précis.

Résumer consistante/stabilité/ordre

1. Bien sûr, on dispose ici d'une formule explicite, et la résolution numérique n'est pas nécessaire. Toutefois, les méthodes présentées s'appliquent à une classe d'EDO plus large.

1.3 Cas purement convectif dans $\mathbb{R}$: l'équation de transport linéaire

On se place dans le cas $\Omega = \mathbb{R}$, et $\nu = \gamma = 0$. Le problème s'écrit alors

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) + c(x,t)\dfrac{\partial u}{\partial x}(x,t) = 0, \\ u|_{t=0} = \varphi_0, \end{cases} \tag{1.3}$$

où φ_0 est une fonction donnée.

1.3.1 Aspect théorique : méthode des caractéristiques

Pour $x_0 \in \mathbb{R}$ fixé, on introduit $t \mapsto X(t)$, solution de l'EDO

$$\begin{cases} X'(t) = c\left(X(t), t\right) \\ X(0) = x_0. \end{cases} \tag{1.4}$$

L'hypothèse de continuité et la condition de Lipschitz globale exigées à la vitesse c garantissent l'existence et l'unicité à ce problème. Alors, on vérifie que la fonction $t \mapsto u\left(X(t), t\right)$ est constante. Ainsi,

$$\forall t > 0, \quad u\left(X(t), t\right) = \varphi_0(x_0).$$

Ainsi, étant donnés $(x,t) \in \mathbb{R} \times \mathbb{R}_+^*$, pour déterminer la valeur de $u(x,t)$, on calcule $x_0 \in \mathbb{R}$ tel que la *caractéristique* $X(t)$ issue de x_0 au temps initial, passe par x au temps t. On appelle x_0 *pied de la caractéristique*.

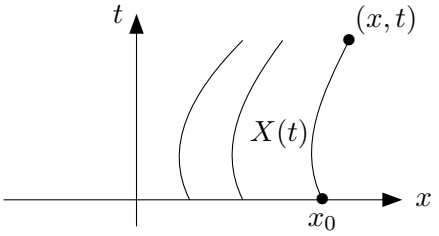


FIGURE 1.1 – Caractéristiques pour l'équation de transport.

1.3.2 Aspect numérique : décentrement

Dans ce paragraphe, on considère que la vitesse c est constante. Ainsi, les caractéristiques sont des droites, et la solution exacte est donnée par $u(x,t) = \varphi_0(x - ct)$. Pour l'approximation numérique, on fixe $\Delta t, \Delta x > 0$, et on propose les schémas suivants [2] pour construire une approximation u_j^n de $u(n\Delta t, j\Delta x)$:

2. Bien entendu, une résolution numérique n'est pas nécessaire puisqu'on dispose d'une formule simple pour la solution exacte. Le but est ici d'éprouver les méthodes numériques dans une situation où, précisément, on dispose de la solution exacte afin d'évaluer l'erreur commise.

◇ *Schéma décentré à gauche.*

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_j^n - u_{j-1}^n}{\Delta x} = 0.$$

◇ *Schéma décentré à droite.*

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_j^n}{\Delta x} = 0.$$

◇ *Schéma centré.*

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + c\frac{u_{j+1}^n - u_{j-1}^n}{2\Delta x} = 0.$$

Les trois schémas sont *explicites* : si les $(u_j^n)_j$ sont connus au rang n, alors on peut calculer les $(u_j^{n+1})_j$ par une formule du type

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

On définit *l'erreur de consistance locale* par

$$\varepsilon_j^n = u(x_j, t_{n+1}) - \Phi\left(u(x_j, t_n), u(x_{j-1}, t_n), u(x_{j+1}, t_n)\right), \tag{1.5}$$

où $x_j = j\Delta x$, $t_n = n\Delta t$ et u désigne la solution exacte de l'équation.

Exercice 2

À l'aide de développements de Taylor, montrer que, sous une hypothèse de régularité sur u_0, il existe une constante C telle que

$$\forall n \in \mathbb{N}, \quad \forall j \in \mathbb{Z}, \quad |\varepsilon_j^n| \leq C\Delta t(\Delta t^p + \Delta x^q), \tag{1.6}$$

avec $p = 1$ pour les trois schémas, et $q = 1$ pour les schémas décentrées, $q = 2$ pour le schéma centré.

Au vu de l'étude de consistance, le schéma centré semble plus précis. Toutefois, l'analyse de stabilité va montrer que cette méthode est inutilisable. Nous envisageons ici deux notions de stabilité :

Stabilité L^∞ La solution exacte satisfait

$$t \longmapsto \|u(\cdot, t)\|_{\mathrm{L}^\infty(\mathbb{R})} \quad \text{est décroissante.}$$

(en fait cette fonction est même constante). On souhaite que cette propriété soit conservée au niveau discret, c'est-à-dire

$$\forall n \in \mathbb{N}, \quad \sup_{j\in\mathbb{Z}} |u_j^{n+1}| \leq \sup_{j\in\mathbb{Z}} |u_j^n|.$$

Pour chaque méthode, u_j^{n+1} apparaît comme une combinaison linéaire de u_j^n, u_{j-1}^n et u_{j+1}^n. Par exemple, pour le schéma décentré à gauche,

$$u_j^{n+1} = (1-\alpha)u_j^n + \alpha u_{j-1}^n \quad \text{avec} \quad \alpha = c\frac{\Delta t}{\Delta x}.$$

Ainsi, cette combinaison est convexe si et seulement si $\alpha \in [0,1]$. On en déduit que le schéma décentré à gauche est (asymptotiquement) stable au sens L^∞ si et seulement si

$$0 \leq c\frac{\Delta t}{\Delta x} \leq 1. \tag{1.7}$$

Cette condition est dite CFL (Courant-Friedrich-Lévy), et impose que le pas de temps soit petit si le pas d'espace est choisi petit, ce d'autant plus que la vitesse est grande.

De la même façon, le schéma décentré à droite est (asymptotiquement) stable au sens L^∞ sous la condition $\alpha \in [-1,0]$. Il n'est donc pas adapté à des vitesses positives.

Enfin, le schéma centré ne correspond jamais à une combinaison linéaire convexe. Il n'est jamais (asymptotiquement) stable au sens L^∞, quels que soient les choix de pas de temps et d'espace.

Exercice 3

Pour le schéma centré, exhiber un exemple de condition initiale qui montre que la stabilité L^∞ n'est pas satisfaite.

Stabilité L^2 De la même façon, la norme L^2 est conservée au cours du temps pour la solution continue. On souhaite donc avoir au niveau discret

$$\forall n \in \mathbb{N}, \quad \sum_{j\in\mathbb{Z}} |u_j^{n+1}|^2 \leq \sum_{j\in\mathbb{Z}} |u_j^n|^2. \tag{1.8}$$

Un calcul direct montre que, pour le schéma décentré à gauche,

$$\sum_{j\in\mathbb{Z}} |u_j^{n+1}|^2 = \left[(1-\alpha)^2+\alpha^2\right] \sum_{j\in\mathbb{Z}} |u_j^n|^2 + 2\alpha(1-\alpha) \sum_{j\in\mathbb{Z}} |u_j^n u_{j-1}^n|$$

Lorsque $\alpha \in [0,1]$, on peut écrire en utilisant la majoration $2ab \leq a^2+b^2$,

$$\begin{aligned}\sum_{j\in\mathbb{Z}} |u_j^{n+1}|^2 &\leq \left[(1-\alpha)^2+\alpha^2\right] \sum_{j\in\mathbb{Z}} |u_j^n|^2 + \alpha(1-\alpha) \sum_{j\in\mathbb{Z}} \left(|u_j^n|^2 + |u_{j-1}^n|^2\right) \\ &= \sum_{j\in\mathbb{Z}} |u_j^n|^2\end{aligned}$$

On peut aussi montrer que la condition $\alpha \in [0,1]$ est nécessaire en exhibant un exemple. Toutefois, la méthode utilisée n'est pas très générale et peut s'avérer inopérante pour l'étude d'autres schémas. On lui préfère la *méthode de Von Neumann*, qui fait appel à l'analyse de Fourier. Définissons, en effet, la fonction w^n constante par morceaux sur $\mathbb{R}$ par

$$w^n(x) = u_j^n \quad \text{pour} \quad x \in \left[x_j - \frac{\Delta x}{2}, x_j + \frac{\Delta x}{2}\right[.$$

Le schéma décentré à gauche s'interprète comme suit à l'aide de la fonction w^n :

$$w^{n+1}(x) = (1-\alpha)w^n(x) + \alpha\left(\tau_{-\Delta x} w^n\right)(x), \tag{1.9}$$

où τ_h désigne l'opérateur de translation : $(\tau_h f)(x) = f(x+h)$. On prend alors la transformée de Fourier [3] de l'égalité (1.9) :

$$\begin{aligned}\widehat{w^{n+1}}(\xi) &= (1-\alpha)\widehat{w^n}(\xi) + \alpha e^{i\Delta x\xi}\widehat{w^n}(\xi) \\ &= \underbrace{\left(1-\alpha+\alpha e^{i\Delta x\xi}\right)}_{=\rho(\xi)}\widehat{w^n}(\xi).\end{aligned}$$

D'après l'égalité de Plancherel,

$$\int_{\mathbb{R}} |\widehat{w^n}(\xi)|^2 \,\mathrm{d}\xi = 2\pi \int_{\mathbb{R}} |w^n(\xi)|^2 \,\mathrm{d}xi = 2\pi \sum_{j\in\mathbb{Z}} |u_j^n|^2.$$

Aussi la stabilité asymptotique au sens L^2 équivaut-elle à la condition

$$\forall \xi \in \mathbb{R} \quad |\rho(\xi)| \leq 1.$$

Or

$$|\rho(\xi)|^2 = 1 - 2\alpha(1-\alpha)\left(1-\cos(\xi\Delta x)\right),$$

et on retrouve bien la condition CFL mise en évidence pour la stabilité L^∞ : $\alpha \in [0,1]$.

Interprétation graphique de la CFL La condition de stabilité exhibée pour le schéma décentré à gauche :

$$c\frac{\Delta t}{\Delta x} \leq 1$$

peut s'interpréter à l'aide de la figure 1.2. En effet, les points du quadrillage intervenant dans le calcul de u_j^{n+1} sont grisés, et *in fine*, la valeur de u_j^{n+1} est une combinaison linéaire des valeurs de la condition initiale φ_0 en les points de la base du triangle. La condition CFL signifie que la caractéristique exacte est située à l'intérieur du *domaine de dépendance numérique*.

Convergence de la méthode À l'aide de l'étude de consistance, cf. Exercice 2, et celle liée à la stabilité, on est en mesure d'établir un résultat de convergence de la méthode. En effet, considérons une méthode de la forme des schémas précédents, à savoir

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

et notons $v_j^n = u(x_j, t_n)$, évaluation de la solution exacte de l'équation sur la grille. Alors, par définition de l'erreur de consistance locale, cf. (1.5), on a bien sûr

$$v_j^{n+1} = \Phi(v_j^n, v_{j-1}^n, v_{j+1}^n) + \varepsilon_j^n.$$

3. Si on suppose, par exemple, la donnée initiale φ_0 à support compact, la transformée de Fourier est à comprendre au sens de fonctions intégrables, donnée par la formule

$$\widehat{f} = \int_{\mathbb{R}} e^{ix\xi} f(x) \,\mathrm{d}x.$$

Si la donnée initiale n'est pas dans $L^1(\mathbb{R})$, on doit recourir à la transformée de Fourier au sens des distributions tempérées

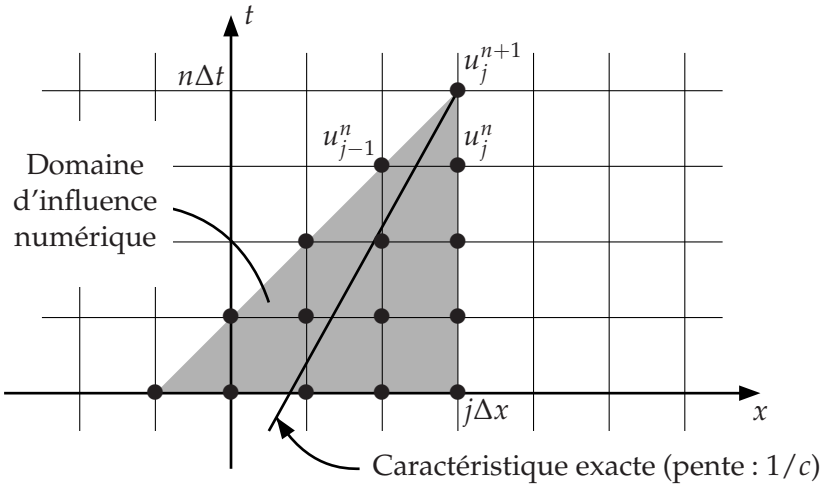


FIGURE 1.2 – Caractéristique exacte et domaine de dépendance numérique

Ainsi, par différence, en exploitant la linéarité de la fonction Φ vis-à-vis de ses différents arguments, en notant $e_j^n = u_j^n - v_j^n$,

$$e_j^{n+1} = \Phi(e_{j-1}^n, e_j^n, e_{j+1}^n) - \varepsilon_j^n.$$

Par ailleurs, la stabilité fournit la majoration suivante :

$$|e_j^{n+1}| \leq |e_j^n| + |\varepsilon_j^n|.$$

Par une récurrence immédiate, en notant que $e_j^0 = 0$, on obtient l'estimation

$$|e_j^n| \leq \sum_{k=0}^{n-1} |\varepsilon_j^k|.$$

Enfin, en fixant un temps final $T > 0$, et choisissant le pas de temps Δt tel que $T = N\Delta t$ avec $N \in \mathbb{N}$, on peut écrire à l'aide de l'estimation de consistance (1.6),

$$\sup_{0\leq n\leq N} \sup_{j\in\mathbb{Z}} \left| u_j^n - u(x_j, t_n) \right| \leq C \sum_{k=0}^{N} \Delta t(\Delta t^p + \Delta x^q) = CT(\Delta t^p + \Delta x^q),$$

qui exprime la convergence d'ordre p en temps et q en espace de la solution discrète vers la solution continue [4].

1.4 Cas purement diffusif : l'équation de la chaleur

On considère le problème

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} = 0, \\ u|_{t=0} = \varphi_0, \\ \Omega = \mathbb{R} \quad \text{ou} \quad u|_{\partial\Omega} = g \quad \text{ou} \quad \dfrac{\partial u}{\partial n}|_{\partial\Omega} = h. \end{cases} \tag{1.10}$$

4. Cette preuve justifie le fait d'avoir isolé le facteur Δt dans l'estimation de consistance.

1.4.1 Cas stationnaire : l'équation de Laplace en domaine borné

Conditions de Dirichlet

Si la solution est supposée indépendante du temps, on est ramené au problème

$$\begin{cases} -\nu \dfrac{\partial^2 u}{\partial x^2} = f(x), \\ u(0) = g(0), \quad u(1) = g(1). \end{cases} \tag{1.11}$$

Notons que l'ajout d'un second membre $f(x)$ ne modifie pas la linéarité de l'équation, et a pour seul but de montrer l'influence d'un éventuel second membre sur les méthodes présentées dans ce paragraphes. Bien sûr, ce problème se résout aisément *à la main*, mais il est instructif d'étudier une méthode d'approximation de ce type d'équation. On discrétise donc l'intervalle $]0,1[$: un entier J étant donné, on note

$$\Delta x = \frac{1}{J+1} \quad \text{et} \quad x_j = j\Delta x \quad (0, 1, \ldots, J+1).$$

On va construire un vecteur $(u_0, u_1, \ldots, u_{J+2})$ dont on espère qu'il approchera la solution u de l'équation aux points $x_0, x_1, \ldots, x_{J+1}$.

Au vu des conditions aux limites, le choix $u_0 = g(0)$ et $u_{J+1} = g(1)$ s'impose. Pour $1 \leq j \leq J+1$, on écrit

$$-u''(x_j) \simeq \frac{-u(x_{j-1}) + 2u(x_j) - u(x_{j+1})}{\Delta x^2},$$

si bien qu'il est naturel d'écrire

$$\frac{-u_{j-1} + 2u_j - u_{j+1}}{\Delta x^2} = f(x_j). \tag{1.12}$$

Avec la notation$f_j = f(x_j)$, on pose

$$A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}, \quad U = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{J-1} \\ u_J \end{bmatrix}, \quad F = \begin{bmatrix} f_1 + \frac{g(0)}{\Delta x^2} \\ f_2 \\ \vdots \\ f_{J-1} \\ f_J + \frac{g(1)}{\Delta x^2} \end{bmatrix},$$

si bien que le problème discret s'écrit sous la forme du système linéaire carré de taille $J \times J$

$$\frac{1}{\Delta x^2} AU = F.$$

On peut montrer que la méthode obtenue est d'ordre 2 dès lors que la solution u est assez régulière. Précisément, on a le résultat suivant : si $f \in \mathscr{C}^2([0,1])$, alors $u \in \mathscr{C}^4([0,1])$ et il existe une constante $C > 0$ telle que

$$\max_{0 \leq j \leq J+1} |u(x_j) - u_j| \leq C\Delta x^2 \max_{x \in [0,1]} |u^{(4)}(x)|.$$

On renvoie à [Cia82] pour une preuve détaillée.

Conditions de Neumann

Les conditions aux limites deviennent $u'(0) = -h(0)$ et $u'(1) = h(1)$. Ainsi, elles ne fournissent plus directement les valeurs de u_0 et u_{J+1}. Il est donc naturel de considérer ici un vecteur de $J+2$ inconnues :

$$V = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_J \\ u_{J+1} \end{bmatrix} \in \mathbb{R}^{J+2}.$$

Aux J équations déjà écrites à travers (1.12), on adjoint deux équation supplémentaires issues des conditions aux limites. Une première idée consiste à écrire

$$\frac{u_1 - u_0}{\Delta x} = -h(0) \quad \text{et} \quad \frac{u_{J+1} - u_J}{\Delta x} = h(1). \tag{1.13}$$

Le système obtenu s'écrit alors $\dfrac{1}{\Delta x^2} BV = G$, avec

$$B = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix}, \quad G = \begin{bmatrix} \frac{h(0)}{\Delta x} \\ f_1 \\ \vdots \\ f_J \\ -\frac{h(1)}{\Delta x} \end{bmatrix}$$

Remarque 1.1

La matrice obtenue n'est pas inversible (la somme des lignes valant 0, le vecteur ne contenant que des 1 apartient à son noyau). Cela est cohérent avec le problème continu, qui n'est pas bien posé : si u, satisfait $-u'' = f$ et $u'(0) = -h(0)$, $u'(1) = h(1)$, alors $u + k$ également, pour toute constante k. On considèrera donc plutôt l'équation $-u'' + u = f$, pour laquelle les conditions de Neumann fournissent une unique solution.

Toutefois, cette méthode n'est pas très performante car le choix d'approximation des conditions de Neumann conduit à une détérioration de l'ordre global :

$$\max_{0 \le j \le J+1} |u(x_j) - u_j| = \mathcal{O}(\Delta x).$$

Il est possible de retrouver l'ordre 2 en écrivant un développement limité à l'ordre 1

$$u(\Delta x) = u(0) + \Delta x\, u'(0) + \frac{\Delta x^2}{2} u''(0) + \mathcal{O}(\Delta x^3),$$

soit encore en utilisant l'équation satisfaite par u

$$u(\Delta x) = u(0) + \Delta x\, u'(0) - \frac{\Delta x^2}{2} f(0) + \mathcal{O}(\Delta x^3).$$

On obtient ainsi les approximations suivantes au lieu de (1.13) :

$$\frac{u_1 - u_0}{\Delta x^2} = -\frac{h(0)}{\Delta x} - \frac{f(0)}{2} \quad \text{et} \quad \frac{u_{J+1} - u_J}{\Delta x^2} = \frac{h(1)}{\Delta x} - \frac{f(1)}{2}. \tag{1.14}$$

Le second membre est alors transformé en

$$G = \begin{bmatrix} \frac{f(0)}{2} + \frac{h(0)}{\Delta x} \\ f_1 \\ \vdots \\ f_J \\ \frac{f(1)}{2} - \frac{h(1)}{\Delta x} \end{bmatrix}.$$

1.4.2 Cas instationnaire en domaine borné

Pour simplifier la présentation, on se restreint ici à des conditions de Dirichlet homogènes. Le problème s'écrit

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} = 0, \\ u|_{t=0} = \varphi_0, \\ u|_{x=0} = u|_{x=1} = 0. \end{cases} \tag{1.15}$$

Remarque 1.2

On rappelle ici l'expression de la solution sous forme d'une série de Fourier : on prolonge φ_0 par imparité, puis par 2-périodicité, et on écrit son développement

$$\varphi(x) = \sum_{n=1}^{+\infty} \beta_n \sin(n\pi x).$$

Alors la solution de (1.15) *s'écrit*

$$u(x,t) = \sum_{n=1}^{+\infty} \beta_n e^{-\nu\pi^2 n^2 t} \sin(n\pi x).$$

Cette formule permet, en particulier, de vérifier que $t \mapsto u(x,t)$ est décroissante et tend vers 0 à l'infini.

Insistons sur le fait que la choix de la base des sinus est dictée par les conditions aux limites. Si l'on avait considéré des conditions de Neumann $u'(0) = u'(1) = 0$, il aurait fallu prolonger par parité, et utiliser la base $(\cos(n\pi x))_n$.

Semi-discrétisation en espace

Comme pour l'équation stationnaire, on fixe un entier J et Δx tel que $(J+1)\Delta x = 1$. On pose aussi

$$U(t) = \begin{bmatrix} u_1(t) \\ u_2(t) \\ \vdots \\ u_{J-1}(t) \\ u_J(t) \end{bmatrix}, \quad \text{et} \quad U_0 = \begin{bmatrix} \varphi_0(x_1) \\ \varphi_0(x_2) \\ \vdots \\ \varphi_0(x_{J-1}) \\ \varphi_0(x_J)(t) \end{bmatrix}$$

où les $u_j(t)$ ont pour but d'approcher $u(j\Delta x)$. L'approximation de la dérivée seconde à trois points, cf. (1.12), fournit l'écriture matricielle suivante après discrétisation en espace :

$$\begin{cases} U'(t) = -\frac{\nu}{\Delta x^2} A U(t), \\ U(0) = U_0 \end{cases} \tag{1.16}$$

Il s'agit d'un système différentiel linéaire pour lequel on dispose des méthodes de résolution numérique des EDO, cf. § 1.2.

Discrétisation en temps

On se concentre ici sur les deux schémas fournis par les méthodes d'Euler.

◇ *Méthode d'Euler explicite.*

$$U^{n+1} = U^n - \nu \frac{\Delta t}{\Delta x^2} A U^n.$$

◇ *Méthode d'Euler implicite.*

$$U^{n+1} = U^n - \nu \frac{\Delta t}{\Delta x^2} A U^{n+1}.$$

Les deux méthodes s'écrivent sous la forme $U^{n+1} = BU^n$ avec $B = (Id - \alpha A)$ dans le cas explicite, et $B = (Id + \alpha A)^{-1}$ dans le cas implicite, si on a posé

$$\alpha = \nu \frac{\Delta t}{\Delta x^2}.$$

Stabilité (asymptotique) au sens L^∞. La question de la stabilité au sens L^∞ se traduit en termes matriciels au travers de la norme subordonnée de B :

$$\|B\|_\infty \leq 1, \quad \text{avec} \quad \|B\|_\infty = \sup_{\|x\|_\infty = 1} \|Bx\|_\infty.$$

Or, la norme infinie de la matrice B s'exprime en fonction de ses coefficients :

$$\|B\|_\infty = \max_i \sum_j |B_{ij}|$$

◇ *Cas explicite.* Le calcul de la norme est immédiat :

$$\|B\|_\infty = |1-2\alpha| + 2|\alpha| \begin{cases} = 1 & \text{si } 0 \leq 1-2\alpha \leq 1, \\ > 1 & \text{sinon.} \end{cases}.$$

Ainsi, la méthode explicite est (asymptotiquement) stable au sens L^∞ si et seulement si $0 \leq 1-2\alpha \leq 1$, ce qui s'écrit aussi

$$\nu \frac{\Delta t}{\Delta x^2} \leq \frac{1}{2}. \tag{1.17}$$

Cette condition CFL est plus contraignante que celle rencontrée dans le cas de l'équation de transport, cf. (1.7).

◇ *Cas implicite.* Il n'est pas immédiat de calculer la norme infinie d'une inverse. On peut montrer que, pour tout choix de $\alpha > 0$,

$$\|(Id + \alpha A\|_\infty \leq 1,$$

ce qui prouve que la méthode implicite est inconditionnellement [5] (asymptotiquement) stable.

Stabilité (asymptotique) au sens L^2. La même étude peut-être effectué en norme 2. Pour une matrice symétrique B, on a

$$\|B\|_2 = \rho(B) = \max\left\{|\lambda| \ ; \ \lambda \in \mathfrak{S}(B)\right\}.$$

On rappelle ici les valeurs propres de la matrice A :

$$\mathfrak{S}(A) = \left\{\lambda_k = 4\sin^2\left(\frac{k\pi}{2(J+1)}\right) \quad ; \quad k = 1, 2, \ldots, J\right\}.$$

◇ *Cas explicite.* Les valeurs propres de la matrice d'itération $B = Id - \alpha A$ s'expriment en fonction de celles de A :

$$\mathfrak{S}(Id - \alpha A) = \{1 - \alpha\lambda_k \ ; \ k = 1, 2, \ldots J\}.$$

Ainsi, la norme 2 de la matrice B est inférieure ou égale à 1 sous la condition $\alpha \leq \frac{1}{2}$. On retrouve la CFL (1.17).

◇ *Cas implicite.* De la même façon

$$\|(Id + \alpha A)^{-1}\|_2 = \max_k \frac{1}{1 + \alpha\lambda_k} \leq 1,$$

ce pour toute valeur de $\alpha > 0$. On retrouve le caractère inconditionnellement stable de la méthode implicite.

5. c'est-à-dire sans condition restrictive sur les pas de temps et d'espace.

1.4.3 Cas instationnaire en domaine non borné

Le problème s'écrit simplement

$$\begin{cases} \dfrac{\partial u}{\partial t} - \nu \dfrac{\partial^2 u}{\partial x^2} = 0, \\ u|_{t=0} = \varphi_0, \end{cases} \tag{1.18}$$

où $x \in \mathbb{R}$ et $t > 0$. Une expression analytique de la solution peut-être déterminée à l'aide de la transformation de Fourier :

$$u(x,t) = \frac{1}{\sqrt{4\pi\nu t}} \int_{\mathbb{R}} u_0(y) \exp\left(-\frac{(x-y)^2}{4\nu t}\right) \mathrm{d}y.$$

Le schéma ne peut plus, ici, s'écrire sous forme matricielle car l'indice d'espace j varie dans $\mathbb{Z}$ tout-entier. Toutefois, on peut transcrire le schéma explicite [6] terme-à-terme :

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

Si l'analyse de consistance est la même que précédemment, la stabilité ne peut plus être étudiée au travers d'une analyse de norme matricielle.

Stabilité (asymptotique) au sens L^∞ Le terme u_j^{n+1} apparaît comme combinaison linéaire des termes au rang précédent :

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

Cette combinaison est convexe si et seulement si $0 \leq 1-2\alpha \leq 1$, qui équivaut à la CFL (1.17).

Stabilité (asymptotique) au sens L^2 On utilise la méthode de Von Neumann, introduite page 9. Avec les mêmes notations, l'équivalent de la formule (1.9) s'écrit ici

$$w^{n+1} = w^n + \alpha\tau_{-\Delta x} w^n - 2\alpha w^n + \alpha\tau_{\Delta x} w^n.$$

En prenant la transformée de Fourier, on obtient

$$\widehat{w^{n+1}}(\xi) = \rho(\xi)\widehat{w^n}(\xi) \quad \text{avec} \quad \rho(\xi) = 1 - 2\alpha + 2\alpha\cos(\Delta x\xi).$$

À nouveau, la condition $|\rho(\xi)| \leq 1$ se ramène à la CFL (1.17).

1.5 Mise en place d'un schéma pour le problème complet

Après avoir étudié un certain nombre de situations particulières, on revient au problème initial (1.1), qu'on étudie dans deux cadres.

6. Le schéma implicite, quant à lui, peut difficilement être exprimé car le système linéaire à résoudre pour déterminer $(u_j^{n+1})_j$ est de dimension infinie. Notons, toutefois, que la mise en place sur machine se ramène nécessairement au cas fini, et que l'on peut alors considérer un schéma implicite.

1.5.1 Cas borné avec conditions de Dirichlet homogènes

Le problème s'écrit :

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} - \nu\dfrac{\partial^2 u}{\partial x^2} = \gamma u\left(1 - \dfrac{u}{K}\right), \\ u|_{t=0} = \varphi_0, \\ u|_{x=0} = u|_{x=1} = 0, \end{cases}$$

pour $x \in]0,1[$ et $t > 0$ (c est supposée constante positive). Comme plus haut, on note U^n le vecteur de discrétisation :

$$U^n = \begin{bmatrix} u_1^n \\ u_2^n \\ \vdots \\ u_{J-1}^n \\ u_J^n \end{bmatrix}.$$

Tirant parti des études faites précédemment, on prend soin de *décentrer* la partie convective à gauche (cf. $c > 0$), de traiter la partie diffusive de manière *implicite*, et le terme de réaction par un schéma *semi-implicite*. On est ainsi conduit à la méthode :

$$\frac{U^{n+1} - U^n}{\Delta t} + \frac{c}{\Delta x} D U^n + \frac{\nu}{\Delta x^2} A U^{n+1} = \gamma U^n (1 - U^{n+1}),$$

le produit $U^n(1 - U^{n+1})$ devant être compris composante à compostante. Les matrices D et A sont données par

$$D = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ -1 & 1 & \ddots & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & -1 & 1 \end{bmatrix} \quad \text{et} \quad A = \begin{bmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & -1 & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{bmatrix}$$

Exercice 4

> *Effectuer une étude de stabilité (asymptotique) au sens* L^2 *dans le cas où* $\gamma = 0$ *(dite* stabilité linéaire*). Exhiber une condition suffisante de type CFL.*

Exercice 5

> *Montrer que, sous la condition CFL déterminée dans l'exercice précédent, on a* $U^n \geq 0$ *(au sens où chaque composante est positive) dès que la donnée initiale* φ_0 *est positive.*

1.5.2 Cas non borné non réactif

Il s'agit d'approcher la solution du problème :

$$\begin{cases} \dfrac{\partial u}{\partial t} + c\dfrac{\partial u}{\partial x} - \nu\dfrac{\partial^2 u}{\partial x^2} = 0, \\ u|_{t=0} = \varphi_0, \end{cases}$$

pour $x \in \mathbb{R}$ et $t > 0$. Le schéma explicite décentré à gauche s'écrit terme-à-terme

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

L'étude de stabilité (asymptotique) au sens L^∞ est aisée car la combinaison linéaire définissant le schéma est convexe si et seulement si

$$0 \leq 1 - c\frac{\Delta t}{\Delta x} - 2\nu\frac{\Delta t}{\Delta x^2} \leq 1,$$

qui fournit la condition CFL à satisfaire.

Exercice 6

Étudier la stabilité (asymptotique) au sens L^2 *du schéma à l'aide de la méthode de Von Neumann.*

1.6 Extensions en dimension supérieure

En dimension supérieure ou égale à 2, la discrétisation s'effectue sur une *grille*, c'est-à-dire un quadrillage. Si l'on considère, par exemple, le rectangle $]a, b[\times]c, d[$, on introduit deux entiers J_x et J_y et les pas d'espace

$$\Delta x = \frac{b-a}{J_x + 1}, \qquad \Delta y = \frac{d-c}{J_y + 1}.$$

et les points de discrétisation $x_{j\ell} = (j\Delta x, \ell\Delta y)$. L'opérateur laplacien est alors approché par la différence finie, dite *laplacien à* 5 *points*

$$-\Delta u(x_{j\ell}) \simeq \frac{-u(x_{j-1,\ell}) + 2u(x_{j\ell}) - u(x_{j+1,\ell})}{\Delta x^2} + \frac{-u(x_{j,\ell-1}) + 2u(x_{j\ell}) - u(x_{j,\ell+1})}{\Delta y^2}.$$

La matrice qui intervient dans l'écriture du schéma est fonction de la numérotation globale que l'on choisit, i.e. de la fonction

$$\begin{array}{ccc} \{0, 1, \ldots, N_x + 1\} \times \{0, 1, \ldots, N_y + 1\} & \longrightarrow & \{0, 1, \ldots, (N_x + 1)(N_y + 1)\} \\ (j, \ell) & \longmapsto & n \end{array}$$

Pour l'ordre lexicographique de gauche à droite, la matrice est tri-diagonale par blocs – elle contient des éléments non-nuls sur 5 diagonales seulement : la diagonale principale, les sur- et sous-diagonales, et celles de rangs $\pm(J_x + 2)$.

Notons que la méthode des différences finies est bien adapté dans le cas où la géométrie du domaine Ω est simple (rectangles ou réunions de rectangles). Pour des cas plus complexes, on utilise plutôt la méthode des élements finis, ou la méthode des volumes finis, basées sur des triangulations du domaine.

Chapitre 2

Méthodes d'éléments finis

2.1 Introduction

On va proposer ici une alternative aux méthode de différences finies pour l'approximation des problèmes elliptiques, dont le prototype s'écrit

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

où Ω désigne un ouvert connexe (un *domaine*) de $\mathbb{R}^d$, et f une fonction de $\mathrm{L}^2(\Omega)$.

Formellement, on note V un espace de fonctions nulles sur le bord $\partial\Omega$, et on fixe une fonction $v \in V$ par laquelle on multiplie l'équation $-\Delta u = f$ par v, et l'intègre sur le domaine Ω :

$$-\int_\Omega \Delta u(x)v(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x.$$

À l'aide d'une intégration par parties (voir Annexe A), on obtient

$$-\int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x)\,\mathrm{d}\sigma_x + \int_\Omega \nabla u(x)\cdot\nabla v(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x.$$

Mais en exploitant le fait que v s'annule sur $\partial\Omega$, on obtient finalement

$$\forall v \in V, \quad \int_\Omega \nabla u(x)\cdot\nabla v(x)\,\mathrm{d}x = \int_\Omega f(x)v(x)\,\mathrm{d}x. \tag{PV}$$

Les questions suivantes se posent :

Question 1

Quel sens donner aux calculs formels précédents ?

Question 2

Les problèmes (P) *et* (PV) *admettent-ils des solutions ? les mêmes ?*

Un des intérêts de considérer le problème (PV) plutôt que le problème (P) réside dans l'approximation qui peut en être construite. En effet, notons V_h un sous-espace de dimension finie de l'espace V (l'indice h se référera en général à un pas de discrétisation par la suite, mais ce n'est pas une obligation), et introduisons le problème

$$\forall v_h \in V, \quad \int_\Omega \nabla u_h(x)\cdot\nabla v_h(x)\,\mathrm{d}x = \int_\Omega f(x)v_h(x)\,\mathrm{d}x, \tag{PV$_h$}$$

dont la solution u_h est recherchée dans le même espace V_h. Il est facile de voir que le problème (PV_h) est en fait un système linéaire carré. En effet, fixons une base $(\varphi_1, \varphi_2, \ldots, \varphi_N)$ de V_h et cherchons u_h sous la forme

$$u_h = \sum_{j=1}^{N} u_j \varphi_j.$$

L'égalité requise dans (PV_h) est linéaire vis-à-vis de v_h, donc il suffit de la vérifier pour les éléments de la base : (PV_h) se récrit donc

$$\forall i = 1, 2, \ldots, N \quad \int_\Omega \nabla \left(\sum_{j=1}^{N} u_j \varphi_j \right)(x) \cdot \nabla \varphi_(x)\, \mathrm{d}x = \int_\Omega f(x) \varphi_i(x)\, \mathrm{d}x,$$

soit encore par linéarité

$$\forall i = 1, 2, \ldots, N \quad \sum_{j=1}^{N} \left[\int_\Omega \nabla \varphi_j(x) \cdot \nabla \varphi_(x)\, \mathrm{d}x \right] u_j = \int_\Omega f(x) \varphi_i(x)\, \mathrm{d}x,$$

qui n'est autre que le système linéaire $\mathbb{A}\mathbb{U} = \mathbb{F}$, si l'on a posé

$$\mathbb{A}_{ij} = \int_\Omega \nabla \varphi_j(x) \cdot \nabla \varphi_(x)\, \mathrm{d}x, \quad \mathbb{U}_i = u_i \quad \text{et} \quad \mathbb{F}_i = \int_\Omega f(x) \varphi_i(x)\, \mathrm{d}x.$$

Ainsi, pour programmer sur machine la méthode de discrétisation proposée, dite *méthode de Galerkine*, il suffit de construite les matrice et second membre, et résoudre le système à l'aide d'une méthode appropriée (on pourra consulter [Cia82], par exemple, pour la résolution numérique des systèmes linéaires). De nouvelles questions surgissent naturellement autour de cette approximation.

Question 3

Le problème (PV_h) *admet-il une unique solution ? (ce qui revient à l'étude de l'inversiblité de la matrice* $\mathbb{A}$*).*

Question 4

Comment choisir le sous-espace V_h *? et la base* $(\varphi_1, \varphi_2, \ldots, \varphi_N)$ *?*

Question 5

Quelle est l'erreur commise entre u *et* u_h*, lorsque le sous-espace* V_h *« remplit »l'espace* V *?*

Dans la suite de ce chapitre, on va donner des réponses à ces questions (ou éléments de réponses pour certaines, dont la portée dépasse le cadre du cours).

2.2 Rappels d'analyse fonctionnelle

2.2.1 Espaces de Sobolev

Espaces de Sobolev $\mathrm{H}^m(\Omega)$

Le cadre des espaces $\mathscr{C}^m$ n'est pas adapté à une formulation de type (PV). On présente ici celui des espaces de Sobolev, qui permet de donner un sens à la discussion faite plus haut. Dans la suite, on considère un domaine Ω de $\mathbb{R}^d$, dont la frontière, $\partial\Omega$, est une courbe régulière.

Définition 2.1

Pour $m \in \mathbb{R}$, on définit l'espace $\mathrm{H}^m(\Omega)$ comme[1]

$$\mathrm{H}^m(\Omega) = \left\{ u \in \mathrm{L}^2(\Omega) \ ; \ \forall \alpha \in \mathbb{N}^d \text{ avec } |\alpha| \leq m, \quad \partial^\alpha u \in \mathrm{L}^2(\Omega) \right\}.$$

On note aussi

$$\|u\|_{\mathrm{H}^m(\Omega)} = \sqrt{\sum_{|\alpha| \leq m} \|\partial^\alpha u\|_{\mathrm{L}^2(\Omega)}}.$$

On définit ainsi une norme sur l'espace vectoriel $\mathrm{H}^m(\Omega)$, qu'on peut noter également $\|\cdot\|_{m,\Omega}$ ou encore $\|\cdot\|_m$ lorsqu'il n'y a pas d'ambigüité.

L'espace $\mathrm{H}^m(\Omega)$ est un espace de Hilbert, sa norme étant associée au produit scalaire :

$$(u, v)_{m,\Omega} = \sum_{|\alpha| \leq m} (\partial^\alpha u, \partial^\alpha v)_{\mathrm{L}^2(\Omega)}.$$

Les dérivées intervenant dans la définition précédente sont à comprendre au sens des distributions, c'est-à-dire au sens des dérivées faibles. Ainsi, par exemple pour l'espace $\mathrm{H}^1(\Omega)$ qu'on utilisera intensivement par la suite, on peut utiliser la définition suivante :

$$\mathrm{H}^1(\Omega) = \left\{ u \in \mathrm{L}^2(\Omega) \ ; \ \forall i \leq d, \ \exists g_i \in \mathrm{L}^2(\Omega), \ \forall \varphi \in \mathscr{C}_c^\infty(\Omega), \ \int_\Omega u \frac{\partial \varphi}{\partial x_i} = -\int_\Omega g_i \varphi \right\}.$$

Proposition 2.1 (*Théorème de Rellich*)

Pour tout $m \geq 0$, on a $\mathrm{H}^m(\Omega) \subset \mathrm{H}^{m-1}$. De plus, si Ω est borné, alors l'inclusion $\mathrm{H}^m(\Omega) \hookrightarrow \mathrm{H}^{m-1}(\Omega)$ est compacte.

La compacité de l'inclusion signifie que que toute suite (u_n) bornée dans $\mathrm{H}^m(\Omega)$ admet une sous suite convergente pour la norme $\mathrm{H}^{m-1}(\Omega)$. Soulignons que, pour $m = 1$, l'inclusion considérée est celle de $\mathrm{H}^1(\Omega)$ dans $\mathrm{L}^2(\Omega)$. Insistons sur le fait que l'hypothèse « Ω borné » est essentielle pour ce résultat de compacité. En effet, la suite de fonctions définie par

$$u_n = \chi_{]n-1,n+1[}(1 - |x - n|)$$

est bornée dans $\mathrm{H}^1(\mathbb{R})$, mais n'admet pas de sous-suite convergeant dans $\mathrm{L}^2(\mathbb{R})$ (car u_n converge simplement vers 0, mais a une norme $\mathrm{L}^2(\mathbb{R})$ constante non nulle).

Le résultat suivant fait le lien entre espaces de Sobolev et espaces classiques.

Proposition 2.2

- ◇ *Pour $m \in \mathbb{N}$ et Ω borné, on a $\mathscr{C}^m(\overline{\Omega}) \subset \mathrm{H}^m(\Omega)$. L'inclusion $\mathscr{C}^m(\overline{\Omega}) \hookrightarrow \mathrm{H}^m(\Omega)$ est continue.*
- ◇ *Si $m > k + \frac{d}{2}$, alors $\mathrm{H}^m(\Omega) \subset \mathscr{C}^k(\overline{\Omega})$. L'inclusion $\mathrm{H}^m(\Omega) \hookrightarrow \mathscr{C}^k(\overline{\Omega})$ est compacte dès que Ω est borné.*

Le premier point est clair car, pour $|\alpha| \leq m$,

$$\|\partial^\alpha u\|_{\mathrm{L}^2(\Omega)} \leq |\Omega|^{\frac{1}{2}} \times \|\partial^\alpha u\|_{\infty,\overline{\Omega}},$$

1. Un d-uplet $\alpha \in \mathbb{N}^d$ est appelé *multi-indice* ; la notation $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ désigne sa *longueur*, et ∂^α la dérivée $\dfrac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \cdots \partial x_d^{\alpha_d}}$.

où $|\Omega|$ désigne la mesure (de Lebesgue) de l'ouvert Ω. Ainsi, si $u \in \mathscr{C}^m(\overline{\Omega})$, alors on a $u \in \mathrm{H}^m(\Omega)$ avec

$$\|u\|_{\mathscr{C}^m(\Omega)} \leq |\Omega|^{\frac{1}{2}} \times \|u\|_{\mathrm{H}^m(\Omega)},$$

qui exprime bien la continuité de l'injection. Notons que le résultat est faux lorsque Ω n'est pas borné : en effet, la fonction constante égale à 1 est de classe $\mathscr{C}^\infty$ sur $\mathbb{R}$ et n'est, pourtant, dans aucun des espaces $\mathrm{H}^m(\Omega)$.

Le second point est un résultat non trivial, qui indique en particulier, pour $k = 0$, que les fonctions de H^m sont continues dès lors que $m > \frac{d}{2}$. Ainsi, en dimension $d = 1$, les fonctions de H^1 sont continues ou, plus précisément, admettent un représentant continu (car elles sont définies comme fonctions de L^2, donc presque partout). En revanche, dès la dimension 2, les fonctions de H^1 ne sont plus nécessairement continues, comme le montre l'exercice suivant.

Exercice 7

Soit $\mathbb{D}$ le disque unité de $\mathbb{R}^2$ et $u : \mathbb{D} \to \mathbb{R}$ la fonction définie par

$$u(x, y) = \ln(x^2 + y^2).$$

Montrer que $u \in \mathrm{H}^1(\mathbb{D})$, mais u n'est pas continue en $(0, 0)$.

Espaces $\mathrm{H}_0^m(\Omega)$ – traces au bord

Définition 2.2

Pour $m \in \mathbb{N}$, on note $\mathrm{H}_0^m(\Omega)$ l'adhérence de $\mathscr{C}_c^\infty(\Omega)$ dans $\mathrm{H}^m(\Omega)$ pour la norme $\|\cdot\|_{m,\Omega}$.

Dans le cas où $\Omega = \mathbb{R}^d$, l'espace $\mathscr{C}_c^\infty(\Omega)$ est dense dans $\mathrm{H}^m(\Omega)$ et on a donc

$$\mathrm{H}_0^m(\mathbb{R}^d) = \mathrm{H}^m(\mathbb{R}^d).$$

On va montrer que, dans le cas d'un domaine Ω général, l'espace $\mathrm{H}_0^m(\Omega)$ est constitué des fonctions de $\mathrm{H}^m(\Omega)$, nulles sur le bord $\partial\Omega$. Toutefois, le défaut de continuité sur $\overline{\Omega}$ des fonctions de $\mathrm{H}^m(\Omega)$ ne permet pas de donner un sens classique à la restriction $u|_{\partial\Omega}$. Ce n'est possible qu'au moyen de la *théorie des traces*, dont nous résumons le résultat principal ci-dessous.

Théorème 2.3 (*traces sur le bord*)

Soit Ω un domaine dont le bord est une courbe régulière. L'application

$$\begin{array}{rcl} \mathscr{C}^0(\overline{\Omega}) & \longrightarrow & \mathscr{C}^0(\partial\Omega) \\ u & \longmapsto & u|_{\partial\Omega} \end{array}$$

admet un prolongement γ_0 défini sur $\mathrm{H}^1(\Omega)$ linéaire. L'image de γ_0 est notée $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$. L'application $\gamma_0 : \mathrm{H}^1(\Omega) \to \mathrm{H}^{\frac{1}{2}}(\partial\Omega)$ est continue, et admet un inverse à droite continu.

Le résultat précédent permet de donner un sens à la restriction au bord d'une fonction de $\mathrm{H}^1(\Omega)$. Bien sûr, lorsque u est régulière (continue sur $\overline{\Omega}$), on a $\gamma_0(u) = u|_{\partial\Omega}$, mais $\gamma_0(u)$ a aussi un sens pour u non nécessairement continue. Par construction, l'application

$\gamma_0 : \mathrm{H}^1(\Omega) \to \mathrm{H}^{\frac{1}{2}}(\partial\Omega)$ est surjective. Ainsi, pour toute fonction $g \in \mathrm{H}^{\frac{1}{2}}(\partial\Omega)$, il existe $u \in \mathrm{H}^1(\Omega)$ telle que $\gamma_0(u) = g$. Le fait que γ_0 admette un inverse à droite exprime qu'il est possible de choisir $u = u_g$ tel que

$$\|u_g\|_{1,\Omega} \leq C\|g\|_{\frac{1}{2},\partial\Omega}, \tag{2.1}$$

la constante C étant indépendante de g.

Remarque 2.1

Il n'est pas anodin de noter $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$ *l'image de l'application* γ_0. *Cet espace a en effet la même structure que les espaces de Sobolev* H^m *déjà définis, mais pour un ordre de dérivation fractionnaire. Précisons les choses dans le cas où* $\partial\Omega = \mathbb{R}$. *Les espaces* $\mathrm{H}^m(\mathbb{R})$ *peuvent être définis par transformation de Fourier à l'aide de l'égalité de Plancherel : par exemple*

$$\mathrm{H}^1(\mathbb{R}) = \left\{u \in \mathrm{L}^2(\mathbb{R}) \ ; \ \xi \mapsto \xi\widehat{u}(\xi) \in \mathrm{L}^2(\mathbb{R})\right\}.$$

On peut montrer que l'espace $\mathrm{H}^{\frac{1}{2}}(\mathbb{R})$ *est alors caractérisé par*

$$\mathrm{H}^{\frac{1}{2}}(\mathbb{R}) = \left\{u \in \mathrm{L}^2(\mathbb{R}) \ ; \ \xi \mapsto |\xi|^{\frac{1}{2}}\widehat{u}(\xi) \in \mathrm{L}^2(\mathbb{R})\right\},$$

ce qui justifie la notation. La norme qui apparaît dans (2.1) *peut alors être définie comme*

$$\|u\|_{\frac{1}{2},\mathbb{R}} = \sqrt{\|u\|^2_{\mathrm{L}^2(\mathbb{R})} + \||\xi|^{\frac{1}{2}}\widehat{u}\|^2_{\mathrm{L}^2(\mathbb{R})}}.$$

Dans le cas où $\partial\Omega$ *n'est pas* $\mathbb{R}$ *tout-entier, on peut procéder par restriction ou définir des normes dites* intrinsèques. *Pour plus de détails, voir par exemple [Ada75].*

Avec la notion de trace, on peut montrer le résultat suivant :

Proposition 2.4

Pour $m \in \mathbb{N}$,

$$\mathrm{H}^m_0(\Omega) = \{u \in \mathrm{H}^m(\Omega) \ ; \ \gamma_0(\partial^\alpha u) = 0 \text{ pour } |\alpha| \leq m-1\}.$$

Par abus de notation, on notera souvent $u|_{\partial\Omega}$ pour $\gamma_0(u)$. Ainsi, on écrira abusivement

$$\mathrm{H}^1_0(\Omega) = \left\{u \in \mathrm{H}^1(\Omega) \ ; \ u = 0 \text{ sur } \partial\Omega\right\}.$$

Inégalité de type Poincaré

On achève cette synthèse sur les espaces de Sobolev avec des inégalités qui nous seront très utiles par la suite. Commençons par prouver un lemme dont découleront tous les autres résultats.

Lemme 2.5

Soit $m \in \mathbb{N}$, Ω *un domaine borné de* $\mathbb{R}^d$, *et* V *un sous-espace fermé de* $\mathrm{H}^m(\Omega)$ *tel que*

$$V \cap \mathbb{P}_{m-1} = \{0\}.$$

Alors il existe une constante $C > 0$ *telle que*

$$\forall v \in V, \quad \|v\|_{m,\Omega} \leq |v|_{m,\Omega},$$

où $|\cdot|_{m,\Omega}$ *désigne la* semi-norme H^m *:*

$$|v|_{m,\Omega} = \sqrt{\sum_{|\alpha|=m} \|\partial^\alpha v\|^2_{\mathrm{L}^2(\Omega)}}.$$

PREUVE. On procède par l'absurde : si l'inégalité voulue n'est pas satisfaite, alors pour chaque $n \in \mathbb{N}$, on peut trouver $v_n \in V$ tel que $\|v_n\|_{m,\Omega} = 1$ et $|v_n|_{m,\Omega} \to 0$.

Comme (v_n) est bornée dans $\mathrm{H}^m(\Omega)$, elle admet une sous-suite $(v_{n_k})_k$ qui converge dans $\mathrm{H}^{m-1}(\Omega)$, cf. Théorème 2.1 ; notons v la limite. Ainsi, pour $|\alpha| \leq m-1$, la suite $(v_{n_k})_k$ est de Cauchy dans $\mathrm{L}^2(\Omega)$. Par ailleurs, puisque $|v_n|_{m,\Omega} \to 0$, c'est encore vrai pour $|\alpha| = m$. On en déduit que la suite $(v_{n_k})_k$ est de Cauchy dans $\mathrm{H}^m(\Omega)$. Elle converge donc. Notons u la limite. Bien sûr, $u = v$ presque-partout par unicité de la limite dans $\mathrm{L}^2(\Omega)$. On déduit aussi de $|v_n|_{m,\Omega} \to 0$ que $\partial^\alpha u = 0$ pour $|\alpha| = m$. Cela impose que u soit un polynôme de degré inférieur ou égal à $m-1$, et comme $u \in V$ (V est fermé), il s'ensuit que $u = 0$. Cette dernière égalité est en contradiction avec $\|v_n\|_{m,\Omega} = 1$. ∎

Le résultat précédent nous sera utile pour l'obtention d'estimations d'erreurs dans les méthodes d'éléments finis (voir § 2.7), mais nous en donnons ici deux conséquences immédiates.

Proposition 2.6 (*Inégalité de Poincaré*)

On suppose le domaine Ω *borné. Il existe une constante* $C > 0$ *telle que*

$$\forall v \in \mathrm{H}^1_0(\Omega), \quad \|v\|_{1,\Omega} \leq C|v|_{1,\Omega}.$$

PREUVE. On applique le lemme 2.5 pour $V = \mathrm{H}^1_0(\Omega)$ et $m = 1$. Par définition de l'espace $\mathrm{H}^1_0(\Omega)$, c'est un sous-espace fermé de $\mathrm{H}^1(\Omega)$. Par ailleurs, la seule fonction constante (i.e. $\mathbb{P}_0$) dans V est la fonction nulle. ∎

Remarque 2.2

Il est possible de montrer l'inégalité de Poincaré « à la main », supposant seulement que Ω *est borné dans une direction. Donnons ici la preuve en dimension* 1 *qui peut être adaptée en dimension supérieure. Soit donc* $\Omega =]0,1[$ *et* $v \in \mathscr{C}^\infty_{\mathrm{c}}(\Omega)$*. On écrit, pour* $x \in \Omega$*, en exploitant* $u(0) = 0$*,*

$$|u(x)|^2 = 2\int_0^x u(t)u'(t)\,dt \leq 2\|u\|_{\mathrm{L}^2(\Omega)} \times \|u'\|_{\mathrm{L}^2(\Omega)}.$$

Il suffit d'intégrer en la variable x *pour obtenir l'inégalité de Poincaré avec* $C = 2$ *pour* v *dans* $\mathscr{C}^\infty_{\mathrm{c}}(\Omega)$*. On conclut par densité de cet espace dans* $\mathrm{H}^1_0(\Omega)$*.*

Proposition 2.7 (*Inégalité de Poincaré-Wirtinger*)

On suppose le domaine Ω *borné. Il existe une constante* $C > 0$ *telle que*

$$\forall v \in \mathrm{H}^1(\Omega), \quad \|v - \langle v\rangle\|_{1,\Omega} \leq C|v|_{1,\Omega}\,,$$

où $\langle v\rangle$ *désigne la moyenne de* v *sur* Ω *:*

$$\langle v\rangle = \frac{1}{|\Omega|}\int_\Omega v(x)\,dx.$$

PREUVE. On va appliquer le lemme 2.5 avec $m = 1$ et

$$V = \left\{ w \in \mathrm{H}^1(\Omega) \; ; \; \langle w \rangle = 0 \right\}.$$

Il est clair que V est fermé dans $\mathrm{H}^1(\Omega)$ (noter que la forme linéaire $w \mapsto \langle w \rangle$ est continue car Ω est borné). Par ailleurs, $V \cap \mathbb{P}_0 = \{0\}$, donc le lemme 2.5 assure l'existence d'une constante $C > 0$ telle que

$$\forall w \in V, \quad \|w\|_{1,\Omega} \leq |w|_{\Omega}.$$

Pour conclure, il suffit de remarque que, pour $v \in \mathrm{H}^1(\Omega)$, $w = v - \langle v \rangle \in V$, et que $|w|_{1,\Omega} = |v|_{1,\Omega}$. ∎

Remarque 2.3

La question de déterminer les meilleurs constantes C (i.e. les plus petites) dans les inégalité des propositions 2.6 et 2.7 peut être traitée au moyen des séries de Fourier en dimension 1. Elle est, plus généralement, reliée à un problème de valeur propre de l'opérateur Laplacien dans Ω.

2.2.2 Le lemme de Lax-Milgram

On revient au problème (PV), qui peut s'écrire

$$\text{Trouver } u \in V, \quad \forall v \in V, \quad a(u,v) = \ell(v), \tag{2.2}$$

si on a posé

$$a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x \quad \text{et} \quad \ell(v) = \int_\Omega f(x) v(x) \, \mathrm{d}x. \tag{PV}$$

Le résultat suivant fournit des conditions suffisantes sur la forme bilinéaire a et la forme linéaire ℓ pour que le problème ait une unique solution.

Théorème 2.8 (*Lax-Milgram*)

Soit V un espace de hilbert, $a : V \times V \to \mathbb{R}$ et $\ell : V \to \mathbb{R}$ des formes respectivement bilinéaire et linéaire, satisfaisant :

- ◇ *a est continue : il existe une constante $C_a > 0$ telle que*

$$\forall u, v \in V, \quad |a(u,v)| \leq C_a \|u\|_V \|v\|_V,$$

- ◇ *a est coercive : il existe une constante $\alpha > 0$ telle que*

$$\forall v \in V, \quad a(v,v) \geq \alpha \|v\|_V^2,$$

- ◇ *ℓ est continue : il existe une constante $C_\ell > 0$ telle que*

$$\forall v \in V, \quad |\ell(v)| \leq C_\ell \|v\|_V.$$

Alors le problème (PV) *admet une unique solution $u \in V$. De plus, u satisfait l'estimation a priori*

$$\|u\|_V \leq \frac{C_\ell}{\alpha}.$$

Dans le cas où la forme bilinéaire a est symétrique, les hypothèses de continuité et de coercivité en font un produit scalaire, dont la norme associée est équivalente à la norme sur l'espace V. Le lemme de Lax-Milgram n'est alors autre que le théorème de représentation de Riesz.

Vérifions que ce résultat s'applique dans le cas du problème de Laplace-Dirichlet.

◇ *Continuité de la forme bilinéaire.* Pour $u, v \in \mathrm{H}^1_0(\Omega)$,

$$\left| \int_\Omega \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x \right| \leq \|\nabla u\|_{\mathrm{L}^2(\Omega)} \times \|\nabla v\|_{\mathrm{L}^2(\Omega)} \leq \|u\|_{1,\Omega} \times \|v\|_{1,\Omega}.$$

Ainsi, la constante $C_a = 1$ convient.

◇ *Coercivité de la forme bilinéaire.* Pour $v \in \mathrm{H}^1_0(\Omega)$,

$$\int_\Omega |\nabla v(x)|^2 \, \mathrm{d}x \geq \frac{1}{C} \|v\|^2_{1,\Omega},$$

d'après l'inégalité de Poincaré. Ainsi $\alpha = \frac{1}{C}$ convient.

◇ *Continuité de la forme liénaire* Pour $v \in \mathrm{H}^1_0(\Omega)$,

$$\left| \int_\Omega f(x) v(x) \, \mathrm{d}x \right| \leq \|f\|_{\mathrm{L}^2(\Omega)} \times \|v\|_{\mathrm{L}^2(\Omega)} \leq \|f\|_{\mathrm{L}^2(\Omega)} \times \|v\|_{1,\Omega}.$$

Ainsi la constante $C_\ell = \|f\|_{\mathrm{L}^2(\Omega)}$ convient.

2.2.3 Autres exemples de formulations variationnelles

Le problème de Neumann avec terme d'ordre 0

On considère le problème

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

De la même façon que pour le problème de Dirichlet, on fixe $v \in \mathrm{H}^1(\Omega)$, et on obtient à l'aide d'une intégration par partie :

$$-\int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x) \, \mathrm{d}\sigma_x + \int_\Omega \nabla u(x) \nabla v(x) \, \mathrm{d}x + \int_\Omega u(x) v(x) \, \mathrm{d}x = \int_\Omega f(x) v(x) \, \mathrm{d}x.$$

La condition de Neumann sur u permet d'annuler le terme de bord et d'obtenir la formulation variationnelle $a(u, v) = \ell(v)$ avec

$$a(u, v) = \int_\Omega \nabla u(x) \cdot \nabla v(x) \, \mathrm{d}x + \int_\Omega u(x) v(x) \, \mathrm{d}x \quad \text{et} \quad \ell(v) = \int_\Omega f(x) v(x) \, \mathrm{d}x.$$

C'est la même égalité que pour le problème de Dirichlet associé à l'opérateur $-\Delta + Id$, mais l'espace V est cette fois $\mathrm{H}^1(\Omega)$ tout-entier. Il est aisé de vérifier que le lemme de Lax-Milgram s'applique ici encore.

Le problème de Neumann pour le Laplacien

On considère le problème

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

De la même manière que plus haut, la formulation variationnelle est associée à

$$a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x)\,\mathrm{d}x \quad \text{et} \quad \ell(v) = \int_\Omega f(x)v(x)\,\mathrm{d}x,$$

et est posée sur l'espace variationnel $V = \mathrm{H}^1(\Omega)$. La forme bilinéaire a n'est ici plus coercive (car il n'est pas possible d'obtenir une inégalité de Poincaré dans $\mathrm{H}^1(\Omega)$ tout-entier). En fait, il est facile de voir que le problème de trouver $u \in V$ satisfaisant $a(u,v) = \ell(v)$ pour tout v dans $\mathrm{H}^1(\Omega)$ n'admet pas existence et unicité. En effet, si u est solution, $u + K$ est encore solution pour toute constante $K \in \mathbb{R}$. Par ailleurs, en choisissant la fonction test $v = 1$, on obtient

$$\int_\Omega f(x)\,\mathrm{d}x = 0, \tag{2.3}$$

qui est une contrainte sur le second membre f. Toutefois, il est possible de montrer que, sous la condition de compatibilité (2.3), il existe une unique solution u de moyenne nulle. En effet, en travaillant dans l'espace

$$V = \{v \in \mathrm{H}^1(\Omega)\ ;\ \langle v \rangle = 0\},$$

l'inégalité de Poincaré-Wirtinger permet d'assurer que la forme a est coercive sur V et le lemme de Lax-Milgram s'applique.

Un problème avec terme d'ordre 1

On considère le problème

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

où $\beta \in \mathbb{R}^d$ est un vecteur donné. L'espace variationnel correspondant aux conditions de Dirichlet est $\mathrm{H}^1_0(\Omega)$. Les formes bilinéaire et linéaire correspondantes s'écrivent

$$a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x)\,\mathrm{d}x + \int_\Omega \beta \cdot \nabla u(x) v(x)\,\mathrm{d}x \quad \text{et} \quad \ell(v) = \int_\Omega f(x)v(x)\,\mathrm{d}x.$$

Si la continuité est claire, la coercivité l'est moins. En effet, grâce aux inégalités de Cauchy-Schwarz et Poincaré, pour $v \in \mathrm{H}^1_0(\Omega)$

$$a(v,v) \geq \frac{1}{C}\|v\|^2_{1,\Omega} - |\beta| \times \|v\|_{1,\Omega} \times \|v\|_{\mathrm{L}^2(\Omega)'} \geq \left(\frac{1}{C} - |\beta|\right) \|v\|^2_{1,\Omega},$$

où $|\beta$ désigne la norme du vecteur β. Ainsi, sous la condition $C|\beta| < 1$, on peut appliquer le théorème de Lax-Milgram.

2.3 Interprétation des formulations variationnelles

S'il est clair que toute solution du problème EDP est aussi solution du problème variationnel, la réciproque ne va pas de soi. Toutefois, si la solution du problème variationnel possède un supplément de régularité, elle est aussi solution du problème EDP initial.

On détaille ce point pour deux problèmes particuliers.

2.3.1 Cas du problème de Dirichlet

On considère le problème

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

dont la formulation variationnelle s'écrit : trouver $u \in H^1_0(\Omega)$ tel que

$$\forall v \in H^1_0(\Omega), \quad \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx = \int_\Omega f(x)v(x)\,dx. \tag{PV_D}$$

Supposons que la solution u du problème (PV_D) appartienne à $H^2(\Omega)$. Alors, pour tout fonction $v \in \mathscr{C}^\infty_c(\Omega) \subset H^1(\Omega)$, on peut effectuer l'intégration par parties dans le sens inverse à ce qui a été fait plus haut, pour obtenir (le terme de bord s'annule car v est à support compact)

$$\int_\Omega \big(-\Delta u(x) - f(x)\big)\,dx = 0.$$

Par densité de $\mathscr{C}^\infty_c(\Omega)$ dans $L^2(\Omega)$, on en conclut que la fonction $-\Delta u(x) - f(x)$ est nulle presque partout. Par ailleurs, comme $u \in H^1_0(\Omega)$, elle résout bien le problème (P_D).

2.3.2 Cas du problème de Neumann

On considère le problème

$$\left\{\begin{array}{rcll} -\Delta u + u & = & f & \text{dans } \Omega, \\ \dfrac{\partial u}{\partial n} & = & 0 & \text{sur } \partial\Omega, \end{array}\right. \tag{P_N}$$

dont la formulation variationnelle s'écrit : trouver $u \in H^1(\Omega)$ tel que

$$\forall v \in H^1(\Omega), \quad \int_\Omega \nabla u(x) \cdot \nabla v(x)\,dx + \int_\Omega u(x)v(x)\,dx = \int_\Omega f(x)v(x)\,dx. \tag{PV_N}$$

Supposons encore que la solution u du problème (P_N) appartienne à $H^2(\Omega)$. À nouveau, prenons $v \in \mathscr{C}^\infty_c(\Omega) \subset H^1(\Omega)$ et écrivons l'intégration par parties (le terme de bord s'annule car v est à support compact) :

$$\int_\Omega \big(-\Delta u(x) + u(x) - f(x)\big)\,dx = 0.$$

Par densité, on en déduit que $-\Delta u + u = f$ presque partout.

Il s'agit maintenant de retrouver la condition aux limites de Neumann. On prend cette fois $v \in \mathscr{C}^\infty(\overline{\Omega}) \subset H^1(\Omega)$ (ici v n'est plus à support compact dans Ω) ; l'intégration par parties fournit

$$\int_{\partial\Omega} \frac{\partial u}{\partial n}(x)v(x)\,d\sigma_x + \int_\Omega \big(-\Delta u(x) + u(x) - f(x)\big)\,dx = 0.$$

Mais on a montré que la fonction $-\Delta u + u - f$ est nulle presque partout. Il s'ensuite donc

$$\int_{\partial\Omega} \frac{\partial u}{\partial n}(x) v(x)\, \mathrm{d}\sigma_x = 0.$$

Cette égalité est valable pour toute fonction v de classe $\mathscr{C}^\infty$ sur le bord $\partial\Omega$. Par densité de cet espace dans $\mathrm{L}^2(\partial\Omega)$, on en déduit la condition aux limites de Neumann.

Remarque 2.4

Les deux exemples précédents montrent que le traitement des conditions aux limites de Dirichlet et de Neumann sont différentes. En effet, la condition de Dirichlet est dite essentielle *pour le laplacien, car elle est imposée explicitement dans l'espace variationnel. La condition de Neumann, quant à elle, est dite* naturelle *pour le laplacien car elle est inscrite dans la formulation variationnelle elle-même.*

Exercice 8

On considère la formulation variationnelle $a(u,v) = \ell(v)$, posée dans $V = \mathrm{H}^1(\Omega)$, où

$$a(u,v) = \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx + \int_\Omega u(x)v(x)\, dx + \gamma \int_\Omega u(x)v(x)\, d\sigma_x$$

et

$$\ell(v) = \int_\Omega f(x)v(x)\, dx,$$

où $\gamma > 0$ est une constante fixée, et $f \in \mathrm{L}^2(\Omega)$ une fonction donnée.

1. *Montrer que $a(u,v)$ est bien définie pour $u, v \in V$.*
2. *Montrer que a est continue et coercive sur V.*
3. *Montrer que ℓ est continue sur V.*
4. *Interpréter le problème en termes d'EDP.*

Remarque 2.5

Le que la solution u du problème variationnel appartienne à $\mathrm{H}^2(\Omega)$ n'est pas toujours réaliste. Si, en dimension $d = 1$, l'hypothèse $f \in \mathrm{L}^2(\Omega)$ implique $u \in \mathrm{H}^2(\Omega)$, ce n'est généralement plus vrai en dimension supérieure. En annexe B, on donne un exemple de fonction f très régulière pour laquelle la solution du problème variationnel de Laplace-Dirichlet n'est pas dans l'espace $\mathrm{H}^2(\Omega)$.

2.4 Le lemme de Céa

On souhaite ici quantifier l'erreur entre u, solution du problème variationnel continu

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

et u_h, solution du problème discret

$$\forall v_h \in V_h, \quad a(u_h, v_n) = \ell(v_h).$$

On a le résultat général suivant (indépendant de l'espace d'approximation choisi).

Lemme 2.9 (*Lemme de Céa*)

Il existe une constante $C > 0$ telle que

$$\|u - u_h\|_V \leq C \inf_{w_h \in V_h} \|u - w_h\|_V.$$

PREUVE. Comme $V_h \subset V$, on a pour tout $v_h \in V_h$,

$$a(u - u_h, v_h) = 0. \tag{2.4}$$

Par ailleurs, par coercivité de a, on peut écrire

$$\alpha \|u - u_h\|_V^2 \leq a(u - u_h, u - u_h).$$

Or, pour tout $w_h \in V_h$, on a d'après (2.4) (prendre $v_h = w_h - u_h$)

$$a(u - u_h, u - u_h) = a(u - u_h, u - w_h).$$

Ainsi, par continuité de a,

$$\|u - u_h\|_V^2 \leq C_a \times \|u - u_h\|_V \times \|u - w_h\|_V,$$

qui fournit le résultat annoncé. ∎

Remarque 2.6

L'égalité (2.4) *est appelée* orthogonalité de Galerkine. *En effet, lorsque la forme bilinéaire a est symétrique, elle définit un produit scalaire $(u, v)_a = a(u, v)$ et* (2.4) *s'écrit $(u - u_h, v_h)_a = 0$. Ainsi, u_h apparaît comme le projeté de u sur V_h pour le produit scalaire $(\cdot, \cdot)_a$. Le lemme de Céa exprime ce fait, et la constante C provient de ce que la mesure des distances est faite dans la norme de V, et non pas dans la norme associée au produit scalaire $(\cdot, \cdot)_a$.*

Le lemme de Céa est un résultat central car il permet de ramener l'estimation de l'erreur de la méthode d'éléments finis à la simple erreur d'approximation de u dans V_h. En particulier, si l'on est capable d'exhiber, connaissant u, une fonction $w_h \in V_h$ simple et proche de u, on aura directement la majoration

$$\|u - u_h\|_V \leq \|u - w_h\|_V.$$

On verra dans le paragraphe suivant que l'interpolation fournit un bon choix pour w_h.

2.5 Interpolation dans des espaces d'éléments finis en dimension 1

2.5.1 Rappels sur l'interpolation de Lagrange en dimension 1

On résume ici quelques résultats bien connus sur l'interpolation de Lagrange en dimension 1. Pour plus de détails, en particulier concernant les preuves, on renvoie à l'annexe C.

On se donne un entier k, un intervalle borné $[a, b]$, qu'on subdivise en

$$a = x_0 < x_1 < \ldots < x_k = b,$$

et une fonction $u \in \mathscr{C}([a,b])$. Alors il existe un unique polynôme dit *interpolateur*) $p \in \mathbb{P}_k$ tel que

$$\forall i = 0, 1, \ldots, k, \quad p(x_i) = u(x_i).$$

Ce polynôme est donné par l'expression (même si cette dernière n'est pas à recommander pour un calcul effectif)

$$p(x) = \sum_{i=0}^{k} u(x_i) L_i(x), \quad \text{avec} \quad L_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

On peut montrer l'inégalité suivante, dès que la fonction u est de classe $\mathscr{C}^{k+1}$ sur $I = [a,b]$:

$$\sup_{x \in I} |u(x) - p(x)| \leq \frac{(b-a)^{k+1}}{(k+1)!} \sup_{x \in I} |u^{(k+1)}(x)|. \tag{2.5}$$

Dans le but d'approcher la fonction u de manière fidèle, il n'est pas, en général, une bonne idée d'augmenter fortement le degré k. Mieux vaut, en effet, découper l'intervalle $[a,b]$ en sous-intervalles et utiliser sur chaque sous-intervalle une interpolation polynomiale de degré modéré.

Fixons donc une subdivision de l'intervalle $[a,b]$:

$$[a,b] = \bigsqcup_j I_j,$$

la réunion étant finie et disjointe, et les I_j étant des intervalles fermés. On note h la taille maximale de la subdivision :

$$h = \max_j |I_j|.$$

On fixe $k \in \mathbb{N}$ et, sur chaque intervalle I_j, on choisit $k+1$ points d'interpolation distincts $x_0^j < x_1^j < \cdots < x_k^j$. On note alors p^j le polynôme d'interpolation de Lagrange de u correspondant, et p_h la fonction polynomiale par morceaux définie par

$$p_h(x) = p^j(x) \quad \text{sur} \quad I_j.$$

Remarque 2.7

L'interpolée p_h n'est a priori *pas continue. Si u est continue et si l'on choisit les points d'interpolation x_0^j et x_k^j comme les extrémités de l'intervalle I_j, alors p est continue. Toutefois, même sous des hypothèses de régularité supplémentaire sur u, la fonction p n'est en général pas mieux que continue. Pour un procédé d'interpolation $\mathscr{C}^1$, voir § 2.5.2.*

On peut montrer l'estimation d'erreur suivante, dès que la fonction u est de classe $\mathscr{C}^{k+1}$ sur l'intervalle $[a,b]$:

$$\sup_{x \in [a,b]} |u(x) - p_h(x)| \leq \frac{h^{k+1}}{(k+1)!} \sup_{x \in [a,b]} |u^{(k+1)}(x)|.$$

Pour cela, il suffit de travailler sur chaque intervalle I_j et d'utiliser l'inégalité (2.5) pour $I = I_j$.

2.5.2 Interpolation cubique de Hermite

Étant donné un intervalle $I = [a, b]$, et une fonction $u \in \mathscr{C}^1(I)$, il existe un unique polynôme de degré inférieur ou égal à 3 tel que

$$p(a) = u(a), \quad p'(a) = u'(a), \quad p(b) = u(b), \quad p'(b) = u'(b).$$

Ce polynôme est donné par l'expression

$$p = u(a)L_1 + u'(a)L_2 + u(b)L_3 + u'(b)L_4,$$

où les polynômes L_i sont donnés par

$$L_1(x) = (2x + b - 3a)\frac{(b-x)^2}{b-a)^3}, \quad L_2(x) = (x-a)\frac{(b-x)^2}{(b-a)^2},$$

$$L_3(x) = (3b - a - 2x)\frac{(x-a)^2}{b-a)^3}, \quad L_4(x) = (x-b)\frac{(x-a)^2}{(b-a)^2}.$$

De la même manière que pour l'interpolation de Lagrange, on peut montrer l'estimation

$$\sup_{x\in I} |u(x) - p(x)| \leq \frac{(b-a)^4}{24} \sup_{x\in I} |u^{(4)}(x)|.$$

Toujours de façon analogue, on peut utiliser le procédé d'interpolation sur une subdivision de pas maximum h et obtenir une approximation polynomiale par morceaux p_h, qui sera proche de u à $\mathcal{O}(h^4)$ près. L'intérêt du procédé de Hermite réside en la régularité $\mathscr{C}^1$ de l'interpolé par morceaux p_h. Il est possible, bien entendu, de généraliser cette construction à des ordres et degrés plus élevés.

2.5.3 Interpolation générale en dimension 1

Nous allons donner ici un cadre général qui permet d'englober les deux procédés d'interpolation par morceaux décrits dans les paragraphes précédents. Remarquons tout d'abord que l'interpolation sur chaque sous-intervalle de la subdivision procède de la même mécanique. En effet, notons $\widehat{L}_i$ ($i = 1, 2, 3, 4$) les fonctions de base de l'interpolation de Hermite sur l'intervalle $[0, 1]$:

$$L_1(\widehat{x}) = (2\widehat{x} - 1)(1 - \widehat{x})^2,\ L_2(\widehat{x}) = \widehat{x}(1 - \widehat{x})\widehat{x}^2,\ L_3(\widehat{x}) = (1 - 2\widehat{x})\widehat{x}^2,\ L_4(\widehat{x}) = (\widehat{x} - 1)\widehat{x}^2.$$

Il est alors facile de voir que les fonctions de base L_i sur un intervalle $[a, b]$ quelconque s'en déduisent : si $x = a + (b - a)\widehat{x}$ parcourt $[a, b]$, alors

$$\widehat{L}_i(\widehat{x}) = L_i(x).$$

On peut effectuer le même raisonnement pour l'interpolation de Lagrange. Ainsi, il suffit de définir le procédé d'interpolation sur un intervalle de référence, ici $\widehat{I} = [0, 1]$, pour être en mesure de construire l'interpolé par morceaux.

Plaçons-nous donc sur $\widehat{I}$. Le procédé d'interpolation est basé sur la donnée de *formes linéaires* :

- ◇ Pour l'interpolation de Lagrange de degré k : il s'agit des $k + 1$ formes linéaires d'évaluation
$$\psi_j : v \mapsto v(\widehat{x}_i) \quad (i = 0, 2, \ldots, k).$$

◇ Pour l'interpolation cubique de Hermite, il s'agit des 4 formes linéaires

$$\psi_1 : v \mapsto v(0), \quad \psi_2 : v \mapsto v'(0), \quad \psi_3 : v \mapsto v(1), \quad \psi_4 : v \mapsto v'(1).$$

Il est nécessaire de préciser l'espace sur lequel on définit ces formes linéaires ($\mathscr{C}^0([0,1])$ pour le cas Lagrange, $\mathscr{C}^1([0,1])$ pour Hermite).

Par ailleurs, les fonctions de base, qui ont été notées $\widehat{L}_i$ peuvent être définies par les relations

$$\psi_j(\widehat{L}_i) = \delta_{ij}.$$

Reste à préciser l'espace d'interpolation, qui doit être choisi de telle sorte que les $\widehat{L}_i$ soient définis de manière unique (ici $\mathbb{P}_k$ pour Lagrange, et $\mathbb{P}_3$ pour Hermite).

La discussion précédente nous conduit à la définition suivante (écrite en dimension quelconque) :

Définition 2.3

Étant donné un espace de fonctions $\mathscr{C}$. Un élément fini est un triplet (K, P, Σ), où

- *$K \subset \mathbb{R}^d$ est un fermé d'intérieur non-vide*[2],
- *P est un sous-espace de $\mathscr{C}$ de dimension finie,*
- *$\Sigma = \{\psi_i \ ; \ i = 0, 1, \ldots, k\}$ un ensemble fini de formes linéaires sur $\mathscr{C}$.,*

tels que Σ soit P-unisolvant, i.e.

$$\forall (\alpha_i) \in \mathbb{R}^{k+1}, \quad \exists! p \in P, \quad \forall i, \quad \psi_i(p) = \alpha_i.$$

Les fonctions de base *ϕ_i sont définies par*

$$\psi_j(\phi_i) = \delta_{ij}, \quad (0 \leq i, j \leq k).$$

*Enfin, pour une fonction $v \in \mathscr{C}$, on définit l'*interpolé local *dans (K, P, Σ) comme étant l'unique élément $\pi v \in P$ satisfaisant*

$$\forall \psi \in \Sigma, \quad \psi(\pi v) = \psi(v).$$

De cette manière, nous avons construit les éléments finis suivants dans les paragraphes précédents :

◇ *Élément de Lagrange 1D d'ordre k.*

$$K = [0,1], \quad P = \mathbb{P}_k, \quad \Sigma = \{v \mapsto v(x_i) \ ; \ i = 0, 1, \ldots k\}.$$

◇ *Élement de Hermite 1D d'ordre* 3.

$$K = [0,1], \quad P = \mathbb{P}_3, \quad \Sigma = \{\widehat{L}_i \ ; \ i = 1, 2, 3, 4\}.$$

2.6 Interpolation en dimension 2

2.6.1 Introduction : interpolation linéaire sur un triangle

En dimension 1, la subdivision d'un intervalle $]a, b[$ ne pose pas de problème, et ne fait intervenir que des sous-intervalles, il n'en va pas de même en dimension supérieure.

2. on supposera en général qu'il est au moins un peu régulier, par exemple de bord lipschitzien.

En effet, si Ω désigne un ouvert quelconque de $\mathbb{R}^d$, il n'est pas édivent d'en obtenir une subdivision en sous-ensembles élémentaires. Toutefois, si Ω est polygonal, il est possible d'en obtenir de manière constructive un *maillage* sous la forme d'une *triangulation* :

$$\overline{\Omega} = \bigsqcup_{T \in \mathscr{T}} T,$$

la réunion étant disjointe, et les triangles fermés au sens topologique. La question se pose alors d'interpoler une fonction sur un triangle. Un procédé simple consiste à utiliser une interpolation linéaire : étant donnée un triangle $T \subset \mathbb{R}^2$ et $v \in \mathscr{C}^0(T)$, on note $\pi_T v$ la fonction affine qui coïncide avec v en les trois sommets du triangle. Précisement, si on note $\mathbf{a}_1$, $\mathbf{a}_2$ et $\mathbf{a}_3$ les sommets, la fonction πv est donnée par

$$\pi_T v(\mathbf{x}) = v(\mathbf{a}_3) + \alpha_1 \big[v(\mathbf{a}_1) - v(\mathbf{a}_3)\big] + \alpha_2 \big[v(\mathbf{a}_2) - v(\mathbf{a}_3)\big],$$

si le vecteur $\mathbf{x} - \mathbf{a}_3$ s'écrit $\alpha_1(\mathbf{a}_1 - \mathbf{a}_3) + \alpha_2(\mathbf{a}_2 - \mathbf{a}_3)$.

Faire un dessin

Il est clair que le calcul de l'interpolé $\pi_T v$ est similaire pour chacun des triangles du maillage. On peut en tirer profit en introduisant un triangle dit *de référence* et en transportant le procédé d'interpolation sur le triangle considéré. Soit donc $\widehat{T}$ le triangle de sommets $\hat{\mathbf{a}}_1 = (1,0)$, $\hat{\mathbf{a}}_2 = (0,1)$ et $\hat{\mathbf{a}}_3 = (0,0)$. Si $\hat{\mathbf{x}} = (x,y)$, alors, pour toute fonction $\hat{v} \in \mathscr{C}^0(\widehat{T})$,

$$\pi_{\widehat{T}} \hat{v}(x,y) = (1-x-y)\hat{v}(0,0) + x\hat{v}(1,0) + y\hat{v}(0,1).$$

Cette formule est tout-à-fait explicite et il ne manque plus qu'un moyen de relier $\pi_{\widehat{T}}\hat{v}$ à π_T pour un triangle quelconque T.

Soit donc T un triangle et $v \in \mathscr{C}^0(T)$. On désigne par F_T l'application affine (bijective si T n'est pas plat) qui envoie le triplet $(\hat{\mathbf{a}}_1, \hat{\mathbf{a}}_2, \hat{\mathbf{a}}_3)$ sur $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$. Pour $\mathbf{x} \in T$, on note $\hat{\mathbf{x}} = F_T^{-1}(\mathbf{x}) \in \widehat{T}$. La relation $\hat{\mathbf{x}} = x(\hat{\mathbf{a}}_1 - \hat{\mathbf{a}}_3) + y(\hat{\mathbf{a}}_2 - \hat{\mathbf{a}}_3)$ se transforme en $\mathbf{x} = F_T(\hat{\mathbf{x}}) = x(\mathbf{a}_1 - \mathbf{a}_3) + y(\mathbf{a}_2 - \mathbf{a}_3)$, si bien que l'on obtient

$$\pi_T v(\mathbf{x}) = v(\mathbf{a}_3) + x\big[v(\mathbf{a}_1) - v(\mathbf{a}_3)\big] + y\big[v(\mathbf{a}_2) - v(\mathbf{a}_3)\big].$$

Si l'on définit la fonction $\hat{v}$ par $\hat{v}(\hat{\mathbf{x}}) = v(F_T(\hat{\mathbf{x}}))$, alors on la relation

$$\pi_T v(x) = \pi_{\widehat{T}} \hat{v}(\hat{\mathbf{x}}).$$

Il est donc commode de se ramener à l'élément de référence pour effectuer les calculs d'interpolation (ce sera également le cas pour l'obtention des estimations d'erreur, cf. § 2.7).

2.6.2 Élements finis en dimension 2

On peut interpréter le travail effectué au paragraphe précédent dans les termes de la définition 2.3 : le triplet $(T, \mathbb{P}_1, \Sigma)$ avec

$$\mathbb{P}_1 = \big\{ p : (x,y) \mapsto ax + by + c \ ; \text{ avec } a,b,c \in \mathbb{R} \big\} \tag{2.6}$$

$$\Sigma = \{ v \mapsto v(\mathbf{a}_1),\ v \mapsto v(\mathbf{a}_2),\ v \mapsto v(\mathbf{a}_3) \} \tag{2.7}$$

est un élément fini. En effet, une fonction affine de $\mathbb{R}^2$ dans $\mathbb{R}$ est uniquement déterminée à partir de la donnée de ses valeurs en trois points distincts, ce qui implique l'unisolvance. Les fonctions de base associées à l'élément fini de référence correspondant sont

$$\psi_1(x,y) = x, \quad \psi_2(x,y) = y, \quad \psi_3(x,y) = 1 - x - y.$$

Cette construction se généralise à tout degré :

Définition 2.4 (*Élément fini de Lagrange d'ordre k*)

Soit $\widehat{T}$ le triangle de référence, dont les sommets sont $(1,0)$, $(0,1)$ et $(0,0)$. On appelle treillis *d'ordre k l'ensemble de points*

$$\mathcal{T}_k = \left\{ x_{ij} = \left(\tfrac{i}{k}, \tfrac{j}{k}\right) \ ; \ 0 \leq i+j \leq k \right\}.$$

L'élément fini de Lagrange d'ordre k est défini par

- $\diamond$ $K = \widehat{T}$,
- $\diamond$ $P = \mathbb{P}_k = \{p : \mathbb{R}^2 \to \mathbb{R}, \text{ de degré total} \leq k\} = \text{Vect}\{x^i y^j \ ; \ 0 \leq i,j \leq k\}$,
- $\diamond$ $\Sigma = \{\psi_{ij} : v \mapsto p(x_{ij})\}$.

Faire un dessin pour $k = 1, 2, 3$

Si ces éléments sont les plus couramment utilisés, ils ne sont pas les seuls, et les deux exemples ci-dessous – vectoriels – sont employés pour la résolution des problèmes de Stokes et de Maxwell, respectivement.

Exemple. [Élement fini de Raviart-Thomas]

- $\diamond$ $K = \widehat{T}$,
- $\diamond$ $P = \mathbb{P}_0^2 \oplus \mathbf{x}\mathbb{P}_0 = \text{Vect}\{\mathbf{x} \mapsto (0,1), \ \mathbf{x} \mapsto (1,0), \ \mathbf{x} \mapsto \mathbf{x}\}$,
- $\diamond$ $\Sigma = \{\psi_1, \psi_2, \psi_3\}$,

où les formes linéaires sont définies par

$$\psi_i(\mathbf{v}) = \frac{1}{|A_i|} \int_{A_i} \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}\sigma.$$

Faire un dessin où on voit que les A_i sont les arêtes

Les fonctions de base sont données par

$$\psi_1(x,y) = (x-1, y), \quad \psi_2(x,y) = (x, y-1), \quad \psi_3(x,y) = \frac{1}{\sqrt{2}}(x,y).$$

$\diamondsuit$

Exemple. [Élement fini d'arête de Nédélec]

- $\diamond$ $K = \widehat{T}$,
- $\diamond$ $P = \mathbb{P}_0^2 \oplus \{\mathbf{p} \in \mathbb{P}_1^2 \ ; \ \mathbf{x} \cdot \mathbf{p} = 0\}$
  $= \text{Vect}\{\mathbf{x} \mapsto (0,1), \ \mathbf{x} \mapsto (1,0), \ \mathbf{x} \mapsto \mathbf{x}^\perp = (-x_2, x_1)\}$,
- $\diamond$ $\Sigma = \{\psi_1, \psi_2, \psi_3\}$,

où les formes linéaires sont définies par

$$\psi_i(\mathbf{v}) = \frac{1}{|A_i|} \int_{A_i} \mathbf{v} \cdot \boldsymbol{\tau} \, \mathrm{d}\sigma.$$

Faire un dessin où on voit les degrés de liberté

$\diamondsuit$

On peut définir de même des éléments finis en dimension 3 (ou même plus) ayant comme supports des tétraèdres, des prismes, etc.

2.6.3 Transformation géométrique d'un élément fini

La plupart du temps, le domaine Ω est maillé en un grand nombre d'éléments du même type (des triangles, par exemple en dimension 2). Les différents calculs à faire sur les différents éléments sont alors très similaires, et il est commode de *transporter* les informations sur un élément unique, dit *de référence*.

Soit donc un élément fini $(\hat{K}, \hat{P}, \hat{\Sigma})$, dit de référence. On se donne $F_K : \hat{K} \to \mathbb{R}^d$ une application injective, et on note

- $K = F_K(\hat{K})$,
- $P_K = \left\{ \hat{p} \circ F_K^{-1} \ ; \ \hat{p} \in \hat{P} \right\}$,
- $\Sigma_K = \left\{ \psi \ ; \ \exists \hat{\psi} \in \hat{\Sigma}, \ \forall p \in P_K, \ \psi(p) = \hat{\psi}(p \circ F_K) \right\}$.

Proposition 2.10

(K, P_K, Σ_K) est un élément fini.

Démonstration : Il s'agit de montrer l'unisolvance : on note $\hat{\Sigma} = \{\hat{\psi}_1, \dots, \hat{\psi}_N\}$ et on fixe $\alpha_1, \dots, \alpha_N \in \mathbb{R}$. Comme $(\hat{K}, \hat{P}, \hat{\Sigma})$ est un élément fini, il existe un unique élément $\hat{p} \in \hat{P}$ tel que $\hat{\psi}_i(\hat{p}) = \alpha_i$ pour tout i. Si l'on pose $p = \hat{p} \circ F_K^{-1} \in P_K$, on a alors $\psi_i(p) = \alpha_i$, d'où l'existence. Pour l'unicité, on procède de même. ■

Il est facile de relier les opérateurs d'interpolation locaux sur l'élément de référence $\hat{K}$ et l'élément courant K.

Proposition 2.11

Si $\hat{\pi}$ et π désignent les opérateurs d'interpolation dans $(\hat{K}, \hat{P}, \hat{\Sigma})$, et (K, P, Σ), respectivement, alors, pour toute fonction v,

$$\hat{\pi}\hat{v} = \widehat{\pi_K v},$$

si $v = \hat{v} \circ F_K^{-1}$ et $\widehat{\pi_K v} = (\pi_K v) \circ F_K$.

Démonstration : Soit $\hat{x} \in \hat{K}$. Par définition, pour $\hat{\psi} \in \hat{\Sigma}$, on a $\hat{\psi}(\hat{\pi}\hat{v}) = \hat{\psi}(\hat{v})$. On en déduit que

$$\psi\left((\hat{\pi}\hat{v}) \circ F_K^{-1}\right) = \psi\left(\hat{v} \circ F_K^{-1}\right) = \psi(v).$$

Ainsi, par définition de $\pi_K v$, on obtient $(\hat{\pi}\hat{v}) \circ F_K^{-1} = \pi_K v$, soit encore

$$\hat{\pi}\hat{v} = (\pi_K v) \circ F_K = \widehat{\pi_K v}.$$

■

Exemple.

1. *Transformation affine d'un élément triangulaire*

Faire un dessin

L'application est $F_K : x \mapsto Ax + b$, avec $A \in \mathbb{R}^{2\times 2}$ et $b \in \mathbb{R}^2$. La connaissance des sommets (a_1, a_2, a_3) est équivalente à celle de F_K.

2. *Transformation $\mathbb{Q}_1$ d'un élément quadrangulaire*

Faire un dessin

L'application F_K est donnée par

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \longmapsto \begin{pmatrix} a + bx_1 + cx_2 + dx_1x_2 \\ a' + b'x_1 + c'x_2 + d'x_1x_2 \end{pmatrix}.$$

Notons que si l'on se restreint à F_K affine ($d = d' = 0$), alors l'image du carré unité par F_K est un parallélogramme.

$\diamondsuit$

2.6.4 Espaces d'approximation par éléments finis triangulaires de Lagrange

Pour mettre en place la méthode de Galerkine (voir § 2.1), on souhaite construire un espace d'approximation V_h de dimension finie.

Définition 2.5

Soit Ω un domaine polygonal borné. On appelle maillage *de Ω une partition*

$$\mathscr{T}_h = \{K \,;\, K \text{ triangles}\}$$

telle que

- $\overline{\Omega} = \bigcup_{K \in \mathscr{T}_h} K$,
- *Si $K, L \in \mathscr{T}_h$, alors $K \cap L$ est soit un élément de $\mathscr{T}_h$, soit une arête commune à K et L, soit un sommet commun à K et L, soit $\emptyset$.*

Faire un dessin d'une situation interdite, et d'une situation permise

Dans la suite, on suppose que tous les triangles du maillage sont issus d'un même élément de référence :

$$\forall K \in \mathscr{T}_h, \quad K = F_K(\hat{K}) \quad \text{avec } F_K \text{ affine inversible.}$$

Naturellement, si $(\hat{K}, \hat{\mathbb{P}}_k, \hat{\Sigma}_k)$ désigne l'élement fini de référence triangulaire de Lagrange d'ordre k, on note $(K, \mathbb{P}_k, \Sigma_k)$ l'élement courant, selon les définitions du paragraphe 2.6.3.

Définition 2.6

L'espace d'approximation par éléments finis associé au maillage $\mathscr{T}_h$ est défini par

$$V_h = \{v_h \in \mathscr{C}(\Omega) \,;\, \forall K \in \mathscr{T}_h,\ v_h|_K \in \mathbb{P}_k\}.$$

Proposition 2.12

On a l'inclusion $V_h \subset \mathrm{H}^1(\Omega)$.

PREUVE. Soit $v_h \in V_h$. Alors pour tout élément $K \in \mathscr{T}_h$, on a $v_h|_K \in \mathbb{P}_k \subset \mathrm{H}^1(K)$. Pour $i = 1, 2$, on définit donc $w_i \in \mathrm{L}^2(\Omega)$ par

$$\forall K \in \mathscr{T}_h, \quad w_i = \frac{\partial (v_h|_K)}{\partial x_i}.$$

Montrons que $w_i = \dfrac{\partial v_h}{\partial x_i}$ au sens des distributions. Soit donc $\varphi \in \mathscr{C}_c^\infty(\Omega)$,

$$\int_\Omega w_i \varphi = \sum_{K \in \mathscr{T}_h} \int_K \frac{\partial (v_h|_K)}{\partial x_i} \varphi = \sum_{K \in \mathscr{T}_h} \int_{\partial K} v_h|_K \varphi n_i - \sum_{K \in \mathscr{T}_h} \int_K v_h|_K \frac{\partial \varphi}{\partial x_i}$$

Comme $\varphi = 0$ sur $\partial\Omega$, seules les arêtes intérieures interviennent dans la première somme, et elles sont comptées deux fois : pour des éléments adjacents, pour lesquels les normales sont opposées. Ainsi, on obtient

$$\int_\Omega w_i \varphi = - \sum_{K \in \mathscr{T}_h} \int_K v_h|_K \frac{\partial \varphi}{\partial x_i} = - \int_\Omega v \frac{\partial \varphi}{\partial x_i},$$

ce qui achève la preuve. ∎

L'espace d'approximation étant maintenant défini, on peut construire un *interpolé* :

Définition 2.7

Soit $u \in \mathscr{C}(\Omega)$. On définit l'interpolé de u dans V_h par

$$\forall K \in \mathscr{T}_h, \quad \pi_h u|_K = \pi_K \left(u|_K \right).$$

Remarque 2.8

On a $\pi_h u \in \mathscr{C}(\Omega)$ car les degrés de liberté sur une arête commune sont les mêmes de part et d'autre de deux éléments adjacents.

2.7 Estimations d'erreur

2.7.1 Estimation de l'erreur d'interpolation pour l'élément fini de Lagrange

Estimation dans l'élément de référence

Proposition 2.13

Il existe une constante $C_k > 0$ telle que, pour tout $m \leq k+1$,

$$\forall \hat{v} \in \mathrm{H}^{k+1}(\hat{K}), \quad \|\hat{v} - \hat{\pi}\hat{v}\|_{\mathrm{H}^m(\hat{K})} \leq C_k |\hat{v}|_{\mathrm{H}^{k+1}(\hat{K})}.$$

PREUVE. On pose $W = \{\hat{v} - \hat{\pi}\hat{v} \ ; \ \hat{v} \in \mathrm{H}^{k+1}(\hat{K})\} = \{\hat{w} \in \mathrm{H}^{k+1}(\hat{K}) \ ; \ \hat{\pi}\hat{w} = 0\}$.

- ◇ W est un sous-espace vectoriel de $\mathrm{H}^{k+1}(\hat{K})$.
- ◇ W est fermé, par continuité de $\hat{\pi}$.
- ◇ $W \cap \mathbb{P}_k = \{0\}$ par unicité de l'interpolé.

Ainsi, d'après le lemme 2.5, on obtient

$$\|\hat{v} - \hat{\pi}\hat{v}\|_{\mathrm{H}^{k+1}(\hat{K})} \leq C_k |\hat{v}|_{\mathrm{H}^{k+1}(\hat{K})},$$

qui permet de conclure. ∎

Transport de l'estimation dans l'élément courant

Afin d'obtenir une estimation de l'élément courant K, on rappelle que $K = F_K(\hat{K})$, avec F_K affine inversible, i.e.

$$\forall \hat{x} \in \hat{K}, \quad F_K(\hat{x}) = B_K \hat{x} + b_K,$$

avec $B_K \in \mathbb{R}^{2\times 2}$ inversible, et $b_K \in \mathbb{R}^2$.

Proposition 2.14

Il existe une constante $C'_k > 0$ telle que, pour tout $m \leq k+1$,

$$\forall v \in \mathrm{H}^{k+1}(K), \quad \|v - \pi_K v\|_{\mathrm{H}^m(K)} \leq C'_k \|B_K\|_2^{k+1} \|B_K^{-1}\|_2^m |v|_{\mathrm{H}^{k+1}(K)}.$$

PREUVE.

◇ Estimation en norme L^2 : il s'agit de majorer la quantité

$$\int_K |v(x) - \pi_K v(x)|^2 \, \mathrm{d}x.$$

On effectue le changement de variable $x = F_K(\hat{x})$, qui fournit

$$\int_K |v(x) - \pi_K v(x)|^2 \, \mathrm{d}x = |\det B_K| \int_{\hat{K}} |\hat{v}(\hat{x}) - \hat{\pi}\hat{v}(\hat{x})|^2 \, \mathrm{d}\hat{x}.$$

La proposition 2.13 permet d'écrire

$$\int_K |v(x) - \pi_K v(x)|^2 \, \mathrm{d}x \leq C_k |\det B_K| \sum_{|\beta|=k+1} \int_{\hat{K}} \left|\partial^\beta \hat{v}(\hat{x})\right|^2 \, \mathrm{d}\hat{x}.$$

Dans cette dernière intégrale, on fait le changement de variable $\hat{x} = F_K^{-1}(x)$, réciproque du précédent. Par dérivation composée, on a

$$\frac{\partial \hat{v}}{\partial \hat{x}_i} = \sum_{j=1}^{2} \frac{\partial v}{\partial x_j} \frac{\partial \hat{x}_j}{\partial x_i} = \left(B^{\mathsf{T}} \nabla v\right)_i.$$

Ainsi $\nabla \hat{v} = B^{\mathsf{T}} \nabla v$. On en déduit par récurrence l'inégalité

$$\left|\partial^\beta \hat{v}(\hat{x})\right|^2 \leq \|B^{\mathsf{T}}\|_2^{2|\beta|} \sum_{|\beta'|=|\beta|} \left|\partial^{\beta'} v(x)\right|^2,$$

qui fournit finalement

$$\int_K |v(x) - \pi_K v(x)|^2 \, \mathrm{d}x \leq C_k |\det B_K| \times |\det B_K^{-1}| \times \|B_K\|^{2(k+1)} \times |\hat{v}|^2_{\mathrm{H}^{k+1}(K)},$$

d'où le résultat pour $m = 0$.

◇ Pour $m \geq 1$, on doit étudier les intégrales

$$\int_K \left|\partial^\beta v(x) - \partial^\beta \pi_K v(x)\right|^2 \, \mathrm{d}x,$$

pour $|\beta| = m$. On procède de même, et un facteur $\|B_K^{-\mathsf{T}}\|_2^m$ apparaît.

■

Estimation globale

On cherche ici à exprimer les quantités intervenant dans la proposition 2.14 en termes géométriques.

Définition 2.8

Soit K une triangle. On note

- $\diamond$ *h_K le diamètre de K :* $h_K = \max\{\|x-y\| \ ; \ x,y \in K\}$.
- $\diamond$ *ρ_k le diamètre du cercle inscrit :* $\rho_K = \sup\{2R \ ; \ \exists x \in K, \ B(x,R) \subset K\}$.

Lemme 2.15

$$\|B_K\|_2 \leq \frac{h_K}{\rho_{\hat{K}}}, \qquad \|B_K^{-1}\|_2 \leq \frac{h_{\hat{K}}}{\rho_K}.$$

PREUVE. Soit $\hat{\xi} \in \mathbb{R}^2$ tel que $\|\hat{\xi}\| = \rho_{\hat{K}}$. Alors, il existe $\hat{x}, \hat{y} \in \hat{K}$ tels que $\hat{x} - \hat{y} = \hat{\xi}$. On a alors $B_K(\hat{x}-\hat{y}) = x - y$ avec $x, y \in K$ et ainsi $\|B_K(\hat{x}-\hat{y})\| \leq h_K$, soit $\|B_K\hat{\xi}\| \leq h_K$. Finalement,

$$\sup_{\|\hat{\xi}\|=\rho_{\hat{K}}} \frac{\|B_K\hat{\xi}\|}{\|\xi\|} \leq \frac{h_K}{\rho_{\hat{K}}},$$

d'où la première inégalité. La seconde est obtenue de même. ∎

Remarque 2.9

La quantité $\det B_K$ *s'interprète également en termes géométriques :*

$$\det B_K = \frac{\mathrm{Vol}(K)}{\mathrm{Vol}(\hat{K})}.$$

En mettant bout-à-bout les résultats des paragraphes précédents, on obtient directement

Proposition 2.16

Il existe une constante $C'_k > 0$ telle que, pour tout $m \leq k+1$,

$$\forall v \in \mathrm{H}^{k+1}(K), \quad \|v - \pi_K v\|_{\mathrm{H}^m(K)} \leq C'_k \left(\frac{h_K}{\rho_K}\right)^m h_K^{k+1-m} |v|_{\mathrm{H}^{k+1}(K)}. \tag{2.8}$$

Définition 2.9

Une famille de maillages $(\mathscr{T}_h)_{h>0}$ est dite régulière s'il existe $\sigma > 0$ tel que

$$\forall h > 0, \quad \forall K \in \mathscr{T}_h, \quad \sigma_K \overset{\text{def.}}{=} \frac{h_K}{\rho_K} \leq \sigma.$$

Le nombre σ_K est appelé rondeur *du triangle K.*

L'hypothèse de régularité sur le maillage impose que les triangles ne soient pas trop aplatis (idéalement, tous équilatéraux !). Précisément, on a, si θ désigne le plus petit angle

du triangle K,

$$\frac{1}{2\tan^2\frac{\theta}{2}} \leq \sigma_K \leq \frac{1}{\tan\frac{\theta}{2}}.$$

La régularité du maillage permet d'obtenir une estimation de l'erreur d'interpolation globale :

Proposition 2.17

Soit $(\mathscr{T}_h)_h$ une famille de maillages régulière, dont tous les éléments sont images affines de l'élément triangulaire de Lagrange d'ordre k. Alors il existe une constante $C > 0$ telle que pour $m \leq k+1$,

$$\forall v \in \mathrm{H}^{k+1}(\Omega), \quad \|v - \pi_h v\|_{\mathrm{H}^m(\Omega)} \leq Ch^{k+1-m}|v|_{\mathrm{H}^{k+1}(\Omega)}. \tag{2.9}$$

PREUVE. C'est immédiat en sommant l'inégalité d'interpolation (2.8) sur tous les éléments du maillage. ∎

2.7.2 Estimation de l'erreur éléments finis pour l'élément fini de Lagrange

La combinaison du lemme de Céa et de l'estimation (2.10) conduit directement à l'estimation de l'erreur d'approximation par la méthode des éléments finis.

Théorème 2.18

Soit $(\mathscr{T}_h)_h$ une famille de maillages régulière, dont tous les éléments sont images affines de l'élément triangulaire de Lagrange d'ordre k. On suppose que la solution u du problème variationnel satisfait

$$u \in \mathrm{H}^{k+1}(\Omega).$$

Alors il existe $C > 0$ telle que

$$\|u - u_h\|_{\mathrm{H}^1(\Omega)} \leq Ch^k|v|_{\mathrm{H}^{k+1}(\Omega)}. \tag{2.10}$$

Notons que le lemme de Céa ne permet d'obtenir qu'une estimation dans la norme d'énergie du problème, i.e. la norme dans l'espace dans lequel la formulation variationnelle est posée. L'estimation d'erreurs en normes H^m pour $m \geq 2$ demande plus de travail. On détaille ici l'estimation d'erreur en norme L^2 pour le problème de Laplace-Dirichlet et l'interpolation de Lagrange d'ordre 1.

Corollaire 2.19

On se place sous les hypothèses du théorème 2.18 avec $k = 1$ et u solution variationnelle du problème $-\Delta u = f$ dans $\mathrm{H}^1_0(\Omega)$. On suppose Ω convexe si bien que $u \in \mathrm{H}^2(\Omega)$. On a alors

$$\|u - u_h\|_{\mathrm{L}^2(\Omega)} \leq Ch^2|v|_{\mathrm{H}^2(\Omega)}.$$

PREUVE. On utilise le *truc* d'Aubin-Nitsché : posant $e = u - u_h$, on peut écrire

$$\|e\|_{\mathrm{L}^2(\Omega)} = \sup\left\{(e,\varphi)_{\mathrm{L}^2} \;;\; \varphi \in \mathrm{L}^2(\Omega),\ \|\varphi\|_{\mathrm{L}^2} = 1\right\}.$$

Soit alors $\varphi \in L^2(\Omega)$ avec $\|\varphi\|_{L^2} = 1$. On note $w \in H_0^1(\Omega)$ l'unique solution du problème variationnel

$$\forall \psi \in H_0^1(\Omega), \quad a(w, \psi) = (\varphi, \psi)_{L^2},$$

(où a est la forme bilinéaire associé au laplacien). On peut montrer (théorème de régularité elliptique) l'estimation *a priori* suivante

$$\|w\|_{H^2} \leq C\|\varphi\|_{L^2}.$$

Par ailleurs, puisque $e \in H_0^1(\Omega)$, on peut écrire $a(w, e) = (\varphi, e)_{L^2}$. Or

$$a(w, e) = a(e, w) = a(u - u_h, w) = a(u - u_h, w - w_h),$$

pour tout $w_h \in V_h$ d'après l'orthogonalité de Galerkine. Ainsi

$$\left|(\varphi, e)_{L^2}\right| \leq C_a\|u - u_h\|_{H^1} \inf_{w_h \in V_h} \|w - w_h\|_{H^1} \leq Ch|u|_{H^2} \times h\|w\|_{H^2} \leq ch^2|u|_{H^2},$$

d'où le résultat. ∎

2.7.3 Remarques et extensions

Les estimations d'erreur ont été montrées seulement pour l'élément fini de Lagrange triangulaire d'ordre k. Toutefois, il est possible d'en obtenir pour des classes beaucoup plus générales d'éléments finis. On renvoie, par exemple, à [RT83] pour plus de détails.

Par ailleurs, lorsque le domaine Ω n'est pas polygonal, une erreur d'approximation géométrique doit être prise en compte. De même, les intégrales intervenant dans les formulations éléments-finis ne peuvent pas toujours être calculées explicitement (même si Ω est polygonal), par exemple dans le cas de seconds membres f quelconques. Une intégration numérique doit être effectuée et l'erreur résultante doit être prise en compte dans l'erreur finale.

2.8 Quelques remarques sur la mise en œuvre de la méthode

La mise en œuvre de la méthode nécessite le calcul de la matrice $\mathbb{A}$ et du vecteur $\mathbb{F}$ définis par

$$\mathbb{A}_{ij} = a(\varphi_j, \varphi_i), \qquad \mathbb{F}_i = \ell(\varphi_i),$$

où les φ_i forment une base de V_h. Elles sont construites à l'aide des fonctions de base locales des éléments du maillage. Précisément, supposons que l'on dispose de l'information suivante sur la numérotation des nœuds du maillage :

$n(K, I) = i$ si le nœud numéroté i dans le maillage est le I^{e} nœud de l'élément K,

voir figure XX.

Faire un dessin avec quelques triangles et des numéros

Alors, φ_i est définie par

$$\forall K \in \mathscr{T}_h, \quad \varphi_i|_K = \phi_I \text{ si } i = n(K, I).$$

il est sous-entendu que φ_i est nulle sur K dès lors que i n'est pas un nœud de K.

Pour construire le vecteur $\mathbb{F}$, on pourrait être tenté de le construire selon l'algorithme suivant,

```
Boucle sur i
𝔽_i = 0
  Boucle sur K
    𝔽_i = 𝔽_i + ℓ_K(φ_i)
  Fin
Fin
```

où $\ell_K(\varphi_i)$ désigne la contribution de l'élément K : typiquement

$$\ell_K(\varphi) = \int_K f\varphi_i \quad \text{si} \quad \ell(\varphi) = \int_\Omega f\varphi_i.$$

Toutefois, dans la méthode des éléments finis, les fonctions de base φ_i ont des supports localisés, si bien que $\ell_K(\varphi_i)$ est nul pour un grand nombre d'éléments K. Afin d'éviter les passages inutiles dans la boucle, on pourrait repérer de tels éléments. En fait, il est plus simple de boucler dans l'autre sens :

```
𝔽 = 0
Boucle sur K
  Boucle sur les nœuds I de K (nombre=♯Σ̂)
    Calcul de i = n(K, I)
    𝔽_i = 𝔽_i + ℓ_K(φ_I)
  Fin
Fin
```

Soulignons que les valeurs des coefficients du vecteur $\mathbb{F}$ sont remplies au gré de la numérotation, et qu'il faut attendre la fin de la boucle sur les éléments pour être certain que toutes les valeurs sont correctement affectées. Signalons par ailleurs que le calcul de la contribution $\ell_K(\phi_I)$ est en fait calculée par passage à l'élément de référence (comme on l'a fait pour l'obtention des estimations d'erreur).

Annexe A

Rappels sur les fonctions de plusieurs variables

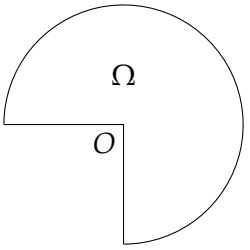
Annexe B

Régularité jusqu'au bord de la solution d'un problème elliptique

On considère le domaine $\Omega \subset \mathbb{R}^2$ suivant

$$\Omega = \left\{(x,y) \in \mathbb{R}^2 \,;\, x^2+y^2 \leq 1\right\} \setminus \left\{(x,y) \in \mathbb{R}^2 \,;\, x<0 \text{ et } y<0\right\}.$$

Il s'agit de trois quarts de disque, comme le montre la figure ci-contre. En utilisant les coordonnées polaires (r,θ), le domaine Ω est décrit par



$$r<1 \text{ et } -\frac{\pi}{2} < \theta < \pi.$$

On pose alors

$$\varphi(x) = r^{\frac{2}{3}} \sin\left(\frac{2\theta}{3}\right).$$

On vérifie aisément que $\varphi \in \mathscr{C}^2(\Omega)$ et que $\Delta\varphi = 0$ dans Ω. Toutefois, φ ne satisfait la condition de Dirichlet que sur les bords droits du domaine.

On introduit donc une fonction de troncature $\chi : \mathbb{R}^2 \to \mathbb{R}$, de classe $\mathscr{C}^\infty$ telle que

$$\chi(x) = 1 \quad \text{si} \quad |x| < \frac{1}{3} \quad \text{et} \quad \chi(x) = 0 \quad \text{si} \quad |x| > \frac{1}{2}.$$

Ainsi, la fonction $u(x) = \chi(x)\varphi(x)$ satisfait $u = 0$ sur $\partial\Omega$. On a, par ailleurs, puisque $\Delta\varphi = 0$,

$$\Delta u(x) = \Delta\chi(x)\varphi(x) + 2\nabla\chi(x)\cdot\nabla\varphi(x),$$

si bien qu'il apparaît clairement que $f = -\Delta u \in \mathscr{C}^\infty(\overline{\Omega})$.

Cet exemple montre qu'on peut trouver une fonction f très régulière pour laquelle la solution du problème le Laplace avec conditions de Dirichlet dans Ω n'est pas régulière. En effet, on vérifie aisément que $u \in \mathrm{H}^1(\Omega)$, mais $u \notin \mathrm{H}^2(\Omega)$.

Annexe C

Interpolation et approximation polynomiales

C.1 Interpolation de Lagrange

C.1.1 Introduction

Le problème de l'interpolation est le suivant : étant donnés $n+1$ *points d'interpolation* $x_0 < \cdots < x_n$ et $n+1$ *valeurs* $y_0, \ldots, y_n$, on cherche un polynôme p_n, de degré au plus n, qui satisfait

$$p_n(x_i) = y_i \quad \text{pour} \quad i = 0, \ldots, n. \tag{C.1}$$

Théorème C.1 (*existence et unicité du polynôme d'interpolation*)

Il existe un unique polynôme p_n, de degré au plus n, satisfaisant la condition (C.1).

PREUVE. L'application $\Phi : \mathbb{P}_n \to \mathbb{R}^{n+1}$ définie par $\Phi(p) = (p(x_i))_{i=0,\ldots,n}$ est clairement linéaire et injective, donc bijective. ∎

Remarque C.1

Si on cherche le polynôme p_n sous la forme $p_n = \sum a_k X^k$, les coefficients $\mathbf{a} = (a_k)_{k=0,\ldots,n}$ satisfont le système linéaire $\mathbf{Va} = \mathbf{y}$ où $\mathbf{V}$ est une matrice de Van der Monde et $\mathbf{y} = (y_k)_{k=0,\ldots,n}$:

$$\mathbf{V} = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix}$$

La matrice $\mathbf{V}$ est inversible puisque les points d'interpolation x_i sont tous distincts. Ce système linéaire fournit un procédé de calcul très simple du polynôme interpolateur. Toutefois la matrice $\mathbf{V}$ est plutôt mal conditionnée, ce qui signifie que pour les grandes valeurs de n (dès quelques dizaines), la résolution du système linéaire peut être entachée d'une erreur importante. Une méthode plus robuste, mais de programmation un peu plus délicate, est l'algorithme de Newton pour lequel on renvoie à [CM84].

On utilise aussi fréquemment la base de Lagrange, *image réciproque de la base canonique de $\mathbb{R}^{n+1}$ par Φ :*

$$\ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}.$$

Le polynôme p_n s'exprime alors simplement comme

$$p_n(x) = \sum_{i=0}^{n} y_i \ell_i(x). \tag{C.2}$$

C.1.2 Stabilité du procédé d'interpolation

On vient d'évoquer la possibilité d'erreurs numériques dans la résolution du système de Van der Monde. Une source d'instabilité réside dans le problème lui-même, indépendamment de la méthode utilisée pour sa mise en œuvre. En effet, considérons que les données (y_i) sont entachées d'erreur : quelle est son influence sur le polynôme p_n ? Le problème étant linéaire, il suffit de considérer la norme de l'application $L_n : \mathbb{R}^{n+1} \to \mathbb{P}_n$ définie par $L_n(\mathbf{y}) = p_n$ (ce n'est autre que l'inverse de l'application Φ de la preuve du paragraphe précédent). À l'aide de l'expression de p_n dans la base de Lagrange, il est facile de calculer la norme de L_n comme application linéaire de $(\mathbb{R}^{n+1}, \|\cdot\|_\infty)$ dans $(\mathbb{P}_n, \|\cdot\|_{\infty,[a,b]})$:

$$\|L_n\| = \sum_{i=0}^{n} \|\ell_i(x)\|_{\infty,[a,b]}.$$

Cette norme, souvent notée Λ_n, est appelée *constante de Lebesgue*. On peut montrer qu'elle tend vers l'infini lorsque n tend vers l'infini à vitesse logarithmique, voir [CM84].

C.1.3 Convergence des polynômes interpolateurs

Plaçons nous désormais dans la situation où les valeurs y_i sont les évaluations d'une fonction *régulière* f aux points x_i. Une question naturelle est de savoir si le polynôme p_n approche correctement la fonction f sur l'intervalle $[x_0, x_n]$. Précisément, on a le résultat suivant.

Théorème C.2 (*estimation de l'erreur d'interpolation*)

Soient f une fonction à valeurs réelles de classe $\mathscr{C}^{n+1}$ sur l'intervalle $[a,b]$ et $x_0, \ldots, x_n$ des réels tels que $a \le x_0 < \cdots < x_n \le b$. Si on note p_n le polynôme interpolateur de f aux abscisses x_i (i.e. $p_n(x_i) = f(x_i)$ pour $i = 0, \ldots, n$), alors

$$\|f - p_n\|_{\infty,[a,b]} \le \frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!} \, \|\pi_{n+1}\|_{\infty,[a,b]}, \tag{C.3}$$

où π_{n+1} est le polynôme (de degré $n+1$) donné par $\pi_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$.

PREUVE. Soit $x \in [a,b]$ fixé, distinct des x_i ; on considère q le polynôme (de degré au plus $n+1$) qui interpole la fonction f en les points $x_0, \ldots, x_n$ et x. On a immédiatement

$$q(t) = p_n(t) + \frac{f(x) - p_n(x)}{\pi_{n+1}(x)} \pi_{n+1}(t). \tag{C.4}$$

Le théorème de Rolle appliqué à la fonction $f(t) - q(t)$, qui s'annule en $n+2$ points fournit l'existence d'un réel $\xi_x \in [a,b]$ tel que

$$q^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x). \tag{C.5}$$

Après insertion de la forme (C.4) dans cette dernière égalité, on obtient immédiatement par dérivation des polynômes p_n et π_{n+1}

$$\frac{f(x) - p_n(x)}{\pi_{n+1}(x)}(n+1)! = f^{n+1}(\xi_x), \tag{C.6}$$

qui conduit au résultat énoncé. ∎

Remarque C.2

L'estimation (C.3) *peut laisser penser, grâce au facteur* $(n+1)!$*, que le polynôme interpolateur* p_n *converge uniformément vers la fonction* f *quand le degré (ou le nombre de points, ce qui revient au même) tend vers* $+\infty$*. C'est en effet le cas pour la fonction exponentielle, par exemple, pour laquelle l'estimation devient*

$$\|f - p_n\|_{\infty,[a,b]} \leq \frac{e^b}{(n+1)!}\,\|\pi_{n+1}\|_{\infty,[a,b]} \leq \frac{e^b(b-a)^{n+1}}{(n+1)!},$$

quantité qui converge bien vers 0 *quand* n *tend vers* $+\infty$*.*

Le cas de la fonction exponentielle est tout-à-fait particulier, puisque ses dérivées sont uniformément bornées. Ce n'est pas toujours le cas : la dérivée d'ordre $n+1$ *de la fonction inverse, par exemple, fait apparaître un facteur* $(n+1)!$ *et on peut montrer que la convergence n'a pas lieu pour des grandes valeurs de* x*. Par ailleurs, il n'y a pas nécessairement convergence, même pour des fonctions très régulières ; un exemple classique est celui de la fonction*

$$f(x) = \frac{1}{1+x^2} \quad \textit{sur} \quad [-5,5], \tag{C.7}$$

qui est une fonction de classe $\mathscr{C}^\infty$*, bornée sur* $\mathbb{R}$*. La figure C.1 montre la fonction et son polynôme interpolateur pour les valeurs* $n = 10, 15$*. Il apparaît que, si la convergence a lieu au centre de l'intervalle, il y a divergence près des extrémités ; le phénomène observé sur les graphes s'accentue pour des plus grandes valeurs du degré : on parle de* phénomène de Runge*, voir [Dem96].*

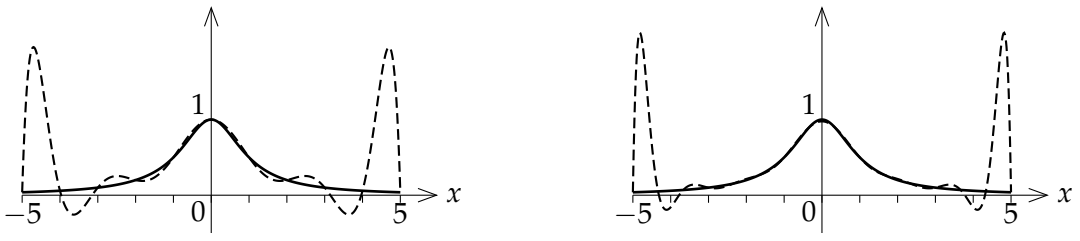


FIGURE C.1 – Interpolation aux abscisses équidistantes de la fonction $\frac{1}{1+x^2}$ (degrés 10 et 15).

Afin d'obtenir la convergence uniforme du polynôme d'interpolation vers la fonction, l'estimation (C.3) suggère deux directions :

- ◇ avoir une « bonne » estimation de la norme de la dérivée $f^{(n+1)}$ sur l'intervalle $[a,b]$ en fonction du degré n ;

◇ diminuer la norme du polynôme π_{n+1}, ce qui revient à choisir intelligemment les points d'interpolation.

La première piste mène naturellement à l'étude des fonctions analytiques, pour lesquelles on peut estimer la norme des dérivées en fonction de l'ordre de dérivation :

Proposition C.3

Si la fonction f est développable en série entière au point $\frac{a+b}{2}$, de rayon de convergence suffisamment grand, alors le polynôme d'interpolation p_n converge uniformément vers f sur l'intervalle $[a,b]$ quand n tend vers $+\infty$ (et ce quel que soit le choix des points d'interpolation).

PREUVE. Il s'agit d'estimer les dérivées successives de f en un point x de l'intervalle $[a,b]$. On note R le rayon de convergence de la série entière en $x_0 = \frac{a+b}{2}$. Alors, pour $\rho < R$, la formule de Cauchy fournit

$$\frac{f^{(n+1)}(x)}{(n+1)!} = \oint_{\mathcal{C}_\rho} \frac{f(z)}{(z-x)^{n+1}}\,\mathrm{d}z,$$

où $\mathcal{C}_\rho$ désigne le cercle de rayon ρ centré en x_0. Si on suppose $\rho = \lambda(b-a)$ ($\lambda > \frac{3}{2}$) alors, avec $M_0 = \sup_{|z-x_0|\leq\rho} |f(z)|$ (le prolongement analytique de f est encore noté f), on obtient directement la majoration

$$\frac{|f^{(n+1)}(x)|}{(n+1)!}\pi_{n+1}(x) \leq 2\pi M_0 \frac{\rho}{\left[(\lambda-\frac{1}{2})(b-a)\right]^{n+1}}(b-a)^{n+1},$$

qui tend vers 0 sous la condition $\lambda > \frac{3}{2}$. On conclut à la convergence uniforme du polynôme interpolateur grâce à l'inégalité (C.3). ∎

Jusqu'ici, on a utilisé la majoration grossière $\pi_{n+1}(x) \leq (b-a)^{n+1}$, valable pour tout jeu de points d'interpolation. On peut toutefois améliorer cette estimation pour des choix particuliers.

Proposition C.4

Le polynôme $\pi_{n+1}(x) = \prod_{i=0}^{n}(x-x_i)$ satisfait l'estimation

$$\|\pi_{n+1}\|_{\infty,[a,b]} \leq C\left(\frac{b-a}{\lambda}\right)^{n+1},$$

où la constante λ est donnée par

Points quelconques	$x_i \in [a,b]$	$\lambda = 1$
Points équidistants	$x_i = i\,\dfrac{b-a}{n}$	$\lambda = e$
Points de Chebychev	$x_i = \dfrac{a+b}{2} + \dfrac{b-a}{2}\cos\left[\dfrac{(2i+1)\pi}{2n+2}\right]$	$\lambda = 4$

PREUVE. Le cas des points quelconques est évident. Considérons tout d'abord celui des points équidistants : pour $x \in [a,b]$, on note $x = s(b-a)/n$,avec $s \in [0,n]$. Maximiser $|\pi_{n+1}(x)|$ sur $[a,b]$ revient ainsi à maximiser sur $[0,n]$ la fonction de $\varphi(s) = \prod_{i=0}^{n} |s-i|$. Il est facile de voir que φ atteint son maximum sur l'intervalle $[0,1]$ d'où

$$|\pi_n(x)| = \left(\frac{b-a}{n}\right)^{n+1} \varphi(s) \leq (b-a)^{n+1} \frac{n!}{n^{n+1}} \sim e\sqrt{\frac{2\pi}{n}} \left(\frac{b-a}{e}\right)^{n+1},$$

d'après la formule de Stirling.

Pour les points de Chebychev, on écrit x sous la forme $x_i = (a+b)/2 + s(b-a)/2$ $(-1 \leq s \leq 1)$, si bien que la polynôme π_{n+1} apparaît comme

$$|\pi_{n+1}(x)| = \left(\frac{b-a}{2n}\right)^{n+1} \varphi(s), \quad \text{avec} \quad \varphi(s) = \prod_{i=0}^{n} \left| s - \cos\left(\frac{(2i+1)\pi}{2n+2}\right)\right| = 2^{-n}|T_{n+1}(s)|,$$

le polynôme de Chebychev de degré $n+1$. Comme le maximum de $|T_{n+1}|$ sur l'intervalle $[-1,1]$ vaut 1, le résultat annoncé s'en déduit immédiatement.

■

Remarque C.3

On peut montrer que les points de Chebychev minimisent la norme uniforme du polynôme π_{n+1} sur l'intervalle $[a,b]$, ce en quoi ils fournissent le choix optimal de points d'interpolation, cf. [Hai]. Toutefois, la convergence uniforme des polynômes interpolateurs n'est pas assurée pour autant. En effet, vu l'explosion de la constante de Lebesgue (voir paragraphe C.1.2), le théorème de Banach-Steinhaus permet de prouver que pour tout choix de points d'interpolation, il existe une fonction continue pour laquelle le polynôme interpolateur ne converge pas uniformément. Ce résultat est à tempérer du point de vue pratique puisque la convergence est uniforme dans le cas des points de Chebychev dès que la fonction à interpoler est de classe $\mathscr{C}^1$ (voir [CM84] pour ces deux derniers points). Bien sûr, dans de nombreuses situations pratiques, il n'est pas possible de choisir les points d'interpolation, qui sont donnés a priori *(par exemple comme suite d'instants régulièrement espacés).*

C.2 Interpolation par morceaux

On a vu dans le paragraphe précédent que l'interpolation de Lagrange ne fournissait pas toujours une suite de polynômes uniformément convergente. Pour pallier ce défaut, on préfère souvent utiliser un procédé d'interpolation *par morceaux* : on subdivise l'intervalle d'intérêt en sous-intervalles sur lesquels on approche la fonction considérée par un polynôme interpolateur de bas degré. Pour obtenir la convergence, on réduit la taille des sous-intervalles, plutôt que d'augmenter le degré.

Précisément, soit $(x^i)_i$ une subdivision de l'intervalle $[a,b]$:

$$a = x^0 < x^1 < \cdots < x^N = b,$$

de pas maximal h :

$$h = \max_{i=1}^{N} |x^i - x^{i-1}|.$$

Sur chaque sous-intervalle $[x^i, x^{i+1}]$ de la subdivision, on introduit des points d'interpolation de Lagrange $(x_\ell^i)_\ell$:

$$x^i = x_0^i < x_1^i < \cdots < x_k^i = x^{i+1}.$$

On note alors p_k^i le polynôme d'interpolation de la fonction f aux points $(x_\ell^i)_\ell$: son degré est au plus ℓ. Il suffit ensuite, à partir des polynômes interpolateurs sur chaque sous-intervalle, de définir une approximation globale $p_{h,k}$ par

$$p_{h,k} = p_k^i \quad \text{sur le sous-intervalle} \quad [x^i, x^{i+1}]. \tag{C.8}$$

Proposition C.5

On suppose la fonction f de classe $\mathscr{C}^{k+1}$ sur l'intervalle $[a, b]$. Le polynôme par morceaux $p_{h,k}$ défini par (C.8) *vérifie*

$$\|f - p_{h,k}\|_{\infty,[a,b]} \leq \frac{1}{(k+1)!}\, h^{k+1}\, \|f^{(k+1)}\|_{\infty,[a,b]}. \tag{C.9}$$

En particulier, $p_{h,k}$ converge uniformément vers f sur $[a, b]$ lorsque le pas h tend vers 0.

PREUVE. D'après l'estimation de l'erreur pour l'interpolation de Lagrange (C.3), on a sur chaque sous-intervalle $[x^i, x^{i+1}]$

$$\|f - p_k^i\|_{\infty,[x^i,x^{i+1}]} \leq \frac{1}{(k+1)!}\, \|f^{k+1}\|_{\infty,[x^i,x^{i+1}]}\, \|\pi_{k+1}^i\|_{\infty,[x^i,x^{i+1}]}, \tag{C.10}$$

où la polynôme $\pi_{k+1}^i = \prod_{\ell=0}^k (x - x_\ell^i)$ peut être majoré par h^{k+1}. On conclut en prenant le maximum pour $0 \leq i \leq N$. ∎

Remarque C.4

- ◇ *on peut choisir un degré d'interpolation différent sur chaque sous-intervalle, i.e. remplacer k par k_i ;*
- ◇ *on a imposé que les extrémités x^i des sous-intervalles soient des points d'interpolation pour assurer la continuité de l'approximation globale $p_{h,k}$;*
- ◇ *même si la fonction f est très régulière, l'approximation $p_{h,k}$ n'est, en général, que continue, il faudrait utiliser une interpolation de type Hermite pour assurer une plus grande régularité. Les fonctions splines en sont un bon exemple (voir [QSS00] ou [SB02] par exemple).*

Références

[Ada75] Robert A. Adams. *Sobolev spaces*. Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[Cia82] Philippe G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1982.

[CM84] Michel Crouzeix and Alain L. Mignot. *Analyse numérique des équations différentielles*. Collection mathématiques appliquées pour la maîtrise. Masson, Paris, 1984.

[DD05] Bruno Després and François Dubois. *Systèmes hyperboliques de lois de conservation - Application à la dynamique des gaz*. École Polytechnique, Paris, 2005.

[Dem96] Jean-Pierre Demailly. *Analyse numérique et équations différentielles*. PUG, Grenoble, 1996.

[EG02] Alexandre Ern and Jean-Luc Guermond. *Éléments finis : théorie, applications, mise en œuvre*, volume 36 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 2002.

[Hai] Ernst Hairer. Introduction à l'analyse numérique. Notes de cours, disponibles à l'adresse `http://www.unige.ch/math/folks/hairer/`

[QSS00] Alfio Quarteroni, Riccardo Sacco, and Fausto Saleri. *Numerical mathematics*, volume 37 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2000.

[RT83] P.-A. Raviart and J.-M. Thomas. *Introduction à l'analyse numérique des équations aux dérivées partielles*. Collection Mathématiques Appliquées pour la Maîtrise. [Collection of Applied Mathematics for the Master's Degree]. Masson, Paris, 1983.

[SB02] J. Stoer and R. Bulirsch. *Introduction to numerical analysis*, volume 12 of *Texts in Applied Mathematics*. Springer-Verlag, New York, third edition, 2002. Translated from the German by R. Bartels, W. Gautschi and C. Witzgall.