Équations aux dérivées partielles

Par définition, une équation aux dérivées partielles (EDP) a pour inconnue une fonction de plusieurs variables (alors qu'une équation différentielle ordinaire a pour inconnue une fonction d'une seule variable). L'analyse (mathématique et) numérique des EDP est un vaste domaine, que nous aborderons ici sous l'angle de trois équations type (linéaires) et de deux méthodes numériques de base : *différences finies* et *éléments finis*.

Le Laplacien On « rappelle » que, pour une fonction $u : \Omega \to \mathbb{R}^n$ deux fois différentiable sur un ouvert Ω de $\mathbb{R}^d$,

$$\Delta u \;=\; \sum_{j=1}^{d} \frac{\partial^2 u}{\partial x_j^2}\,.$$

L'opérateur différentiel Δ est appelé *Laplacien*.

Trois EDP-type

Équation de Poisson Étant donnée une fonction $f : \Omega \to \mathbb{R}^n$, on appelle équation de Poisson de *terme source* f :

$$-\Delta u \;=\; f\,.$$

Cette équation est qualifiée d'*elliptique* (par analogie avec l'équation générale d'une ellipse $\xi_1^2/a^2 + \xi_2^2/b^2 = 1$).

Équation de la chaleur Étant donné un réel positif κ, on appelle équation de la chaleur, ou équation de diffusion pour le coefficient de diffusion κ :

$$\partial_t u = \kappa \Delta u\,.$$

Cette équation est qualifiée de *parabolique* (par analogie avec l'équation générale d'une parabole $y \;=\; \xi^2/a^2$).

Équation des ondes Étant donné un réel positif c, on appelle équation des ondes pour la vitesse c :

$$\partial_{tt}^2 u \;-\; c^2\,\Delta u \;=\; 0\,.$$

Cette équation est qualifiée d'*hyperbolique* (par analogie avec l'équation générale d'une hyperbole $\xi_1^2/a^2 - \xi_2^2/b^2 = 1$).

Les équations des ondes et de la chaleur sont dites *d'évolution* car elles modélisent en général un phénomène instationnaire, évoluant avec le temps t. L'équation de Poisson est quant à elle *stationnaire* : elle modélise en général un phénomène à l'équilibre dans l'espace $\mathbb{R}^d$. Les trois équations sont *linéaires*, c'est-à-dire qu'elles dépendent linéairement de l'inconnue u. Nous n'étudierons pas ici d'équation non-linéaire. Les équations de Poisson et de la chaleur modélisent des phénomènes de *diffusion*, comme celle de la chaleur (!), de la matière (par exemple un polluant dans une rivière, ou des bactéries dans un organe, etc.), ou encore d'une charge électrique. L' équation de Poisson pour $f = 0$, aussi appelée *équation de Laplace*, peut être vue comme un cas particulier d''equation de la chaleur lorsque l'équilibre est atteint, c'est-à-dire lorsque l'inconnue u ne dépend plus de t. L'équation des ondes modélise des phénomènes de *propagation*, comme celle du son, de la lumière.

TAB. 1 – Quelques unités physiques

quantité	unité S.I.
masse	kilogramme (kg)
longueur	mètre (m)
temps	seconde (s)
température	Kelvin (K)
vitesse	$m.s^{-1}$
force	Newton, $1\,N = 1\,kg.m.s^{-2}$
pression	Pascal, $1\,Pa = 1\,kg.m^{-1}.s^{-2}$
énergie	Joule, $1\,J = 1\,kg.m^{2}.s^{-2}$
intensité électrique	Ampère (A)
charge électrique	Coulomb, $1\,C = 1\,A.s$
potentiel électrique	Volt (V)

Éléments de modélisation Les modèles mentionnés ci-après utilisent divers types de quantités physiques. Pour mémoire, quelques unités du système international (unités S.I.) sont « rappelées » dans le tableau 1.

Diffusion Les modèles de phénomènes diffusifs ont en commun d'être régis par une loi empirique exprimant que le flux j par unité de surface et de temps d'une certaine quantité n est proportionnel au gradient de cette quantité. Par ailleurs, la conservation de cette quantité n s'exprime dans ce que l'on appelle l'*équation de continuité* :

$$\partial_t n + \operatorname{div} j = 0\,, \tag{1}$$

où div désigne la divergence en espace :

$$\operatorname{div} j = \sum_{k=1}^{3} \partial_k j_k\,.$$

On comprend facilement l'équation de continuité si n est par exemple une densité de particules (c'est-à-dire le nombre de particules par unité de volume) : le nombre de particules dans un domaine (régulier, borné) Ω de $\mathbb{R}^3$ à l'instant t est égal à

$$\iiint_{\Omega} n(t, x_1, x_2, x_3)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\,,$$

tandis qu'à l'instant $t + \delta t$ il est égal à

$$\iiint_\Omega n(t+\delta t, x_1, x_2, x_3)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3$$

$$\simeq \iiint_\Omega n(t, x_1, x_2, x_3)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3 \;-\; \iint_{\partial\Omega} \sum_{k=1}^{3} j_k(t, x_1, x_2, x_3)\, n_k(x_1, x_2, x_3)\mathrm{d}\sigma(x_1, x_2, x_3)\,,$$

où σ désigne la mesure sur la surface $\partial\Omega$, n le vecteur normal unitaire extérieur à Ω, et l'approximation est valable si δt est assez petit pour que j ne varie pas trop entre t et $t + \delta t$. Or par la formule de Stokes

$$\iint_{\partial\Omega} \sum_{k=1}^{3} j_k(t, x_1, x_2, x_3)\, n_k(x_1, x_2, x_3)\mathrm{d}\sigma(x_1, x_2, x_3) \;=\; \iiint_\Omega \sum_{k=1}^{3} \partial_k j_k(t, x_1, x_2, x_3)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\,.$$

En divisant par δt et en passant à la limite $\delta t \to 0$ on obtient

$$\iiint_\Omega (\partial_t n \;+\; \mathrm{div}\, j)(t, x_1, x_2, x_3)\,\mathrm{d}x_1\mathrm{d}x_2\mathrm{d}x_3\,.$$

Ceci étant vrai quel que soit Ω, on en déduit l'équation locale (1). Voici trois exemples de modèles de diffusion.

Particules Comme expliqué ci-dessus, si n est une densité de particules et j est son flux, l'évolution de ces quantités est régi par l'équation (1). De plus, la *loi de Fick* s'écrit $j \;=\; -\,D\,\nabla n$, où $D > 0$ est un coefficient empirique, appelé *coefficient de diffusion*, de sorte que n doit satisfaire l'EDP :

$$\partial_t n \;=\; D\,\Delta n\,.$$

Charges électriques Si n est une densité volumique de charges électriques et j est la densité de courant associée on a encore l'équation (1). De plus, la *loi d'Ohm* exprime que $j = \sigma E$, où σ est la conductivité du matériau et E est le champ électrostatique. Si ce dernier dérive d'un potentiel V, c'est-à-dire si $E \;=\; -\nabla V$, on a alors $j \;=\; -\sigma\,\nabla V$. En particulier à l'équilibre, c'est-à-dire pour $\partial_t n \;=\; 0$, l'équation de continuité implique l'équation de Laplace (c'est-à-dire l'équation de Poisson avec terme source nul) pour V : $\Delta V = 0$. Si f est une source extérieure de charge électrique, on a l'équation de Poisson $\Delta V = f/\sigma$.

Chaleur Si $u(t, x_1, x_2, x_3)$ désigne la température dans un matériau à l'instant t au point de coordonnées (x_1, x_2, x_3), la *loi de Fourier* exprime que le flux de chaleur par unité de surface et par unité de temps est proportionnel au gradient de température $\nabla u = (\partial_1 u, \partial_2 u, \partial_3 u)^{\mathrm{t}}$, le coefficient de proportionnalité λ étant appelé *conductivité thermique* du matériau :

$$j \;=\; -\lambda\,\nabla u\,.$$

Par ailleurs, si Q est la chaleur par unité de volume, $\partial_t Q = \rho\, c_p\, \partial_t u$, où ρ désigne la masse volumique du matériau et c_p sa chaleur massique. Ainsi u doit vérifier l'équation

$$\partial_t u \;=\; \kappa\,\Delta u$$

avec $\kappa := \lambda/(\rho c_p)$. Si f représente une source extérieure de chaleur par unité de volume, l'équation de continuité doit être modifiée en

$$\partial_t Q \;+\; \mathrm{div}\, j \;=\; f\,.$$

Propagation

Cordes vibrantes Considérons une corde tendue entre ses deux extrémités, suffisamment fine pour pouvoir négliger les variations de tension dans sa section (et les variations de section dûes à son élasticité !). Les paramètres physiques mis en jeu sont sa section σ_0 (aire), sa densité linéaire ρ_0, reliée à sa densité volumique ω_0 par

$$\rho_0 \,=\, \sigma_0\,\omega_0\,,$$

et T_0 sa tension « au repos », nombre positif homogène à une force. Soit $u(t,x) \in \mathbb{R}^3$ le déplacement *transversal* de cette corde à l'instant t, par rapport à une position de référence $x\,\mathrm{e}_1 \in \mathbb{R}^3$, $x \in \mathbb{R}$. On suppose le déplacement longitudinal négligeable. Autrement dit, le point situé en $x\,\mathrm{e}_1$ dans la position de référence se retrouve en $w(t,x) = x\,\mathrm{e}_1 + u(t,x)$, et $u(t,x) \perp \mathrm{e}_1$. La tension de la corde $T(t,x)$ au point $w(t,x)$ est un nombre positif tel qu'un morceau de corde correspondant au segment $[x, x+\delta x]$ ($\delta x > 0$) soit soumis à la force

$$T(t,x+\delta x)\,\theta(t,x+\delta x) \,-\, T(t,x)\,\theta(t,x)\,,$$

où $\theta(t,x) := \partial_x w(t,x)$ (noter que $\theta(t,x)$ est tangent à la corde en $w(t,x)$). L'accélération de la corde au point $w(t,x)$ est simplement $\partial^2_{tt} w(t,x) = \partial^2_{tt} u(t,x)$. La relation fondamentale de la mécanique, ou *loi de Newton* ($F = m\,\gamma$) appliquée au morceau de corde $[x, x+\delta x]$ s'écrit donc, pour la composante parallèle à e_1 :

$$T(t,x+\delta x) \,-\, T(t,x) \,=\, 0\,,$$

et pour la composante orthogonale à e_1 :

$$T(t,x+\delta x)\,\partial_x u(x+\delta x,t) \,-\, T(t,x)\,\partial_x u(t,x) \,=\, \int_x^{x+\delta x} \rho_0\,\partial^2_{tt} u(y,t)\,\mathrm{d}y\,.$$

Par suite, $T(t,x) = T_0(t)$ est indépendant de x, et en faisant tendre δx vers 0 dans la seconde équation, on obtient

$$T_0\,\partial^2_{xx} u \,=\, \rho_0\,\partial^2_{tt} u\,.$$

Si T_0 est indépendant de t (ce qui revient à supposer que la tension exercée aux extrémités est fixe), on a bien une équation des ondes pour u, en une dimension d'espace, avec

$$c \,=\, \sqrt{\frac{T_0}{\rho_0}}\,,$$

à condition que T_0 soit effectivement positif (une corde qui n'est pas en tension s'affaisse et ne peut pas vibrer !).

Barres élastiques À l'inverse d'un corde, dans une barre élastique rigide, on peut ne considérer que les déplacements longitudinaux, c'est-à-dire qu'un point situé en $x\,\mathrm{e}_1$ dans la position de référence se retrouve après compression ou étirement en $w(t,x) = x\,\mathrm{e}_1 + u(t,x)$ avec $u(t,x) \parallel \mathrm{e}_1$. La tension de la barre $T(t,x)$ en $w(t,x)$, n'a pas de signe défini (la barre pouvant être indifféremment en compression ou en étirement). Une loi de l'élasticité affirme que pour faire varier de δl un morceau de longueur l_0 il faut une variation de

tension δT proportionnelle à $\delta l/l_0$. Quantitativement, on définit E_0 le *module d'Young* du matériau tel que

$$\delta T = E_0\,\sigma_0\,\frac{\delta l}{l_0}\,.$$

Par définition, E_0 est un nombre positif homogène à une pression. En appliquant cette loi à un morceau $[x, x+\delta x]$, qui devient $[x+u(t,x), x+\delta x+u(t,x+\delta x)]$, on obtient

$$T(t,x) - T_0(x) = E_0\,\sigma_0\,\frac{u(x+\delta x,t)-u(t,x)}{\delta x}\,,$$

d'où à la limite lorsque δx tend vers 0 :

$$T(t,x) = T_0(x) + E_0\,\sigma_0\,\partial_x u\,.$$

D'autre part, d'après la loi de Newton appliquée au morceau $[x, x+\delta x]$:

$$T(t,x+\delta x) - T(t,x) = \int_x^{x+\delta x} \rho_0\,\partial_{tt}^2 u(y,t)\,\mathrm{d}y\,,$$

d'où à la limite lorsque δx tend vers 0 :

$$\partial_x T = \rho_0\,\partial_{tt}^2 u\,.$$

En supposant la tension de référence T_0 homogène, c'est-à-dire indépendante de x, on en déduit que u (ainsi que T d'ailleurs, par dérivation) satisfait l'équation des ondes de vitesse

$$c = \sqrt{\frac{E_0}{\omega_0}}\,.$$

Si l'on s'intéresse à la densité ρ le long de la barre, on voit assez facilement qu'elle est donnée par

$$\rho(t,x) = \rho_0\,(\,1 - \partial_x u\,)\,.$$

En effet, pour chaque morceau de longueur initiale l_0 on a

$$\rho\,l = \rho_0\,l_0 \quad \text{d'où} \quad \frac{\delta\rho}{\rho_0} + \frac{\delta l}{l_0} = 0\,.$$

En appliquant cette relation au morceau $[x, x+\delta x]$ et en faisant tendre δx vers 0, on en déduit

$$\frac{\rho-\rho_0}{\rho_0} + \partial_x u = 0\,.$$

Par suite, en supposant la densité initiale ρ_0 homogène, on voit par dérivation que ρ satisfait la même équation des ondes que u (et T).

Tuyaux sonores Pour un fluide, un peu d'intuition physique montre que la tension T est reliée à la pression p par $p = -T/\sigma_0$. D'où,

$$\frac{\delta T}{T_0} = \frac{\delta p}{p_0} = -\frac{1}{\chi_0}\,\frac{\delta v}{v_0}\,,$$

où χ_0 est le coefficient de compressibilité (sans dimension), et v le volume. Or dans un tube de section constante,

$$\frac{\delta v}{v_0} = \frac{\delta l}{l_0}\,.$$

Donc on a une loi analogue à celle de l'élasticité, avec

$$E_0 \,=\, \frac{p_0}{\chi_0}\,.$$

En particulier, pour un gaz parfait adiabatique,

$$p\,v^\gamma \,=\, \text{cte}\,,$$

d'où $\chi_0 \,=\, 1/\gamma$ et $E_0 \,=\, \gamma\,p_0$. On trouve comme vitesse de propagation

$$c \,=\, \sqrt{\frac{\gamma\,p_0}{\omega_0}}\,.$$

C'est l'expression bien connue de la vitesse du son.

Application numérique. Dans l'air, assimilé à un gaz di-atomique, on a approximativement $\gamma \,=\, 7/5$ (on obtient ce nombre en raisonnant sur le nombre n de degrés de liberté des molécules ; de façon générale, $\gamma \,=\, (5+n)/(3+n)$). La loi des gaz parfaits

$$p \,=\, \frac{\omega\,R\,T}{M}\,,\quad R \,=\, 8,3144\,J.K^{-1}.mol^{-1}\,,\quad M \,=\, 28,8.10^{-3}\,kg.mol^{-1}\,,$$

permet de calculer

$$c \,=\, \sqrt{\gamma\,\frac{R\,T}{M}} \,\simeq\, 332\,m.s^{-1}$$

à une température de $273\,K$, ce qui correspond très bien à la réalité !

Acoustique Les équations de l'acoustique portent sur les variations de pression p et de vitesse v du gaz ambiant (généralement l'air !) par rapport à son état au repos. Elles s'écrivent

$$\begin{cases} \partial_t\,p \,+\, \omega_0\,c_0^2\,\text{div}\,v \,=\, 0\,, \\[2ex] \partial_t\,v \,+\, \dfrac{1}{\omega_0}\,\nabla p \,=\, 0\,, \end{cases}$$

où ω_0 la densité volumique du gaz et c_0 est la vitesse du son ($c_0 \,=\, \sqrt{\gamma p_0/\omega_0}$), $\text{div}\,v \,=\, \sum_{j=1}^3 \partial_j v_j$. En prenant la divergence de la seconde équation et en éliminant $\text{div}\,v$ grâce à la première, on obtient l'équation des ondes pour p :

$$\partial_{tt}^2\,p \,-\, c_0^2\,\Delta p \,=\, 0\,.$$

Schémas aux différences finies

Le principe des schémas aux différences finies repose sur la formule de Taylor, permettant d'approcher les dérivées de la fonction inconnue par des « dérivées discrètes ».

Si u est une fonction régulière de x et $\Delta x > 0$ est un *pas d'espace* (destiné à tendre vers 0), on peut discrétiser de plusieurs manières différentes sa dérivée :

- différence décentrée avant :
$$\frac{u(x+\Delta x)-u(x)}{\Delta x} \;=\; \partial_x u(x) \;+\; \mathcal{O}(\Delta x)\,,$$
- différence décentrée arrière :
$$\frac{u(x)-u(x-\Delta x)}{\Delta x} \;=\; \partial_x u(x) \;+\; \mathcal{O}(\Delta x)\,,$$
- différence centrée :
$$\frac{u(x+\Delta x)-u(x-\Delta x)}{2\Delta x} \;=\; \partial_x u(x) \;+\; \mathcal{O}(\Delta x^2)\,.$$

Pour la dérivée seconde, une discrétisation « naturelle » est
$$\frac{u(x+\Delta x)-2u(x)+u(x-\Delta x)}{(\Delta x)^2} \;=\; \partial^2_{xx} u(x) \;+\; \mathcal{O}(\Delta x^2)\,.$$

La formule de Taylor avec reste intégral donne en effet, pour une fonction de classe $\mathscr{C}^4$,
$$u(x\pm\Delta x) \;=\; u(x) \;\pm\; \Delta x\,\partial_x u(x) \;+\; \frac{1}{2}\,(\Delta x)^2\,\partial^2_{xx} u(x) \;\pm\; \frac{1}{6}\,(\Delta x)^3\,\partial^3_{xxx} u(x)$$
$$+\,\frac{1}{6}\;(\Delta x)^4\int_0^1 (1-\theta)^3\,\partial^4_{xxxx} u(x+\theta\Delta x)\,\mathrm{d}\theta\,.$$

Problème aux limites elliptique Considérons pour commencer le problème aux limites suivant, où $\kappa > 0$, α, $\beta \in \mathbb{R}$, et $f : x \in [a,b] \to f(x) \in \mathbb{R}$ sont des données :

$$\left\{\begin{array}{l} -\,\kappa\,\partial^2_{xx} u \;=\; f\,, \quad x \in [a,b]\,, \\ u(a) = \alpha\,, \quad u(b) = \beta\,, \end{array}\right. \tag{2}$$

Noter que même si la première équation est une équation différentielle ordinaire, ce n'est pas ce que l'on appelle un problème de Cauchy (pour lequel il faudrait prescrire u et sa dérivée première en un même point, par exemple en a ou en b). C'est un *problème de Dirichlet*, car l'équation dans le domaine $[a,b]$ est complétée par la donnée des valeurs au bord ($\{a\} \cup \{b\}$) de la solution : on parle de *conditions au bord de Dirichlet*.

Bien sûr on a ici une équation différentielle réduite à sa plus simple expression, que l'on peut résoudre analytiquement si f est continue au prix du calcul de primitives, et les deux conditions au bord permettent de déterminer les constantes d'intégration. Le problème (2) va néanmoins servir de modèle pour présenter les schémas aux différences finies.

Une façon simple d'approcher ce problème est de diviser l'intervalle $[a,b]$ en $(N+1)$ *mailles* $[x_j, x_{j+1}]$ pour $j \in \{0,\ldots,N\}$, avec $x_0 = a$, $x_{N+1} = b$, et $x_{j+1} - x_j = \Delta x$, et de résoudre le système discret

$$\left\{\begin{array}{l} -\,\kappa\,\dfrac{u_{j+1}-2u_j+u_{j-1}}{(\Delta x)^2} \;=\; f(x_j)\,, \quad j \in \{1,\ldots,N\}, \\ u_0 = \alpha\,, \quad u_{N+1} = \beta\,. \end{array}\right. \tag{3}$$

Ce système peut se mettre sous forme matricielle (et c'est comme cela qu'on le résout avec un logiciel comme Scilab[1], ou Matlab[2]) : si U désigne le vecteur de $\mathbb{R}^N$ de composantes $u_1, \dots, u_N$,

$$\kappa\, A\, U \,=\, (\Delta x)^2\, F\,, \quad A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \quad F \,=\, \begin{pmatrix} f(x_1) + \frac{\kappa}{(\Delta x)^2}\,\alpha \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) + \frac{\kappa}{(\Delta x)^2}\,\beta \end{pmatrix}.$$

Dans la matrice A, les coefficients qui ne sont pas précisés sont nuls. On voit facilement que A est définie positive : en effet, si l'on note $U_- \in \mathbb{R}^{N-1}$ le vecteur de composantes $(u_1, \dots, u_{N-1})$ et $U_+ \in \mathbb{R}^{N-1}$ celui de composantes $(u_2, \dots, u_N)$,

$$\langle U, AU\rangle \,=\, 2\,\|U\|^2 \,-\, 2\,\langle U_-, U_+\rangle \,=\, u_1^2 \,+\, u_N^2 \,+\, \|U_+ - U_-\|^2 \,\geq\, 0$$

avec égalité si et seulement si

$$u_1 \,=\, 0,\; u_N \,=\, 0\,,\; U_- \,=\, U_+\,,$$

ce qui implique $U = 0$. Par suite, le système linéaire $\kappa\, A\, U \,=\, (\Delta x)^2\, F$ admet une solution U unique, que l'on peut calculer numériquement par diverses méthodes que l'on détaillera pas ici.

Attachons nous à montrer que cette approximation fournit une famille de solutions numériques, notées $u_{\Delta x}$, qui converge, dans un sens à préciser, vers la solution exacte. Nous allons nous placer dans l'espace $L^2([a,b])$ des fonctions de carré intégrable sur $[a,b]$, muni de la norme définie par

$$\|u\|_{L^2} \,=\, \left(\int_a^b |u(x)|^2\,\mathrm{d}x\right)^{1/2}.$$

(Si u est à valeurs vectorielles, dans $\mathbb{R}^p$ disons, $|\cdot|$ désigne la norme euclidienne.) Sachant qu'une solution numérique $u_{\Delta x}$ est en fait donnée par un vecteur de $\mathbb{R}^{N+2}$ (de composantes $u_0, \dots, u_{N+1}$) il faut naturellement munir $\mathbb{R}^{N+2}$ d'une norme dépendant de N et de $\Delta x = (b-a)/(N+1)$, disons $\|\cdot\|_{N,\Delta x}$, de sorte que, pour une fonction suffisamment régulière,

$$\|(u(a), u(a+\Delta x), \dots, u(b-\Delta x), u(b))\|_{N,\Delta x} \,\to\, \|u\|_{L^2}$$

lorsque N tend vers $+\infty$ et $\Delta x = (b-a)/(N+1)$ tend vers 0. Rappelons pour cela la *formule des trapèzes*, pour une fonction $f : [a,b] \to \mathbb{R}^d$ de classe $\mathscr{C}^2$:

$$\int_a^b f(x)\,\mathrm{d}x \,= (b-a)\,\frac{f(a)+f(b)}{2} \,+\, \mathcal{O}((b-a)^3)\,.$$

La démonstration repose évidemment sur la formule de Taylor. Considérons p la fonction affine définie par $p(x) := f(a) + \frac{f(b)-f(a)}{b-a}\,(x-a)$. Alors

$$\int_a^b f(x)\,\mathrm{d}x \,-(b-a)\,\frac{f(a)+f(b)}{2} \,=\, \int_a^b (f(x) \,-\, p(x))\,\mathrm{d}x\,,$$

[1] logiciel libre développé à l'INRIA

[2] logiciel commercialisé par Mathworks

et, si l'on note $c = (a+b)/2$, la formule de Taylor avec reste intégral appliquée à $f - p$ donne

$$f(x) - p(x) \;=\; f(c) - p(c) + (f'(c) - p'(c))\,(x-c) + \int_0^1 (1-\theta)\, f''(c+\theta(x-c))\,(x-c)^2 \,\mathrm{d}\theta\,.$$

Notons $r(x) := \int_0^1 (1-\theta)\, f''(c+\theta(x-c))\,(x-c)^2 \,\mathrm{d}\theta$. En appliquant la formule ci-dessus à $x = a$ et $x = b$ en particulier, on en déduit (puisque $f(a) = p(a)$, $f(b) = p(b)$, et $a - c + b - c = 0$)

$$f(c) \;-\; p(c) \;=\; -\,\frac{r(a) + r(b)}{2}\,.$$

Par suite,

$$f(x) \;-\; p(x) \;=\; r(x) \;-\; \frac{r(a) + r(b)}{2}\,,$$

d'où

$$\int_a^b (f(x) \;-\; p(x))\,\mathrm{d}x \;=\; \int_a^b r(x)\,\mathrm{d}x \;-\; (b-a)\,\frac{r(a) + r(b)}{2}\,.$$

Si l'on note $C := \max_{\xi\in[a,b]} |f''(\xi)|$, on a donc

$$\left| \int_a^b r(x)\,\mathrm{d}x \;-\; (b-a)\,\frac{r(a) + r(b)}{2} \right| \;\leq\; \frac{C}{2} \int_a^b (x-c)^2\,\mathrm{d}x \;+\; \frac{C}{8}\,(b-a)^3 \;=\; C\,\frac{(b-a)^3}{6}\,.$$

Par suite,

$$\left| \int_a^b f(x)\,\mathrm{d}x \;-\; (b-a)\,\frac{f(a) + f(b)}{2} \right| \;\leq\; \frac{(b-a)^3}{6} \max_{\xi\in[a,b]} |f''(\xi)|\,.$$

(En fait on peut améliorer cette inégalité en remplaçant $1/6$ par $1/12$ lorsque f est à valeurs scalaires, par application du théorème de Rolle à la dérivée seconde de $f - p_2$, où p_2 est le polynôme du second degré coïncidant avec f en a, b et x, pour tout $x \in [a,b]$.)

Ainsi, pour une fonction u de classe $\mathscr{C}^2$ sur $[a,b]$, d'après la formule des trapèzes appliquée à la fonction $|u|^2$ sur chaque maille $[x_j, x_{j+1}]$ de taille $\Delta x = (b-a)/(N+1)$, on a

$$\int_a^b |u(x)|^2\,\mathrm{d}x \;=\; \sum_{j=0}^{N} \left(\frac{|u(x_j)|^2 + |u(x_{j+1})|^2}{2} \;+\; \mathcal{O}(\Delta x^3) \right)$$

$$=\; \Delta x\,\frac{|u(x_0)|^2 + |u(x_{N+1})|^2}{2} \;+\; \Delta x \sum_{j=1}^{N} |u(x_j)|^2 \;+\; \mathcal{O}(\Delta x^2)\,.$$

Cette expression nous conduit à définir

$$\|(u_0, \ldots, u_{N+1})\|_{N,\Delta x} \;=\; \left(\Delta x\,\frac{|u_0|^2 + |u_{N+1}|^2}{2} \;+\; \Delta x \sum_{j=1}^{N} |u_j|^2 \right)^{1/2}\,.$$

En admettant l'existence d'une solution pour le problème exact (2), nous sommes maintenant en mesure de montrer le résultat de convergence suivant.

Théorème 1 *Soient $\kappa > 0$, $f \in \mathscr{C}^2([a,b];\mathbb{R}^d)$, α, $\beta \in \mathbb{R}^d$, et $u \in \mathscr{C}^4([a,b];\mathbb{R}^d)$ solution de* (2)*. Alors les solutions $u_{\Delta x}$, de composantes $u_0, \ldots, u_{N+1}$, de* (3) *sont telles que*

$$\|(u(a), u(a+\Delta x), \ldots, u(b-\Delta x), u(b)) \;-\; (u_0, u_1, \ldots, u_N, u_{N+1})\|_{N,\Delta x} \;\to\; 0$$

lorsque N tend vers $+\infty$.

Dém. La démonstration repose (comme pour le problème de Cauchy et le schéma d'Euler) sur une majoration de l'*erreur de consistance* (locale)

$$\mathcal{R}(x,u,\Delta x) \;:=\; -\,\kappa\,\frac{u(x+\Delta x)-2u(x)+u(x-\Delta x)}{(\Delta x)^2} \;-\; f(x)\,,$$

et sur la *stabilité* de la méthode numérique. Pour l'erreur de consistance, on a (comme déjà vu)

$$\mathcal{R}(x,u,\Delta x) \;=\; -\,\kappa\,\partial^2_{xx}u(x) \;-\; f(x) \;+\; \mathcal{O}(\Delta x^2) \;=\; \mathcal{O}(\Delta x^2)$$

puisque u est solution du problème exact. Pour la stabilité, on va avoir recours au lemme 1 ci-après. L'*erreur globale*

$$\mathcal{E}(\Delta x) \;:=\; \|(u(a),u(a+\Delta x),\dots,u(b-\Delta x),u(b)) \;-\; (u_0,u_1,\dots,u_N,u_{N+1})\|_{N,\Delta x}$$

vaut par définition

$$\mathcal{E}(\Delta x) \;=\; \left(\Delta x\,\sum_{j=1}^{N}\,|u(x_j)-u_j|^2\right)^{1/2}\,.$$

Or on a pour tout $j\in\{1,\dots,N\}$, en notant pour simplifier $v_j := u(x_j)-u_j$,

$$-\,\kappa\,(v_{j+1}-2v_j+v_{j-1}) \;=\; (\Delta x)^2\,\mathcal{R}(x_j,u,\Delta x)\,,$$

d'où en prenant le produit scalaire dans $\mathbb{R}^d$ par $v_j/\Delta x$ et en sommant sur j,

$$-\,\kappa\,\sum_{j=1}^{N}\,\frac{\langle v_j,v_{j+1}-v_j\rangle+\langle v_j,v_{j-1}-v_j\rangle}{\Delta x} \;=\; \Delta x\,\sum_{j=1}^{N}\langle v_j,\mathcal{R}(x_j,u,\Delta x)\rangle\,.$$

En faisant une translation d'indice et en utilisant que $v_0=0$ et $v_{N+1}=0$, on remarque que le membre de gauche s'écrit encore

$$\kappa\,\sum_{j=0}^{N}\,\frac{|v_{j+1}-v_j|^2}{\Delta x}\,,$$

et est donc minoré par

$$\frac{2\kappa}{(b-a)^2}\,\sum_{j=1}^{N}\Delta x\,|v_j|^2 \;=\; \frac{2\kappa}{(b-a)^2}\,\mathcal{E}(\Delta x)^2$$

d'après le lemme 1. Par ailleurs, l'inégalité de Cauchy-Schwarz permet de majorer le membre de droite :

$$\Delta x\,\sum_{j=1}^{N}\langle v_j,\mathcal{R}(x_j,u,\Delta x)\rangle \;\le\; \mathcal{E}(\Delta x)\,\left(\Delta x\,\sum_{j=1}^{N}\,|\mathcal{R}(x_j,u,\Delta x)|^2\right)^{1/2} \;\le\; C\,\sqrt{b-a}\,(\Delta x)^2\,\mathcal{E}(\Delta x)\,,$$

où la constante C provient de la majoration $|\mathcal{R}(x_j,u,\Delta x)| \;\le\; C\,(\Delta x)^2$, uniforme pour $x_j\in[a,b]$. Par suite,

$$\mathcal{E}(\Delta x) \;\le\; C\,\frac{(b-a)^{5/2}}{2\kappa}\,(\Delta x)^2\,.$$

Lemme 1 (« inégalité de Poincaré discrète ») *Si $\Delta x = (b-a)/(N+1)$, quels que soient $v_1, \dots, v_N \in \mathbb{R}^d$*

$$\sum_{j=1}^{N} \Delta x \, |v_j|^2 \;\leq\; \frac{(b-a)^2}{2} \sum_{j=0}^{N-1} \frac{|v_{j+1} - v_j|^2}{\Delta x} \,,$$

où l'on a posé $v_0 = 0$. De façon symétrique, en posant $v_{N+1} = 0$,

$$\sum_{j=1}^{N} \Delta x \, |v_j|^2 \;\leq\; \frac{(b-a)^2}{2} \sum_{j=1}^{N} \frac{|v_{j+1} - v_j|^2}{\Delta x} \,.$$

Dém. Puisque $v_0 = 0$ on a quel que soit $j \in \{1, \dots, N\}$,

$$v_j \;=\; \sum_{k=0}^{j-1} (v_{k+1} - v_k) \,,$$

d'où par l'inégalité triangulaire dans $\mathbb{R}^d$,

$$|v_j|^2 \;\leq\; \left(\sum_{k=0}^{j-1} |v_{k+1} - v_k| \right)^2 \,,$$

puis par l'inégalité de Cauchy-Schwarz dans $\mathbb{R}^j$,

$$|v_j|^2 \;\leq\; j \, \left(\sum_{k=0}^{j-1} |v_{k+1} - v_k|^2 \right) \,,$$

et donc

$$\sum_{j=1}^{N} \Delta x \, |v_j|^2 \;\leq\; (\Delta x) \left(\sum_{j=1}^{N} j \right) \left(\sum_{k=0}^{N-1} |v_{k+1} - v_k|^2 \right) \;=\; (\Delta x)^2 \, \frac{N(N+1)}{2} \sum_{k=0}^{N-1} \frac{|v_{k+1} - v_k|^2}{\Delta x}$$

$$\leq\; \frac{(b-a)^2}{2} \sum_{k=0}^{N-1} \frac{|v_{k+1} - v_k|^2}{\Delta x} \,.$$

La méthode des différences finies décrite précédemment s'étend assez facilement à d'autres problèmes, et pour commencer au problème avec des conditions aux limites dites « mixtes » :

$$\left\{ \begin{array}{l} -\kappa \, \partial^2_{xx} u \;=\; f \,, \quad x \in [a,b] \,, \\ u'(a) = \gamma \,, \quad u(b) = \beta \,. \end{array} \right. \tag{4}$$

Noter que si $f \in \mathscr{C}^2([a,b])$ et $\kappa > 0$ le problème (4) admet (comme (2)) une solution unique $u \in \mathscr{C}^2([a,b])$, qui s'exprime sous forme intégrale :

$$u(x) = \; \beta \; - \; (b-x) u'(a) \; + \; \int_x^b \int_a^y \; f(z) \, \mathrm{d}z \, \mathrm{d}y \,.$$

Dans l'approximation par différences finies, il faut cependant discrétiser avec soin la nouvelle condition $u'(a) = \gamma$. Considérons d'abord une simple discrétisation décentrée

$$\frac{u_1 - u_0}{\Delta x} \;=\; \gamma \,.$$

Le schéma utilisé précédemment aux nœuds x_j pour $j \in \{1, \ldots, N\}$ et la discrétisation de la condition au bord comme ci-dessus reviennent au système, si U désigne maintenant le vecteur de $\mathbb{R}^{N+1}$ de composantes $u_0, u_1, \ldots, u_N$,

$$A\,U \;=\; \frac{(\Delta x)^2}{\kappa}\,F\,, \quad A = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}, \quad F \;=\; \begin{pmatrix} -\frac{\kappa}{\Delta x}\,\gamma \\ f(x_2) \\ \vdots \\ f(x_{N-1}) \\ f(x_N) + \frac{\kappa}{(\Delta x)^2}\,\beta \end{pmatrix}.$$

Cette nouvelle matrice A (de taille $N+1$) est telle que

$$\langle U, AU \rangle \;=\; u_N^2 \;+\; \sum_{j=0}^{N-1} (u_{j+1} - u_j)^2 \;\geq\; 0$$

avec égalité si et seulement si $u_N \;=\; 0$ et $u_j \;=\; u_{j+1}$ quel que soit j, ce qui implique $U \;=\; 0$. On obtient donc sans problème une solution numérique. Cependant, la discrétisation décentrée de la condition au bord introduit une erreur de consistance trop importante : en reprenant la démonstration du théorème 1 on voit qu'il faut tenir compte de l'erreur locale de consistance supplémentaire

$$\widetilde{\mathcal{R}}(x_0, u, \Delta x) \;:=\; -\,\frac{\kappa}{\Delta x}\,\left(\,\frac{u(x_0 + \Delta x) - u(x_0)}{\Delta x} \;-\; \gamma\,\right),$$

car

$$\kappa \sum_{j=0}^{N} \frac{|v_{j+1} - v_j|^2}{\Delta x} \;=\; -\,\kappa \sum_{j=1}^{N} \frac{\langle v_j, v_{j+1} - v_j \rangle + \langle v_j, v_{j-1} - v_j \rangle}{\Delta x} \;-\; \frac{\kappa}{\Delta x}\,\langle v_0, v_1 - v_0 \rangle \;=$$

$$\Delta x \sum_{j=1}^{N} \langle v_j, \mathcal{R}(x_j, u, \Delta x) \rangle \;+\; \Delta x\,\langle v_0, \widetilde{\mathcal{R}}(x_0, u, \Delta x) \rangle\,.$$

Or si u est solution exacte de (4) (on rappelle que $x_0 = a$),

$$\widetilde{\mathcal{R}}(x_0, u, \Delta x) \;=\; -\,\frac{\kappa}{2}\,u''(a) \;+\; \mathcal{O}(\Delta x) \;=\; \frac{1}{2}\,f(a) \;+\; \mathcal{O}(\Delta x)\,,$$

ce qui ne tend donc pas vers 0 lorsque Δx tend vers 0 (sauf dans le cas particulier $f(a) = 0$). Une façon de remédier à ce problème est de définir une « valeur fictive » u_{-1} par la discrétisation centrée

$$\frac{u_1 - u_{-1}}{2\Delta x} \;=\; \gamma\,,$$

et d'utiliser cette valeur dans le schéma « intérieur »

$$-\,\kappa\,\frac{u_{j+1} - 2u_j + u_{j-1}}{(\Delta x)^2} \;=\; f(x_j)$$

en $j = 0$, ce qui donne

$$\frac{\kappa}{\Delta x}\,\frac{u_1 - u_0}{\Delta x} \;=\; \frac{1}{2}\,f(a) \;-\; \frac{\kappa}{\Delta x}\,\gamma\,.$$

Alors la nouvelle erreur locale de consistance

$$\widetilde{\mathcal{R}}(x_0, u, \Delta x) \;:=\; -\,\frac{\kappa}{\Delta x}\,\left(\,\frac{u(x_0 + \Delta x) - u(x_0)}{\Delta x} \;-\; \gamma\,\right) \;-\; \frac{1}{2}\,f(a)$$

est en $\mathcal{O}(\Delta x)$, ce qui est moins bon qu'à l'intérieur mais tend quand même vers zéro.

Éléments finis

L'élaboration d'une *méthode d'éléments finis* (terme souvent abrégé en FEM dans les références anglophones) repose sur des ingrédients

- d'analyse fonctionnelle, permettant de donner une *formulation variationnelle* du problème exact, que l'on ramène alors (formellement) à un problème en dimension finie par la *méthode de Galerkin* ;
- de géométrie et d'algèbre, consistant à construire un *maillage* du domaine physique, associé à des *fonctions de base* définissant l'espace d'approximation de Galerkin.

L'implémentation d'une méthode d'éléments finis nécessite par ailleurs un algorithme de résolution de grands systèmes linéaires dont on ne parlera pas ici. On abordera en revanche l'analyse de la convergence des méthodes d'éléments finis, sur un exemple simple.

Considérons le problème modèle suivant (problème de Dirichlet homogène) :

$$\left\{ \begin{array}{l} -\kappa\,\Delta u \,=\, f\,, \\ u_{|\partial\Omega} = 0\,, \end{array} \right. \tag{5}$$

où Ω est un ouvert borné de $\mathbb{R}^n$, de bord $\partial\Omega$ régulier (c'est-à-dire qu'au voisinage de chaque point, $\partial\Omega$ est l'image d'une fonction régulière $\mathbb{R}^{n-1} \to \mathbb{R}^n$). Pour définir une formulation variationnelle de ce problème, on se contentera d'un terme source $f \in L^2(\Omega)$. Notons $\mathscr{D}(\Omega)$ l'ensemble des fonctions $\mathscr{C}^\infty$ sur Ω, à support compact inclus dans Ω (ayant donc un prolongement continu nul sur $\partial\Omega$). Si $u \in \mathscr{D}(\Omega)$ est solution de (5), alors quel que soit $v \in \mathscr{D}(\Omega)$ on a $-\kappa \int_\Omega v\Delta u \,=\, \int_\Omega fv$, d'où par la formule de Green (ce qui revient à intégrer par parties dans chaque direction, les termes de bord étant nuls puisque v est nulle sur $\partial\Omega$)

$$\kappa \int_\Omega \nabla u \cdot \nabla v \,=\, \int_\Omega fv\,. \tag{6}$$

Rappelons que ∇u (gradient de u) désigne la fonction à valeurs dans $\mathbb{R}^n$ dont les composantes sont les dérivées partielles $\partial_i u$ de u par rapport aux coordonnées x_i. Ainsi $\nabla u \cdot \nabla v = \sum_{i=1}^n (\partial_i u)(\partial_i v)$. Il se trouve que l'équation (6) a un sens pour u et $v \in H^1(\Omega) := \{v \in L^2(\Omega)\,;\nabla v \in L^2(\Omega)\}$. Dans cette définition de l'espace $H^1(\Omega)$ il faut entendre ∇v comme le gradient au sens faible, défini par $\int_\Omega \phi\,\partial_i v \,=\, -\int_\Omega v\,\partial_i\phi$ pour tout $\phi \in \mathscr{D}(\Omega)$. L'espace vectoriel $H^1(\Omega)$ est un espace de Hilbert pour la norme définie par

$$\|v\|_{H^1}^2 \,=\, \|v\|_{L^2}^2 \,+\, \|\nabla v\|_{L^2}^2 \,=\, \int_\Omega |v|^2 \,+\, \sum_{i=1}^n \int_\Omega (\partial_i v)^2\,.$$

Cependant, si l'on se contente de chercher u dans $H^1(\Omega)$, on perd la condition au bord $u_{|\partial\Omega} = 0$. C'est pourquoi il est naturel de chercher u dans $H_0^1(\Omega)$, défini comme l'adhérence de $\mathscr{D}(\Omega)$ dans $H^1(\Omega)$ (cette définition apparemment compliquée est liée au fait que l'on ne peut pas « brutalement » imposer la condition $u_{|\partial\Omega} = 0$ à un élément u de $H^1(\Omega)$, car il n'a pas nécessairement de trace au bord, sauf en dimension $n = 1$). Un outil essentiel pour la suite est l'*inégalité de Poincaré* (qui se démontre en dimension 1 de façon tout à fait analogue au lemme 1, et que l'on admettra en dimension supérieure) : l'ouvert Ω étant borné, il existe $C > 0$ tel que pour tout $u \in H_0^1(\Omega)$,

$$\|u\|_{L^2(\Omega)} \leq C\|\nabla u\|_{L^2(\Omega)}\,.$$

Comme conséquence, on voit que

$$\|u\|_{H_0^1(\Omega)} := \|\nabla u\|_{L^2(\Omega)}$$

définit une norme sur $H^1_0(\Omega)$, équivalente à celle de $H^1(\Omega)$. En effet, pour tout $u \in H^1_0(\Omega)$,

$$\frac{1}{\sqrt{1+C^2}} \|u\|_{H^1(\Omega)} \leq \|\nabla u\|_{L^2(\Omega)} \leq \|u\|_{H^1(\Omega)} .$$

Ce cadre fonctionnel étant posé, $a(u,v) := \kappa \int_\Omega \nabla u \cdot \nabla v$ définit une forme bilinéaire (symétrique) continue sur $H^1_0(\Omega)$, car d'après l'inégalité de Cauchy-Schwarz,

$$|a(u,v)| \leq \kappa \|\nabla u\|_{L^2} \|\nabla v\|_{L^2} = \kappa \|u\|_{H^1_0(\Omega)} \|v\|_{H^1_0(\Omega)} ,$$

De plus, comme $a(u,u) = \kappa \|u\|^2_{H^1_0(\Omega)}$, si $\kappa > 0$ on dit que a est *coercive*. Ces propriétés vont permettre d'appliquer le résultat fondamental suivant.

Théorème 2 *Soit a une forme bilinéaire continue et coercive sur un espace de Hilbert V. Si ℓ est une forme linéaire continue sur V, il existe un unique $u \in V$ tel que*

$$a(u,v) = \ell(v) \quad \textit{quel que soit } v \in V . \tag{7}$$

Dém. Le théorème de représentation de Riesz-Fréchet affirme qu'il existe $f \in V$ tel que $\ell(v) = \langle f, v\rangle$ quel que soit $v \in V$, où $\langle \cdot, \cdot \rangle$ est le produit scalaire dans V. (Si $\ell \equiv 0$ alors $f = 0$. Sinon, on prend $f = \ell(g)g$, où g est un vecteur unitaire orthogonal à l'hyperplan $\ell^\perp := \{w \in V\,;\ \ell(w) = 0\}$, lui-même obtenu comme $(g_0 - g_1)/\|g_0 - g_1\|$ avec $g_0 \notin \ell^\perp$ et g_1 l'image de g_0 par la projection orthogonale sur $\ell^\perp$.) Pour la même raison, quel que soit $u \in V$, l'application $v \mapsto a(u,v)$ étant une forme linéaire continue, il existe un vecteur Au tel que $a(u,v) = \langle Au, v\rangle$ quel que soit $v \in V$. De plus, l'application $A : u \mapsto Au$ est linéaire continue, sa norme étant

$$|||A||| := \sup_{u \neq 0} \frac{\|Au\|}{\|u\|} \leq |||a||| := \sup_{u,v \neq 0} \frac{|a(u,v)|}{\|u\|\|v\|} .$$

Par suite, le problème (7) est équivalent au problème $Au = f$. Montrer que ce dernier a une solution unique revient à montrer que A est un isomorphisme de V. Or, d'après la coercivité de a, il existe $\kappa > 0$ tel que $\langle Au, u\rangle = a(u,u) \geq \kappa \|u\|^2$ quel que soit $u \in V$. Ceci implique que A est injective (si $Au = 0$ alors $u = 0$). Par ailleurs, cette inégalité et celle de Cauchy-Schwarz impliquent que $\|Au\| \geq \kappa \|u\|$ quel que soit $u \in V$. On en déduit que l'image de A est fermée (si $(Au_p)_{p\in\mathbb{N}}$ est une suite de Cauchy alors $(u_p)_{p\in\mathbb{N}}$ aussi). On peut ensuite conclure que cette image est égale à V tout entier : sinon il existerait $w_0 \in V \backslash \mathrm{Im}A$, que l'on pourrait projeter orthogonalement en $w_1 \neq w_0$ sur le sous-espace vectoriel fermé $\mathrm{Im}A$; on aurait alors $\langle Au, w_0 - w_1\rangle = 0$ quel que soit $u \in V$, et en particulier $\langle A(w_0 - w_1), w_0 - w_1\rangle = 0$, ce qui contredirait (puisque $w_0 \neq w_1$) l'inégalité $\langle A(w_0 - w_1), w_0 - w_1\rangle \geq \kappa \|w_0 - w_1\|^2$.

L'idée de la méthode de *Galerkin* consiste alors à résoudre le problème approché

$$a(u_h, v_h) = \ell(v_h) \quad \text{quel que soit } v_h \in V_h , \tag{8}$$

dans V_h, où $(V_h)_{h>0}$ est une famille de sous-espaces fermés (et en pratique de dimension finie) de V, supposés « tendre » vers V lorsque le paramètre h tend vers 0. D'après le théorème de Lax-Milgram appliqué dans V_h, on sait que le problème (8) a une solution unique. Si de plus V_h est de dimension finie N_h et engendré par une famille $(\varphi_1, \ldots, \varphi_{N_h})$, alors le problème (8) se réduit (en cherchant $u_h = \sum_{j=1}^{N_h} X_j \phi_j$) à la résolution du système linéaire $M_h X_h = Y_h$, où

$$M_h := (a(\varphi_j, \varphi_i))_{1 \leq i,j \leq N_h} , \quad Y_h := (\ell(\varphi_1), \ldots, \ell(\varphi_{N_h}))^{\mathrm{t}} .$$

Puisqu'il y a une solution unique, la matrice M_h est inversible. Comme annoncé plus haut, on ne discutera pas la façon de résoudre le système $M_h X_h = Y_h$.

Nous allons maintenant nous concentrer sur deux questions essentielles :

i). comment définir V_h : ce sera l'occasion, après avoir vu un exemple simple en dimension $n = 1$, de définir les *éléments finis de Lagrange* de manière formelle puis sur deux grandes classe déléments finis (dits P_k et Q_k) ;

ii). comment montrer que la famille de solutions « approchées » u_h tend vers la solution exacte.

Grâce au lemme suivant, la seconde question se ramène en fait à l'estimation de $\|u - \Pi_h u\|$, où Π_h est un projecteur (par exemple orthogonal) de V sur V_h. (Rappelons qu'un projecteur Π_h est par définition un opérateur linéaire idempotent, c'est-à-dire tel que $\Pi_h \circ \Pi_h = \Pi_h$. Il est orthogonal si $\langle \Pi_h u, u - \Pi_h u \rangle = 0$ pour tout $u \in V$.)

Lemme 2 (Céa) *Soit a une forme bilinéaire continue et coercive sur un espace de Hilbert V. Il existe $C > 0$ tel que, pour toute forme linéaire continue ℓ, pour tout sous-espace vectoriel V_h de V, si u est la solution de* (7) *et u_h la solution de* (8)*,*

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

Dém. Par définition de u_h et u, $a(u_h, v_h) = \ell(v_h)$ quel soit $v_h \in V_h$, et comme $V_h \subset V$, $a(u, v_h) = \ell(v_h)$. En faisant la différence on en déduit par bilinéarité de a que $a(u - u_h, v_h) = 0$ quel soit $v_h \in V_h$. Par suite, puisque V_h est un sous-espace vectoriel ($u_h - v_h \in V_h$), en utilisant à nouveau la bilinéarité de a on peut écrire $a(u - u_h, u - u_h) = a(u - u_h, u - v_h)$. D'où

$$\kappa \|u - u_h\|^2 \leq a(u - u_h, u - u_h) = a(u - u_h, u - v_h) \leq |||a||| \, \|u - u_h\| \, \|u - v_h\| ,$$

et par conséquent

$$\|u - u_h\| \leq \frac{|||a|||}{\kappa} \inf_{v_h \in V_h} \|u - v_h\| .$$

Voyons maintenant un exemple d'espace d'approximation $V_h \subset H_0^1(]a, b[)$ (adapté au problème modèle (6) en dimension $n = 1$). On définit des nœuds $x_j \in [a, b]$ avec $x_0 = a$, $x_{N_h+1} = b$ et $0 < x_{j+1} - x_j \leq h$ pour $j \in \{0, \ldots, N_h\}$, et $V_h = \mathrm{Vect}(\varphi_1, \ldots, \varphi_{N_h})$ avec

$$\varphi_j(x) = \begin{cases} \dfrac{x - x_{j-1}}{x_j - x_{j-1}} & \text{si } x \in [x_{j-1}, x_j] , \\ \dfrac{x - x_{j+1}}{x_j - x_{j+1}} & \text{si } x \in [x_j, x_{j+1}] , \\ 0 & \text{sinon .} \end{cases}$$

Ces fonctions affines par morceaux (leur graphe étant en forme de « chapeau chinois ») sont continues, s'annulent en $x_0 = a$ et $x_{N_h+1} = b$, et l'on vérifie aisément que ce sont des éléments de $H_0^1(]a, b[)$. On remarque de plus que $\varphi_j(x_i) \neq 0$ si et seulement si $i = j$, et $\varphi_j(x_j) = 1$. Ceci permet de définir le projecteur Π_h par

$$\Pi_h u = \sum_{j=1}^{n} u(x_j) \, \varphi_j .$$

Proposition 1 *Si $u \in \mathscr{C}^2([a, b])$, $u(a) = u(b) = 0$, et $\Pi_h u$ est défini comme ci-dessus, alors $\|u - \Pi_h u\|_{H_0^1(]a,b[)}$ tend vers 0 lorsque h tend vers 0.*

Dém. On a par définition de la norme sur H_0^1,

$$\|u - \Pi_h u\|^2_{H_0^1(]a,b[)} \;=\; \int_a^b |(u - \Pi_h u)'(x)|^2 \,\mathrm{d}x \;= \sum_{j=0}^{N_h} \int_{x_j}^{x_{j+1}} |(u - \Pi_h u)'(x)|^2 \,\mathrm{d}x \,.$$

Or par définition de Π_h, pour $x \in [x_j, x_{j+1}]$,

$$\Pi_h u(x) = u(x_j)\frac{x - x_{j+1}}{x_j - x_{j+1}} + u(x_{j+1})\frac{x - x_j}{x_{j+1} - x_j} \,,$$

donc

$$(\Pi_h u)'(x) = \frac{u(x_{j+1}) - u(x_j)}{x_{j+1} - x_j} \;=\; \int_0^1 u'(x_j + \theta(x_{j+1} - x_j)) \,\mathrm{d}\theta$$

(d'après la formule de Taylor avec reste intégral à l'ordre 1, ce qui évite d'avoir recours à la formule des accroissements finies, limitée aux fonctions à valeurs réelles). Ainsi pour $x \in [x_j, x_{j+1}]$,

$$(u - \Pi_h u)'(x) \;=\; \int_0^1 (u'(x) - u'(x_j + \theta(x_{j+1} - x_j)) \,\mathrm{d}\theta \;=\; \int_0^1 \int_{x_j + \theta(x_{j+1} - x_j)}^{x} u''(y) \,\mathrm{d}y \,\mathrm{d}\theta \,,$$

d'où

$$|(u - \Pi_h u)'(x)| \;\leq\; h \max_{y \in [x_j, x_{j+1}]} |u''(y)| \,.$$

On en déduit

$$\int_a^b |(u - \Pi_h u)'(x)|^2 \,\mathrm{d}x \;\leq\; (b - a)\, h^2 \max_{y \in [a,b]} |u''(y)|^2 \,,$$

ce qui tend bien vers 0 lorsque h tend vers 0.

Lorsque $\Omega \subset \mathbb{R}^n$ avec $n \geq 2$, il y a diverses façons de le découper géométriquement en « mailles », et donc de définir V_h. Bien que l'on ait supposé au départ le bord de Ω régulier, on va ici supposer que Ω est polyhédral, de sorte que l'on puisse le découper en mailles elles aussi polyhédrales. Dans le cas $n = 2$ on considèrera par exemple des mailles triangulaires ou rectangulaires. Avant cela, donnons une définition générale.

Définition 1 *On appelle* élément fini de Lagrange *dans $\mathbb{R}^n$ la donnée d'un compact $K \subset \mathbb{R}^n$, d'un ensemble $\Sigma = \{a_1, \ldots, a_N\}$ de points de K et d'un espace P de fonctions $K \to \mathbb{R}$ telles que pour tout $(\alpha_1, \ldots, \alpha_N) \in \mathbb{R}^N$, il existe une unique fonction $p \in P$ telle que $p(a_j) = \alpha_j$ pour tout $j \in \{1, \ldots, N\}$. On appelle* fonctions de base *les éléments $p_1, \ldots, p_N$ (formant une base) de P tels que $p_j(a_i) = 0$ si $i \neq j$ et $p_i(x_i) = 1$.*

Exemples en dimension $n = 1$

- Les ensembles $K = [a, b]$, $\Sigma = \{x_1, \ldots, x_N\}$ et $P = \text{Vect}\,(\varphi_1, \ldots, \varphi_N)$ où les fonctions φ_j sont celles définies précédemment, définissent un élément fini de Lagrange.
- Les ensembles $K = [a, b]$, $\Sigma = \{x_1, \ldots, x_N\}$ et

  $$P := \{\text{fonctions polynomiales de degré au plus } N - 1\}$$

  définissent un élément fini de Lagrange, de fonctions de base les *polynômes d'interpolation de Lagrange* :

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

En dépit des apparences, le premier exemple est un cas particulier du second, appliqué pour $N = 1$ dans chaque maille $[x_j, x_{j+1}]$. C'est en fait l'idée pour construire un espace d'approximation V_h en toute dimension : définir un élément fini sur une maille K de référence, et obtenir un élément fini sur toutes les autres mailles par transformation affine.

En dimension n arbitraire Comme on l'a déjà dit, le maillage peut revêtir diverses formes. Par ailleurs, il n'y a plus de notion « absolue » de degré pour les fonctions polynomiales : on peut notamment parler de *degré total*, ou de *degré partiel*.

Lorsque la maille de référence est le cube unité $[0,1]^n$, il est (sur le papier) assez facile de définir un élément fini. Pour $k \in \mathbb{N}^*$ on considère

$$Q_k := \{\text{fonctions polynomiales de degré partiel au plus } k \text{ dans chaque variable}\}\,.$$

Sa dimension $N := (k+1)^n$ est exactement le cardinal de l'ensemble

$$\Sigma_k := \{a \in [0,1]^n\,;\ a_j \in \{0, 1/k, 2/k, \dots, 1\} \text{ pour tout } j \in \{1,\dots,n\}\ \}\,.$$

(Attention, ici a_j désigne la j-ième composante du point a, et non un point de $\mathbb{R}^n$.) On montre que $([0,1]^n, \Sigma_k, Q_k)$ est un élément fini de Lagrange. Par exemple pour $n = 2$, les éléments de Σ_1 sont les quatre sommets du carré $[0,1]^2$, et les fonctions de base sont

$$p_1(x) = (1-x_1)(1-x_2)\,,\ p_2(x) = x_1(1-x_2)\,,\ p_3(x) = x_1x_2\,,\ p_4(x) = x_2(1-x_1)\,.$$

Considérons maintenant comme maille de référence un *simplexe* de $\mathbb{R}^n$, c'est-à-dire l'enveloppe convexe de $(n+1)$ points $b_1, \dots, b_{n+1} \in \mathbb{R}^n$:

$$S = \left\{ x = \sum_{j=1}^{n+1} \lambda_j b_j\,;\ \lambda_j \in [0,1] \text{ et } \sum_{j=1}^{n+1} \lambda_j = 1 \right\}\,.$$

Pour $x \in S$, on appelle *coordonnées barycentriques* les nombres $\lambda_j = \lambda_j(x) \in [0,1]$ de somme égale à 1 tels que $x = \sum_{j=1}^{n+1} \lambda_j b_j$. Les fonctions $x \mapsto \lambda_j(x)$ sont affines, car on peut voir par exemple $\lambda_1(x), \dots, \lambda_n(x)$ comme les coordonnées de $g\vec{x}$ dans la base $g\vec{b}_1, \dots, g\vec{b}_n$, où g est le *centre de gravité* de S, de coordonnées barycentriques $1/(n+1)$, et $\lambda_{n+1}(x) = 1 - \sum_{j=1}^{n} \lambda_j(x)$. On montre que

$$P_k := \{\text{fonctions polynomiales de degré total au plus } k\}$$

est de dimension $(n+k)!/(n!k!)$, c'est-à-dire exactement le cardinal de

$$\Lambda_k := \{x \in S\,;\ \lambda_j(x) \in \{0, 1/k, 2/k, \dots, 1\} \text{ pour tout } j \in \{1,\dots,n+1\}\}\ .$$

et que (S, Λ_k, P_k) est un élément fini de Lagrange. Par exemple pour $n = 2$, les éléments de Λ_1 sont les sommets du triangle et les fonctions de base sont les coordonnées barycentriques λ_1, λ_2 et λ_3.