Séance n°8
Approximation des équations paraboliques

2 Novembre 2004

Exercice 1. Analyse de stabilité L^2 du θ-schéma (1)

On s'intéresse à la résolution du problème d'évolution :

$$(1) \qquad \begin{cases} \dfrac{\partial u}{\partial t} - \dfrac{\partial^2 u}{\partial x^2} = 0, \quad x \in \mathbb{R}, \quad t > 0, \\ \\ u(x,0) = u_0(x) \quad x \in \mathbb{R}, \quad (u_0(x) \in L^2(\mathbb{R})). \end{cases}$$

On considère le schéma aux différences finies qui s'écrit, avec les notations habituelles :

$$(2) \qquad \frac{u_j^{n+1} - u_j^n}{\Delta t} - \theta \, \frac{u_{j+1}^{n+1} - 2u_j^{n+1} + u_{j-1}^{n+1}}{h^2} - (1-\theta) \, \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0$$

θ désignant un paramètre réel entre 0 et 1.

1) Discuter l'ordre et le caractère explicite du schéma en fonction de θ.

De façon évidente, le schéma est d'ordre 1 en temps et d'ordre 2 en espce sauf pour $\theta = 1/2$ pour lequel il est d'ordre 2 en espace et en temps. Par ailleurs, le schéma est implicite sauf pour $\theta = 0$.

2) Etudier par Fourier la stabilité L^2 du schéma suivant les valeurs de θ.

On cherche les soluions de la forme :

$$u_j^n = \hat{u}^n \, e^{ikx_j}, \quad k \in \mathbb{R}.$$

On obtient alors aisément :

$$\hat{u}^{n+1} = G_h(k, \Delta t) \, \hat{u}^n,$$

où le coefficient d'amplification $G_h(k,\Delta t)$ est réel et donné par :

$$G_h(k,\Delta t) = \frac{1-4(1-\theta)\beta\sin^2(\frac{kh}{2})}{1+4\theta\beta\sin^2(\frac{kh}{2})},$$

où on a posé $\beta = \Delta t/h^2$. L'inégalité $-1 \leq G_h(k,\Delta t) \leq 1$ se réécrit :

$$-1-4\theta\beta\sin^2(\frac{kh}{2}) \leq 1-4(1-\theta)\beta\sin^2(\frac{kh}{2}) \leq 1+4\theta\beta\sin^2(\frac{kh}{2}).$$

L'inégalité de droite est toujours satisfaite pour $\theta > 0$. L'autre se réécrit :

$$(1-2\theta)\beta\sin^2(\frac{kh}{2}) \leq \frac{1}{2}. \tag{3}$$

Il faut alors distinguer deux cas :

- $\theta \geq 1/2$. *Dans ce cas, le membre de gauche de (3) étant toujours négatif, (3) est donc toujours satisfaite et le schéma est inconditionnellement stable.*

- $\theta < 1/2$. *Dans ce cas, le membre de gauche de (3) est positif et atteint son maximum (par rapport à k) pour $\sin^2(\frac{kh}{2}) = 1$. (3) est donc équivalent à*

  $$(1-2\theta)\beta \leq \frac{1}{2},$$

  c'est à dire la condition de stabilité

  $$\frac{\Delta t}{h^2} \leq \frac{1}{2(1-2\theta)}.$$

Exercice 2. Analyse de stabilité L^2 du θ-schéma (2)

Dans ce qui suit, Ω désigne un ouvert polygonal de $\mathbb{R}^2$, $a(x)$ et $q(x)$ deux fonctions définies sur Ω à valeurs réelles positives, bornées avec en outre

$$0 < a_- \leq a(x), \quad \text{p. p. } x \in \Omega.$$

On s'intéresse à la résolution du problème d'évolution

$$\left\{ \begin{array}{lll} \dfrac{\partial u}{\partial t} - \operatorname{div}(a(x)\nabla u) + q(x)\, u = 0, & x \in \Omega, & t > 0, \\ u(x,0) = u_0(x), & x \in \Omega, & (u_0(x) \in L^2(\Omega)), \\ \dfrac{\partial u}{\partial n} = 0, & \text{sur } \Gamma = \partial\Omega. & \end{array} \right. \tag{4}$$

1) Ecrire une formulation variationnelle en espace du problème sous la forme :

(5)
$$\left\{\begin{array}{l} \text{Trouver } u(t) : \mathbb{R}^+ \to V \text{ tel que } : \\ \\ \dfrac{d}{dt}(u(t), v) + a(u(t), v) = 0 \end{array}\right.$$

où on précisera l'espace de Hilbert V, le produit scalaire $(\cdot,\cdot)$ (celui qui intervient dans (5) et non pas celui de V) et la forme bilinéaire $a(\cdot,\cdot)$.

If faut travailler avec $V = H^1(\Omega)$. Par ailleurs :

$$\left\{\begin{array}{l} a(u,v) = \displaystyle\int_\Omega (a\nabla u \cdot \nabla v + q\ u\ v)\ dx \\ (u,v) = \displaystyle\int_\Omega u\ v\ dx, \quad \textit{(le produit scalaire } L^2\textit{)}. \end{array}\right.$$

2) On approche (4) par éléments finis en espace (V_h désigne ci-dessous une famille de sous-espaces de dimension finie de V) et par un θ-schéma en temps :

(6)
$$\left\{\begin{array}{l} \text{Trouver } u_h^n \in V_h, n \geq 1 \text{ tel que } : \\ \\ (\dfrac{u_h^{n+1} - u_h^n}{\Delta t}, v_h) + a(\theta\ u_h^{n+1} + (1-\theta)\ u_h^n, v_h) = 0, \quad \forall v_h \in V_h. \end{array}\right.$$

Démontrer l'existence et l'unicité de la suite u_h^n, étant donné u_h^0, une approximation “convenable” de u^0 dans V_h.

Par récurrence, il suffit de montrer que l'on peut calculer u_h^{n+1} à partir de u_h^n. Or, lorsque u_h^n est connu, le problème à résoudre pour calculer u_h^{n+1} s'ecrit :

(7)
$$\left\{\begin{array}{l} \textit{Trouver } u_h^n \in V_h, \textit{ tel que } : \\ \\ (u_h^{n+1}, v_h) + \theta\ \Delta t\ a(u_h^{n+1}, v_h) = (u_h^n, v_h) - (1-\theta)\ \Delta t\ a(u_h^n, v_h). \end{array}\right.$$

Pour conclure, soit on invoque Lax-Milgram (dans V_h) soit on remarque que ce problème équivaut à résoudre un système linéaire symétrique défini positif. C'est à ce niveau qu'interviennent les hypothèses sur les fonctions $a(x)$ et $q(x)$.

3) On suppose que $\theta \geq 1/2$. Montrer que, $\|\cdot\|$ désignant la norme associée à $(\cdot,\cdot)$:

(8)
$$\|u_h^{n+1}\| \leq \|u_h^n\|.$$

Indication : on prendra $v_h = \theta\ u_h^{n+1} + (1-\theta)u_h^n$.

En déduire que le schéma est inconditionnellement L^2 stable.

Remarquons que :

$$v_h = \theta\ u_h^{n+1} + (1-\theta)\ u_h^n = \frac{u_h^{n+1}+u_h^n}{2} + (\theta - \frac{1}{2})(u_h^{n+1} - u_h^n).$$

Par conséquent, pour ce v_h, il vient :

$$(\frac{u_h^{n+1}-u_h^n}{\Delta t}, v_h) = \frac{1}{2\Delta t}\left(\|u_h^{n+1}\|^2 - \|u_h^n\|^2\right) + (\theta - \frac{1}{2})\ \Delta t\ \|\frac{u_h^{n+1}-u_h^n}{\Delta t}\|^2.$$

Par ailleurs, si on pose $u_h^{n+\theta} = \theta\ u_h^{n+1} + (1-\theta)u_h^n$, il vient :

$$a(\theta\ u_h^{n+1} + (1-\theta)u_h^n, v_h) = a(u_h^{n+\theta}, u_h^{n+\theta}).$$

Nous obtenons donc l'identité :

$$\frac{1}{2\Delta t}\left(\|u_h^{n+1}\|^2 - \|u_h^n\|^2\right) + (\theta - \frac{1}{2})\ \Delta t\ \|\frac{u_h^{n+1}-u_h^n}{\Delta t}\|^2 + a(u_h^{n+\theta}, u_h^{n+\theta}) = 0. \tag{9}$$

Pour $\theta \geq 1/2$, comme $a(\cdot,\cdot)$ est une forme bilinéaire positive, on en déduit $\|u_h^{n+1}\| \leq \|u_h^n\|$, soit de proche en proche :

$$\|u_h^n\| \leq \|u_h^0\|,$$

ce qui correspond bien à un résultat de stabilité inconditionnelle dans L^2.

4) On suppose que $\theta < 1/2$ et on introduit :

$$\|a_h\| = \sup_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|^2}.$$

Montrer que l'inégalité (8) reste vraie si

$$\Delta t\ \|a_h\| \leq \frac{1}{1-2\theta}\ .$$

Conclure.

D'après (9), le résultat reste vrai si on arrive à démontrer que

$$a(u_h^{n+\theta}, u_h^{n+\theta}) - (\frac{1}{2} - \theta)\ \Delta t\ \|\frac{u_h^{n+1}-u_h^n}{\Delta t}\|^2 \geq 0.$$

Or, en choisissant maintenant $v_h = \frac{u_h^{n+1}-u_h^n}{\Delta t}$, nous obtenons :

$$\|\frac{u_h^{n+1}-u_h^n}{\Delta t}\|^2 = -\ a(u_h^{n+\theta}, \frac{u_h^{n+1}-u_h^n}{\Delta t}).$$

En utilisant l'inégalité de Cauchy-Schwartz (valable pour $a(\cdot,\cdot)$ est une forme bilinéaire symétrique et positive) puis la définition de $\|a_h\|$, il vient :

$$\begin{array}{|lcl} \|\dfrac{u_h^{n+1}-u_h^n}{\Delta t}\|^2 & \leq & a(u_h^{n+\theta},u_h^{n+\theta})^{\frac{1}{2}} \;\; a(\dfrac{u_h^{n+1}-u_h^n}{\Delta t},\dfrac{u_h^{n+1}-u_h^n}{\Delta t})^{\frac{1}{2}} \\ & \leq & \|a_h\|^{\frac{1}{2}} \; \|\dfrac{u_h^{n+1}-u_h^n}{\Delta t}\| \;\; a(u_h^{n+\theta},u_h^{n+\theta})^{\frac{1}{2}} \end{array}$$

Par conséquent :

$$\|\frac{u_h^{n+1}-u_h^n}{\Delta t}\|^2 \leq \|a_h\| \; a(u_h^{n+\theta},u_h^{n+\theta}).$$

Nous en déduisons

$$a(u_h^{n+\theta},u_h^{n+\theta}) - (\frac{1}{2}-\theta)\;\Delta t\;\|\frac{u_h^{n+1}-u_h^n}{\Delta t}\|^2 \geq \Big(1-(\frac{1}{2}-\theta)\;\|a_h\|\Big)\; a(u_h^{n+\theta},u_h^{n+\theta})$$

qui est bien positif si on a la condition de stabilité :

$$\big(\frac{1}{2}-\theta\big)\;\|a_h\| \leq 1$$

qui est bien l'inégalité annoncée.

5) Facultatif. Montrer que le calcul de $\|a_h\|$ équivaut à la recherche de la plus grande valeur propre d'un problème de valeur propres généralisé. Effectuer le calcul de $\|a_h\|$ lorsque :

- L'ouvert Ω est un carré et les fonctions $a(x)$ et $q(x)$ sont constantes.
- Le carré est discrétisé à l'aide d'un maillage carré régulier de coté h.
- On utilise les éléments finis P_1 sur le maillage obtenu en découpant chaque petit carré en deux triangles rectangles.

Si A_h et M_h désignent respectivement les matrices de masse et de rigidité, $\|a_h\|$ n'est autre que la plus grande valeur propre λ du problème de valeurs propres :

$$A_h\,U_h = \lambda\,M_h\,U_h.$$

Le reste est essentiellement calculatoire. On commence par réécrire l´equation aux valeurs propres précédent sous la forme d'un schéma aux différences finies d'inconnues $u_{ij}, 0 \leq i,j \leq N$ si le carré a été divisé en N^2 petits carrés. On cherche les solutions à variables séparées sous la forme :

$$u_{ij} = f_i\;g_j.$$

On obtient pour chacune des suites f_i et g_j une relation de récurrence (la même) à 3 niveaux et à coefficients constants. Je n'ai pas le courage de faire les calculs. C'est long mais pas trop difficile.

Exercice 3. Analyse L^∞ du schéma explicite

Pour approcher la solution de (1), on considère le schéma numérique suivant :

$$\frac{u_j^{n+1}-u_j^n}{\Delta t}-\frac{u_{j+1}^n-2u_j^n+u_{j-1}^n}{h^2}=0 \tag{10}$$

complété par la condition initiale :

$$u_j^0=u_0(x_j).$$

1) Donner l'algorithme de calcul de la solution (on introduira $\beta=\dfrac{\Delta t}{h^2}$)). Expliquer pourquoi la solution numérique u_j^n se propage "à vitesse finie".

En résolvant par rapport à u_j^{n+1}, on obtient :

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

Il est clair que en un pas de temps, le support de la solution numérique ne peut progresser que d'une maille vers la droite où vers la gauche. en d'autres termes la vitesse de propagation numérique est bornée par $h/\Delta t$.

2) Dans la suite on adoptera la notation :

$$u_h=(u_j),\qquad \|u_h\|_\infty=\sup_j|u_j|$$

Si $\beta=\dfrac{\Delta t}{h^2}\leq\dfrac{1}{2}$, démontrer le principe du maximum discret :

$$a\leq u_j\leq b\Rightarrow a\leq u_j^n\leq b.$$

En particulier $\|u_h^n\|_\infty\leq\|u_h^0\|_\infty$ (on dit que le schéma est donc L^∞ stable).

La formule définissant u_j^{n+1} peut se réécrire sous la forme :

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

Autrement dit, si $\beta\leq\frac{1}{2}$, u_j^{n+1} apparaît comme une combinaison linéaire convexe de u_{j+1}^n, u_j^n et u_{j-1}^n. On en déduit le résultat demandé de façon immédiate.

Remarque ; en utilisant la convexité de la fonction x^p pour $p\geq 1$, on obtient la décroissance de toutes les normes l^p de la solution.

3) On définit l'erreur de troncature

$$\varepsilon_j^{n+\frac{1}{2}}=\frac{U_j^{n+1}-U_j^n}{\Delta t}-\frac{U_{j+1}^n-2U_j^n+U_{j-1}^n}{h^2}$$

où par définition $U_j^n = u(x_j, t^n)$. On suppose que u est de classe C^2 en temps et C^4 en espace et que ses dérivées sont uniformément bornées sur $\mathbb{R} \times [0, T]$. A l'aide d'un développement de Taylor, montrer que, pour tout entier n tel que $(n+1)\ \Delta t\ \leq\ T$, on a l'estimation :

$$\|\varepsilon_h^{n+\frac{1}{2}}\|_\infty \leq C\ (\ \Delta t\ \|\frac{\partial^2 u}{\partial t^2}\|_{L^\infty} + h^2\ \|\frac{\partial^4 u}{\partial x^4}\|_{L^\infty})$$

où par définition :

$$\|f\|_{L^\infty} = \sup_{(x,t)\,\in\,\mathbb{R}\times[0,T]} |f(x,t)|.$$

Il suffit de remarquer que, par Taylor :

$$\left|\begin{array}{ll} \dfrac{U_j^{n+1} - U_j^n}{\Delta t} = \dfrac{\partial u}{\partial t}(x_j, t^n) + \dfrac{\Delta t}{2}\dfrac{\partial^2 u}{\partial t^2}(x_j, t^*), & t^* \in [t^n, t^{n+1}] \\ \dfrac{U_{j+1}^n - 2U_j^n + U_{j-1}^n}{h^2} = \dfrac{\partial^2 u}{\partial x^2}(x_j, t^n) + \dfrac{h^2}{12}\dfrac{\partial^4 u}{\partial x^4}(x_*, t^n), & x_* \in [x_{j-1}, x_{j-1}]. \end{array}\right.$$

4) On introduit l'erreur sur la solution :

$$e_j^n = U_j^n - u_j^n \quad \text{(erreur commise).}$$

Ecrire les équations qui relient e_h^n et $\varepsilon_h^{n+\frac{1}{2}}$. Démontrer l'estimation d'erreur

$$\sup_{(n+1)\Delta t \leq T} \|e_h^n\|_\infty \leq C\ T\ (\ \Delta t\ \|\frac{\partial^2 u}{\partial t^2}\|_{L^\infty} + h^2\|\frac{\partial^4 u}{\partial t^4}\|_{L^\infty}).$$

On obtient de façon évidente :

$$\frac{e_j^{n+1} - e_j^n}{\Delta t} - \frac{e_{j+1}^n - 2e_j^n + e_{j-1}^n}{h^2} = \varepsilon_j^{n+\frac{1}{2}}, \tag{11}$$

avec en outre $e_j^0 = 0$ *pour tout* j. *Nous avons donc :*

$$e_j^{n+1} = \beta\ e_{j+1}^n + (1-2\beta)\ e_j^n + \beta\ e_{j-1}^n + \Delta t\ \varepsilon_j^{n+\frac{1}{2}}$$

En passant aux valeurs absolues et au supremum sur j, *il vient :*

$$\|e_h^{n+1}\|_\infty \leq \|e_h^n\|_\infty + \Delta t\ \|\varepsilon_h^{n+\frac{1}{2}}\|_\infty$$

De proche en proche :

$$\|e_h^n\|_\infty \leq \Delta t\ \sum_{k=0}^{n-1} \|\varepsilon_h^{n+\frac{1}{2}}\|_\infty$$

c'est à dire, compte tenu de l'estimation obtenue pour $\|\varepsilon_h^{n+\frac{1}{2}}\|_\infty$ *:*

$$\|e_h^n\|_\infty \leq C\ t^n\ (\ \Delta t\ \|\frac{\partial^2 u}{\partial t^2}\|_{L^\infty} + h^2\|\frac{\partial^4 u}{\partial t^4}\|_{L^\infty}),$$

ce qui permet de conclure.

Exercice 4. Schéma d'Euler implicite et principe du maximum

1) Question préliminaire. On dit qu'une matrice carrée M (d'ordre N) est une M-matrice si et seulement si :

$$
\left\{
\begin{array}{l}
M_{ii} > 0, \quad 1 \le i \le N, \quad M_{ij} \le 0, \quad \text{pour } i \ne j. \\
\\
\inf\limits_{1 \le i \le N} \left[M_{ii} - \sum\limits_{j \ne i} |M_{ij}| \right] > 0.
\end{array}
\right. \tag{12}
$$

Montrer qu'une telle matrice est nécessairement inversible et que, si b est un vecteur dont toutes les composantes sont positives, la solution x du systèmr $Mx = b$ a toutes ses composantes positives (Indication : regarder là où x_i est minimum).

Soit $x \ne 0$ dans le noyau de M et x_i tel que $|x_i| = max\, |x_j|$. L'équation $n^o i$ de $Mx = 0$ s'écrit :

$$x_i = -\frac{1}{M_{ii}} \sum_{j \ne i} M_{ij} x_j$$

Par conséquent

$$|x_i| \le \frac{1}{M_{ii}} \sum_{j \ne i} |M_{ij}|\, |x_j| \le \frac{1}{M_{ii}} \sum_{j \ne i} |M_{ij}|\, |x_i| < |x_i|,$$

ce qui est impossible.

Soit x solution de $Mx = b$ et x_i tel que $x_i = min\, x_j$, nous avons

$$x_i = b_i - \frac{1}{M_{ii}} \sum_{j \ne i} M_{ij} x_j$$

L'inégalité $x_j \ge x_i$ jointe à $-M_{ij} \ge 0$ (pour $i \ne j$) entraine $-M_{ij} x_j \ge -M_{ij} x_i$ et par conséquent :

$$x_i \ge \ b_i - \left[\frac{1}{M_{ii}} \sum_{j \ne i} M_{ij} \right] x_i,$$

ce qui se réécrit

$$x_i \left(1 - \left[\frac{1}{M_{ii}} \sum_{j \ne i} |M_{ij}| \right] \right) \ge b_i.$$

Finalement $b_i > 0$ entraine $x_i > 0$ et donc $x_j > 0$ pour tout j.

On s'intéresse à la résolution du problème d'évolution

$$
(13) \qquad \begin{cases} \dfrac{\partial u}{\partial t} - \Delta u = 0, & x \in \Omega, \quad t > 0, \\ u(x,0) = u_0(x), & x \in \Omega, \quad (u_0(x) \in L^2(\Omega)), \\ \dfrac{\partial u}{\partial n} = 0, & \text{sur } \Gamma = \partial\Omega. \end{cases}
$$

2) Montrer que si u_0 est positive alors

$$\forall\, t \geq 0, \quad u(x,t) \geq 0, \quad \text{p.p. } x \in \Omega.$$

Indication : multiplier l'équation par $u^- = \inf(u, 0)$.

Il est "facile" de voir que si $u \in H^1(\Omega \times [0,T]$, alors $u^- \in H^1(\Omega \times [0,T]$ avec en outre :

$$
(14) \qquad \begin{cases} \dfrac{\partial u^-}{\partial t} = \chi_{u \leq 0} \, \dfrac{\partial u}{\partial t}, \\ \nabla u^- = \chi_{u \leq 0} \, \nabla u. \end{cases}
$$

Nous avons donc :

$$\int_\Omega u^- \, \frac{\partial u}{\partial t} \, dx - \int u^- \Delta u \, dx = 0.$$

En utilisant (14) on voit que :

$$
\begin{cases} \displaystyle\int_\Omega u^- \, \frac{\partial u}{\partial t} \, dx = \frac{1}{2} \frac{d}{dt} \int_\Omega |u^-|^2 \, dx, \\ \displaystyle -\int u^- \Delta u \, dx = \int_\Omega |\nabla u^-|^2 \, dx. \end{cases}
$$

Par conséquent, on a l'identité :

$$\frac{1}{2} \frac{d}{dt} \int_\Omega |u^-|^2 \, dx + \int_\Omega |\nabla u^-|^2 \, dx. = 0$$

En particulier, pour tout $t > 0$:

$$\int_\Omega |u^-(x,t)|^2 \, dx \leq \int_\Omega |u^-(x,0)|^2 \, dx = 0$$

puisque, u_0 étant positive, $u_0^- = 0$. Par conséquent, $u^- = 0$, c'est à dire $u \geq 0$.

3) On discrétise (13) en espace par éléments finis P^1 (on appelle V_h l'espace d'approximation correspondant). Si on calcule les intégrales exactement, on aboutit au système

linéaire de dimension N_h (U_h désigne ci-dessous le vecteur des degrés de libertés habituels de u_h) :

$$M_h \frac{dU_h}{dt} + A_h\, U_h = 0. \tag{15}$$

Montrer que la matrice de rigidité A_h vérifie :

$$\forall\, 1 \leq i \leq N_h, \quad \sum_{j=1}^{M_h} A_{ij} = 0.$$

Montrer que, si tous les angles du maillage triangulaire de Ω sont strictement inférieurs à $\pi/2$,

$$\forall\, i \neq j, \quad A_{ij} < 0.$$

Désignons par w_i la base canonique habituelle de V_h. Nous avons

$$\sum_{j=1}^{N_h} A_{ij} = \sum_{j=1}^{N_h} \int_\Omega \nabla w_i \cdot \nabla w_j \; dx = \int_\Omega \nabla w_i \cdot \nabla \big(\sum_{j=1}^{N_h} w_j \big) \; dx = 0,$$

puisque $\sum_{j=1}^{N_h} w_j = 1$. *Par ailleurs, pour $i \neq j$,*

$$A_{ij} = \sum_{K \in \mathcal{T}_h} \int_K \nabla w_i \cdot \nabla w_j \; dx.$$

Or, sur chaque triangle K, les vecteurs ∇w_i et ∇w_i sont constants. Soient ils sont nuls (lorsque K n'admet pas le noeud i somme sommet) soient ils sont diriges selon les normales extérieures aux arêtes qui sont opposées respectivement aux noeuds i et j dans K. Si tous les angles de sont strictement inférieurs à $\pi/2$, l'angle entre ces deux normales est strictement supérieur à $\pi/2$ et le produit scalaire $\nabla w_i \cdot \nabla w_j$ est donc strictement négatif. On conclut aisément.

4) Au lieu de calculer les intégrales sur Ω de façon exacte, on les calcule de façon approchée par la formule :

$$\int_\Omega f(x)\; dx = \sum_{K \in \mathcal{T}_h} I_K(f), \quad I_K(f) = \frac{mes(K)}{3} \sum_{j=1}^{3} f(S_j(K)).$$

où les $S_j(K)$ sont les sommets du triangle K. On obtient alors le système différentiel :

$$\widetilde{M}_h \frac{dU_h}{dt} + \widetilde{A}_h\, U_h = 0. \tag{16}$$

Montrer que $I_K(f) = \displaystyle\int_K f(x)\,dx$ pour tout $f \in P_1(K)$.

En déduire que $\widetilde{A}_h = A_h$. Montrer que $\widetilde{M}_h$ est la matrice diagonale définie par (c'est ce qu'on appelle la condensation de masse) :

$$\widetilde{M}_{ii} = \sum_{j=1}^{N_h} M_{ij}$$

Le fait que la formule de quadrature soit exacte pour les polynômes de degré 1 se vérifie aisément par le calcul sur le triangle de référence. Pour le cas général, il suffit de passer par la transformation qui envoie le triangle de référence dans le triangle courant.

Le fait que $\widetilde{A}_h = A_h$ résulte alors du fait que les produits $\nabla w_i \cdot \nabla w_j$ sont constants dans les triangles et du fait que la formule de quadrature est en particulier exacte pour les constantes.

Pour $i \neq j$, nous avons :

$$\widetilde{M}_{ij} = \sum_{K \in \mathcal{T}_h} I_K(w_i\, w_j).$$

Or, pour chaque K, $I_K(w_i\, w_j) = 0$ puisque chaque fonction w_i n'etant non nulle qu'en un sommet du maillage, pour $i \neq j$, le produit $w_i\, w_j$ s'annule sur tous les sommets. Enfin

$$\widetilde{M}_{ii} = \sum_{K \in \mathcal{T}_h} I_K(w_i^2).$$

Comme $w_i(S_l(K)) = 0$ ou 1, $w_i(S_l(K))^2 = w_i(S_l(K))$. Par conséquent :

$$I_K(w_i^2) = mes(K) \sum_{\ell=1}^{3} w_i(S_l(K))^2 = mes(K) \sum_{\ell=1}^{3} w_i(S_l(K)) = \int_K w_i\, dx$$

puisque w_i appartient à P_1 et la formule de quadrature est exacte pour les polynômes de degré 1. Par suite :

$$\widetilde{M}_{ii} = \sum_{K \in \mathcal{T}_h} \int_K w_i\, dx = \int_\Omega w_i\, dx.$$

Comme $\displaystyle\sum_{j=1}^{N_h} w_j = 1$, cette dernière égalité se réécrit

$$\widetilde{M}_{ii} = \int_\Omega w_i\, \Big(\sum_{j=1}^{N_h} w_j\Big)\, dx = \sum_{j=1}^{N_h} \int_\Omega w_i\, w_j\, dx = \sum_{j=1}^{N_h} M_{ij},$$

ce que nous voulions démontrer.

5) On approche (16) en temps à l'aide du schéma d'Euler implicite. On désigne par $u_h^n \in V_h$ l'approximation de la solution au temps $t^n = n\Delta t$. Montrer que, si tous les angles du maillage triangulaire de Ω sont strictement inférieurs à $\pi/2$, on a le principe du maximum discret :

$$u_h^0(x) \geq 0 \Rightarrow u_h^n(x) \geq 0, \quad \forall\, n \geq 1.$$

Le schéma d'Euler implicite s'écrit, sous forme matricielle :

$$\widetilde{M}_h \,\frac{U_h^{n+1} - U_h^n}{\Delta t} + A_h\, U_h^{n+1} = 0,$$

soit encore

$$(\widetilde{M}_h + \Delta t\ A_h)\, U_h^{n+1} = U_h^n$$

Comme nous travaillons avec les éléments finis P_1, montrer que $u_h^n(x) \geq 0$ équivaut à montrer que toutes les composantes du vecteur U_h^n sont positives, ce que nous écrirons $U_h^n \geq 0$. De proche en proche, nous allons montrer que :

$$U_h^n \geq 0 \Longrightarrow U_h^{n+1} \geq 0.$$

D'après la question 1), if suffit de vérifier que la matrice :

$$B = \widetilde{M}_h + \Delta t\ A_h,$$

que l'on inverse pour passer de U_h^n à U_h^{n+1}, est une M matrice. Or, pour $i \neq j$,

$$B_{ij} = \Delta t\ A_{ij} \leq 0 \quad \textit{(d'après la question 3)).}$$

De toute évidence $B_{ii} > 0$ et, d'après la question 3),

$$B_{ii} - \sum_{j\neq i} |B_{ij}| = B_{ii} + \sum_{j\neq i} B_{ij} = \widetilde{M}_{ii} + \sum_{j=1}^{N_h} A_{ij} = \widetilde{M}_{ii} > 0,$$

ce qui achève la démonstration.

6) Montrer que, si on ne fait pas la condensation de masse, on récupère le principe du maximum discret à condition que le pas de temps soit assez petit.

Je n'entre pas dans les détails. La matrice B à inverser à chaque pas de temps est cette fois $B = M_h + \Delta t\ A_h$. Le seul point non évident est que B_{ij} soit négatif pour $i \neq j$. En effet, une fois ce point acquis, on a :

$$B_{ii} - \sum_{j\neq i} |B_{ij}| = B_{ii} + \sum_{j\neq i} B_{ij} = \sum_{j=1}^{N_h} M_{ij} + \sum_{j=1}^{N_h} A_{ij} = \widetilde{M}_{ii} > 0,$$

Pour que B_{ij} soit négatif, il suffit que, pout tout couple (i,j) et tout triangle K :

$$\Delta t \int_K \nabla w_i \cdot \nabla w_j \, dx \leq \int_K w_i \, w_j \, dx$$

ce qui sera surement vérifié sous une contrainte du type :

$$\frac{\Delta t}{h^2} \leq C$$

où la constante C est liée au plus petit angle du maillage.

7) Les raisonnements précédents fonctionnent-t-ils avec un θ schéma quelconque ?

Non car cette fois, même si on fait la condensation de masse, le passage de n à $n+1$ se fait par l'intermédiaire du système :

$$(\widetilde{M}_h + \theta \Delta t \, A_h) \, U_h^{n+1} = (\widetilde{M}_h - (1-\theta)\Delta t \, A_h) U_h^n.$$

Il est facile de vérifier que $\widetilde{M}_h + \theta \Delta t \ A_h$ reste une M-matrice. En revanche, $U_h^n \geq 0$ n'implique plus a priori $(\widetilde{M}_h - (1-\theta)\Delta t \ A_h) \ U_h^n \geq 0$. Ceci ne sera vrai que si la matrice $\widetilde{M}_h - (1-\theta)\Delta t \ A_h$ est à coefficients positifs. Un tel résultat sera obtenu sous une contrainte du type :

$$(1-\theta) \, \frac{\Delta t}{h^2} \leq C$$

où la constante C est liée au plus petit angle du maillage.

Exercice 5. Schémas de Richardson et de DuFort-Frankel.

On s'intéresse à l'équation de la chaleur en dimension 1 :

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \qquad x \in \mathbb{R}, \quad t > 0 \tag{17}$$

que l'on approche à l'aide du schéma de Richardson :

$$\frac{u_j^{n+1} - u_j^{n-1}}{2\Delta t} - \frac{u_{j+1}^n - 2u_j^n + u_{j-1}^n}{h^2} = 0 \tag{18}$$

1) Quel est a priori l'inconvénient pratique de ce schéma ? Montrer qu'il s'agit d'un schéma du second ordre en espace et en temps mais inconditionnellement instable.

Le schéma étant complètement centré est au moins d'ordre 2. Le calcul le confirme. L'inconvénient pratique de ce schéma est qu'il s'agit d'un schéma à trois pas de temnps qui nécessite donc un schéma de démarrage. Pour étudier sa stabilité, étudions les solutions *particulières de la forme :* $u_j^n = \hat{u}^n\ e^{ikx_j}, \quad k \in \mathbb{R}$. *Nous aboutissons à la relation de récurrence à deux niveaux (posons* $\beta = \Delta t/h^2$*) :*

$$\hat{u}^{n+1} + 2\beta\ \sin^2(\frac{kh}{2})\ \hat{u}^n + \hat{u}^{n-1} = 0$$

dont la solution générale est donnée par :

$$\hat{u}^n = A\ r_1^n + B\ r_2^n,$$

où r_1 *et* r_2 *sont les deux racines de l'équation caractéristique :*

$$r^2 + 2\beta \sin^2(\frac{kh}{2})\ r - 1 = 0$$

Le produit des deux racines de cette équation est égal à 1. Le discriminant réduit, donné par :

$$\Delta = 1 + \beta^2 \sin^4(\frac{kh}{2})$$

étant strictement positif, les deux racines sont réelles dictinctes. L'une d'entre elles est donc en module strictement supérieure à 1, d'où l'instabilité inconditionnelle du schéma.

2) Dans (18), on remplace u_j^n par $\frac{u_j^{n+1}+u_j^{n-1}}{2}$. On obtient ainsi le schéma de DuFort-Frankel. Montrer que ce schéma est explicite et inconditionnellement stable.

Le schéma reste bien explicite comme le montre la formule :

$$u_j^{n+1} = \frac{1-\beta}{1+\beta}\ u_j^{n-1} + \frac{\beta}{1+\beta}\ \Big(u_{j+1}^n + u_{j-1}^n\Big).$$

Pour la stabilité on procède comme à la question précédente. L´équation caractéristique devient :

$$P(r) \equiv (1+\beta)\ r^2 - 2\beta\ \cos kh\ r - (1-\beta) = 0.$$

Le produit des racines, égal à $-(1-\beta)/(1+\beta)$, *est en valeur absolue strictement plus petit que 1. Par ailleurs, on observe que :*

$$P(\cos kh) = \beta \sin^2 kh > 0, \quad P(-1) = 2\beta\ (1 - \cos kh) > 0.$$

Le polynôme P *étant positif à l'infini, si ses deux racines sont réelles, elles sont donc comprises entre* -1 *et* $\cos kh$, *donc de module plus petit que 1. Sinon, ces deux racines sont complexes conjuguées et donc toutes de de module strictement inférieur à 1 puisque le module de leur produit l'est.*

Le schéma est bien inconditionnellement stable.

3) Analyser l'erreur de troncature du schéma de DuFort-Frankel et conclure (le schéma

n'est consistant que si $\Delta t/h$ tend vers 0). Ce résultat était-il prévisible ?

On peut réécrire le schéma de DuFort-Frankel comme une perturbation du schéma de Richardson sous la forme :

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

Par conséquent si $(\varepsilon_j^n)_R$ et $(\varepsilon_j^n)_{DF}$ désignent respectivement les erreurs de trancature associées aux schémas de Richardson et DuFort-Frankel, on voit que :

$$(\varepsilon_j^n)_{DF} = (\varepsilon_j^n)_R + \frac{\Delta t^2}{h^2}\,\frac{\partial^2 u}{\partial t^2}(x_j, t^n) + O(\frac{\Delta t^4}{h^2}).$$

Autrement dit, pour que le schéma de DuFort-Frankel soit consistant, il faut que $\Delta t/h$ tende vers 0 et pour qu'il soit d'ordre 2, on retombe sur une condition du type $\Delta t < C\ h^2$.

Un tel résultat était prévisible. En effet, à cause du caractère explicite du schéma la solution numérique se propage à vitesse finie, égale en l'occurrence à $h/\Delta t$. Comme la solution exacte se propage à vitesse infinie, pour espérer avoir convergence du schéma il faut que cette vitesse tende vers l'infini, c'est à dire que $\Delta t/h$ tende vers 0.