1) Soit $t_1 = -\alpha$, $t_2 = \beta$, $t_3 = \alpha$. On veut $\omega_1, \omega_2, \omega_3$ t.q. $J(g) = \sum_{j=1}^{3} \omega_j \cdot g(t_j)$.

a) Trouver $\omega_1, \omega_2, \omega_3$ en fct. de α, β. t.q. la formule de quadrature soit exacte pour des poly. de degré 2. i.e. $J(p) = \int_{-1}^{1} p(t)\,dt$ $\forall$ polynôme p de degré 2.

b) Démontrer que si $\beta = 0$, alors la formule est exacte pour $p \in \mathbb{P}_3$

Prendre $\alpha = 1$ et vérifier qu'on a la formule de Simpson.

c) Vérifier que si $\beta = 0$, $\alpha = \sqrt{3/5}$, alors la formule est exacte pour des poly. de degré 5.

Vérifier que c'est une formule de Gauss-Legendre à 3 points

---

2) Soit $t_1 = \alpha$, $t_2 = 0$, $t_3 = \beta$. On veut chercher $\omega_1, \omega_2, \omega_3$ qui définissent la formule de quadrature : $J(g) = \sum_{j=1}^{3} \omega_j \cdot g(t_j)$

a) Trouver $\omega_1, \omega_2, \omega_3$ en fonction de α, β t.q. $J(p) = \int_{-1}^{1} p(t)\,dt$ $\forall p \in \mathbb{P}_2$.

b) Montrer que si $J(p) = \int_{-1}^{1} p(t)\,dt$, $\forall p \in \mathbb{P}_3$, on a nécessairement $\alpha = -\beta$.

c) Existe-t-il α, β t.q. $J(p) = \int_{-1}^{1} p(t)\,dt$, $\forall p \in \mathbb{P}_4$? Si oui, que valent α, β ?

---

3) Soit t_1, t_2 donnés. $f(t)$ une fonction continue sur $I = [0, \infty[$, on veut définir une formule de quadrature de la forme $J(f) = \omega_1 \cdot f(t_1) + \omega_2 \cdot f(t_2)$ pour approcher numériquement $\int_0^\infty e^{-t} \cdot f(t)\,dt$.

a) Trouver ω_1, ω_2 en fct. de t_1, t_2 t.q. $J(p) = \int_0^\infty e^{-t} \cdot p(t)\,dt$, $\forall p \in \mathbb{P}_1$

× b) Montrer que si $t_1 = 2 - \sqrt{2}$, $t_2 = 2 + \sqrt{2}$, alors la formule de quadrature est exacte $\forall p \in \mathbb{P}_3$.

---

4) Soient M points $t_1 < t_2 < \ldots < t_M$ entre -1 et 1 et soit $\{\varphi_i\}_{i=1}^{M}$ la base de Lagrange des polynômes de degré $(M-1)$ associés à ces M points. On pose

$$\omega_j = \int_{-1}^{1} \varphi_j(t)\,dt, \; j = 1, 2, \ldots, M, \quad \text{et} \quad J(g) = \sum_{j=1}^{M} \omega_j \cdot g(t_j), \; g\,[-1,+1] \to \mathbb{R}.$$

a) Démontrer que si $g \in \mathbb{P}_{M-1}$, alors $J(g) = \int_{-1}^{1} g(t)\,dt$.

b) Démontrer que on a la relation $\sum_{j=1}^{M} \omega_j \cdot t_j^s = \dfrac{1 + (-1)^s}{s+1}$ $\forall s = 0, 1, \ldots, M-1$

5) Soit $J(p) = \int_{-1}^{1} \varphi(t)\, dt$, $\forall p \in \mathbb{P}_M$. Trouver ω_j.

Remarque : en partant de $\varphi_j \in \mathbb{P}_M$:

$$\int_{-1}^{1} \varphi_j(t)\, dt \overset{\text{hyp.}}{=} J(\varphi_i) = \sum_{i=1}^{M} \omega_i \cdot \underbrace{\varphi_j(t_i)}_{\delta_{ij}} = \omega_j$$

---

6) $J(p) = \int_0^\infty e^{-t} . p(t)\, dt \quad \forall p \in \mathbb{P}_M$. Trouver ω_j.

---

7) Si $\omega_j = \int_{-1}^{1} \varphi_j(t)\, dt$, alors vérifier que $J(p) = \int_{-1}^{1} p(t)\, dt \quad \forall p \in \mathbb{P}_M$.

---

8) Si ω_j, t_j sont donnés, chercher M maximum t.q.

$$J(p) = \int_{-1}^{1} p(t)\, dt \quad \forall p \in \mathbb{P}_M$$

principe : travailler avec une base, expliquer pourquoi, utiliser les éléments déjà connus.

Exercices types examen sur séries 3 - 5

Chapitre 3

1) – si $J(P) = \int_{-1}^{1} \varphi(t)\, dt$, $\forall P \in \mathbb{R}_M$, trouver ω_j.

Remarque : en partant de $\varphi_j \in \mathbb{R}_M$:

$$\int_{-1}^{1} \varphi_j(t)\, dt \overset{hyp.}{=\!=} J(\varphi_j) = \sum_i \omega_i \underbrace{\varphi_j(t_i)}_{=\delta_{ij}} = \omega_j$$

– $J(P) = \int_0^{\infty} e^{-t} p(t)\, dt$, $\forall P \in \mathbb{R}_M$, trouver ω_j v.c.f. examen 94.

2) si $\omega_j = \int_{-1}^{1} \varphi_j(t)\, dt$, alors vérifier que $J(P) = \int_{-1}^{1} p(t)\, dt$ $\forall P \in \mathbb{R}_n$

3) si ω_j, t_j sont donnés, chercher M (nb. de points d'intégration) maximum t.q.

$$J(P) = \int_{-1}^{1} p(t)\, dt \quad \forall P \in \mathbb{R}_m$$

$m = M - 1$
$\Rightarrow M = m + 1$

principe : travailler avec une base (expliquer pourquoi), utiliser les éléments déjà connus.

Chapitre ~~4~~ 5 : pas adapter l'algorithme du cours

1) Supposer une forme
2) Multiplier
3) identifier terme à terme

II) justifier la forme : – thm. de conservation Flèche / Tridiagonale
– thm. d'unicité de la décomposition

III) écrire en résolvant successivement les termes

Symétrique définie positive : (dém)

I) signaler que A est Symétrique

II) Démontrer que $X^t . A . X \geq 0$, i.e. $X^t . A . X = \sum_i d_i x_i^2 + 2 \sum c_i x_i$ ...etc.

III) constater que $X^t . A . X = 0 \iff x = 0$

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 11 mars 1997

Série 1

Exercice 1

Soit $t_0 < t_1$ deux nombres réels distincts et soit ε tel que $0 < \varepsilon < \frac{t_1 - t_0}{2}$.

a) Expliciter un polynôme p_ε de degré 3 tel que

$$p_\varepsilon(t_0) = p_\varepsilon(t_0 + \varepsilon) = 1,$$
$$p_\varepsilon(t_1) = p_\varepsilon(t_1 + \varepsilon) = 0.$$

b) Si $\phi_0(t) = \lim_{\varepsilon \to 0} p_\varepsilon(t)$, montrer que $\phi_0(t_0) = 1$, $\phi_0(t_1) = \phi_0'(t_0) = \phi_0'(t_1) = 0$, et ainsi ϕ_0 est une fonction de base des polynômes de degré 3 pour l'interpolation d'Hermite.
↳ 2 pts, et dérivées.

c) Peut-on obtenir toutes les fonctions de base d'Hermite par un procédé analogue? Si oui, lequel?

Exercice 2

Ecrire explicitement un polynôme $p(t)$ de degré 3 tel que

2 pts, et dérivées => Hermite

$$p(1) = 4, \qquad p(2) = 5,$$
$$p'(1) = 3, \qquad p'(2) = 2.$$

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 11 mars 1997

Corrigé 1

Exercice 1

Point a)

Le polynôme p_ε, exprimé dans la base de Lagrange $(\varphi_j)_{j=0}^3$ des polynômes de degré 3 associée aux points $r_0 = t_0$, $r_1 = t_0 + \varepsilon$, $r_2 = t_1$ et $r_3 = t_1 + \varepsilon$, est

$$p_\varepsilon(t) = \sum_{j=0}^{3} p_\varepsilon(r_j)\varphi_j(t).$$

Il se simplifie $p_\varepsilon(t) = \varphi_0(t) + \varphi_1(t)$ puisque $p_\varepsilon(r_0) = p_\varepsilon(r_1) = 1$ et $p_\varepsilon(r_2) = p_\varepsilon(r_3) = 0$.

Rappelons que ces deux fonctions valent

$$\varphi_0(t) = \frac{t-(t_0+\varepsilon)}{t_0-(t_0+\varepsilon)} \cdot \frac{t-t_1}{t_0-t_1} \cdot \frac{t-(t_1+\varepsilon)}{t_0-(t_1+\varepsilon)} \quad \text{et} \quad \varphi_1(t) = \frac{t-t_0}{(t_0+\varepsilon)-t_0} \cdot \frac{t-t_1}{(t_0+\varepsilon)-t_1} \cdot \frac{t-(t_1+\varepsilon)}{(t_0+\varepsilon)-(t_1+\varepsilon)}.$$

En utilisant ces fonctions de base et en regroupant les fractions sur un dénominateur commun, p_ε s'écrit

$$p_\varepsilon(t) = \frac{(t-t_1)(t-t_1-\varepsilon)}{\varepsilon(t_0-t_1)} \left[\frac{-(t-t_0-\varepsilon)(t_0+\varepsilon-t_1) + (t-t_0)(t_0-t_1-\varepsilon)}{(t_0-t_1-\varepsilon)(t_0+\varepsilon-t_1)}\right].$$

Si nous développons le numérateur de la fraction entre crochets, nous remarquons que les termes sans facteur ε disparaissent. Nous pouvons donc simplifier par ε et obtenir

$$p_\varepsilon(t) = \frac{(t-t_1)(t-t_1-\varepsilon)(-2t+3t_0-t_1+\varepsilon)}{(t_0-t_1)(t_0-t_1-\varepsilon)(t_0-t_1+\varepsilon)}.$$

Point b)

Avec p_ε ainsi exprimé, nous voyons facilement que

$$\phi_0(t) = \lim_{\varepsilon \to 0} p_\varepsilon(t) = \frac{(t-t_1)^2(-2t+3t_0-t_1)}{(t_0-t_1)^3}.$$

Le calcul de la dérivée donne

$$\phi_0'(t) = -6\frac{(t-t_1)(t-t_0)}{(t_0-t_1)^3}$$

Evidemment, ϕ_0 a les propriétés suivantes : $\phi_0(t_0) = 1$ et $\phi_0'(t_0) = \phi_0(t_1) = \phi_0'(t_1) = 0$, ce qui correspond à celles d'un polynôme de base d'Hermite (des polynômes de degré 3). (Voir la fonction $\varphi_0(t)$ du cours polycopié.)

Point c)

Les autres fonctions de base d'Hermite peuvent être obtenues par des procédés analogues.

Par exemple, la fonction $\psi_0(t)$ ayant les propriétés $\psi_0(t_0) = \psi_0(t_1) = \psi_0'(t_1) = 0$ et $\psi_0'(t_0) = 1$ est construite à partir du polynôme q_ε de degré 3 tel que

$$q_\varepsilon(t_0) = q_\varepsilon(t_1) = q_\varepsilon(t_1+\varepsilon) = 0 \quad \text{et} \quad q_\varepsilon(t_0+\varepsilon) = \varepsilon.$$

En effet, dans la base de Lagrange, avec le même raisonnement qu'au point a), q_ε s'écrit

$$q_\varepsilon(t) = \sum_{j=0}^{3} q_\varepsilon(r_j)\varphi_j(t) = \varepsilon\varphi_1(t) = \varepsilon\frac{(t-t_0)}{\varepsilon}\frac{(t-t_1)}{(t_0+\varepsilon-t_1)}\frac{(t-t_1-\varepsilon)}{(t_0-t_1)}$$

ou, simplifié, $q_\varepsilon(t) = \frac{(t-t_0)(t-t_1)(t-t_1-\varepsilon)}{(t_0-t_1+\varepsilon)(t_0-t_1)}$. Nous trouvons que

$$\psi_0(t) = \lim_{\varepsilon \to 0} q_\varepsilon(t) = \frac{(t-t_0)(t-t_1)^2}{(t_0-t_1)^2}$$

et nous reconnaissons une autre fonction de base d'Hermite.

Les deux dernières fonctions de base pour l'interpolation d'Hermite par des polynômes de degré 3 sont obtenues par symétrie, c'est-à-dire en échangeant t_0 et t_1.

Exercice 2

De manière générale, dans la base de Hermite $(\phi_0, \phi_1, \psi_0, \psi_1)$ des polynômes de degré 3 associée aux points $r_0 = 1$ et $r_1 = 2$, un polynôme p s'exprime

$$p(t) = p(r_0)\phi_0(t) + p(r_1)\phi_1(t) + p'(r_0)\psi_0(t) + p'(r_1)\psi_1(t).$$

Les fonctions de la base sont obtenues ci-dessus (ou données dans le cours).

Dans notre cas particulier, le polynôme $p(t)$ s'écrit

$$p(t) = 4(t-2)^2(2t-1) - 5(t-1)^2(2t-5) + 3(t-2)^2(t-1) + 2(t-1)^2(t-2).$$

Le polynôme cherché, simplifié, est donc $p(t) = 3t^3 - 14t^2 + 22t - 7$.

Point c)

Si $x \in [x_1, x_0]$, nous trouvons, en utilisant l'indication, que

$$|r(x)| = \left|\int_{x_1}^{x} r'(s)ds\right| \leqslant \int_{x_1}^{x_0} |r'(s)|\, ds \leqslant h \max_{s\in[x_1,x_0]} |r'(s)| \tag{1}$$

$$\text{et } |r'(x)| = \left|\int_{\xi_0}^{x} r''(s)ds\right| \leqslant \int_{\xi_0}^{x_0} |r''(s)|\, ds \leqslant h \max_{s\in[x_1,x_0]} |r''(s)|\,. \tag{2}$$

De b), nous savons que si $x \in [x_1, x_0]$ nous avons

$$|r''(x)| = \left|\int_{\eta}^{x} f'''(s)ds\right|.$$

Alors

$$|r''(x)| \leqslant \left|\int_{\eta}^{x} |f'''(s)|\, ds\right| \leqslant \int_{x_2}^{x_0} |f'''(s)|\, ds, \quad \text{car } \eta \in [\xi_1, \xi_0] \subset [x_2, x_0],$$

et donc

$$|r''(x)| \leqslant 2h \max_{s\in[x_2,x_0]} |f'''(s)|\,.$$

Il découle de (2) que, pour $x \in [x_0, x_1]$, on a

$$|r'(x)| \leqslant 2h^2 \max_{s\in[x_2,x_0]} |f'''(s)|\,.$$

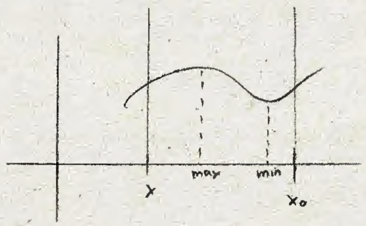
Donc $\max\limits_{x\in[x_1,x_0]} |r'(x)| \leqslant 2h^2 \max\limits_{s\in[x_2,x_0]} |f'''(s)|$. Alors de (1), nous pouvons conclure que

$$|r(x)| \leqslant 2h^3 \max_{s\in[x_2,x_0]} |f'''(s)|\,, \forall x \in [x_1, x_0].$$

Idem pour l'intervalle $[x_2, x_1]$ en prenant ξ_1 à la place de ξ_0.

Comparaison avec la formule de Taylor:

Soit G le polynôme de Taylor, de degré 2, de la fonction f autour du point x_0 i.e.



$$G(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2.$$

Nous avons l'égalité suivante

$$f(x) = G(x) + \frac{f'''(\xi)}{3!}(x - x_0)^3, \quad \text{où } \xi \in [\min(x_0, x), \max(x_0, x)].$$

Si $x \in [x_2, x_0]$, nous obtenons l'inégalité suivante

$$|f(x) - G(x)| \leqslant \frac{1}{3!} \max_{s\in[x_2,x_0]} \left|(x - x_0)^3\right| \cdot \max_{s\in[x_2,x_0]} |f'''(s)| \leqslant \frac{4h^3}{3} \max_{s\in[x_2,x_0]} |f'''(s)|\,.$$

Nous pouvons en déduire que les deux polynômes, g et G, donnent une approximation de f qui est de l'ordre de h^3. La formule de Taylor est légèrement plus précise, puisque $\frac{4}{3} < 2$, mais elle utilise par contre les dérivées de f qui ne sont pas toujours aisées à obtenir. (Par exemple, lorsque la fonction f n'est donnée que sous forme tabulée ou s'obtient par un calcul compliqué.)

Le 18 mars 1997

$(tg\,x)'' = 2tg(x)\left(1 + tg^2(x)\right)$

$(tg\,x)' = 1 + tg(x)^2$

$\nabla_h f(x_0) = f(x_0) - f(x_0 - h)$

$\delta_h f(x_0) = f(x_0 + \frac{h}{2}) - f(x_0 - \frac{h}{2})$

Série 2

Exercice 1

$tg(1)' = 3{,}425518$, $tg(1)'' = 10{,}669858$

On considère la fonction $f(x) = \tan(x)$. On cherche à évaluer la dérivée $f'(x)$ en x valant $x_0 = 1$. Pour ce faire, on utilise les formules de différences finies rétrograde : $t_h = \dfrac{\nabla_h f(x_0)}{h}$, centrée : $r_h = \dfrac{\delta_h f(x_0)}{h}$ et de Richardson : $R_h = \dfrac{8f(x_0 + h/4) - f(x_0 + h/2) + f(x_0 - h/2) - 8f(x_0 - h/4)}{3h}$.
Remplir le tableau suivant et conclure.

dérivée numérique Erreur totale E^h

h	t_h	$\lvert f'(x_0) - t_h\rvert$	$\frac{\lvert f'(x_0)-t_h\rvert}{h^1}$	r_h	$\lvert f'(x_0) - r_h\rvert$	$\frac{\lvert f'(x_0)-r_h\rvert}{h^2}$	R_h	$\lvert f'(x_0) - R_h\rvert$	$\frac{\lvert f'(x_0)-R_h\rvert}{h^4}$
10^{-1}	2,97249	$4{,}530237 \cdot 10^{-1}$	$4{,}530237 \cdot 10^{0}$	3,44932	$2{,}380839 \cdot 10^{-2}$	$2{,}380839 \cdot 10^{0}$	3,42547	$4{,}562180 \cdot 10^{-5}$	$4{,}562180 \cdot 10^{-1}$
10^{-2}	3,37309	$5{,}242046 \cdot 10^{-2}$	$5{,}242046 \cdot 10^{0}$	3,42575	$2{,}362802 \cdot 10^{-4}$	$2{,}362802 \cdot 10^{0}$	3,42551	$4{,}13 \cdot 10^{-9}$	$4{,}13 \cdot 10^{-1}$
10^{-3}	3,42019	$5{,}325501 \cdot 10^{-3}$	$5{,}325501 \cdot 10^{0}$	3,42552	$2{,}3592 \cdot 10^{-6}$	$2{,}3592 \cdot 10^{0}$	3,42551	$2{,}08 \cdot 10^{-8}$	$2{,}08 \cdot 10^{4}$
10^{-4}	3,42498	$5{,}334208 \cdot 10^{-4}$	$5{,}334208 \cdot 10^{0}$	3,42551	$2{,}08 \cdot 10^{-8}$	$2{,}08 \cdot 10^{0}$	3,42551	$1{,}5413 \cdot 10^{-7}$	$1{,}5413 \cdot 10^{9}$
10^{-5}	3,42546	$5{,}38208 \cdot 10^{-5}$	$5{,}38208 \cdot 10^{0}$	3,42551	$1{,}792 \cdot 10^{-7}$	$1{,}792 \cdot 10^{3}$	3,42552	$1{,}1792 \cdot 10^{-6}$	$1{,}1792 \cdot 10^{14}$
10^{-6}	3,42551	$8{,}8208 \cdot 10^{-6}$	$8{,}8208 \cdot 10^{0}$	3,42552	$1{,}1792 \cdot 10^{-6}$	$1{,}1792 \cdot 10^{6}$	3,42550	$1{,}88208 \cdot 10^{-5}$	$1{,}88208 \cdot 10^{19}$
10^{-8}	3,42500	$5{,}188208 \cdot 10^{-4}$	$5{,}188208 \cdot 10^{4}$	3,425	$5{,}188208 \cdot 10^{-4}$	$5{,}188208 \cdot 10^{12}$	3,42666	$1{,}147846 \cdot 10^{-3}$	$1{,}147846 \cdot 10^{29}$
10^{-10}	3,4	$2{,}551882 \cdot 10^{-2}$	$2{,}551882 \cdot 10^{8}$	3,5	$7{,}448118 \cdot 10^{-2}$	$7{,}448118 \cdot 10^{18}$	3,00000	$4{,}255188 \cdot 10^{-1}$	$4{,}255188 \cdot 10^{39}$

Exercice 2

Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction trois fois continûment dérivable donnée, soient $x_0 \in \mathbb{R}$ et $h > 0$ donnés. On pose $x_j = x_0 - jh$ avec $j = 1, 2$, et on définit

$$g(x) = f(x_0) + \frac{\nabla_h f(x_0)}{h}(x - x_0) + \frac{\nabla_h^2 f(x_0)}{2!h^2}(x - x_0)(x - x_1).$$

a) Vérifier que $g(x_j) = f(x_j)$ si $j = 0, 1, 2$ et en déduire qu'il existe $\xi_0 \in [x_1, x_0]$ et $\xi_1 \in [x_2, x_1]$ tels que $f'(\xi_0) = g'(\xi_0)$, $f'(\xi_1) = g'(\xi_1)$.

b) Si on pose $r(x) = f(x) - g(x)$, déduire de a) qu'il existe $\eta \in [\xi_1, \xi_0]$ tel que $r''(\eta) = 0$ et donc $r''(x) = \displaystyle\int_\eta^x r'''(t)dt = \int_\eta^x f'''(t)dt$.

c) Déduire de b) que $|f(x) - g(x)| \leqslant 2h^3 \displaystyle\max_{t \in [x_2, x_0]} |f'''(t)|$ si $x \in [x_2, x_0]$. Comparer avec le développement de Taylor.

Indication: il suffit de remarquer que pour $x \in [x_{j+1}, x_j]$, $j = 0, 1$, on peut écrire $r(x) = \int_{x_j}^x r'(s)ds$ et $r'(x) = \int_{\xi_j}^x r''(s)ds$.

TB

Corrigé 2

Exercice 1

Les calculs ont été faits sur une station SG IP20 en simple et double précisions, i.e. avec 8 et 16 chiffres significatifs. Seuls les 5 premiers chiffres sont affichés dans les tableaux.

h	t_h	$\lvert f'(x_0)-t_h\rvert$	$\frac{\lvert f'(x_0)-t_h\rvert}{h^1}$	r_h	$\lvert f'(x_0)-r_h\rvert$	$\frac{\lvert f'(x_0)-r_h\rvert}{h^2}$	R_h	$\lvert f'(x_0)-R_h\rvert$	$\frac{\lvert f'(x_0)-R_h\rvert}{h^4}$
10^{-1}	2.9725e+00	4.5302e-01	4.5302e+00 ▼	3.4493e+00	2.3807e-02	2.3807e+00▼	3.4255e+00	5.1022e-05	5.1022e-01
10^{-2}	3.3731e+00	5.2420e-02	5.2420e+00	3.4258e+00	2.3437e-04	2.3437e+00▲	3.4256e+00	4.7684e-05★	4.7684e+03
10^{-3}	3.4201e+00	5.4040e-03	5.4040e+00	3.4255e+00	3.9577e-05★	3.9577e+01	3.4253e+00	2.3818e-04	2.3818e+08
10^{-4}	3.4261e+00	5.5647e-04★	5.5647e+00 ▲	3.4237e+00	1.8277e-03	1.8277e+05	3.4269e+00	1.3514e-03	1.3514e+13
10^{-5}	3.4332e+00	7.7093e-03	7.7093e+02	3.4332e+00	7.7093e-03	7.7093e+07	3.4332e+00	7.7093e-03	7.7093e+17
10^{-6}	3.4571e+00	3.1551e-02	3.1551e+04	3.3379e+00	8.7658e-02	8.7658e+10	3.4968e+00	7.1287e-02	7.1287e+22
10^{-7}	4.7684e+00	1.3429e+00	1.3429e+07	2.3842e+00	1.0413e+00	1.0413e+14	0.0000e+00▽	3.4255e+00	3.4255e+28
10^{-8}	0.0000e+00▽	3.4255e+00	3.4255e+08	0.0000e+00▽	3.4255e+00	3.4255e+16	0.0000e+00	3.4255e+00	3.4255e+32
10^{-9}	0.0000e+00	3.4255e+00	3.4255e+09	0.0000e+00	3.4255e+00	3.4255e+18	0.0000e+00	3.4255e+00	3.4255e+36
10^{-10}	0.0000e+00	3.4255e+00	3.4255e+10	0.0000e+00	3.4255e+00	3.4255e+20	0.0000e+00	3.4255e+00	3.4255e+40

h	t_h	$\lvert f'(x_0)-t_h\rvert$	$\frac{\lvert f'(x_0)-t_h\rvert}{h^1}$	r_h	$\lvert f'(x_0)-r_h\rvert$	$\frac{\lvert f'(x_0)-r_h\rvert}{h^2}$	R_h	$\lvert f'(x_0)-R_h\rvert$	$\frac{\lvert f'(x_0)-R_h\rvert}{h^4}$
10^{-1}	2.9725e+00	4.5302e-01	4.5302e+00 ▼	3.4493e+00	2.3808e-02	2.3808e+00 ▼	3.4255e+00	4.5622e-05	4.5622e-01 ▼
10^{-2}	3.3731e+00	5.2420e-02	5.2420e+00	3.4258e+00	2.3628e-04	2.3628e+00	3.4255e+00	4.5190e-09	4.5190e-01
10^{-3}	3.4202e+00	5.3255e-03	5.3255e+00	3.4255e+00	2.3626e-06	2.3626e+00	3.4255e+00	2.9088e-13★	2.9088e-01 ▲
10^{-4}	3.4250e+00	5.3340e-04	5.3340e+00	3.4255e+00	2.3632e-08	2.3632e+00	3.4255e+00	8.5807e-12	8.5807e+04
10^{-5}	3.4255e+00	5.3348e-05	5.3348e+00	3.4255e+00	2.7268e-10	2.7268e+00 ▲	3.4255e+00	1.5661e-10	1.5661e+10
10^{-6}	3.4255e+00	5.3348e-06	5.3348e+00	3.4255e+00	2.1582e-10★	2.1582e+02	3.4255e+00	6.7236e-10	6.7236e+14
10^{-7}	3.4255e+00	5.3401e-07	5.3401e+00 ▲	3.4255e+00	1.1164e-09	1.1164e+05	3.4255e+00	1.1040e-09	1.1040e+19
10^{-8}	3.4255e+00	3.3369e-09★	3.3369e-01 ◁	3.4255e+00	3.3369e-09	3.3369e+07	3.4255e+00	1.0738e-08	1.0738e+24
10^{-9}	3.4255e+00	6.3276e-08	6.3276e+01	3.4255e+00	1.5877e-07	1.5877e+11	3.4255e+00	1.0738e-08	1.0738e+28
10^{-10}	3.4255e+00	1.0469e-06	1.0469e+04	3.4255e+00	1.1735e-06	1.1735e+14	3.4255e+00	1.7871e-06	1.7871e+34

La valeur exacte de $f'(x_0)$ est $f'(1) = 1 + \tan^2 1 = 3.4255188208147597609\ldots$

Les conclusions sont les suivantes :

1) La formule décentrée est moins précise que la formule centrée, qui est elle-même moins précise que celle de Richardson.

2) Avant que les erreurs d'arrondi ne prennent le dessus sur les erreurs de troncature (entre ▼ et ▲), les valeurs sont presque constantes dans les dernières colonnes, c.-à-d. nous observons l'ordre de convergence en h, h^2 ou h^4.

3) Pour les petites valeurs de h les erreurs d'arrondi deviennent importantes vis-à-vis des erreurs de troncature. Il existe une valeur optimale de h (★) qui minimise l'erreur totale, c.-à-d. qui fait un compromis entre l'erreur de troncature et l'erreur d'arrondi. Ces h optimaux augmentent lorsque l'ordre de convergence augmente.

4) La dérivation numérique n'est pas très stable vis-à-vis des erreurs d'arrondi. Par contre, il peut arriver que les erreurs se compensent et que le résultat soit meilleur que prévu (◁).

5) Dès que h est inférieur à la précision-machine, les valeurs de x_0, $x_0 \pm h/4$, $x_0 \pm h/2$, $x_0 - h$ ne sont plus différenciées. Les approximations sont alors nulles (depuis ▽).

Remarque : le support de cours sur WWW contient des commentaires complémentaires sur les erreurs.

Exercice 2

Point a)

Nous avons

$$g(x) = f(x_0) + \frac{\nabla_h f(x_0)}{h}(x - x_0) + \frac{\nabla_h^2 f(x_0)}{2!h^2}(x - x_0)(x - x_1),$$

où $\nabla_h f(x_0) = f(x_0) - f(x_0 - h)$ et $\nabla_h^2 f(x_0) = f(x_0) - 2f(x_0 - h) + f(x_0 - 2h)$. Alors, en tenant compte des égalités $x_1 = x_0 - h$ et $x_2 = x_0 - 2h$, nous obtenons

$$\begin{aligned} g(x_0) &= f(x_0), \\ g(x_1) &= f(x_0) - \nabla_h f(x_0) = f(x_1) \quad \text{et} \\ g(x_2) &= f(x_0) - 2\nabla_h f(x_0) + \nabla_h^2 f(x_0) = f(x_2). \end{aligned}$$

Ceci montre que g est l'interpolant, de degré 2, de f passant par les points x_0, x_1, et x_2.

Considérons la fonction $r = f - g$. Nous observons que

i) r est (au moins) trois fois continûment dérivable et

ii) $r(x_0) = r(x_1) = r(x_2) = 0$.

Par le théorème de Rolle, nous savons que

i) $\exists \xi_0 \in [x_1, x_0]$ t.q. $r'(\xi_0) = 0$, i.e. $f'(\xi_0) = g'(\xi_0)$ et

ii) $\exists \xi_1 \in [x_2, x_1]$ t.q. $r'(\xi_1) = 0$, i.e. $f'(\xi_1) = g'(\xi_1)$.

Point b)

De a), nous remarquons que $r'(\xi_0) = r'(\xi_1) = 0$ et que r' est une fonction (au moins) deux fois continûment dérivable. En utilisant encore le théorème de Rolle, mais sur r', nous trouvons que

$$\exists \eta \in [\xi_1, \xi_0] \quad \text{t.q.} \quad r''(\eta) = 0.$$

En utilisant le théorème fondamental du calcul intégral et le fait que g est un polynôme de degré 2, nous obtenons

$$r''(x) = \int_\eta^x r'''(s)ds = \int_\eta^x \{f'''(s) - g'''(s)\}ds = \int_\eta^x f'''(s)ds.$$

Point c)

Si $x \in [x_1, x_0]$, nous trouvons, en utilisant l'indication, que

$$|r(x)| = \left|\int_{x_1}^{x} r'(s)ds\right| \leqslant \int_{x_1}^{x_0} |r'(s)|\,ds \leqslant h \max_{s\in[x_1,x_0]} |r'(s)| \qquad (1)$$

$$\text{et } |r'(x)| = \left|\int_{\xi_0}^{x} r''(s)ds\right| \leqslant \int_{\xi_0}^{x_0} |r''(s)|\,ds \leqslant h \max_{s\in[x_1,x_0]} |r''(s)|\,. \qquad (2)$$

De b), nous savons que si $x \in [x_1, x_0]$ nous avons

$$|r''(x)| = \left|\int_{\eta}^{x} f'''(s)ds\right|.$$

Alors

$$|r''(x)| \leqslant \left|\int_{\eta}^{x} |f'''(s)|\,ds\right| \leqslant \int_{x_2}^{x_0} |f'''(s)|\,ds, \quad \text{car } \eta \in [\xi_1, \xi_0] \subset [x_2, x_0],$$

et donc

$$|r''(x)| \leqslant 2h \max_{s\in[x_2,x_0]} |f'''(s)|\,.$$

Il découle de (2) que, pour $x \in [x_0, x_1]$, on a

$$|r'(x)| \leqslant 2h^2 \max_{s\in[x_2,x_0]} |f'''(s)|\,.$$

Donc $\max_{x\in[x_1,x_0]} |r'(x)| \leqslant 2h^2 \max_{s\in[x_2,x_0]} |f'''(s)|$. Alors de (1), nous pouvons conclure que

$$|r(x)| \leqslant 2h^3 \max_{s\in[x_2,x_0]} |f'''(s)|\,, \forall x \in [x_1, x_0].$$

Idem pour l'intervalle $[x_2, x_1]$ en prenant ξ_1 à la place de ξ_0.

Comparaison avec la formule de Taylor:

Soit G le polynôme de Taylor, de degré 2, de la fonction f autour du point x_0 i.e.

$$G(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2.$$

Nous avons l'égalité suivante

$$f(x) = G(x) + \frac{f'''(\xi)}{3!}(x - x_0)^3, \quad \text{où } \xi \in [\min(x_0, x), \max(x_0, x)].$$

Si $x \in [x_2, x_0]$, nous obtenons l'inégalité suivante

$$|f(x) - G(x)| \leqslant \frac{1}{3!} \max_{s\in[x_2,x_0]} \left|(x - x_0)^3\right| \cdot \max_{s\in[x_2,x_0]} |f'''(s)| \leqslant \frac{4h^3}{3} \max_{s\in[x_2,x_0]} |f'''(s)|\,.$$

Nous pouvons en déduire que les deux polynômes, g et G, donnent une approximation de f qui est de l'ordre de h^3. La formule de Taylor est légèrement plus précise, puisque $\frac{4}{3} < 2$, mais elle utilise par contre les dérivées de f qui ne sont pas toujours aisées à obtenir. (Par exemple, lorsque la fonction f n'est donnée que sous forme tabulée ou s'obtient par un calcul compliqué.)

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 1 avril 1997

Série 3

Exercice 1

On considère deux nombres α et β vérifiant la relation $-1 \leqslant -\alpha < \beta < \alpha \leqslant 1$.

On pose $t_1 = -\alpha$, $t_2 = \beta$ et $t_3 = \alpha$, on cherche à trouver trois nombres $\omega_1, \omega_2, \omega_3$ qui définissent la formule de quadrature

$$J(g) = \sum_{j=1}^{3} \omega_j g(t_j),$$

où g est une fonction continue quelconque donnée sur $[-1, +1]$.

a) Trouver ω_1, ω_2, ω_3, en fonction de α et β, pour que la formule de quadrature soit exacte pour les polynômes de degré 2, c'est-à-dire pour que $J(p) = \int_{-1}^{+1} p(t)dt$ pour tout polynôme p de degré 2.

b) Démontrer que si $\beta = 0$, alors la formule de quadrature est exacte pour les polynômes de degré 3.

Prendre $\alpha = 1$ et vérifier que l'on obtient la formule de Simpson.

c) Vérifier, par le calcul, que si $\beta = 0$ et si $\alpha = \sqrt{\frac{3}{5}}$, alors la formule de quadrature est exacte pour les polynômes de degré 5.

Vérifier que l'on obtient la formule de Gauss-Legendre à 3 points.

Exercice 2

Pour intégrer une fonction f continue sur l'intervalle $[a, b]$, on considère la formule d'intégration composite suivante (formule de Gauss à 3 points) :

$$L_h(f) = \sum_{i=0}^{N-1} \frac{h}{2} \overbrace{\left(\underset{\omega_1}{\frac{5}{9}} f\left(x_i + h\frac{1-\sqrt{\frac{3}{5}}}{2}\right) + \underset{\omega_2}{\frac{8}{9}} f\left(x_i + h\frac{1}{2}\right) + \underset{\omega_3}{\frac{5}{9}} f\left(x_i + h\frac{1+\sqrt{\frac{3}{5}}}{2}\right)\right)}^{J(g) = \sum_{i=1}^{3} \omega_i g(t_i)}$$

où $h = \frac{b-a}{N}$ et $x_i = a + ih$ avec $i = 0, 1, 2, \ldots, N$.

Utiliser cette formule de quadrature pour calculer numériquement $\int_0^4 e^x x^2 dx$ en divisant l'intervalle $[0, 4]$ en N parties égales. On prendra successivement $N = 10, 15, 20, 25, \ldots$ et on observera la convergence d'ordre 6. $\leq C \cdot h^6 = C \cdot \left(\frac{4}{N}\right)^6$

$h = \frac{b-a}{N} = \frac{4}{N}$

$x_i = a + i \cdot h = 0 + i \cdot \frac{4}{N} = i \cdot \frac{4}{N}$

ici, $f = e^x \cdot x^2$
alors il suffit de remplacer.
C'est connu.

)

Quelques conseils concernant l'utilisation de votre Mac :

Netscape :

- Allumez le Mac, `username:student` sans mot de passe, tapez votre nom ainsi que votre date de naissance. Lancez l'application `Netscape`, menu `bookmark`, sélectionnez `analyse numérique`. Vous êtes alors sur la page d'accueil du groupe d'analyse et simulation numériques du prof. J. Rappaz. Vous pouvez également aboutir à cette page depuis la page d'accueil du DMA, à l'adresse `http://dmawww.epfl.ch`.
- Récupérez l'exercice de programmation de la semaine sur votre Mac. L'application `Think Pascal` s'ouvre automatiquement. Le fichier est automatiquement placé dans le dossier `Netscape f`.

Think Pascal:

- Complétez le programme. Vérifier la syntaxe menu `run`, sous-menu `check syntax`.
- menu `project`, sous-menu `new project`, allez dans le dossier `Netscape f` et donnez le nom `integr` à votre projet.
- menu `project`, sous-menu `add file`, ajoutez le fichier `integr.pas`.
- menu `run`, sous-menu `build`, le programme `integr` est construit.
- menu `run`, sous-menu `go`, le programme `integr` est éxécuté.

Avant de partir :

- Sélectionnez à nouveau l'application `Netscape`.
- Retournez sur la page d'accueil du prof. J. Rappaz.
- Indiquez votre présence.

)

ANALYSE NUMÉRIQUE le 1er avril 1997

Exercice de programmation

Série 3

Intégration numérique

Position du problème :

Soit f une fonction réelle définie sur un intervalle $[a, b]$. Le but de cet exercice est de compléter et d'utiliser un programme permettant d'approximer

$$\int_a^b f(x)\, dx \tag{1}$$

en utilisant la formule du trapèze ainsi que la formule de Gauss-Legendre à 2 points d'intégration.

Pour ce faire l'intervalle $[a, b]$ est partitionné en N intervalles $[x_i, x_{i+1}]$, $i = 0, 1, 2, ..., N-1$ de même longueur $h = (b-a)/N$. Notons $\alpha = (1 - 1/\sqrt{3})/2$, $\beta = (1 + 1/\sqrt{3})/2$ et

$$\begin{aligned} L_h^{trap}(f) &= \frac{h}{2}\sum_{i=0}^{N-1}\Big(f(x_i) + f(x_{i+1})\Big), \\ L_h^{Gau2}(f) &= \frac{h}{2}\sum_{i=0}^{N-1}\Big(f(x_i + h\alpha) + f(x_i + h\beta)\Big) \end{aligned} \tag{2}$$

les deux approximations de (1), correspondant à la formule du trapèze et à la formule de Gauss-Legendre à 2 points, respectivement. Le théorème p. 22 du polycopié nous assure que, si f est 4 fois continûment dérivable, alors il existe une constante C indépendante de h telle que

$$\begin{aligned} &|\int_a^b f(x)\, dx - L_h^{trap}(f)| \le Ch^2, \\ &|\int_a^b f(x)\, dx - L_h^{Gau2}(f)| \le Ch^4. \end{aligned} \tag{3}$$

Travail demandé :

Le programme `integr.pas` est à votre disposition. A partir de la donnée de f, a, b, N ce programme permet d'approcher (1). Ce programme est incomplet, en particulier vous devez programmer les formules (2).

Outils informatiques :

Les logiciels à votre disposition sont :

- `Netscape` pour aller chercher le fichier `integr.pas` sur la page WWW du prof. J. Rappaz.
- `Think Pascal` pour compléter le programme, le compiler et l'éxécuter.

Questions :

Soit $a = 0$, $b = 1$ et f la fonction définie par $f(x) = x^2e^x$. Vérifiez les ordres de convergence annoncés en (3). Par exemple, pour la formule des trapèzes, vérifiez que l'erreur est approximativement divisée par 4 chaque fois que N est multiplié par 2.

vérifié.

```
program Integration_numerique;


{fonction a integrer numeriquement}

        function f (x: double): double;
        begin

                f := x * x * exp(x);        f(x) = x²·e^x
        end;

        procedure trapeze (N: integer; a, b: double; var lh: double);

                var
                        i: integer;
                        h, xi: double;
                             ,xi2
        begin

                h := (b - a) / N;
                lh := 0;

                for i := 0 to (N - 1) do
                        begin
                                xi := a + (i * h);  xi2 := a + ((i+1)·h)
                                lh := lh + {************ A COMPLETER *************}
                        end;                 f(xi) + f(xi2)

                lh := lh * h / 2;

        end;
        procedure gauss (N: integer; a, b: double; var lh: double);

                var
                        i: integer;
                        h, alpha, beta, xi: double;

        begin

                alpha := (1.0 - 1.0 / sqrt(3.0)) / 2.0;
                beta := (1.0 + 1.0 / sqrt(3.0)) / 2.0;
                h := (b - a) / N;
                lh := 0;

                for i := 0 to (N - 1) do
                        begin
                                xi := a + (i * h);
                                lh := lh + {************ A COMPLETER *************}
                        end;                 f(xi + h·alpha) + f(xi + h·beta)

                lh := lh * h / 2;

        end;

        type
                reponse = (s, c, t, g);

        var
                N: integer;
                a, b, lh, exact: double;
                rep: reponse;

begin{Integration_numerique}
        a := 0;
        b := 1;
        exact := exp(1) - 2;

        repeat
                write('Entrez N :');
                readln(N);
                write('tapez t (trapeze) ou g (Gauss) : ');
                readln(rep);
                if rep = t then
                        trapeze(N, a, b, lh)
                else
                        gauss(N, a, b, lh);
                writeln;
                writeln('Erreur : ', abs(exact - lh) : 16 : 15);
                writeln('tapez s pour stopper, c pour continuer');
                readln(rep);
        until rep = s;

end.{Integration_numerique}
```

ANALYSE NUMÉRIQUE
INFO, PII, PII-Fac
Prof. J. Rappaz

Le 1 avril 1997

Corrigé 3

Exercice 1

Point a)

Si la formule de quadrature est exacte pour tout polynôme de degré 2, elle l'est en particulier pour les polynômes φ_i de la base de Lagrange $(\varphi_i)_{j=1}^{3}$ de degré 2 associée aux points t_1, t_2 et t_3.

En utilisant la propriété $\varphi_i(t_j) = \delta_{ij}$ de ces polynômes, avec δ_{ij} le symbole de Kronecker, nous obtenons que

$$J(\varphi_i) \overset{\text{déf.}}{=} \sum_{j=1}^{3} \omega_j \varphi_i(t_j) = \omega_i.$$

De plus, par hypothèse, nous devons avoir $J(\varphi_i) = \int_{-1}^{1} \varphi_i(t)dt$. Ainsi, nous trouvons

$$\omega_i = J(\varphi_i) = \begin{cases} \displaystyle\int_{-1}^{1} \frac{t-\beta}{\mp\alpha-\beta} \cdot \frac{t \mp \alpha}{\mp 2\alpha} dt = \frac{1}{3} \cdot \frac{1 \pm 3\beta\alpha}{\alpha(\alpha \pm \beta)} & \text{si } i = \begin{matrix} 1, \\ 3, \end{matrix} \\ \displaystyle\int_{-1}^{1} \frac{t+\alpha}{\beta+\alpha} \cdot \frac{t-\alpha}{\beta-\alpha} dt = \frac{2}{3} \cdot \frac{3\alpha^2 - 1}{(\alpha+\beta)(\alpha-\beta)} & \text{si } i = 2, \end{cases}$$

puisque que les polynômes φ_i sont définis par $\varphi_i(t) = \prod_{\substack{j=1 \\ j \neq i}}^{3} \frac{t - t_j}{t_i - t_j}, \quad i = 1, 2, 3.$

Remarque:

Puisque la formule de quadrature et l'intégration sont linéaires, il suffit de vérifier que l'égalité $J(\zeta_i) = \int_{-1}^{1} \zeta_i(t)dt$ est vraie pour chaque élément d'une base $(\zeta_i)_{i=0}^{r}$ quelconque des polynômes de degré r, pour qu'elle le soit pour tous les polynômes de degré r.

Point b)

Choisissons la base $(t^3, \varphi_1(t), \varphi_2(t), \varphi_3(t))$ des polynômes de degré 3, où les φ_i, avec $i = 1, 2, 3$, sont définis ci-dessus.

Au point a), nous avons déjà démontré que $J(\varphi_i) = \int_{-1}^{1} \varphi_i(t)dt$, pour $i = 1, 2, 3$. Si $\beta = 0$, c'est-à-dire si $t_1 = -\alpha$, $t_2 = 0$ et $t_3 = \alpha$, les poids valent $\omega_1 = \omega_3 = \frac{1}{3\alpha^2}$ et $\omega_2 = \frac{2}{3}\frac{3\alpha^2-1}{\alpha^2}$. La formule de quadrature J, appliquée au polynôme t^3, se développe

$$J(t^3) \overset{\text{déf.}}{=} \sum_{j=1}^{3} \omega_j t_j^3 = -\omega_1\alpha^3 + \omega_3\alpha^3 = 0, \quad \text{qui est bien égale à } \int_{-1}^{1} t^3 dt = 0.$$

D'après la remarque, nous concluons qu'elle est exacte pour les polynômes de degré 3.

En particulier, si de plus $\alpha = 1$, c'est-à-dire si $t_1 = -1$, $t_2 = 0$ et $t_3 = 1$, les poids valent $\omega_1 = \omega_3 = \frac{1}{3}$ et $\omega_2 = \frac{4}{3}$. La formule de quadrature est alors

$$J(g) \overset{\text{déf.}}{=} \sum_{j=1}^{3} \omega_j g(t_j) = \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1),$$

dans laquelle nous reconnaissons la formule de Simpson.

Point c)

Si $\beta = 0$ et $\alpha = \sqrt{\frac{3}{5}}$, c'est-à-dire si $t_1 = -\sqrt{\frac{3}{5}}$, $t_2 = 0$, $t_3 = \sqrt{\frac{3}{5}}$, les poids valent $\omega_1 = \omega_3 = \frac{5}{9}$ et $\omega_2 = \frac{8}{9}$. La formule de quadrature est alors $J(g) = \frac{5}{9}g\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}g(0) + \frac{5}{9}g\left(\sqrt{\frac{3}{5}}\right)$.

Choisissons la base $(\xi_i)_{i=0}^{5}$ des polynômes de degré 5, où $\xi_i(t) = t^i$ pour $i = 3, 4, 5$ et où $\xi_i(t) = \varphi_{i+1}(t)$ pour $i = 0, 1, 2$, avec les φ_i, $i = 1, 2, 3$, définis au point a).

Aux points a) et b), nous avons déjà démontré que $J(\xi_i) = \int_{-1}^{1} \xi_i(t)dt$, pour $i = 0, 1, 2, 3$. Nous vérifions facilement que

$$J(t^4) = \frac{5}{9}\left(-\sqrt{\tfrac{3}{5}}\right)^4 + \frac{8}{9}0^4 + \frac{5}{9}\left(\sqrt{\tfrac{3}{5}}\right)^4 = 2 \cdot \frac{5}{9} \cdot \frac{9}{25} \quad \text{qui est égale à } \int_{-1}^{1} t^4 dt = \frac{2}{5} \text{ et que}$$

$$J(t^5) = \frac{5}{9}\left(-\sqrt{\tfrac{3}{5}}\right)^5 + \frac{8}{9}0^5 + \frac{5}{9}\left(\sqrt{\tfrac{3}{5}}\right)^5 = 0 \quad \text{qui est égale à } \int_{-1}^{1} t^5 dt = 0.$$

D'après la remarque, nous concluons que la formule est exacte pour les polynômes de degré 5.

Pour vérifier que nous obtenons la formule de Gauss-Legendre à 3 points, il suffit de voir que les points d'intégration $-\sqrt{\frac{3}{5}}$, 0, $\sqrt{\frac{3}{5}}$ correspondent bien aux points de Gauss, c'est-à-dire qu'ils sont bien les zéros du polynôme de Legendre $\frac{1}{2^3 3!}\frac{d^3}{dt^3}(t^2-1)^3 = \frac{1}{2}t\left(5t^2 - 3\right)$. Les poids sont évidemment les mêmes que ceux de la formule de Gauss-Legendre puisque nous utilisons les mêmes formules pour les calculer.

Exercice 2

La formule d'intégration composite utilisée, basée sur la formule de Gauss à 3 points est, d'après l'exercice 1.c), exacte pour des polynômes de degré 5. Puisque la fonction $e^x x^2$ appartient à l'espace $C^\infty(\mathbb{R})$, un théorème du cours s'applique. Il nous dit

$$\exists C \in \mathbb{R} \quad \text{t.q.} \quad \left| L_h(e^x x^2) - \int_a^b e^x x^2 dx \right| \leqslant Ch^6, \quad \text{ou} \quad \frac{\left| L_h(e^x x^2) - \int_a^b e^x x^2 dx \right|}{h^6} \leqslant C, \quad \forall a, b \in \mathbb{R}.$$

L'intégrale exacte, utilisée comme référence, se calcule par intégration par parties. Elle vaut

$$\int_0^4 x^2 e^x dx = \left(e^x(x^2 - 2x + 2)\right)\Big|_0^4 = 10e^4 - 2.$$

Les résultats obtenus par la formule composite en divisant l'intervalle $[0, 4]$ en N parties égales, avec N valant successivement 5, 10, 15, 20, ..., 50 sont reportés dans le tableau ci-dessous.

N	h^*	$L_h(e^x x^2)^\star$	$\left\lvert L_h(e^x x^2) - \int_0^4 e^x x^2 dx \right\rvert^\star$	$\frac{\lvert L_h(e^x x^2) - \int_0^4 e^x x^2 dx \rvert}{h^6}^\star$
5	$8.000000e-01$	$5.4398098046070e+02$	$5.1987073857163e-04$	$1.9831494849076e-03$
10	$4.000000e-01$	$5.4398149200940e+02$	$8.3220402302686e-06$	$2.0317481030929e-03$
15	$2.666667e-01$	$5.4398149959753e+02$	$7.3391061050643e-07$	$2.0409425165527e-03$
20	$2.000000e-01$	$5.4398150020062e+02$	$1.3082694749755e-07$	$2.0441710546493e-03$
25	$1.600000e-01$	$5.4398150029712e+02$	$3.4320692066103e-08$	$2.0456726590456e-03$
30	$1.333333e-01$	$5.4398150031994e+02$	$1.1498627827677e-08$	$2.0465087124943e-03$
35	$1.142857e-01$	$5.4398150032688e+02$	$4.5611159293912e-09$	$2.0470074767187e-03$
40	$1.000000e-01$	$5.4398150032940e+02$	$2.0472725736909e-09$	$2.0472725736909e-03$
45	$8.888889e-02$	$5.4398150033043e+02$	$1.0097664926434e-09$	$2.0470860085178e-03$
50	$8.000000e-02$	$5.4398150033091e+02$	$5.3682924772147e-10$	$2.0478410633906e-03$
∞		$5.4398150033144e+02$	formule exacte	

Calcul à 16 chiffres significatifs, impression de 14 (⋆), respectivement 7 (∗) chiffres.

La dernière colonne met en évidence la constante C du théorème et montre ainsi la convergence d'ordre 6.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 8 avril 1997

Série 4

Exercice 1 (Exercice 1, examen propédeutique du 3 juillet 1996, section INFO.)

Soient a, b, c et d quatre nombres non nuls donnés. Si N est un entier positif, on considère la $N \times N$-matrice A et le N-vecteur $\vec{b}$ définis par

$$A = \begin{pmatrix} a & d & & & & \\ & a & d & & 0 & \\ & & \ddots & \ddots & & \\ & 0 & & \ddots & \ddots & \\ & & & & a & d \\ c & c & \dots & \dots & c & c \end{pmatrix} \quad \text{et} \quad \vec{b} = \begin{pmatrix} b \\ b \\ b \\ \vdots \\ b \\ b \end{pmatrix}.$$

En supposant que A soit une matrice régulière, on demande d'écrire un algorithme de résolution du système $A\vec{x} = \vec{b}$. Cet algorithme aura

- pour entrées, les valeurs a, c, d, N et le vecteur $\vec{b}$ de composantes $b_j = b$, $1 \leqslant j \leqslant N$;
- pour sortie, le vecteur $\vec{b}$ de composantes $b_j = x_j$, $1 \leqslant j \leqslant N$, où les valeurs x_j sont les composantes de la solution $\vec{x}$.

Important: l'algorithme ne devra **pas utiliser d'autres tableaux** *(tableau = variable indicée)* que le vecteur $\vec{b}$.

Exercice 2

La matrice $\tilde{A} = \begin{pmatrix} a_1 & d_1 & & & \\ & a_2 & d_2 & 0 & \\ & & \ddots & \ddots & \\ & & & \ddots & \ddots & \\ & 0 & & a_{N-1} & d_{N-1} \\ & & & & a_N \end{pmatrix}$ où $a_1, a_2, \dots, a_N$ et $d_1, d_2, \dots, d_{N-1}$ sont des réels et N est un entier positif, est une généralisation de la matrice A de l'exercice 1, avec $c = 0$. Dans le cas particulier où $a_i = \sqrt{\frac{i+1}{i}}$ pour $i = 1, 2, \dots, N$, $d_i = -\sqrt{\frac{i}{i+1}}$ pour $i = 1, 2, \dots, N-1$, calculer $\chi(\tilde{A})$.

pas tout à fait. $a_n \neq 0$.

↳ norme spectrale

Indications:

i) Vérifier que $\lambda_j = 2(1 - \cos \frac{\pi}{N+1} j)$, $j = 1, 2, \dots, N$ est valeur propre de la $N \times N$-matrice M pour le vecteur propre $\vec{x}_j$ de composantes $(x_j)_k = \sin \frac{j\pi k}{N+1}$, $1 \leqslant k \leqslant N$, avec

$$M = \begin{pmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & 0 & \\ & -1 & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & \ddots & 2 & -1 \\ & & & & -1 & 2 \end{pmatrix}.$$

$$\chi(A) = \sqrt{\frac{\max_{\lambda \in \sigma(A^T A)} \lambda}{\min_{\lambda \in \sigma(A^T A)} \lambda}}$$

$$(A - \lambda \cdot I_n) \cdot X = 0$$

ii) Utiliser les valeurs données de λ_j, $j = 1, 2, \dots, N$.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 8 avril 1997

Corrigé 4

Exercice 1

Point a)

Principe de la résolution par élimination de Gauss

- étape 0 : initialement, le système linéaire $A\vec{x} = \vec{b}$ s'écrit

$$ax_i + dx_{i+1} = b_i,\ 1 \leqslant i \leqslant N-1, \qquad \text{(i)}$$

$$\sum_{j=1}^{N} cx_j = b_N. \qquad \text{(N)}$$

- étape 1a : normalisation de la $1^{\text{ère}}$ ligne.
  Diviser la $1^{\text{ère}}$ équation par le 1^{er} pivot, i.e. multiplier par $p = a^{-1}$, pour obtenir

$$x_1 + \frac{d}{a}x_2 = \frac{b_1}{a}, \qquad \left(1,\ \frac{d}{a},\ 0,\ 0 \dots 0\right)\vec{x} = \frac{b_1}{a} \qquad \text{(1')}$$

- étape 1b : élimination de x_1 dans la $N^{\text{ème}}$ ligne.
  Soustraire c fois l'équation (1') de la $N^{\text{ème}}$ pour obtenir

$$\left(c - c\frac{d}{a}\right)x_2 + \sum_{j=3}^{N} cx_j = \left(b_N - c\frac{b_1}{a}\right). \qquad \text{(N')}$$

- étape 2a : normalisation de la $2^{\text{ème}}$ ligne.
  Diviser l'équation (i) par le $2^{\text{ème}}$ pivot, i.e. multiplier par $p = a^{-1}$, pour obtenir

$$x_2 + \frac{d}{a}x_3 = \frac{b_2}{a}, \qquad \left(0, 1,\ \frac{d}{a},\ 0 \dots ,0\right)\cdot\vec{x} = \frac{b_2}{a} \qquad \text{(2')}$$

- étape 2b : élimination de x_2 dans la $(N')^{\text{ème}}$ ligne.
  Soustraire $(c - c\frac{d}{a})$ fois l'équation (2') de la $(N')^{\text{ème}}$ pour obtenir

$$\left(c - \left(c - c\frac{d}{a}\right)\frac{d}{a}\right)x_3 + \sum_{j=4}^{N} cx_j = \left(b_N - c\frac{b_1}{a} - \left(c - c\frac{d}{a}\right)\frac{b_2}{a}\right). \qquad \text{(N'')}$$

- étape 3a : normalisation de la $3^{\text{ème}}$ ligne.
- étape 3b : élimination de x_3 dans la $(N'')^{\text{ème}}$ ligne.
- etc.
- étape Na : normalisation de la $N^{\text{ème}}$ ligne.

Finalement, nous obtenons un système d'équations triangulaire supérieur, dont les coefficients sont nuls, sauf ceux des $i^{\text{ème}}$ et $(i+1)^{\text{ème}}$ inconnues de la $i^{\text{ème}}$ ligne qui valent respectivement 1 et $\frac{d}{a}$ pour les lignes 1 à $N-1$, et celui de la $N^{\text{ème}}$ inconnue de la $N^{\text{ème}}$ ligne qui vaut 1.

ici on normalise les premiers coefficients des lignes.

la subtilité ici c'est d'avoir choisi une variable supplémentaire C qui est celle dans laquelle on stocke les valeurs des premiers coeff. non nuls de la dernière ligne.

Commentaire: Nous voyons que seuls les coefficients du second membre et celui de la $(i+1)^{\text{ème}}$ inconnue de la $N^{\text{ème}}$ ligne à la $i^{\text{ème}}$ étape doivent être stockés.

Algorithme

Initialisation de a, c, d; $b_i = b,\ i = 1$ à N;	Initialisation.
$d := \frac{d}{a}$;	Calcul des $(N-1)$ coefficients sur-diagonaux.
$C := c$;	Initialisation du coefficient de la $1^{\text{ère}}$ inconnue de la $N^{\text{ème}}$ ligne.
Faire pour $i = 1$ à $N-1$	
$\quad (b_i := \frac{b_i}{a};)$ on aurait pu le mettre dans une boucle avant.	Normalisation de la $i^{\text{ème}}$ ligne.
$\quad b_N := b_N - C\ b_i$; $\quad \underbrace{C := c - C\ d;}_{\text{c'est ça.}}$	Elimination de la $i^{\text{ème}}$ inconnue dans la $N^{\text{ème}}$ ligne.
$b_N := \frac{b_N}{C}$;	Normalisation de la $N^{\text{ème}}$ ligne et calcul de la $N^{\text{ème}}$ inconnue.
Faire pour $i = N-1$ à 1 par pas de -1 $\quad b_i := b_i - d\ b_{i+1}$;	Calcul de la $i^{\text{ème}}$ inconnue.

Exercice 2

Nous noterons $\sigma(\cdot)$ l'ensemble des valeurs propres d'une $N \times N$-matrice.
Par définition, le nombre de condition spectral de la $N \times N$-matrice B est

$$\chi(B) = |||B||| \cdot |||B^{-1}|||,$$

avec $|||B|||$ la norme spectrale de B. D'une remarque du cours, nous savons que la norme spectrale vaut

$$|||B||| = \sqrt{\max_{\omega \in \sigma(B^T B)} |\omega|}.$$

Nous savons aussi que les matrices symétriques S ont la propriété

$$\max_{\lambda \in \sigma(S^{-1})} |\lambda| = \left(\min_{\lambda \in \sigma(S)} |\lambda|\right)^{-1}$$

et que $B^T B$ est symétrique. La norme spectrale $|||B^{-1}|||$ peut alors se calculer* ainsi

$$|||B^{-1}||| = \sqrt{\max_{\omega \in \sigma(B^{-T}B^{-1})} |\omega|} = \sqrt{\max_{\omega \in \sigma((BB^T)^{-1})} |\omega|} = \sqrt{\left(\min_{\omega \in \sigma(BB^T)} |\omega|\right)^{-1}}.$$

Si nous démontrons que $\sigma(B^T B) = \sigma(BB^T)$, nous obtenons que

$$|||B^{-1}||| = \sqrt{\left(\min_{\omega \in \sigma(B^T B)} |\omega|\right)^{-1}}.$$

En résumé, il suffit de calculer $B^T B$, de démontrer que $\sigma(B^T B) = \sigma(BB^T)$, de trouver l'ensemble de valeurs propres $\sigma(B^T B)$, d'en tirer les valeurs propres maximale et minimale, pour finalement obtenir le nombre de condition spectral :

$$\chi(B) = \sqrt{\frac{\max_{\omega \in \sigma(B^T B)} |\omega|}{\min_{\omega \in \sigma(B^T B)} |\omega|}}.$$

Appliquons ce raisonnement à la matrice $\tilde{A}$.

- **Calculons $\tilde{A}^T\tilde{A}$.**

Le produit $\tilde{A}^T\tilde{A}$ donne $\begin{pmatrix} a_1^2 & a_1d_1 & & & & \\ a_1d_1 & a_2^2+d_1^2 & \ddots & & 0 & \\ & \ddots & \ddots & a_id_i & & \\ & & a_id_i & a_{i+1}^2+d_i^2 & \ddots & \\ & 0 & & \ddots & \ddots & a_{N-1}d_{N-1} \\ & & & & a_{N-1}d_{N-1} & a_N^2+d_{N-1}^2 \end{pmatrix}$. Nous calculons facilement que $a_1^2 = 2$ et que $a_id_i = -1$ et $a_{i+1}^2+d_i^2 = \frac{i+2}{i+1}+\frac{i}{i+1} = 2$ pour $i = 1, \ldots, N-1$, i.e. que $\tilde{A}^T\tilde{A} = M$.

- **Démontrons que $\sigma\left(\tilde{A}^T\tilde{A}\right) = \sigma\left(\tilde{A}\tilde{A}^T\right)$.**

trop compliqué

Soit une valeur propre λ de $\tilde{A}^T\tilde{A}$ associée au vecteur propre $\vec{x}$, i.e.

$$\tilde{A}^T\tilde{A}\vec{x} = \lambda\vec{x}.$$

Comme $\tilde{A}$ est régulière, puisque son déterminant vaut $\sqrt{N+1}$, nous pouvons multiplier l'égalité ci-dessus par $\tilde{A}$ et obtenir

$$\tilde{A}\tilde{A}^T\left(\tilde{A}\vec{x}\right) = \lambda\left(\tilde{A}\vec{x}\right).$$

Autrement dit, la valeur propre λ de $\tilde{A}^T\tilde{A}$ est aussi valeur propre de $\tilde{A}\tilde{A}^T$, mais associée au vecteur propre $\tilde{A}\vec{x}$.
Puisque $\operatorname{card}\left(\sigma\left(\tilde{A}^T\tilde{A}\right)\right) = \operatorname{card}\left(\sigma\left(\tilde{A}\tilde{A}^T\right)\right) = N$, nous concluons que $\sigma\left(\tilde{A}^T\tilde{A}\right) = \sigma\left(\tilde{A}\tilde{A}^T\right)$.

- **Vérifions que $\sigma\left(\tilde{A}^T\tilde{A}\right) = \left\{\lambda_i = 2\left(1-\cos\frac{\pi}{N+1}j\right),\ \text{avec } i = 1, 2, \ldots, N\right\}$.**

Définissons $(x_j)_0 = \sin\frac{0j\pi}{N+1} = 0$ et $(x_j)_{N+1} = \sin\frac{(N+1)j\pi}{N+1} = 0$, $1 \leqslant j \leqslant N$. Nous obtenons que la $k^{\text{ème}}$ composante du N-vecteur $M\vec{x}_j$ s'écrit

$$\begin{aligned}
\left(M\vec{x}_j\right)_k &= -(x_j)_{k-1} + 2(x_j)_k - (x_j)_{k+1} \\
&= -\sin\frac{(k-1)j\pi}{N+1} + 2\sin\frac{kj\pi}{N+1} - \sin\frac{(k+1)j\pi}{N+1} \\
&\overset{(1)}{=} -\sin\frac{kj\pi}{N+1}\cos\frac{j\pi}{N+1} + \sin\frac{j\pi}{N+1}\cos\frac{kj\pi}{N+1} + 2\sin\frac{kj\pi}{N+1} \\
&\quad - \sin\frac{kj\pi}{N+1}\cos\frac{j\pi}{N+1} - \sin\frac{j\pi}{N+1}\cos\frac{kj\pi}{N+1} \\
&= 2\left(1-\cos\frac{j\pi}{N+1}\right)\sin\frac{kj\pi}{N+1} \\
&= \lambda_j\,(x_j)_k,
\end{aligned}$$

pour $1 \leqslant k \leqslant N$. Ce qui prouve que les λ_j sont bien les valeurs propres de $\tilde{A}^T\tilde{A}$.

- **Trouvons le minimum et le maximum de $\sigma\left(\tilde{A}^T\tilde{A}\right)$.**

Les indices j valent $1, 2, \ldots, N$. Les $\frac{j\pi}{N+1}$ appartiennent donc à l'intervalle $]0, \pi[$ sur lequel le cosinus est strictement décroissant. En conséquence,

$$\lambda_N = \max_{\lambda\in\sigma(\tilde{A}^T\tilde{A})}|\lambda| \quad \text{et} \quad \lambda_1 = \min_{\lambda\in\sigma(\tilde{A}^T\tilde{A})}|\lambda|.$$

1. Rappel: $\sin(\alpha\pm\beta) = \sin\alpha\cos\beta \pm \cos\alpha\sin\beta$.

- **Conclusion.**

Finalement, nous obtenons que

$$\begin{aligned}
\chi(\tilde{A}) &= \sqrt{\frac{\lambda_N}{\lambda_1}} = \sqrt{\frac{1-\cos\frac{N\pi}{N+1}}{1-\cos\frac{\pi}{N+1}}} \\
&\overset{(2)}{=} \frac{\sin\frac{N\pi}{2(N+1)}}{\sin\frac{\pi}{2(N+1)}} \overset{(1)}{=} \frac{\sin\frac{\pi}{2}\cos\frac{\pi}{2(N+1)} - \cos\frac{\pi}{2}\sin\frac{\pi}{2(N+1)}}{\sin\frac{\pi}{2(N+1)}} = \cot\frac{\pi}{2(N+1)}.
\end{aligned}$$

- **Commentaire.**

Si N devient grand, le nombre de condition spectral se comporte comme

$$\chi(\tilde{A}) = \frac{1}{\tan\frac{\pi}{2(N+1)}} \overset{(3)}{=} \frac{1}{\frac{\pi}{2(N+1)} + \frac{1}{3}\left(\frac{\pi}{2(N+1)}\right)^3 + \ldots} \overset{(4)}{=} \frac{2(N+1)}{\pi} - \frac{\pi}{6(N+1)} + \ldots$$

***Remarque**

Une autre manière de calculer $|||B^{-1}|||$, qui nous évite de démontrer que $\sigma\left(B^TB\right) = \sigma\left(BB^T\right)$, est basée sur la propriété $|||B^T||| = |||B|||$, pour toute $N\times N$-matrice B. En effet, nous avons alors

$$|||B^{-1}||| = |||B^{-T}||| = \sqrt{\max_{\omega\in\sigma\left((B^{-T})^TB^{-T}\right)}|\omega|} = \sqrt{\max_{\omega\in\sigma\left((B^TB)^{-1}\right)}|\omega|} = \sqrt{\left(\min_{\omega\in\sigma(B^TB)}|\omega|\right)^{-1}}.$$

Propriété: $|||B^T||| = |||B|||$, pour toute $N\times N$-matrice B :
Preuve :

Rappelons la propriété suivante de la norme euclidienne: $\|x\| = \max_{\vec{y}\in\mathbb{R}^N}\frac{|y^Tx|}{\|y\|}, \quad \forall\vec{x}\in\mathbb{R}^N.$

Nous pouvons alors écrire

$$|||B||| \overset{\text{déf.}}{=} \max_{\vec{x}\in\mathbb{R}^N}\frac{\|Bx\|}{\|x\|} = \max_{\vec{x}\in\mathbb{R}^N}\max_{\vec{y}\in\mathbb{R}^N}\frac{|y^TBx|}{\|x\|\,\|y\|} = \max_{\vec{y}\in\mathbb{R}^N}\max_{\vec{x}\in\mathbb{R}^N}\frac{|x^TB^Ty|}{\|x\|\,\|y\|} = \max_{\vec{y}\in\mathbb{R}^N}\frac{\|B^Ty\|}{\|y\|} \overset{\text{déf.}}{=} |||B^T|||.$$

□

2. Rappel: $2\sin^2\left(\frac{\alpha}{2}\right) = 1-\cos\alpha$.
3. Rappel: le développement limité autour de 0 de $\tan x = x + \frac{1}{3}x^3 + \ldots$.
4. Rappel: le développement limité autour de 0 de $\frac{1}{1+x} = 1 - x + \ldots$.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 15 avril 1997

Série 5

Exercice 1

Soit a, c, d, e, f cinq nombres réels donnés. On définit une $N \times N$ matrice A par

$$A = \begin{pmatrix} a & c & & & & & d \\ e & a & c & & & 0 & d \\ & e & \ddots & \ddots & & & \vdots \\ & & \ddots & \ddots & \ddots & & \vdots \\ & 0 & & \ddots & \ddots & c & d \\ & & & & e & a & c \\ f & f & \dots & \dots & f & e & a \end{pmatrix}$$

a) En supposant que les nombres a, c, d, e, f, sont tels que toutes les sous-matrices principales de A soient régulières, écrire un algorithme de décomposition LU de la matrice A.

(On prendra garde à **ne pas remplir** un tableau a_{ij}, $1 \leqslant i, j \leqslant N$, à N^2 éléments.)

Pour le faire, on notera
- $\vec{a}$ le N-vecteur des coefficients diagonaux,
- $\vec{c}$ le $(N-1)$-vecteur des coefficients sur-diagonaux,
- $\vec{e}$ le $(N-1)$-vecteur des coefficients sous-diagonaux,
- $\vec{d}$ le $(N-2)$-vecteur formé des $(N-2)$ premières composantes de la dernière colonne,
- $\vec{f}$ le $(N-2)$-vecteur formé des $(N-2)$ premières composantes de la dernière ligne

et on montrera que L et U peuvent être entièrement mémorisées dans ces vecteurs !
$\Longrightarrow$ la forme est conservée

b) Ecrire un algorithme de résolution du système $A\vec{x} = \vec{b}$ en utilisant a).

c) *(Facultatif)* Programmer le point b).

Exercice 2

On définit la $N \times N$ matrice A comme dans l'exercice 1, en posant $e = c$ et $d = f = 0$.

a) Montrer que si $a \geqslant 2\,|c|$, $a > 0$, alors A est symétrique définie positive.

b) Ecrire un algorithme de Cholesky de A.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 15 avril 1997

Corrigé 5

Exercice 1

Point a)

Profils de L et U

Supposons que les matrices L et U aient les profils suivants :

$$L = \begin{pmatrix} a_1 & & & & & 0 \\ e_1 & a_2 & & & & \\ & e_2 & \ddots & & & \\ & & \ddots & \ddots & & \\ 0 & & & e_{N-2} & a_{N-1} & \\ f_1 & f_2 & \dots & f_{N-2} & e_{N-1} & a_N \end{pmatrix} \text{ et } U = \begin{pmatrix} 1 & c_1 & & & 0 & d_1 \\ & 1 & c_2 & & & d_2 \\ & & \ddots & \ddots & & \vdots \\ & & & \ddots & c_{N-2} & d_{N-2} \\ 0 & & & & 1 & c_{N-1} \\ & & & & & 1 \end{pmatrix}.$$

Nous remarquons en exécutant le produit LU que les régions sous-diagonales nulles de L, respectivement sur-diagonales de U, restent nulles dans la matrice LU. Puisque toutes les sous-matrices principales de A sont régulières par hypothèse, nous savons par le théorème du cours concernant la décomposition LU que les matrices L et U existent et sont uniques. Elles devront donc avoir les profils supposés ci-dessus.

Résolution

En multipliant L par la $1^{\text{ère}}$ colonne de U, nous obtenons directement la $1^{\text{ère}}$ colonne de A. Ainsi

$$a_1 = a, \quad e_1 = e \quad \text{et} \quad f_1 = f.$$

En multipliant la $1^{\text{ère}}$ ligne de L par les $2^{\text{ème}}$ et $N^{\text{ème}}$ colonnes de U, nous obtenons respectivement les $2^{\text{ème}}$ et $N^{\text{ème}}$ termes de la $1^{\text{ère}}$ ligne de A. Ainsi, nous avons

$$a_1c_1 = c \quad \text{et} \quad a_1d_1 = d,$$

où, puisque a_1 est connu,

$$c_1 = \frac{c}{a_1} \quad \text{et} \quad d_1 = \frac{d}{a_1}.$$

Nous avons ainsi trouvé la $1^{\text{ère}}$ colonne de L et la $1^{\text{ère}}$ ligne de U, ou les $1^{\text{ères}}$ composantes de $\vec{a}$, $\vec{c}$, $\vec{d}$, $\vec{e}$ et $\vec{f}$.

En multipliant, les $2^{\text{ème}}$, $(2+1)^{\text{ème}}$ et $N^{\text{ème}}$ lignes de L par la $2^{\text{ème}}$ colonne de U, nous obtenons

$$e_{2-1}c_{2-1} + a_2 = a, \quad e_2 = e \quad \text{et} \quad f_{2-1}c_{2-1} + f_2 = f.$$

Puisque les $(2-1)^{\text{ème}}$ composantes de $\vec{c}$, $\vec{e}$ et $\vec{f}$ sont connues, nous pouvons déterminer

$$a_2 = a - e_{2-1}c_{2-1}, \quad e_2 = e \quad \text{et} \quad f_2 = f - f_{2-1}c_{2-1}.$$

En multipliant $2^{\text{ème}}$ ligne de L par les $(2+1)^{\text{ème}}$ et $N^{\text{ème}}$ colonnes de U, nous obtenons

$$a_2c_2 = c \quad \text{et} \quad e_{2-1}d_{2-1} + a_2d_2 = d.$$

Puisque nous connaissons, les $(2-1)^{\text{ème}}$ composantes de $\vec{e}$ et $\vec{d}$, et la $2^{\text{ème}}$ de $\vec{a}$, nous pouvons déterminer

$$c_2 = \frac{c}{a_2} \quad \text{et} \quad d_2 = \frac{d - e_{2-1}d_{2-1}}{a_2}.$$

Nous avons ainsi trouvé la $2^{\text{ème}}$ colonne de L et la $2^{\text{ème}}$ ligne de U, ou les $2^{\text{ème}}$ composantes de $\vec{a}$, $\vec{c}$, $\vec{d}$, $\vec{e}$ et $\vec{f}$. Le même raisonnement peut s'appliquer pour calculer toutes les composantes encore inconnues, sauf les $(N-1)^{\text{ème}}$ de $\vec{a}$, $\vec{c}$ et $\vec{e}$ et la $N^{\text{ème}}$ de $\vec{a}$.

Ces dernières inconnues se déterminent facilement en utilisant toujours le même principe. En multipliant, les $(N-1)^{\text{ème}}$ et $N^{\text{ème}}$ lignes de L par la $(N-1)^{\text{ème}}$ colonne de U, nous obtenons

$$e_{N-2}c_{N-2} + a_{N-1} = a \quad \text{et} \quad f_{N-2}c_{N-2} + e_{N-1} = e,$$

qui implique que

$$a_{N-1} = a - e_{N-2}c_{N-2} \quad \text{et} \quad e_{N-1} = e - f_{N-2}c_{N-2}.$$

En multipliant la $(N-1)^{\text{ème}}$ ligne de L par la $N^{\text{ème}}$ colonne de U, nous obtenons

$$e_{N-2}d_{N-2} + a_{N-1}c_{N-1} = c, \quad \text{i.e.} \quad c_{N-1} = \frac{c - e_{N-2}d_{N-2}}{a_{N-1}}.$$

Finalement, la multiplication de la $N^{\text{ème}}$ ligne de L avec la $N^{\text{ème}}$ colonne de U donne

$$\sum_{i=1}^{N-2} f_id_i + e_{N-1}c_{N-1} + a_N = a, \quad \text{i.e.} \quad a_N = a - \sum_{i=1}^{N-2} f_id_i - e_{N-1}c_{N-1}.$$

Algorithme de décomposition LU :

$a_1 := a;$
$e_1 := e;$
$f_1 := f;$ } Calcul de la $1^{\text{ère}}$ colonne de L.
$c_1 := \frac{c}{a_1};$
$d_1 := \frac{d}{a_1};$ } Calcul de la $1^{\text{ère}}$ ligne de U.
Faire pour $i = 2$ à $N-2$
  $a_i := a - e_{i-1}c_{i-1};$
  $e_i := e;$
  $f_i := f - f_{i-1}c_{i-1};$ } Calcul de la $i^{\text{ème}}$ colonne de L.
  $c_i := \frac{c}{a_i};$
  $d_i := \frac{d - e_{i-1}d_{i-1}}{a_i};$ } Calcul de la $i^{\text{ème}}$ ligne de U.
$a_{N-1} := a - e_{N-2}c_{N-2};$
$e_{N-1} := e - f_{N-2}c_{N-2};$ } Calcul de la $(N-1)^{\text{ème}}$ colonne de L.
$c_{N-1} := \frac{c - e_{N-2}d_{N-2}}{a_{N-1}};$ Calcul de la $(N-1)^{\text{ème}}$ ligne de U.
$a_N := a - e_{N-1}c_{N-1};$
Faire pour $i = 1$ à $N-2$
  $a_N := a_N - f_id_i;$ } Calcul de a_N.

Point b)

Après avoir trouvé la décomposition LU, il reste à résoudre successivement deux systèmes triangulaires, premièrement le système triangulaire inférieur $L\vec{y} = \vec{b}$, deuxièmement le système supérieur $U\vec{x} = \vec{y}$. Ces types de systèmes ont été vu dans la série 4, exercice 1. Nous savons qu'ils donnent les algorithmes suivants, en stockant les vecteurs intermédiaires $\vec{x}$ et $\vec{y}$ dans le vecteur initial $\vec{b}$:

$b_1 := \frac{b_1}{a_1};$
Faire pour $i = 2$ à $N-1$
  $b_i := b_i - e_{i-1}b_{i-1};$
  $b_i := b_i/a_i;$
$b_N := b_N - e_{N-1}b_{N-1};$
Faire pour $i = 1$ à $N-2$
  $b_N := b_N - f_ib_i;$
$b_N := \frac{b_N}{a_N};$ } Algorithme de résolution $L\vec{y} = \vec{b}$.
$b_{N-1} := b_{N-1} - c_{N-1}b_N;$
Faire pour $i = N-2$ à 1 par pas de -1
  $b_i := b_i - c_ib_{i+1} - d_ib_N;$ } Algorithme de résolution $U\vec{x} = \vec{y}$.

Point c)

Le programme en C correspondant à l'algorithme ci-dessus, mais avec les indices décalés de 1 puisque l'indexation des vecteurs en C commence à 0 et avec une initialisation préalable, peut être :

```
#include <stdio.h>
#define  N    5

void main(void)
{
 int    i;
 double a[N], b[N], c[N-1], d[N-2], e[N-1], f[N-2];

 for (i=0; i<N; i++)    { a[i] = 2.; b[i] = 1.;}
 for (i=0; i<N-1; i++) { c[i] =-1.; e[i] =-1.;}
 for (i=0; i<N-2; i++) { d[i] = 0.; f[i] = 0.;}

 c[0] /= a[0];
 d[0] /= a[0];

 for(i=1; i<N-2; i++)
  {
   a[i]-= e[i-1]*c[i-1];
   f[i]-= f[i-1]*c[i-1];
   c[i]/= a[i];
   d[i] = (d[i]-e[i-1]*d[i-1])/a[i];
  }

 a[N-2]-= e[N-3]*c[N-3];
 e[N-2]-= f[N-3]*c[N-3];
 c[N-2] = (c[N-2]-e[N-3]*d[N-3])/a[N-2];

 a[N-1]-= e[N-2]*c[N-2];
 for(i=0; i<N-2; i++)
   a[N-1]-= f[i]*d[i];

 b[0]/= a[0];
 for (i=1; i<N-1; i++)
  {
   b[i]-= e[i-1]*b[i-1];
   b[i]/= a[i];
  }

 b[N-1]-= e[N-2]*b[N-2];
 for (i=0; i<N-2; i++)
   b[N-1]-= f[i]*b[i];
 b[N-1]/= a[N-1];

 b[N-2]-= c[N-2]*b[N-1];
 for (i=N-3; i>=0; i--)
   b[i]-= c[i]*b[i+1]+d[i]*b[N-1];

 for (i=0; i<N; i++)
   printf("%13.6le \n", b[i]);
}
```

Exercice 2

Point a)

De façon évidente, A est symétrique. Vérifions qu'elle est bien définie positive. Pour le faire, nous distinguons deux cas.

- Si $c = 0$, nous avons $\vec{x}^T A \vec{x} = a\|\vec{x}\|^2 \geqslant 0$, $\forall \vec{x}$ et $\vec{x}^T A \vec{x} = 0 \Rightarrow \vec{x} = 0$.
- Si $c \neq 0$:

$$\begin{aligned}
\vec{x}^T A \vec{x} &= a x_1^2 + 2c x_1 x_2 + a x_2^2 + 2c x_2 x_3 + a x_3^2 + \cdots + a x_{N-1}^2 + 2c x_{N-1} x_N + a x_N^2 \\
&\geqslant 2\,|c|\, x_1^2 + 2c x_1 x_2 + 2\,|c|\, x_2^2 + 2c x_2 x_3 + \cdots + 2\,|c|\, x_{N-1}^2 + 2c x_{N-1} x_N + 2\,|c|\, x_N^2 \\
&\geqslant |c| \left(x_1^2 + (x_1 \pm x_2)^2 + (x_2 \pm x_3)^2 + \cdots + (x_{N-1} \pm x_N)^2 + x_N^2 \right) \qquad \begin{cases} + & \text{si } c > 0 \\ - & \text{si } c < 0 \end{cases} \\
&\geqslant 0.
\end{aligned}$$

Si $\vec{x}^T A \vec{x} = 0$, alors $x_1 = x_N = 0$ et $(x_{j+1} \pm x_j)^2 = 0$, $\forall j = 1, 2, \ldots, (N-1)$, donc $\vec{x} = 0$.

Point b)

Puisque A est symétrique définie positive, le théorème du cours concernant les matrices de bande dit que la matrice L sera, comme A, une matrice de bande de demi-largeur 2. Donc, $A = LL^T$ s'écrira

$$\begin{pmatrix} a & c & & & \\ c & a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & a & c \\ & & & c & a \end{pmatrix} = \begin{pmatrix} a_1 & & & & \\ c_1 & a_2 & & & \\ & \ddots & \ddots & & \\ & & \ddots & a_{N-1} & \\ & & & c_{N-1} & a_N \end{pmatrix} \begin{pmatrix} a_1 & c_1 & & & \\ & a_2 & \ddots & & \\ & & \ddots & \ddots & \\ & & & a_{N-1} & c_{N-1} \\ & & & & a_N \end{pmatrix}$$

$$= \begin{pmatrix} a_1^2 & * & & & \\ a_1 c_1 & c_1^2 + a_2^2 & \ddots & & \\ & a_2 c_2 & \ddots & * & \\ & & \ddots & c_{N-2}^2 + a_{N-1}^2 & * \\ & & & c_{N-1} a_{N-1} & c_{N-1}^2 + a_N^2 \end{pmatrix}.$$

Remarquons qu'il n'est pas nécessaire d'utiliser les termes sur-diagonaux du produit LL^T - nous les avons remplacés par des $*$ - et rappelons que les termes diagonaux de L sont positifs. En égalant terme à terme A et LL^T :

- nous trouvons que $a_1 = \sqrt{a}$,
- ensuite, en connaissant le coefficient diagonal d'une ligne, nous calculons que le coefficient sous-diagonal de la ligne suivante vaut $c_i = \frac{c}{a_i}$,
- puis, en connaissant le coefficient sous-diagonal, nous trouvons que le coefficient diagonal de la $(i+1)^{\text{ème}}$ ligne vaut $a_{i+1} = \sqrt{a - c_i^2}$.

Algorithme de Cholesky :

$a_1 := \sqrt{a}$;	Calcul du coefficient a_1.
Faire pour $i = 1$ à $N-1$	
$c_i := \frac{c}{a_i}$;	Calcul du coefficient sous-diagonal de la $(i+1)^{\text{ème}}$ ligne.
$a_{i+1} := \sqrt{a - c_i^2}$;	Calcul du coefficient diagonal de la $(i+1)^{\text{ème}}$ ligne.

```
program Analyse_Numerique_Exercice_1c;

const
  max = 100;

type
  tableau = array[1..max] of ^real;

var
  a,b,c,e,d,f: tableau;
  val_a,val_b,val_c,val_e,val_f,val_d: real;
  n: integer;
```

mieux : val : record
a,b,c,e,f,f : real
end;

```
procedure Demande_les_indications;
  begin
    writeln('a c             d');
    writeln('e a c           d');
    writeln('  e a c         d');
    writeln('    . . .       .');
    writeln('      . . .     .');
    writeln('              c d');
    writeln('            e a c');
    writeln('f f . . . .  f e a');
    writeln;
    writeln('Entrez la dimension de la matrice carrée.');
    readln(n);
    writeln('Entrez la valeur de la constante a.');
    readln(val_a);
    writeln('Entrez la valeur de la constante c.');
    readln(val_c);
    writeln('Entrez la valeur de la constante e.');
    readln(val_e);
    writeln('Entrez la valeur de la constante f.');
    readln(val_f);
    writeln('Entrez la valeur de la constante d.');
    readln(val_d);
    writeln('Entrez la valeur des (bi) du vecteur solutions.');
    readln(val_b);
  end;

procedure Initialise_les_vecteurs;

    i: integer;
  begin
    for i := 1 to n do
      begin
        new(a[i]);
        new(b[i]);
        b[i]^ := val_b;
      end;
    for i := 1 to n-1 do
      begin
        new(c[i]);
        new(e[i]);
      end;
    for i := 1 to n-2 do
      begin
        new(d[i]);
        new(f[i]);
      end;
```

commentaires ?

```
  end;

procedure Factorisation_LU;
  var
    i: integer;
  begin
    a[1]^ := val_a;
    e[1]^ := val_e;
    f[1]^ := val_f;
    c[1]^ := val_c/a[1]^;
    d[1]^ := val_d/a[1]^;
    for i := 2 to n-2 do
      begin
        a[i]^ := val_a - e[i-1]^ * c[i-1]^;
        e[i]^ := val_e;
        f[i]^ := val_f - f[i-1]^ * c[i-1]^;
        c[i]^ := val_c/a[i]^;
        d[i]^ := (val_d - e[i-1]^ * d[i-1]^)/a[i]^;
      end;
    a[n-1]^ := val_a - c[n-2]^ * e[n-2]^;
    e[n-1]^ := val_e - c[n-2]^ * f[n-2]^;
    c[n-1]^ := (val_c - e[n-2]^ * d[n-2]^)/a[n-1]^;
    a[n]^ := val_a - e[n-1]^ * c[n-1]^;
    for i := 1 to n-2 do
      a[n]^ := a[n]^ - f[i]^ * d[i]^;
  end;

procedure Resolution_du_systeme;
  var
    i: integer;
  begin
    b[1]^ := b[1]^/a[1]^;
    for i := 2 to n-1 do
      begin
        b[i]^ := b[i]^ - e[i-1]^ * b[i-1]^;
        b[i]^ := b[i]^/a[i]^;
      end;
    for i := 1 to n-2 do
      b[n]^ := b[n]^ - f[i]^ * b[i]^;
    b[n]^ := b[n]^ - e[n-1]^ * b[n-1]^;
    b[n]^ := b[n]^/a[n]^;
    b[n-1]^ := b[n-1]^ - c[n-1]^ * b[n]^;
    for i := n-2 downto 1 do
      b[i]^ := b[i]^ - c[i]^ * b[i+1]^ - d[i]^ * b[n]^;
  end;

procedure Affiche_les_solutions;
  var
    i: integer;
  begin
    for i := 1 to n do
      writeln('b',i,' = ',b[i]^ :7:16,'   ');
  end;

begin
  Demande_les_indications;
  Initialise_les_vecteurs;
  Factorisation_LU;
  Resolution_du_systeme;
  Affiche_les_solutions;
  readln;
end.
```

Exemple

```
          c d
        e a c
f f . . . . f e a
```

Entrez la dimension de la matrice carrée.
6
Entrez la valeur de la constante a.
1
Entrez la valeur de la constante c.
2
Entrez la valeur de la constante e.
3
Entrez la valeur de la constante f.
4
Entrez la valeur de la constante d.
5
Entrez la valeur des (bii) du vecteur solutions.
6
b1 = 0.72124756336
b2 = -0.26510721248
b3 = -0.85380116960
b4 = 0.92007797272
b5 = 0.91617933724
b6 = 1.16179337230

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 22 avril 1997

Série 6

Exercice 1

On définit la matrice $A = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$ et les vecteurs $\vec{x}^0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ et $\vec{\varphi} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ où $\vec{\varphi}$ correspond au vecteur propre de A relatif à la valeur propre $\lambda = 2$.

Si, pour $n = 1, 2, \ldots$, on pose

$$\vec{x}^{(n)} = A\vec{x}^{(n-1)} \qquad \text{et} \qquad \mu_n = \frac{\vec{x}^{(n)T}\vec{x}^{(n+1)}}{\vec{x}^{(n)T}\vec{x}^{(n)}},$$

$\| x^{(n)} \| = \| x^{(n)t} \|$

démontrer directement (sans utiliser le cours) que

a) la suite de vecteurs $\left(\frac{\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|}\right)_{n=0}^{\infty}$ converge vers le vecteur propre $\vec{\varphi}$ lorsque $n \to \infty$.

Indication: on utilisera la relation $\frac{1}{(1+x)^{1/2}} = 1 - \frac{x}{2} + O(x^2)$ si $x \to 0$.

b) la suite des valeurs $(\mu_n)_{n=1}^{\infty}$ converge vers la valeur propre $\lambda = 2$ lorsque $n \to \infty$.

Exercice 2

Si d est une droite du plan Ox_1x_2 et si P est un point fixé, on définira la distance de P à d par

$$\text{dist}(P, d) = \min_{Q \in d} \left\|\overrightarrow{PQ}\right\|.$$

Soit d_1, d_2, d_3 les trois droites données par

$$\begin{aligned} d_1 &: \quad x_1 + 2x_2 - 2 = 0, \\ d_2 &: \quad 5x_1 + x_2 - 5 = 0, \\ d_3 &: \quad x_1 - 4x_2 + 4 = 0. \end{aligned}$$

$\delta P_{d_1} = \frac{1}{\sqrt{1+4}} \left| x_1 + 2x_2 - 2 \right|$

Trouver un point P tel que la quantité

$$\left(\text{dist}(P, d_1)^2 + \text{dist}(P, d_2)^2 + \text{dist}(P, d_3)^2\right)^{1/2}$$

$\delta = \frac{1}{\sqrt{a^2+b^2}} \cdot |as + bt + c|$

soit la plus petite possible.
Indication: si d est une droite d'équation $ax + by + c = 0$ avec $a^2 + b^2 = 1$, alors la distance δ entre le point (s, t) et la droite d est donnée par $\delta = |as + bt + c|$.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 22 avril 1997

Corrigé 6

Exercice 1

Rappel: une suite $\{\vec{x}^{(k)}\}_{k=0}^{\infty}$ de $\mathbb{R}^n$ converge vers la limite $\vec{y} \in \mathbb{R}^n$, si $\lim_{k\to\infty} \left\|\vec{x}^{(k)} - \vec{y}\right\| = 0$, i.e. si $\vec{x}^{(k)} = (x_1^{(k)}, \dots, x_n^{(k)})^T$ et $\vec{y} = (y_1, \dots, y_n)^T$, alors $\{\vec{x}^{(k)}\}_{k=0}^{\infty}$ converge vers $\vec{y}$ si et seulement $\lim_{k\to\infty} x_j^{(k)} = y_j$, $\forall j = 1, \dots, n$.

Point a)

En calculant $\vec{x}^{(1)} = A\vec{x}^{(0)} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, $\vec{x}^{(2)} = A\vec{x}^{(1)} = \begin{pmatrix} 2^2 \\ 1 \end{pmatrix}$, ..., on voit que $\vec{x}^{(n)} = \begin{pmatrix} 2^n \\ 1 \end{pmatrix}$.
La norme de $\vec{x}^{(n)}$ est $\|\vec{x}^{(n)}\| = \sqrt{1 + 2^{2n}}$. Ainsi,

$$\frac{\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|} = \begin{pmatrix} \frac{2^n}{(1+2^{2n})^{1/2}} \\ \frac{1}{(1+2^{2n})^{1/2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{(1+\frac{1}{2^{2n}})^{1/2}} \\ \frac{1}{2^n(1+\frac{1}{2^{2n}})^{1/2}} \end{pmatrix}.$$

En utilisant l'indication, on obtient

$$\frac{\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|} - \vec{\varphi} = \begin{pmatrix} -\frac{1}{2^{2n+1}} + O\left(\frac{1}{2^{4n}}\right) \\ \frac{1}{2^n} - \frac{1}{2^{3n+1}} + O\left(\frac{1}{2^{5n}}\right) \end{pmatrix} = \begin{pmatrix} O\left(\frac{1}{2^{2n+1}}\right) \\ O\left(\frac{1}{2^n}\right) \end{pmatrix}.$$

Ceci implique que la suite de vecteurs $\left(\frac{\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|}\right)_{n=0}^{\infty}$ converge vers $\vec{\varphi}$ lorsque $n \to \infty$.

Si on veut avoir une idée sur la "rapidité de convergence", on calcule

$$\left\|\frac{\vec{x}^{(n)}}{\|\vec{x}^{(n)}\|} - \vec{\varphi}\right\| = \left(O\left(\frac{1}{2^{2(2n+1)}}\right) + O\left(\frac{1}{2^{2n}}\right)\right)^{\frac{1}{2}} = O\left(\frac{1}{2^n}\right).$$

Point b)

Le quotient de Rayleigh vaut $\mu_n = \dfrac{\vec{x}^{(n)T}\vec{x}^{(n+1)}}{\vec{x}^{(n)T}\vec{x}^{(n)}} = \dfrac{2^n 2^{n+1} + 1}{2^n 2^n + 1} = 2 - \dfrac{1}{2^n 2^n + 1}$. Son comportement asymptotique est

$$\mu_n = 2 + O\left(\frac{1}{2^{2n}}\right)$$

et donc $|\mu_n - 2| = O\left(\frac{1}{2^{2n}}\right)$ lorsque $n \to \infty$. Ceci implique que la suite des valeurs $(\mu_n)_{n=1}^{\infty}$ converge vers la valeur propre $\lambda = 2$ lorsque $n \to \infty$.

Exercice 2

Point a)

Notons $\sqrt{z}$ la quantité à minimiser, i.e.

$$z = (\mathrm{dist}(P, d_1))^2 + (\mathrm{dist}(P, d_2))^2 + (\mathrm{dist}(P, d_3))^2.$$

Un minimum de z correspond à un minimum de $\sqrt{z}$, car si $0 \leqslant z_1 < z_2$, alors $0 \leqslant \sqrt{z}_1 < \sqrt{z}_2$.
Dans notre cas, si P a les coordonnées $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$, z s'écrit comme

$$z = \frac{1}{5}(p_1 + 2p_2 - 2)^2 + \frac{1}{26}(5p_1 + p_2 - 5)^2 + \frac{1}{17}(p_1 - 4p_2 + 4)^2 = \left\|A\vec{p} - \vec{b}\right\|^2,$$

où $A = \begin{pmatrix} \frac{1}{\sqrt{5}} & \frac{2}{\sqrt{5}} \\ \frac{5}{\sqrt{26}} & \frac{1}{\sqrt{26}} \\ \frac{1}{\sqrt{17}} & -\frac{4}{\sqrt{17}} \end{pmatrix}$, $\vec{b} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{5}{\sqrt{26}} \\ -\frac{4}{\sqrt{17}} \end{pmatrix}$ et $\vec{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$.

Par conséquent :

$$\begin{aligned} \text{minimiser } \sqrt{z} &\iff \text{minimiser } z\,, \\ &\iff \text{minimiser } \left\|A\vec{p} - \vec{b}\right\|^2, \\ &\iff \text{minimiser } \left\|A\vec{p} - \vec{b}\right\|, \\ &\iff \text{trouver } \vec{p} \text{ solution de } A\vec{p} = \vec{b} \text{ au sens des moindres carré.} \end{aligned}$$

Comme A est une 3×2-matrice de rang 2, nous pouvons utiliser le théorème du cours concernant les systèmes surdéterminés. Nous résolvons alors l'équation $A^T A\vec{p} = A^T\vec{b}$, c'est-à-dire

$$\frac{1}{2210}\begin{pmatrix} 2697 & 789 \\ 789 & 3933 \end{pmatrix}\begin{pmatrix} p_1 \\ p_2 \end{pmatrix} = \frac{1}{2210}\begin{pmatrix} 2489 \\ 4273 \end{pmatrix},$$

qui donne $p_1 = \frac{484}{753} \simeq .6427622842$ et $p_2 = \frac{721}{753} \simeq .9575033201$.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 29 avril 1997

Série 7

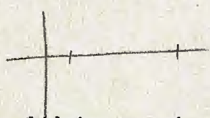
Exercice 1

Soit $a < b$ deux nombres réels et soit $g : x \in [a, b] \mapsto g(x) \in \mathbb{R}$ une fonction une fois continûment dérivable. On posera $\alpha = \max_{x \in [a,b]} |g'(x)|$ et on supposera que $g(x) \in [a, b]$, $\forall x \in [a, b]$.

a) Démontrer que g a au moins un point fixe $\overline{x}$ dans $[a, b]$.

b) Démontrer que si $\alpha < 1$, alors ce point fixe $\overline{x}$ est unique.

c) Démontrer que si $x_0 \in [a, b]$ et si on définit la suite $(x_n)_{n=0}^{\infty}$ par $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, alors on a



$$|x_{n+m} - x_n| \leqslant \alpha^n \left(\sum_{j=0}^{m-1} \alpha^j \right) |x_1 - x_0| , \; \forall n, m \geqslant 0.$$

En déduire que si $\alpha < 1$, alors la suite $(x_n)_{n=0}^{\infty}$ converge vers l'unique point fixe $\overline{x}$.
donc il s'agit de l'unicité

d) Expliciter un exemple avec $\alpha = 1$ tel que la suite définie par $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, ne converge pas.

Exercice 2

Soit λ un nombre réel positif donné. On pose ~~l'équation~~

$$f(x) = e^x - \lambda x.$$

on cherche $f(x) = 0$
pb. équivalent : $g(x) = x$
$g(x) = x + \alpha \cdot f(x)$
$= x + \alpha (e^x - \lambda \cdot x)$

a) Montrer graphiquement que
si $\lambda < e$, alors $f(x) = 0$ n'a pas de solution,
si $\lambda = e$, alors $f(x) = 0$ a une solution unique $\overline{x}_0$ et
si $\lambda > e$, alors $f(x) = 0$ a deux solutions $\overline{x}_1$ et $\overline{x}_2$ telles que $0 < \overline{x}_1 < 1 < \overline{x}_2$.

b) On suppose $\lambda > e$. Montrer qu'il existe $\varepsilon > 0$ tel que si $x_0 \in \mathbb{R}$ satisfait $|x_0 - \overline{x}_1| \leqslant \varepsilon$, alors la suite définie par $x_{n+1} = \frac{1}{\lambda} e^{x_n}$, $n = 0, 1, 2, \ldots$, converge vers $\overline{x}_1$.
Est-ce encore vrai pour $\overline{x}_2$?

$|\overline{x}_0 - \overline{x}_1| \leq \varepsilon \Rightarrow \exists\, \varepsilon > 0$ t.q. $\lim_{n \to \infty} x_{n+1} = \overline{x}_1$

c) On suppose toujours $\lambda > e$. Montrer qu'il existe $\varepsilon > 0$, tel que si $x_0 \in \mathbb{R}$ satisfait $|x_0 - \overline{x}_2| < \varepsilon$, alors la suite définie par par $x_{n+1} = \ln \lambda + \ln x_n$, $n = 0, 1, 2, \ldots$, converge vers $\overline{x}_2$.
Est-ce encore vrai pour $\overline{x}_1$?

ANALYSE NUMÉRIQUE Le 29 avril 1997
INFO, PH, PH-Fac
Prof. J. Rappaz

Corrigé 7

Exercice 1

Point a)

Si $g(a) = a$, respectivement si $g(b) = b$, nous avons a, respectivement b, comme point fixe de g.

Sinon, nous savons $g(a) > a$ et $g(b) < b$ puisque $g(x) \in [a, b]$, $\forall x \in [a, b]$. Puisque la différence $(g(x) - x)$ est $C^0([a, b])$ et que $g(a) - a > 0$ et $g(b) - b < 0$, il existe $\eta \in]a, b[$ tel que $g(\eta) - \eta = 0$. Nous avons alors le point fixe η de g.

Point b)

Soient deux points fixes de g, $\overline{x}_1$ et $\overline{x}_2$. En utilisant la propriété du point fixe $g(\overline{x}) = \overline{x}$ et le théorème des accroissements finis sur g – permis puisque $g \in C^1([a, b])$ –, nous trouvons qu'il existe ζ appartenant à l'intervalle d'extrémités $\overline{x}_1$ et $\overline{x}_2$ tel que

$$|\overline{x}_1 - \overline{x}_2| = |g(\overline{x}_1) - g(\overline{x}_2)| = |g'(\zeta)|\,|\overline{x}_1 - \overline{x}_2| \leqslant \alpha\,|\overline{x}_1 - \overline{x}_2|$$

Puisque $\alpha < 1$, nous avons nécessairement $|\overline{x}_1 - \overline{x}_2| = 0$, c'est-à-dire que les points fixes $\overline{x}_1$ et $\overline{x}_2$ sont confondus.

Point c)

Utilisons le théorème des accroissements finis, nous obtenons que

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| = |g'(\xi)||x_n - x_{n-1}| \leqslant \alpha|x_n - x_{n-1}|$$

où ξ appartient à l'intervalle d'extrémités x_{n-1} et x_n. Par conséquent, nous avons

$$|x_{n+1} - x_n| \leqslant \alpha^n|x_1 - x_0|.$$

Soit $m > 0$ quelconque, nous avons

$$|x_{n+m} - x_n| \leqslant \sum_{j=n}^{n+m-1} |x_{j+1} - x_j| \leqslant \sum_{j=n}^{n+m-1} \left(\alpha^j\,|x_1 - x_0|\right) = \alpha^n \left(\sum_{j=0}^{m-1} \alpha^j\right) |x_1 - x_0|.$$

Si nous remarquons que $\sum_{j=0}^{m-1} \alpha^j$ est la somme des m premiers termes d'une suite géométrique de raison α, nous pouvons écrire, lorsque $\alpha = 1$,

$$|x_{n+m} - x_n| = \alpha^n \frac{1 - \alpha^m}{1 - \alpha}|x_1 - x_0|.$$

Comme $0 \leqslant \alpha < 1$, nous pouvons majorer $\frac{1-\alpha^m}{1-\alpha}$ par $\frac{1}{1-\alpha}$. Ainsi,

$$|x_{n+m} - x_n| \leqslant \frac{\alpha^n}{1 - \alpha}|x_1 - x_0|, \forall n, m \geqslant m.$$

Cette dernière expression tend vers 0 lorsque n tend vers l'infini. Nous avons donc montré que x_n est une suite de Cauchy dans $[a, b]$, dès que $\alpha < 1$. Par conséquent, la suite $(x_n)_{n=0}^{\infty}$ converge vers un $\overline{x} \in [a, b]$.

Montrons maintenant que $\overline{x}$ est le point fixe de g dans $[a, b]$. Nous avons $x_{n+1} = g(x_n)$. Faisons tendre n vers l'infini dans cette égalité. Puisque g est continue, nous obtenons $\overline{x} = g(\overline{x})$ et par conséquent, $\overline{x}$ est un point fixe de g, unique par le point b).

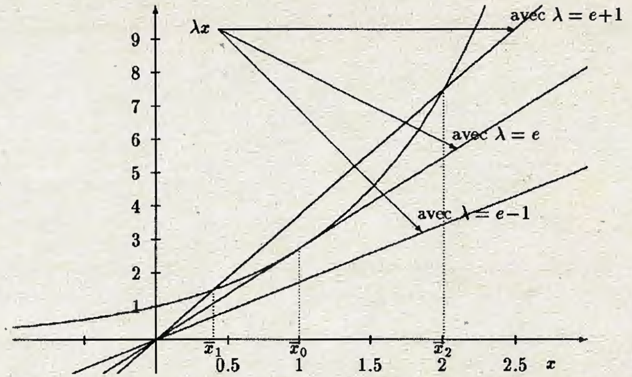
Point d)

Si la fonction g est définie par $g(x) = (a + b - x)$, les hypothèses d'être continûment dérivable sur $\mathbb{R}$ et d'avoir une image incluse dans l'intervalle $[a, b]$, sont satisfaites. Le α qui lui correspond vaut 1. Mais la suite définie par $x_{n+1} = g(x_n)$, $n = 0, 1, 2, \ldots$, oscille entre x_0 et $(a + b - x_0)$, quel que soit le point initial $x_0 \in [a, b]$. Elle ne converge donc pas.

Exercice 2

Point a)

La figure ci-dessus montre les graphes de $f(x) = e^x$ et de $f(x) = \lambda x$ avec λ valant $(e - 1)$, e et $(e + 1)$. Naturellement, les intersection de la courbe $y = e^x$ avec les différentes droites $y = \lambda x$ vont correspondre aux zéros de $f(x)$.



Nous savons que la fonction e^x a une courbure positive et qu'elle vaut 1 en 0. En conséquence, trois situations sont possibles en fonction de λ :

- soit λ est inférieur à la valeur critique λ_c, la pente de la droite est trop faible et nous n'avons pas d'intersection,
- soit λ atteind la valeur critique λ_c, la pente de la droite est telle que les graphes de la fonction e^x et de la droite $y = \lambda_c x$ sont tangents et nous avons une intersection $\overline{x}_0$,
- soit λ est supérieur à la valeur critique λ_c et nous avons deux intersections, $\overline{x}_1$ et $\overline{x}_2$. Puisque $\lambda x \leqslant 0$ si $x \leqslant 0$ et que $\lambda x > \lambda_c x$ si $x > 0$, ces deux points satisfont la relation suivante

$$0 < \overline{x}_1 < \overline{x}_0 < \overline{x}_2.$$

La valeur critique doit satisfaire le système d'équations

$$\begin{aligned} e^{\overline{x}_0} &= \lambda_c \overline{x}_0, && \text{(point de tangence appartient à la droite et à l'exponentielle)} \\ e^{\overline{x}_0} &= \lambda_c, && \text{(pentes de la droite et de l'exponentielle sont égales au point de tangence)} \end{aligned}$$

ce qui implique que $\overline{x}_0 = 1$ et que $\lambda_c = e$.

Point b)

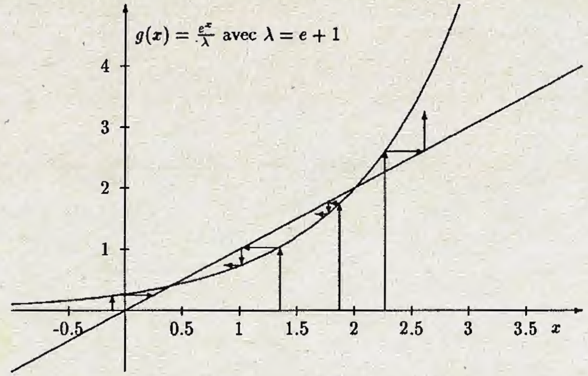
Chercher la limite de la suite $x_{n+1} = \frac{1}{\lambda}e^{x_n}$, $n = 0, 1, 2, \ldots$, peut être vu comme un problème de point fixe $g(x) = x$ avec $g(x) = \frac{e^x}{\lambda}$.

La fonction g est continûment dérivable, de dérivée première valant g. Au point fixe $\overline{x}_1 < 1$, nous avons $|g'(\overline{x}_1)| = |g(\overline{x}_1)| = \overline{x}_1 < 1$. Nous pouvons conclure en utilisant un théorème du cours que

$$\exists\, \varepsilon > 0 \quad \text{t.q.} \quad \text{si } |\overline{x}_1 - x_0| \leqslant \varepsilon, \quad \text{alors } \lim_{n\to\infty} x_n = \overline{x}_1.$$

Le théorème ne s'applique pas pour le point fixe $\overline{x}_2 > 1$, car alors $|g'(\overline{x}_2)| = |g(\overline{x}_2)| = \overline{x}_2 > 1$.

La figure ci-dessous illustre que si le point x_0 est proche du point $\overline{x}_1$, la suite converge vers $\overline{x}_1$, tandis que si il est proche du point $\overline{x}_2$, elle ne converge pas vers $\overline{x}_2$.



Point c)

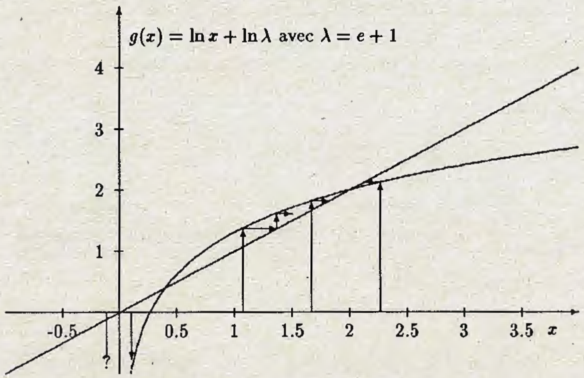
Chercher la limite de la suite $x_{n+1} = \ln\lambda + \ln x_n$, $n = 0, 1, 2, \ldots$, peut être vu comme un problème de point fixe $g(x) = x$ avec $g(x) = \ln\lambda + \ln x$.

La fonction g est continûment dérivable sur $]0, \infty[$, de dérivée première valant $\frac{1}{x}$. Au point fixe $\overline{x}_2 > 1$, nous avons $|g'(\overline{x}_2)| = \frac{1}{\overline{x}_2} < 1$. Nous pouvons conclure en utilisant un théorème du cours que

$$\exists\, \varepsilon > 0 \quad \text{t.q.} \quad \text{si } |\overline{x}_2 - x_0| \leqslant \varepsilon, \quad \text{alors } \lim_{n\to\infty} x_n = \overline{x}_2.$$

Le théorème ne s'applique pas pour le point fixe $\overline{x}_1 < 1$, car alors $|g'(\overline{x}_1)| = \frac{1}{\overline{x}_1} > 1$.

La figure ci-dessous illustre que si le point x_0 est proche du point $\overline{x}_2$, la suite converge vers $\overline{x}_2$, tandis que si il est proche du point $\overline{x}_1$, elle ne converge pas vers $\overline{x}_1$.



ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 13 mai 1997

Série 8

Exercice 1 (Méthode d'Adams-Bashforth à 3 pas)

Soit $f : (x,t) \in \mathbb{R} \times \mathbb{R}_+ \mapsto f(x,t) \in \mathbb{R}$ une fonction continue donnée. On considère le problème de Cauchy

trouver une fonction $u : t \in [0, \infty[\mapsto u(t) \in \mathbb{R}$ qui satisfait

$$(P_1) \qquad \begin{cases} \dot{u}(t) = f(u(t),t) & t > 0, \\ u(0) = u_0, \end{cases}$$

où $\dot{u} = \frac{du}{dt}$ et u_0 est un nombre réel donné.
Si h est un paramètre positif donné, si $t_j = jh$ avec $j = 0,\, 1,\, 2, \ldots$, on suppose avoir obtenu, pour $n \geq 2$, des approximations $u^0, u^1, \ldots, u^n$ de $u(t_0), u(t_1), \ldots, u(t_n)$ respectivement.
Pour calculer u^{n+1}, on procède ainsi

i) on calcule le polynôme $\varphi(t)$ de degré 2 tel que

$$\varphi(t_{n-2}) = f(u^{n-2}, t_{n-2}), \quad \varphi(t_{n-1}) = f(u^{n-1}, t_{n-1}) \quad \text{et} \quad \varphi(t_n) = f(u^n, t_n)$$

ii) on pose $u^{n+1} = u^n + \int_{t_n}^{t_{n+1}} \varphi(t)\, dt.$ ici c'est le nom donné au polynôme.

Expliciter et justifier cette méthode.

Exercice 2

On considère le problème de Cauchy :

trouver une fonction $u : t \in [0, \infty) \mapsto u(t) \in \mathbb{R}$ qui satisfait

$$(P_2) \qquad \begin{cases} \dot{u}(t) = e^{-(u(t))^2}, & t > 0, \\ u(0) = 0. \end{cases}$$

a) Montrer que le problème (P_2) a une solution unique notée $u(t)$.

b) Si h est un nombre positif donné, si $t_n = nh$, $n = 0,\, 1,\, 2, \ldots$, donner le schéma de Heun pour calculer numériquement les approximations u^n de $u(t_n)$.

c) Calculer la valeur u^1 qui résulte du schéma de Heun.

d) Montrer que si v^1 est l'approximation de $u(t_1)$ par le schéma d'Euler progressif, alors

$$|u^1 - v^1| \leqslant \frac{h^3}{2}.$$

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 13 mai 1997

Corrigé 8

Exercice 1

On suppose connues les valeurs $u^0 \simeq u(t_0)$, $u^1 \simeq u(t_1)$, $\ldots$, $u^n \simeq u(t_n)$. Pour calculer u^{n+1}, on procède en 2 étapes.

i) On calcule le polynôme d'interpolation de f, de degré 2, qui passe par les 3 points suivants

$$(t_{n-2}, f(u^{n-2}, t_{n-2})), \quad (t_{n-1}, f(u^{n-1}, t_{n-1})), \quad (t_n, f(u^n, t_n)).$$

La base de Lagrange associée à ces trois points est

$$\varphi_0(t) = \frac{(t-t_{n-1})(t-t_n)}{(t_{n-2}-t_{n-1})(t_{n-2}-t_n)},$$
$$\varphi_1(t) = \frac{(t-t_{n-2})(t-t_n)}{(t_{n-1}-t_{n-2})(t_{n-1}-t_n)} \text{ et}$$
$$\varphi_2(t) = \frac{(t-t_{n-2})(t-t_{n-1})}{(t_n-t_{n-2})(t_n-t_{n-1})}.$$

$$u^{n+1} - u^n = u_0 \cdot \int_{n\cdot h}^{(n+1)h} \varphi_0 \, dt + u_1 \int_{n\cdot h}^{(n+1)\cdot h} \varphi_1 \, dt + u_2 \int_{n\cdot h}^{(n+1)h} \varphi_2 \, dt$$

Exprimé dans cette base, l'interpolant de f passant par les 3 points ci-dessus est

$$\varphi(t) = f(u^{n-2}, t_{n-2})\varphi_0(t) + f(u^{n-1}, t_{n-1})\varphi_1(t) + f(u^n, t_n)\varphi_2(t).$$

ii) On pose $u^{n+1} = u^n + \int_{t_n}^{t_{n+1}} \varphi(t)dt$, où l'intégrand est un polynôme de $\boxed{\text{degré 2}}$. On a vu dans la série 3 que la formule de quadrature de Simpson est exacte pour des polynômes de degré 3. On peut donc l'utiliser pour intégrer $\varphi(t)$ et approximer l'intégrale $\int_a^b f(x)dx$ par $J(f) = \frac{b-a}{6}(f(a) + 4f(\frac{a+b}{2}) + f(b))$. En calculant séparément les intégrales pour chaque fonction de base, on obtient que

$$J(g) = \frac{1}{3} \cdot [g(-1) + 4 \cdot g(0) + g(1)]$$

Applique les définitions $\Longrightarrow$
Comme la formule est exacte, il suffit de dilater sur 1 intervalle, et la formule de Simpson

$$J_h = \frac{h}{2} \cdot \frac{1}{3} [g(-1) + 4g(0) + g(1)]$$

donne

$$\frac{h}{6} \cdot \Big[f(x_j) + 4f(\tfrac{h}{2} + x_j) + f(\underbrace{h + x_j}_{x_{j+1}}) \Big]$$

$$x_j = j \cdot h + a, \quad h = \frac{b-a}{n} = h$$

1 intervale $\Longrightarrow j = 0$.
$\Longrightarrow x_j = a = t_n$

$$\Longrightarrow \frac{h}{6} \cdot \left[f(t_n) + 4f\left(\frac{t_{n+1} + t_n}{2}\right) + f(t_{n+1}) \right]$$

avec f qui représente φ_k.

$$\int_{t_n}^{t_{n+1}} \varphi_0(t) = \frac{h}{6}\left(\varphi_0(t_n) + 4\varphi_0(\tfrac{t_{n+1}+t_n}{2}) + \varphi_0(t_{n+1})\right) = \frac{5}{12}h,$$
$$\int_{t_n}^{t_{n+1}} \varphi_1(t) = \frac{h}{6}\left(\varphi_1(t_n) + 4\varphi_1(\tfrac{t_{n+1}+t_n}{2}) + \varphi_1(t_{n+1})]\right) = -\frac{4}{3}h = -\frac{16}{12}h \text{ et}$$
$$\int_{t_n}^{t_{n+1}} \varphi_2(t) = \frac{h}{6}\left(\varphi_2(t_n) + 4\varphi_2(\tfrac{t_{n+1}+t_n}{2}) + \varphi_2(t_{n+1})]\right) = \frac{23}{12}h,$$

puisque

$$\varphi_0(t_n) = 0, \quad \varphi_0(\tfrac{t_{n+1}+t_n}{2}) = \varphi_0(nh + \tfrac{h}{2}) = \tfrac{3}{8}, \quad \varphi_0(t_{n+1}) = 1,$$
$$\varphi_1(t_n) = 0, \quad \varphi_1(\tfrac{t_{n+1}+t_n}{2}) = \varphi_1(nh + \tfrac{h}{2}) = -\tfrac{5}{4}, \quad \varphi_1(t_{n+1}) = -3,$$
$$\varphi_2(t_n) = 1, \quad \varphi_2(\tfrac{t_{n+1}+t_n}{2}) = \varphi_2(nh + \tfrac{h}{2}) = \tfrac{15}{8} \text{ et } \quad \varphi_2(t_{n+1}) = 3. \quad = \frac{3h-2h}{2h-h} = 3.$$

Finalement, on a $u^{n+1} = u^n + \frac{h}{12}\big(5\underbrace{f(u^{n-2}, t_{n-2})}_{u_{n-2}} - 16\underbrace{f(u^{n-1}, t_{n-1})}_{u_{n-1}} + 23\underbrace{f(u^n, t_n)}_{u_n}\big)$.

Justification de la méthode d'Adams-Bashforth.

En intégrant l'équation différentielle $\dot{u}(t) = f(u(t), t)$ entre t_n et t_{n+1}, on obtient

$$u(t_{n+1}) = u(t_n) + \int_{t_n}^{t_{n+1}} f(u(s), s)ds.$$

Une approximation de $u(t_{n+1})$, notée u^{n+1}, peut être obtenue comme $u^{n+1} = u^n + J(f)$, où $u^n \simeq u(t_n)$ et $J(f)$ est une approximation de l'intégrale $\int_{t_n}^{t_{n+1}} f(u(s), s)ds$. L'idée de la méthode d'Adams-Bashforth à 3 pas consiste à :

i) calculer le polynôme φ d'interpolation de degré 2 passant par les 3 points suivants :
$$(t_{n-2}, f(u^{n-2}, t_{n-2})), \quad (t_{n-1}, f(u^{n-1}, t_{n-1})), \quad (t_n, f(u^n, t_n)),$$
en supposant connues les valeurs de u^{n-2}, u^{n-1}, u^n.

ii) et utiliser l'extrapolation de ce polynôme sur l'intervalle $[t_n, t_{n+1}]$ à la place de f pour approximer l'intégrale $\int_{t_n}^{t_{n+1}} f(u(s), s)ds$, i.e. calculer $J(f)$ comme $\int_{t_n}^{t_{n+1}} \varphi(t)dt$.

Remarque : la méthode d'Adams-Bashforth à 3 pas donne une précision plus grande que celles des méthodes simples comme les schémas d'Euler progressif ou rétrograde, ou même de Runge-Kutta d'ordre 2.

Exercice 2

Définissons $f(x, t) = e^{-x^2}$, indépendant de t. Notre problème de Cauchy se note :

trouver une fonction $u : t \in [0, \infty[\mapsto u(t) \in \mathbb{R}$ qui satisfait

$$(P_2') \quad \begin{cases} \dot{u}(t) = f(u(t), t), & t > 0, \\ u(0) = 0. \end{cases}$$

Point a)

Puisque le maximum de la fonction $\left|se^{-s^2}\right|$ vaut $\frac{e^{-\frac{1}{2}}}{\sqrt{2}}$ $\left(\text{maximum atteint en } s = \pm\frac{1}{\sqrt{2}}\right)$, nous avons, pour y, $z \in \mathbb{R}$ quelconques,

$$|f(y,t) - f(z,t)| = \left|e^{-y^2} - e^{-z^2}\right| = \left|\int_z^y -2se^{-s^2}ds\right| \leqslant \sqrt{2}e^{-\frac{1}{2}} \left|\int_y^z ds\right| = \sqrt{2}e^{-\frac{1}{2}}|y - z|.$$

Comme $f(x,t)$ est, en plus, continue sur $\mathbb{R} \times \mathbb{R}_+$, les hypothèses du théorème de Cauchy-Lipschitz sont remplies et le problème (P_2'), et donc (P_2), a une solution unique.

Point b)

Si u^n est l'approximation de $u(t)$ par le schéma de Heun, on a

$$p^1 = f(u^n, t_n),$$
$$p^2 = f(u^n + h_n p^1, t_{n+1}),$$
$$u^{n+1} = u^n + h_n\left(\frac{1}{2}p^1 + \frac{1}{2}p^2\right),$$

pour $n = 0, 1, 2, \ldots$, avec $u^0 = u(t_0) = 0$ et $h_n = h$. Le schéma devient donc

$$p^1 = e^{-(u^n)^2},$$
$$p^2 = e^{-(u^n + hp^1)^2},$$
$$u^{n+1} = u^n + \frac{h}{2}(p^1 + p^2),$$

pour $n = 0, 1, 2, \ldots$, avec $u^0 = 0$.

Point c)

Le premier pas de ce schéma donne

$$p^1 = e^{-(u^0)^2} = 1,$$
$$p^2 = e^{-\left(u^0+hp^1\right)^2} = e^{-h^2},$$
$$u^1 = u^0 + \frac{h}{2}\left(p^1 + p^2\right) = \frac{h}{2}\left(1 + e^{-h^2}\right).$$

Point d)

Le schéma d'Euler progressif est (cf. cours §8.3a)

$$v^{n+1} = v^n + h_n f\left(v^n, t_n\right), \quad \text{pour } n = 0, 1, 2, \ldots,$$

avec $v^0 = u(t_0)$. Dans notre cas, $v^0 = u(0) = 0$, donc

$$v^1 = v^0 + hf\left(v^0, t_0\right) = v^0 + he^{-(v^0)^2} = h.$$

La différence $\left|u^1 - v^1\right|$ vaut

$$\left|u^1 - v^1\right| = \left|\frac{h}{2}\left(1 + e^{-h^2}\right) - h\right| = \left|\frac{h}{2}\left(1 - e^{-h^2}\right)\right| = \frac{h}{2}\left(1 - e^{-h^2}\right).$$

Comme la fonction $1 - e^{-x^2} - x^2$ est strictement décroissante sur $\mathbb{R}_+$ et s'annule en 0, nous avons que $\left|u^1 - v^1\right| \leqslant \frac{h^3}{2}$.

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 20 mai 1997

Série 9

Exercice 1

Soit $f : x \in [0,1] \mapsto f(x) \in \mathbb{R}$ une fonction continue donnée; on se propose de chercher $u : x \in [0,1] \mapsto u(x) \in \mathbb{R}$ qui satisfait

$$(P) \qquad \begin{cases} -\frac{d}{dx}\left((1+x)\frac{du(x)}{dx}\right) = f(x), & \text{si } x \in]0,1[, \\ u(0) = u(1) = 0. \end{cases}$$

a) Donner une formulation ~~faible~~ du problème (P).
ou variationnelle

b) Si N est un entier positif, si $h = \frac{1}{N+1}$ et si $x_j = jh$ avec $j = 0, 1, 2, \ldots, N+1$, on définit les fonctions de base de type "éléments finis" $\varphi_1, \varphi_2, \ldots, \varphi_N$ comme dans le cours. Utiliser (a) pour donner une approximation du problème (P) par la méthode des éléments finis.

Remarque :
- on calculera explicitement la matrice A du système à résoudre,
- on calculera le second membre en utilisant la méthode des trapèzes.

c) Ecrire explicitement le système linéaire obtenu suite à (b) lorsque $f(x) = x^2$.

d) *Les plus courageux pourront résoudre le système et comparer la solution obtenue à la solution exacte.*

$x^3 \mid x+1$
$x^3+x^2 \mid x^2 - x + 1$
$-x^2$
$-x^2-x$
x
$x+1$
-1

$$(1+x)\,u'(x) = -\tfrac{1}{3}x^3 + C_1$$

$$u'(x) = -\frac{x^3}{1+x} + \frac{C_1}{(1+x)}$$

$$u(x) = C_1 \cdot \mathrm{Arctg}(x) - \int x^2 - x + 1 - \frac{1}{1+x}\,dx$$

$$= C_1 \cdot \mathrm{Arctg}(x) - \tfrac{1}{3}x^3 + \tfrac{1}{2}x^2 - x + \ln(x+1) + C_2$$

$u(0) = 0 \Rightarrow \ln(1) + C_2 = 0 \Rightarrow C_2 = 0$

$u(1) = 0 \Rightarrow C_1 \cdot \frac{\pi}{4} - \frac{1}{3} + \frac{1}{2} - 1 + \ln(2) = 0$

$\Rightarrow -\frac{5}{6} + C_1 \cdot \frac{\pi}{4} + \ln(2) = 0$

$\Rightarrow C_1 \frac{\pi}{4} = \frac{5}{6} - \ln(2)$

$\Rightarrow C_1 = \frac{4}{\pi} \cdot \frac{5}{6} - \frac{4}{\pi} \cdot \ln(2) = \frac{10}{3\pi} - \frac{4}{\pi} \cdot \ln(2)$

$= \frac{1}{\pi}\left(\frac{10}{3} - 4 \cdot \ln(2)\right)$

Ainsi :

$$u(x) = \frac{1}{\pi}\left(\frac{10}{3} - 4 \cdot \ln(2)\right) \cdot \mathrm{Arctg}(x) - \tfrac{1}{3}x^3 + \tfrac{1}{2}x^2 - x + \ln(x+1)$$

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 20 mai 1997

Corrigé 9

Exercice 1

Point a)

Soit v une fonction continue qui a les deux propriétés suivantes

i) sa première dérivée est continue par morceaux sur $[0, 1]$,

ii) $v(0) = v(1) = 0$.

Multiplions l'équation différentielle par v et intégrons entre 0 et 1. Nous obtenons

$$\int_0^1 -v(x)\frac{d}{dx}\left((1+x)\frac{du}{dx}(x)\right)dx = \int_0^1 f(x)v(x)dx.$$

L'intégration par parties du terme de gauche donne

$$\int_0^1 -v(x)\frac{d}{dx}\left((1+x)\frac{du}{dx}(x)\right)dx = \int_0^1 \frac{dv}{dx}\left((1+x)\frac{du}{dx}(x)\right)dx,$$

puisque, grâce à la propriété (ii), le terme de bord est nul.

Notons $u'(x) = \frac{du(x)}{dx}$ et $v'(x) = \frac{dv(x)}{dx}$ et définissons V comme l'ensemble de toutes les fonctions continues, de dérivée première continue par morceaux, s'annulant en 0 et 1. Le problème faible est :

chercher une fonction $u \in V$ qui satisfasse

$$\int_0^1 (1+x)u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx \quad \forall v \in V. \tag{PF}$$

Point b)

Soient $\{\varphi_i\}_{i=1}^N$ les fonctions de base de type "éléments finis" définies par

$$\varphi_i(x) = \begin{cases} \frac{x-x_{i-1}}{x_i-x_{i-1}}, & \text{si } x_{i-1} \leqslant x \leqslant x_i, \\ \frac{x-x_{i+1}}{x_i-x_{i+1}}, & \text{si } x_i \leqslant x \leqslant x_{i+1}, \\ 0, & \text{si } x \leqslant x_{i-1} \text{ ou si } x_{i+1} \leqslant x. \end{cases}$$

L'espace V_h engendré par ces fonctions est un sous-espace vectoriel de V. Une approximation du problème (PF) est :

trouver une fonction $u_h \in V_h$ qui satisfasse

$$\int_0^1 (1+x)u_h'(x)v_h'(x)dx = \int_0^1 f(x)v_h(x)dx \quad \forall v_h \in V_h. \tag{P$_h$F}$$

Puisque $(\varphi_j)_{j=1}^N$ est une base de V_h, dire que $(\mathrm{P}_h\mathrm{F})$ est valable pour toute fonction v_h dans V_h est équivalent à dire que $(\mathrm{P}_h\mathrm{F})$ est valable pour toute fonction φ_j, $\forall j = 1, 2, \ldots, N$. De plus, comme la fonction u_h est un élément de V_h, $u_h(x)$ peut se développer $u_h(x) = \sum_{i=1}^N u_i\varphi_i(x)$ où $u_1, u_2, \ldots, u_N$ sont N nombres réels. En effectuant ces substitutions, le problème $(\mathrm{P}_h\mathrm{F})$ devient :

déterminer les inconnues $u_i \in \mathbb{R}$, $i = 1, 2, \ldots, N$ telles que

$$\sum_{i=1}^N u_i\left(\int_0^1 (1+x)\varphi_j'(x)\varphi_i'(x)dx\right) = \int_0^1 f(x)\varphi_j(x)dx \quad \forall j = 1, 2, \ldots, N. \tag{(P$_h$F)$_\varphi$}$$

En définissant la $N \times N$ matrice A de coefficients $a_{ji} = \int_0^1 (1+x)\varphi_j'(x)\varphi_i'(x)dx$, le N-vecteur $\vec{f}$ dont la j-ième composante est $f_j = \int_0^1 f(x)\varphi_j(x)dx$ et le N-vecteur $\vec{u}$ de composantes $u_1, u_2, \ldots, u_N$, $((\mathrm{P}_h\mathrm{F})_\varphi)$ peut s'écrire sous forme matricielle

déterminer le vecteur $\vec{u} \in \mathbb{R}^N$ tel que $A\vec{u} = \vec{f}$. $\quad (\mathrm{P}_M\mathrm{F})$

Calcul de A

Rappelons que $x_j = jh$ avec $j = 0, 1, \ldots, N+1$, que $h = \frac{1}{N+1}$ et donc que

$$\varphi_i(x) = \begin{cases} \frac{x-x_{i-1}}{h}, & \text{si } x_{i-1} \leqslant x \leqslant x_i, \\ -\frac{x-x_{i+1}}{h}, & \text{si } x_i \leqslant x \leqslant x_{i+1}, \\ 0, & \text{sinon}, \end{cases} \quad \text{et} \quad \varphi_i'(x) = \begin{cases} \frac{1}{h}, & \text{si } x_{i-1} \leqslant x \leqslant x_i, \\ -\frac{1}{h}, & \text{si } x_i \leqslant x \leqslant x_{i+1}, \\ 0, & \text{sinon}. \end{cases}$$

Puisque $\varphi_j(x) = 0$ si $x \notin]x_{j-1}, x_{j+1}[$, il suffit de remarquer que la matrice A est symétrique et de distinguer alors 3 cas

i) si $|i-j| > 1$, alors $a_{ji} = 0$. La matrice A est tridiagonale.

ii) si $|i-j| = 1$, alors

$$a_{ji} = a_{j+1,j} = a_{j,j+1} = \int_0^1 (1+x)\varphi_{j+1}'(x)\varphi_j'(x)dx = -\frac{1}{h^2}\int_{jh}^{(j+1)h}(1+x)dx = -\frac{1}{h} - \frac{1}{2}(2j+1).$$

iii) si $i = j$, alors

$$a_{jj} = \frac{1}{h^2}\int_{(i-1)h}^{(i+1)h}(1+x)dx = 2(i+\frac{1}{h}).$$

Calcul du second membre

Par la méthode des trapèzes, l'intégrale $\int_{x_0}^{x_{N+1}} g(x)dx$ est approximée par L_hg donnée par $L_hg \overset{\text{déf.}}{=} \sum_{k=0}^N \frac{g(x_k)+g(x_{k+1})}{2}h$. Dans notre cas, $g(x) = f(x)\varphi_j(x)$. Notons $\tilde{f}_j$ l'approximation $L_h(f_j)$ de f_j. Alors

$$\tilde{f}_j = \frac{h}{2}\left(f(x_0)\varphi_j(x_0) + 2\sum_{k=1}^N f(x_k)\varphi_j(x_k) + f(x_{N+1})\varphi_j(x_{N+1})\right) = hf(x_j),$$

pour $j = 1, 2, \ldots, N$,

car $\varphi_j(x_k) = 1$ si $j = k$, et $\varphi_j(x_k) = 0$ sinon.

Point c)

Le système linéaire à résoudre lorsque $f(x) = x^2$ est

$$\begin{pmatrix} \frac{2}{h}+2 & -\frac{1}{h}-\frac{3}{2} & & & \\ -\frac{1}{h}-\frac{3}{2} & \frac{2}{h}+4 & -\frac{1}{h}-\frac{5}{2} & & \\ & \ddots & \ddots & \ddots & \\ & -\frac{1}{h}-j+\frac{1}{2} & \frac{2}{h}+2i & -\frac{1}{h}-j-\frac{1}{2} & \\ & & \ddots & \ddots & \ddots \\ & & & -N+\frac{1}{2}-\frac{1}{h} & 2N+\frac{1}{h} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_i \\ \vdots \\ u_N \end{pmatrix} = h^3 \begin{pmatrix} 1 \\ 4 \\ \vdots \\ i^2 \\ \vdots \\ N^2 \end{pmatrix}.$$

Il sera résolu en utilisant une décompostion LL^T ou avec des routines *toutes faites* fournies dans des librairies numériques.

Point d)

Lorsque $f(x) = x^2$, $\forall x \in [0,1]$, l'équation différentielle devient

$$-\frac{d}{dx}\left((1+x)\frac{du(x)}{dx}\right) = x^2, \quad \forall x \in [0,1],$$

et sa solution générale est

$$u : x \in [0,1] \longmapsto u(x) = -\frac{1}{3}x + \frac{1}{6}x^2 - \frac{1}{9}x^3 + C_1 + C_2 \ln(1+x).$$

ou C_1 et C_2 sont deux constantes à déterminer. Pour satisfaire aux conditions de bord, il faut (et il suffit) que $C_1 = 0$ et $C_2 = \frac{5}{18 \ln 2}$. La solution est donc

$$u : x \in [0,1] \longmapsto u(x) = -\frac{1}{3}x + \frac{1}{6}x^2 - \frac{1}{9}x^3 + \frac{5}{18 \ln 2}\ln(1+x) \in \mathbb{R}.$$

Le calcul, en double précision pour la partie numérique, donne

x	Solution exacte $u(x)$	Solution calculée $u_h(x)$ avec N=7	avec N=3	avec N=1
0.00	0.00000000	0.00000000	0.00000000	0.00000000
1.25e-01	7.92187540e-03	7.82264454e-03		
2.50e-01	1.47716930e-02	1.46162607e-02	1.41543016e-02	
3.75e-01	2.01980191e-02	2.00188182e-02		
5.00e-01	2.36006946e-02	2.34230555e-02	2.28941847e-02	2.08333333e-02
6,25e-01	2.42106856e-02	2.40549538e-02		
7.50e-01	2.11402561e-02	2.10231466e-02	2.06740857e-02	
8.75e-01	1.34157904e-02	1.33512917e-02		
1.00	0.00000000	0.00000000	0.00000000	0.00000000
	$h\left(\sum_i \left(u(x_i)-u_h(x_i)\right)^2\right)^{\frac{1}{2}}$	4.67375848e-05	2.61921198e-04	1.38368069e-03

Le graphique ci-après montre d'une autre manière la convergence des solutions approchées u_h vers la solution u, en fonction de N.

<u>Remarque</u>

Le passage entre u, la solution exacte, et $\vec{u}$, le vecteur calculé, se fait à l'aide de 2 approximations :

i) approcher u par une méthode d'éléments finis faites avec des fonctions linéaires par morceaux. L'erreur $e = u - u_h$ est de l'ordre $O(h^2)$.

ii) utiliser les formules de quadrature, pour effectuer les intégrations, dans le calcul de la matrice $\underline{A}$ et du vecteur $\vec{f}$. Dans notre cas les éléments de A sont calculés exactement et les éléments de $\vec{f}$ sont calculés à l'aide de la formule des trapèzes. Il est possible de montrer que l'intégration numérique proposée dans b) ne donne pas lieu à des perturbations plus grandes que celles obtenues en n'utilisant pas d'intégration numérique.

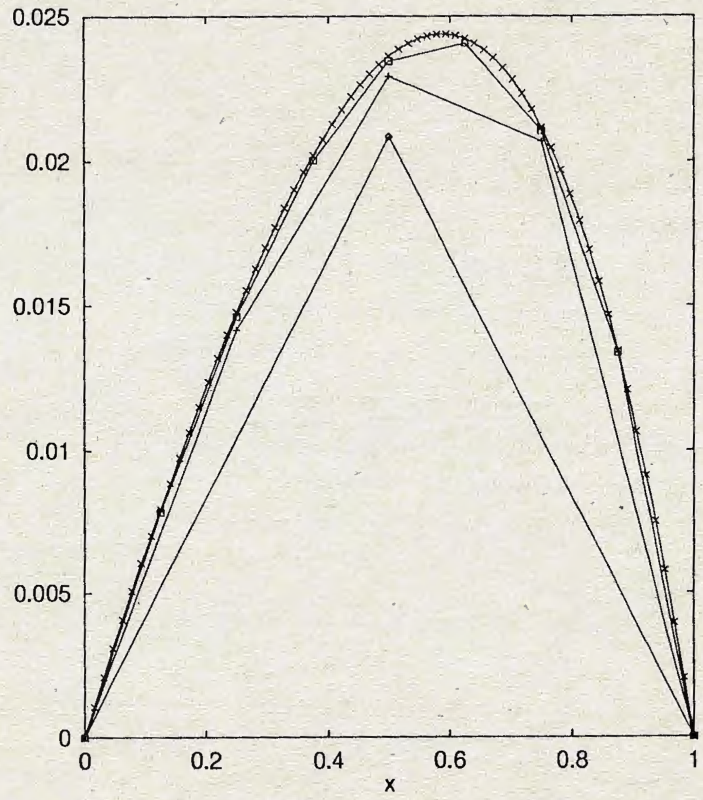


FIG. – *Graphes des fonctions $u(x)$ et $u_h(x)$ avec $N = 1$ $^\diamond$, 3 $^+$, 7 $^\square$ et 64 $^\times$.*

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 3 juin 1997

Série 10

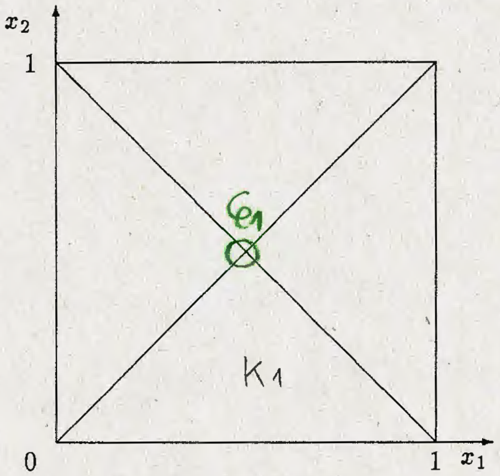
Exercice 1

Soit Ω le carré unité de $\mathbb{R}^2$ de frontière $\partial\Omega$. On se propose de chercher une fonction $u : \overline{\Omega} \longrightarrow \mathbb{R}^2$ qui satisfait

$$(P) \qquad \begin{cases} -\dfrac{\partial}{\partial x_1}\left((1+x_1+x_2)\dfrac{\partial}{\partial x_1}u(x_1,x_2)\right) - \dfrac{\partial^2}{\partial x_2^2}u(x_1,x_2) = 1, & \forall (x_1,x_2) \in \Omega, \\ u(x_1,x_2) = 0, & \forall (x_1,x_2) \in \partial\Omega. \end{cases}$$

a) Etablir une formulation faible du problème (P).

b) Si Ω est subdivisé en 4 triangles comme sur la figure ci-dessous,



construire explicitement la méthode d'éléments finis triangulaires de degré 1 pour approcher numériquement la solution du problème (P).

Indication: Si K est un triangle de sommets P_1, P_2, P_3 et si q est un polynôme de degré 1, i.e. $q(x) = a + bx_1 + cx_2$ avec a, b, $c \in \mathbb{R}$, alors on vérifie que

$$\iint_K q(x)dx = \text{aire de } K \cdot \frac{q(P_1)+q(P_2)+q(P_3)}{3}.$$

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 3 juin 1997

Corrigé 10

Exercice 1

Point a)

Calcul préliminaire

Comme nous l'avons fait dans la série 9, nous pouvons multiplier la première équation de (P) par une fonction $v : \overline{\Omega} \to \mathbb{R}$ "suffisamment régulière" et intégrer sur Ω. Nous obtenons

$$-\iint_\Omega \frac{\partial}{\partial x_1}\left((1+x_1+x_2)\frac{\partial u}{\partial x_1}(x)\right) v(x)\,dx - \iint_\Omega \frac{\partial^2 u}{\partial x_2^2}(x) v(x)\,dx = \iint_\Omega v(x)\,dx,$$

où $x = (x_1, x_2)$ et $dx = dx_1\,dx_2$. Posons $\vec{w}(x) = \left((1+x_1+x_2)\frac{\partial u}{\partial x_1}(x), \frac{\partial u}{\partial x_2}(x)\right)^T$. L'équation se récrit alors, en omettant pour alléger l'écriture la dépendance en x des fonctions,

$$-\iint_\Omega \left(\frac{\partial w_1}{\partial x_1} + \frac{\partial w_2}{\partial x_2}\right) v\,dx = \iint_\Omega v\,dx, \quad \text{ou} \quad -\iint_\Omega (\operatorname{div}\vec{w})\, v\,dx = \iint_\Omega v\,dx.$$

Cette formule, qui possède des dérivées partielles d'ordre 2, se transforme de la même manière que dans le polycopié (section 10.1) en une formule qui n'a que des dérivées d'ordre 1. Pour le faire, nous utilisons d'abord la formule d'analyse

$$\boxed{\operatorname{div}(v\vec{w}) = (\operatorname{div}\vec{w})v + \vec{w}\cdot\overrightarrow{\operatorname{grad}}\, v}$$

et nous obtenons

$$\iint_\Omega \vec{w}\cdot\overrightarrow{\operatorname{grad}}\, v\,dx - \iint_\Omega \operatorname{div}(v\vec{w})\,dx = \iint_\Omega v\,dx.$$

Le théorème de la divergence nous permet de déduire que

$$\iint_\Omega \vec{w}\cdot\overrightarrow{\operatorname{grad}}\, v\,dx - \int_{\partial\Omega} v\vec{w}\cdot\vec{n}\,ds = \iint_\Omega v\,dx,$$

où $\vec{n}$ est la normale unité extérieure à la frontière $\partial\Omega$ et ds est l'élément de longueur de $\partial\Omega$.

Formulation faible

Soit maintenant V l'ensemble de toutes les fonctions $g : \overline{\Omega} \to \mathbb{R}$ qui sont continues sur $\overline{\Omega}$, nulles sur $\partial\Omega$, et dont les premières dérivées partielles $\frac{\partial g}{\partial x_1}$, $\frac{\partial g}{\partial x_2}$ sont continues par morceaux.

En remplaçant $\vec{w}$ par la valeur définie ci-dessus et en tenant compte du fait que l'intégrale sur $\partial\Omega$ est nulle si $v \in V$ (car $v \equiv 0$ sur $\partial\Omega$), les calculs préliminaires nous permettent d'écrire que le problème faible correspondant au problème (P) est

chercher $u \in V$ qui satisfait

$$\iint_\Omega \left((1+x_1+x_2)\frac{\partial u}{\partial x_1}\frac{\partial v}{\partial x_1} + \frac{\partial u}{\partial x_2}\frac{\partial v}{\partial x_2}\right) dx = \iint_\Omega v\,dx, \quad \forall v \in V. \tag{PF}$$

Point b)

Approximation de Galerkin

Elle consiste à utiliser un sous-espace V_h de dimension finie de V et à résoudre le problème (PF) dans V_h au lieu de V, c'est-à-dire

chercher $u_h \in V_h$ qui satisfait

$$\iint_\Omega \left((1+x_1+x_2)\frac{\partial u_h}{\partial x_1}\frac{\partial v_h}{\partial x_1} + \frac{\partial u_h}{\partial x_2}\frac{\partial v_h}{\partial x_2}\right) dx = \iint_\Omega v_h\,dx, \quad \forall v_h \in V_h. \tag{P$_h$F}$$

Si $\varphi_1, \varphi_2, \ldots, \varphi_N$ est une base de V_h, nous pouvons écrire $u_h(x) = \sum_{i=1}^N u_i\varphi_i(x)$ et choisir $v_h = \varphi_j$, $j = 1, \ldots, N$ dans (P_hF). Soit $\vec{u}$ le N-vecteur de composantes $u_1, \ldots, u_N$, soit A la $N \times N$ matrice de coefficients

$$A_{ji} = \iint_\Omega \left((1+x_1+x_2)\frac{\partial\varphi_i}{\partial x_1}\frac{\partial\varphi_j}{\partial x_1} + \frac{\partial\varphi_i}{\partial x_2}\frac{\partial\varphi_j}{\partial x_2}\right) dx,$$

et $\vec{f}$ le N-vecteur de composantes $f_1, \ldots, f_N$ définies par

$$f_j = \iint_\Omega \varphi_j\,dx.$$

Pour obtenir la solution de (P_hF), il suffit donc de

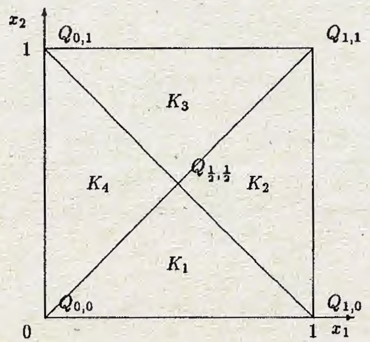
chercher $u_1, \ldots, u_N$ tels que $\displaystyle\sum_{i=1}^N A_{ji}u_i = f_j, \quad j = 1, \ldots, N,$ $\qquad ((\mathrm{P}_h\mathrm{F})_b)$

ou, de façon équivalente, de

résoudre le système linéaire $A\vec{u} = \vec{f}$. $\qquad ((\mathrm{P}_h\mathrm{F})_M)$

Cas particulier du maillage à 4 triangles

Nous construisons le sous-espace V_h de V de type **éléments finis triangulaires de degré 1.** La triangulation donnée se compose de quatre triangles, définis sur la figure ci-dessous, et des noeuds Q_{ij} de coordonnées $x_i = i$ et $x_j = j$, $0 \leqslant i, j \leqslant 1$ et $i = j = \frac{1}{2}$.



Nous n'avons qu'un point intérieur $Q_{\frac{1}{2},\frac{1}{2}}$ et donc qu'une fonction de base φ_1, définie dans le cours. La fonction φ_1 est polynômiale de degré 1 sur chaque triangle et prend les valeurs $\varphi_1(Q_{ij}) = 0$ si $(i, j) \neq (\frac{1}{2}, \frac{1}{2})$ et $\varphi_1(Q_{\frac{1}{2},\frac{1}{2}}) = 1$. Son support est $\bigcup_{i=1}^4 K_i$ et son graphe est une pyramide à base carrée dans $\mathbb{R}^3$. En conséquence, son gradient vaut

$$\overrightarrow{\text{grad}}\,\varphi_1 = \begin{pmatrix} 0 \\ 2 \end{pmatrix} \text{ sur } K_1, \quad \overrightarrow{\text{grad}}\,\varphi_1 = \begin{pmatrix} -2 \\ 0 \end{pmatrix} \text{ sur } K_2,$$

$$\overrightarrow{\text{grad}}\,\varphi_1 = \begin{pmatrix} 0 \\ -2 \end{pmatrix} \text{ sur } K_3, \quad \overrightarrow{\text{grad}}\,\varphi_1 = \begin{pmatrix} 2 \\ 0 \end{pmatrix} \text{ sur } K_4.$$

Avec cette triangulation, la matrice A et le vecteur $\vec{f}$ appartiennent respectivement à $M_1(\mathbb{R})$ et $\mathbb{R}^1$ et valent

$$A_{11} = \sum_{i=1}^{4} \iint_{K_i} \left((1 + x_1 + x_2) \left(\frac{\partial \varphi_1}{\partial x_1} \right)^2 + \left(\frac{\partial \varphi_1}{\partial x_2} \right)^2 \right) dx$$

$$= \iint_{K_1} 4\, dx + \iint_{K_2} 4(1 + x_1 + x_2)\, dx + \iint_{K_3} 4\, dx + \iint_{K_4} 4(1 + x_1 + x_2)\, dx \qquad \text{Ok.}$$

et

$$f_1 = \sum_{i=1}^{4} \iint_{K_i} \varphi_1\, dx.$$

Remarquons premièrement que les aires de tous les triangles K_i, $i = 1$ à 4, sont égales et valent $\mathcal{A} = \frac{1}{4}$. Le calcul des intégrales se fait alors facilement en utilisant l'indication donnée. Nous obtenons que la matrice A se réduit à $A = (A_{11}) = (\frac{4}{3}\mathcal{A} \cdot 18)$ et le vecteur $\vec{f}$ vaut $\vec{f} = (f_1) = (\frac{4}{3}\mathcal{A})$.
Le système linéaire se résume donc à $\frac{4}{3}\mathcal{A} \cdot 18 u_1 = \frac{4}{3}\mathcal{A}$ et sa solution est $u_1 = \frac{1}{18}$.

Donc nous obtenons l'approximation par éléments finis u_h de la solution exacte u du problème faible (PF):

$$u_h(x) = \frac{1}{18} \cdot \varphi_1(x).$$

Liste des errata du corrigé de la série 9.

- Au point c), lire "$\frac{2}{h} + 2N$" au lieu de "$2N + \frac{1}{h}$", pour le coefficient $a_{N,N}$.
- Dans la légende du graphe, lire " ... *et* 63 ×." au lieu de " ... *et* 64 ×."

$$b_1 = \sum_i \iint_{K_i} \varphi_1(x)\, dK_i \quad \text{, avec : } \iint_{K_1} \varphi_1(x)\, dK_1 = \text{Aire} \cdot \frac{1}{3} (0 + 0 + 1) = \frac{1}{12}$$

idem pour les autres $\Rightarrow$ $b_1 = 4 \cdot \frac{1}{12} = \frac{4}{12} = \boxed{\frac{1}{3}}$

$$\iint_{K_1} 4\, dK_1 = 4 \iint_{K_1} dK_1 = 4 \cdot \text{Aire}(K_1) = 4 \cdot \frac{1}{4} = 1$$

(pas besoin de formule de quadrature pour ce calcul.)

$$\iint_{K_3} 4\, dK_3 = 1$$

$$\iint_{K_2} 4(1 + x_1 + x_2)\, dK_2 = 4 \iint_{K_2} 1 + x_2 + x_1\, dK_2 = (1).$$

avec l'indication, on a : $Q_1 = (1,0)$, $Q_2 = (1,1)$, $Q_3 = (1/2, 1/2)$

$$(1) = \underset{1/4}{\overset{\text{Aire}}{\shortparallel}} \cdot 4 \cdot \frac{1}{3} \Big(\underbrace{1 + 1 + 0}_{f(Q_1)} + \underbrace{1 + 1 + 1}_{f(Q_2)} + \underbrace{1 + 1/2 + 1/2}_{f(Q_3)} \Big) = \frac{7}{3}$$

$$\iint_{K_4} 4(1 + x_1 + x_2)\, dK_4 = 4 \iint_{K_4} 1 + x_1 + x_2\, dK_4 = (2)$$

avec : $Q_1 = (0,0)$, $Q_2 = (0,1)$, $Q_3 = (1/2, 1/2)$

$$(2) = \underset{1/4}{\overset{\text{Aire}}{\shortparallel}} \cdot 4 \cdot \frac{1}{3} \left(1 + 0 + 0 + 1 + 0 + 1 + 1 + \frac{1}{2} + \frac{1}{2} \right) = \frac{5}{3}$$

$$A_{11} = 1 + 1 + \frac{7}{3} + \frac{5}{3} = \frac{6}{3} + \frac{12}{3} = \frac{18}{3} = 6$$

$$\Longrightarrow \quad 6 \cdot u_1 = \frac{1}{3}$$

$$\Longrightarrow \quad u_1 = \frac{1}{18} \quad \#$$

ANALYSE NUMÉRIQUE
INFO, PH, PH-Fac
Prof. J. Rappaz

Le 10 juin 1997

Série 11

Exercice 1

On considère le problème 1D de la chaleur non linéaire suivant :

Trouver une fonction u dépendant de x et de t telle que

$$(P)\quad \begin{cases} \dfrac{\partial u}{\partial t}(x,t) - \dfrac{\partial^2 u}{\partial x^2}(x,t) = u(x,t)^3, & \forall x \in]0,1[,\ \forall t > 0, \\ u(0,t) = u(1,t) = 0, & \forall t > 0, \\ u(x,0) = w(x), & \forall x \in]0,1[, \end{cases}$$

où $w(x)$ est une condition initiale donnée.

Pour établir une approximation numérique de (P), on utilise une méthode de **<u>différence finies</u> en espace et en temps**. Si $\tau > 0$ est le pas de temps fixé et si $h = \frac{1}{N+1}$ est le pas spatial (N entier positif), on pose $t_n = n\tau$, $n = 1, 2, \ldots$ et $x_j = j\,h$, $j = 0, 1, 2, \ldots, N+1$. On note u_j^n l'approximation de $u(x_j, t_n)$.

a) Etablir le schéma d'approximation du problème (P) en utilisant la méthode de (Crank-Nicholson)
(Remarquer le caractère implicite et non-linéaire du schéma !)
Trapèze

b) Pour calculer u_j^{n+1}, $1 \leqslant j \leqslant N$, à partir des valeurs u_j^n, $1 \leqslant j \leqslant N$, on utilisera la méthode de Newton.

Etablir explicitement cette méthode.

Corrigé 11

Exercice 1

Point a)

Nous commençons la résolution numérique du problème (P) en discrétisant par rapport à la variable x. La solution est donnée dans le cours (formules (11.10-11)), si nous remplaçons $f_i(t)$ par $(u_i(t))^3$, $1 \leqslant i \leqslant N$.

Nous utiliserons les notations suivantes :

$\vec{u}(t)$	pour le N-vecteur de composantes	$u_i(t)$,	$1 \leqslant i \leqslant N$,
$(\vec{u}(t))^3$	pour le N-vecteur de composantes	$u_i(t)^3$,	$1 \leqslant i \leqslant N$,
$\dot{\vec{u}}(t)$	pour le N-vecteur de composantes	$\frac{du_i(t)}{dt}$,	$1 \leqslant i \leqslant N$
$\vec{w}$	pour le N-vecteur de composantes	$w(x_i)$,	$1 \leqslant i \leqslant N$,

et A pour la $N \times N$-matrice $\frac{1}{h^2}\begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix}$. La semi-discrétisation en espace du problème (P) par la méthode des différences finies, s'écrit alors

$$\begin{cases} \dot{\vec{u}}(t) = -A\vec{u}(t) + (\vec{u}(t))^3, & \forall t > 0, \\ \vec{u}(0) = \vec{w}. \end{cases} \qquad (P_h)$$

Les fonctions $u_i(t)$, $1 \leqslant i \leqslant N$, sont les inconnues du problème.

La méthode de Crank-Nicholson (formule (11.36) du cours) appliquée à ce système donne le schéma

$$\begin{cases} \dfrac{\vec{u}^{n+1} - \vec{u}^n}{\tau} = \dfrac{(-A\vec{u}^{n+1} + (\vec{u}^{n+1})^3) + (-A\vec{u}^n + (\vec{u}^n)^3)}{2} & n = 0, 1, 2, \dots, \\ \vec{u}^0 = \vec{w}, \end{cases}$$

ou, de façon équivalente

$$\begin{cases} (I + \frac{\tau}{2}A)\,\vec{u}^{n+1} - \frac{\tau}{2}(\vec{u}^{n+1})^3 = (I - \frac{\tau}{2}A)\,\vec{u}^n + \frac{\tau}{2}(\vec{u}^n)^3, & n = 0, 1, 2, \dots, \\ \vec{u}^0 = \vec{w}. \end{cases} \qquad (P_{h\tau})$$

Pour obtenir $\vec{u}^{n+1}$ à partir de $\vec{u}^n$, nous devons **résoudre** un système **cubique** de N équations à N inconnues : ce schéma est donc **implicite** et **non linéaire**.

Point b)

Définissons la fonction $\vec{F}$ par

$$\begin{aligned} \vec{F} &: \mathbb{R}^N \longrightarrow \mathbb{R}^N, \\ \vec{x} &\longmapsto (I + \tfrac{\tau}{2}A)\,\vec{x} - \tfrac{\tau}{2}(\vec{x})^3 - (I - \tfrac{\tau}{2}A)\,\vec{u}^n - \tfrac{\tau}{2}(\vec{u}^n)^3, \end{aligned}$$

où $\vec{x}^3$ est le N-vecteur de composantes x_i^3, $1 \leqslant i \leqslant N$. Le schéma $(P_{h\tau})$ devient alors à

chercher $\vec{y} \in \mathbb{R}^N$ satisfaisant $\vec{F}(\vec{y}) = 0$,

où $\vec{y}$ est l'inconnue, c'est-à-dire le vecteur $\vec{u}^{n+1}$.

Si nous appliquons la généralisation de la méthode de Newton à des systèmes d'équations (formule (7.22) du cours), pour résoudre ce problème de recherche de zéro, nous obtenons la formule de récurrence suivante

$$D\vec{F}(\vec{y}^k)(\vec{y}^{k+1} - \vec{y}^k) = -\vec{F}(\vec{y}^k), \qquad k = 1, 2, \dots,$$

où $D\vec{F}$ est la matrice jacobienne de la fonction $\vec{F}$. Explicitement, elle vaut $D\vec{F}(\vec{x}) = I + \frac{\tau}{2}A - 3\frac{\tau}{2}B(\vec{x})$ où $B(\vec{x})$ est la $N \times N$-matrice diagonale, de coefficients diagonaux $(x_i)^2$, $1 \leqslant i \leqslant N$.

Cette formule converge quadratiquement vers l'inconnue $\vec{y}$ si nous partons suffisamment proche de la solution. Si la fonction $\vec{u}(t)$ varie peu durant un intervalle de temps, il est logique de choisir $\vec{y}^0$ valant $\vec{u}^n$.

Erratum du corrigé de la série 10.

Lire "*Le théorème de la divergence permet de déduire que*

$$\iint_\Omega \vec{w} \cdot \overrightarrow{\mathrm{grad}}\, v \, dx - \int_{\partial\Omega} v\vec{w} \cdot \vec{n} \, ds = \iint_\Omega v \, dx,\text{"}$$

plutôt que "*Le théorème de la divergence permet de déduire que*

$$\iint_\Omega \vec{w} \cdot \overrightarrow{\mathrm{grad}}\, v \, dx - \iint_{\partial\Omega} v\vec{w} \cdot \vec{n} \, ds = \iint_\Omega v \, dx,\text{"}.$$

ANALYSE NUMÉRIQUE
Informatique 2ème année
Prof. J. Rappaz

Le 3 juillet 1996

Examen propédeutique

4 exercices à faire entre 8h15 et 10h00

Rappels : **Seul un formulaire** sur une feuille A4 recto-verso est autorisé.
La **calculatrice** est **interdite**.

Remarques : Déposez vos **cartes d'étudiants** sur la table.
Les **cartes d'étudiants** et les **formulaires** seront contrôlés.
Laissez les feuilles **agrafées**.

Exercice 1 (2½ points)

c.f. série 4

Soient a, b, c et d quatre nombres non nuls donnés. Si N est un entier positif, on considère la $N \times N$-matrice A et le N-vecteur $\vec{b}$ définis par

$$A = \begin{pmatrix} a & d & & & & 0 \\ & a & d & & & \\ & & \ddots & \ddots & & \\ & 0 & & \ddots & \ddots & \\ & & & & a & d \\ c & c & \dots & \dots & c & c \end{pmatrix} \quad \text{et} \quad \vec{b} = \begin{pmatrix} b \\ b \\ b \\ \vdots \\ b \\ b \end{pmatrix}.$$

En supposant que A soit une matrice régulière, on demande d'écrire un algorithme de résolution du système $A\vec{x} = \vec{b}$. Cet algorithme aura

- pour entrées, les valeurs a, c, d, N et le vecteur $\vec{b}$ de composantes $b_j = b$, $1 \leqslant j \leqslant N$;
- pour sortie, le vecteur $\vec{b}$ de composantes $b_j = x_j$, $1 \leqslant j \leqslant N$, où les valeurs x_j sont les composantes de la solution $\vec{x}$.

Important : l'algorithme ne devra **pas utiliser d'autres tableaux** *(tableau = variable indicée)* que le vecteur $\vec{b}$.

Exercice 2 (2½ points)

Soient M points $t_1 < t_2 < \ldots < t_M$ tels que $-1 \leqslant t_1$ et $t_M \leqslant 1$ et soit $\{\varphi_i\}_{i=1}^M$ la base de Lagrange des polynômes de degré $(M-1)$ associée à ces M points (M est un entier, $M \geqslant 1$). On pose $\omega_j = \int_{-1}^{1} \varphi_j(t)dt$, $j = 1, 2, \ldots, M$ et $J(g) = \sum_{j=1}^{M} \omega_j g(t_j)$ où g est une fonction continue sur l'intervalle $[-1, +1]$.

a) Démontrer que si g est un polynôme de degré $M-1$, alors on a $J(g) = \int_{-1}^{1} g(t)dt$.

TOURNEZ LA PAGE, S.V.P.

b) Démontrer qu'on a la relation $\sum_{j=1}^{M} \omega_j t_j^s = \dfrac{1+(-1)^s}{s+1}$, pour tout $s = 0, 1, 2, \ldots, M-1$.

×Exercice 3 (2 points)

On considère le problème de Cauchy:

$$(\mathcal{P}) \quad \begin{cases} \dot{u}(t) = -(u(t))^3 + \cos t, & \text{si} \quad t > 0, \\ u(0) = 0, \end{cases}$$

a) Montrer, en utilisant un théorème du cours, que $(\mathcal{P})$ a une et une seule solution u.

b) Si h est un paramètre positif donné, si $t_j = jh$ avec $j = 0, 1, 2, \ldots$, écrire le schéma d'Euler rétrograde qui permet de calculer $u^{n+1} \simeq u(t_{n+1})$ à partir de $u^n \simeq u(t_n)$.

c) Pour calculer effectivement u^1 à partir de $u^0 = 0$ avec le schéma trouvé en b), on utilisera la méthode de Newton. Effectuer un seul pas de cette méthode.

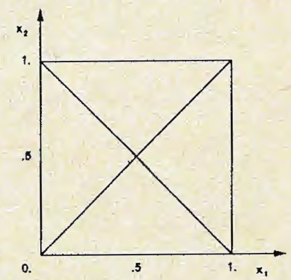
Exercice 4 (3 points)

Soit Ω le carré unité de $\mathbb{R}^2$ de frontière $\partial\Omega$. On se propose de chercher une fonction $u : \overline{\Omega} \longrightarrow \mathbb{R}$ qui satisfait

$$(\mathcal{P}) \quad \begin{cases} -\dfrac{\partial}{\partial x_1}\left((1 + x_1 + x_2)\dfrac{\partial}{\partial x_1}u(x_1, x_2)\right) - \dfrac{\partial^2}{\partial x_2^2}u(x_1, x_2) = 1, & \forall (x_1, x_2) \in \Omega, \\ u(x_1, x_2) = 0, & \forall (x_1, x_2) \in \partial\Omega. \end{cases}$$

a) Etablir une formulation faible du problème $(\mathcal{P})$.

b) Si Ω est subdivisé en 4 triangles comme sur la figure ci-dessous,



construire explicitement la méthode d'éléments finis triangulaires de degré 1 pour approcher numériquement la solution du problème $(\mathcal{P})$.

Indication : Si K est un triangle de sommets P_1, P_2, P_3 et si q est un polynôme de degré 1, alors on a

$$\iint_K q(x)dx = \text{aire de } K \cdot \frac{q(P_1) + q(P_2) + q(P_3)}{3}.$$

Quelques conseils concernant l'utilisation de votre Mac :

Netscape :

- Allumez le Mac, `username:student` sans mot de passe, tapez votre nom ainsi que votre date de naissance. Lancez l'application `Netscape`, menu `bookmark`, sélectionnez `analyse numérique`. Vous êtes alors sur la page d'accueil du groupe d'analyse et simulation numériques du prof. J. Rappaz. Vous pouvez également aboutir à cette page depuis la page d'accueil du DMA, à l'adresse `http://dmawww.epfl.ch`.
- Récupérez l'exercice de programmation de la semaine sur votre Mac. L'application `Think Pascal` s'ouvre automatiquement. Le fichier est automatiquement placé dans le dossier `Divers` puis `Netscape f`.

Think Pascal :

- Complétez le programme. Vérifier la syntaxe menu `run`, sous-menu `check syntax`.
- menu `project`, sous-menu `new project`, allez dans le dossier `Divers` puis `Netscape f` et donnez le nom `eltfini` à votre projet.
- menu `project`, sous-menu `add file`, ajoutez le fichier `eltfini.pas`.
- menu `run`, sous-menu `build`, le programme `eltfini` est construit.
- menu `run`, sous-menu `go`, le programme `eltfini` est éxécuté.

Avant de partir :

- Sélectionnez à nouveau l'application `Netscape`.
- Retournez sur la page d'accueil du prof. J. Rappaz.
- Indiquez votre présence.

$$\begin{pmatrix} 2/h & -1/h & & \\ -1/h & \ddots & \ddots & \\ & \ddots & \ddots & \\ & & 2/h & -1/h \end{pmatrix} \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} = \begin{pmatrix} h\cdot f(1,h) \\ \vdots \\ h\cdot f(n,h) \end{pmatrix}$$

$u(x) = \sin(\alpha x)$

$u''(x) = -\alpha^2 \sin(\alpha x)$

$-u''(x) = \alpha^2 \sin(\alpha x)$

π^2

$-u''(x) = \pi^2 \sin(\pi x)$

$u(x) = \sin(\pi x)$

$-u''(x) = f(x)$

On recherche donc la fonction tq.

$u(x) = \sin(\alpha \cdot x)$

$u''(x) = \sin(\alpha \cdot x) \cdot \alpha^2$

ANALYSE NUMÉRIQUE le 20 mai 1997

Exercice de programmation

Série 9

Position du problème :

Le but de cet exercice est de compléter et d'utiliser un programme permettant d'approcher la solution d'un problème aux limites du deuxième ordre par une méthode d'éléments finis.

Soit f une fonction continue sur l'intervalle $[0,1]$. Nous cherchons une fonction u deux fois continûment dérivable sur $[0,1]$ telle que

$$\begin{aligned} -u''(x) &= f(x) \qquad \text{si } 0 < x < 1, \\ u(0) &= u(1) = 0. \end{aligned} \tag{1}$$

Approximons le problème ci-dessus par une méthode de Galerkin (voir polycopié p. 83). Multiplions la première équation de (1) par une fonction v une fois continûment dérivable, nulle en $x = 0$ et $x = 1$, et intégrons par parties. Nous obtenons

$$\int_0^1 u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx. \tag{2}$$

Soit V l'ensemble de toutes les fonctions g continues, de première dérivée g' continue par morceaux et telle que $g(0) = g(1) = 0$. Le problème faible consiste donc à chercher une fonction $u \in V$ telle que l'équation (2) soit satisfaite pour toute fonction $v \in V$.

Pour tout entier N strictement positif, divisons l'intervalle $[0,1]$ en $(N+1)$ parties de longueur $h = 1/(N+1)$ et notons $x_i = ih$, $i = 0,1,\dots,N+1$. Soit φ_1, ..., φ_N les N fonctions de base définies p. 86 du polycopié, et soit V_h le sous-espace vectoriel de V, engendré par les fonctions φ_i. Une approximation de Galerkin de (2) consiste donc à chercher une fonction $u_h \in V_h$ telle que l'équation

$$\int_0^1 u_h'(x)v_h'(x)dx = \int_0^1 f(x)v_h(x)dx \tag{3}$$

soit satisfaite pour toute fonction $v_h \in V_h$. Décomposons u_h dans la base de V_h, i.e. $u_h(x) = u_1\varphi_1(x) + \cdots + u_N\varphi_N(x)$, et choisissons $v_h = \varphi_1,\dots,\varphi_N$. Le problème (3) est équivalent à chercher les coefficients $u_1,\dots,u_N$ tels que

$$\sum_{i=1}^N u_i \left(\int_0^1 \varphi_i'(x)\varphi_j'(x)dx\right) = \int_0^1 f(x)\varphi_j(x)dx \qquad j = 1,\dots,N. \tag{4}$$

Ecrivons (4) sous forme d'un système linéaire de N équations à N inconnues, nous obtenons

$$A\vec{u} = \vec{f}, \tag{5}$$

où A est la $N \times N$ matrice et $\vec{f}$ est le N-vecteur de coefficients

$$A_{ji} = \int_0^1 \varphi_i'(x)\varphi_j'(x)dx, \qquad f_j = \int_0^1 f(x)\varphi_j(x)dx, \tag{6}$$

$i,j = 1,\dots,N$. La deuxième intégrale de (6) sera évaluée numériquement en utilisant la formule des trapèzes, voir la formule (9.16) p. 87 du polycopié. Finalement, le système linéaire (5) sera résolu en utilisant la décomposition de Choleski d'une matrice tridiagonale.

Travail demandé:

Le programme `eltfini.pas` est à votre disposition. A partir de la donnée de f et N ce programme fournit une approximation de u, solution de (1). Ce programme est incomplet, en particulier vous devez calculer les coefficients de la matrice A et du second membre $\vec{f}$.

Outils informatiques:

Les logiciels à votre disposition sont :

- `Netscape` pour aller chercher le fichier `eltfini.pas` sur la page WWW du prof. J. Rappaz.
- `Think Pascal` pour compléter le programme, le compiler et l'éxécuter.

Questions:

- Complétez le programme.
- La fonction f est définie par $f(x) = -2$, vérifiez que la solution du problème (1) est donnée par $u(x) = x(x-1)$. Vérifier que la solution numérique u_h est exacte aux noeuds, c'est-à-dire que $u_h(x_i) = u(x_i)$, $i = 1, ..., N$.
- La fonction f est définie par $f(x) = \pi^2 sin(\pi x)$, vérifier que la solution du problème (1) est donnée par $u(x) = sin(\pi x)$. Vérifier que l'erreur maximale est approximativement quatre fois plus petite chaque fois que N est multiplié par deux.

$-u''(x) = f(x)$

diag[1] = 2;
secmem[1] = h · h · f(1)
diag[i] = 2
ss diag[i-1] = -1
secmem[i] := $h^2 \cdot f(i \cdot h)$

```
program eltfini;
        type
                vecteur = array[1..100] of double;

        function f (x: double): double;
{-u''(x)=f(x)}
        begin
                f := -2;
        end;

        function uexact (x: double): double;
{la solution de -u''(x)=f(x), u(0)=u(1)=0}
        begin
                uexact := x * (x - 1);
        end;

{remplissage de la matrice de diagonale diag et sous-diagonale}
{ ssdiag et du second membre secmem}
        procedure remplissage (N: integer; h: double; var diag, ssdiag, secmem: vecteur);
                var
                        i: integer;
        begin{remplissage}
                diag[1] := { *********** A COMPLETER **************};
                secmem[1] := { *********** A COMPLETER **************};
                for i := 2 to N do
                        begin
                                ssdiag[i - 1] :={ *********** A COMPLETER **************};
                                diag[i] := { *********** A COMPLETER **************};
                                secmem[i] :={ *********** A COMPLETER **************};
                        end;{for}
        end;{remplissage}

{decomposition de choleski A=LL^T. }
{En entree diag et ssdiag sont}
{les diagonales et sous-diagonales de la matrice A. }
{En sortie diag et ssdiag sont}
{les diagonales et sous-diagonales de la matrice L}
        procedure choleski (N: integer; var diag, ssdiag: vecteur);
                var
                        k: integer;
        begin{choleski}
                diag[1] := sqrt(diag[1]);
                for k := 1 to N - 1 do
                        begin
                                ssdiag[k] := ssdiag[k] / diag[k];
                                diag[k + 1] := sqrt(diag[k + 1] - ssdiag[k] * ssdiag[k])
                        end;
        end;{choleski}

{resolution des deux systemes lineaires}
{En entree secmem est le second membre du systeme lineaire}
{En sortie secmem est la solution du systeme lineaire}
        procedure resol (N: integer; diag, ssdiag: vecteur; var secmem: vecteur);
                var
                        i: integer;
        begin
                secmem[1] := secmem[1] / diag[1];
                for i := 1 to N - 1 do
                        begin
                                secmem[i + 1] := (secmem[i + 1] - ssdiag[i] * secmem[i])
                        end;
                secmem[N] := secmem[N] / diag[N];
                for i := N - 1 downto 1 do
                        secmem[i] := (secmem[i] - ssdiag[i] * secmem[i + 1]) / diag[i];
        end;{resol}

        var
                N, j, jmax: integer;
                h, err, errmax: double;
                diag, ssdiag, secmem: vecteur;

begin{eltfini}

        WRITE('Entrer N : ');
        READLN(N);
        h := 1 / (N + 1);


        remplissage(N, h, diag, ssdiag, secmem);
        choleski(N, diag, ssdiag);
        resol(N, diag, ssdiag, secmem);

        errmax := 0;
        for j := 1 to N do
                begin
                        WRITELN('j = ', j, ' x[ j ] = ', j * h, ' u[ j ] = ', secmem[j],
                        err := abs(secmem[j] - uexact(j * h));
                        if (err > errmax) then
                                begin
                                        errmax := err;
                                        jmax := j;
                                end;
                end;
        WRITELN('erreur max : ', errmax, ' au point x= ', jmax * h);

end.{eltfini}
```

 Le 13 avril 1995

Solution de la série 2

Exercice 1

Explicitement, $r'_h(x_0)$, $r''_h(x_0)$ et $r'''_h(x_0)$ valent

$$r'_h(x_0) = \frac{\delta_h f(x_0)}{h} = \frac{f(x_0+\frac{h}{2}) - f(x_0-\frac{h}{2})}{h},$$
$$r''_h(x_0) = \frac{\delta_h^2 f(x_0)}{h^2} = \frac{f(x_0+h) - 2f(x_0) + f(x_0-h)}{h^2} \quad \text{et}$$
$$r'''_h(x_0) = \frac{\delta_h^3 f(x_0)}{h^3} = \frac{f(x_0+\frac{3}{2}h) - 3f(x_0+\frac{h}{2}) + 3f(x_0-\frac{h}{2}) - f(x_0-\frac{3}{2}h)}{h^3}.$$

En double précision, i.e. avec 16 chiffres significatifs, les calculs donnent

h	r'_h	$\lvert f'(x_0) - r'_h\rvert$	$\frac{\lvert f'(x_0)-r'_h\rvert}{h^2}$	
1.00e-01	2.719414587473179e+00	1.132759014133899e-03	1.132759014133899e-01	
5.00e-02	2.718564991664882e+00	2.831632058373224e-04	1.132652823349289e-01	
1.00e-02	2.718293154647444e+00	1.132618839916333e-05	1.132618839916333e-01	
5.00e-03	2.718284660003434e+00	2.8315443887[illegible]228e-06	1.132617755494891e-01	
1.00e-03	2.718281941720413e+00	1.13261367[illegible]74e-07	1.132613678400674e-01	
1.00e-04	2.718281829592328e+00	1.13328[illegible]41e-09	1.133283245025041e-01	
1.00e-05	2.718281828517632e+00	5.858691309867936e-11	5.858691309867935e-01	◀
1.00e-06	2.718281828517632e+00	5.858735718788921e-11	5.858735718788921e+01	
1.00e-08	2.718281821856294e+00	6.602750790563050e-09	6.602750790563049e+07	

h	r''_h	$\lvert f''(x_0) - r''_h\rvert$	$\frac{\lvert f''(x_0)-r''_h\rvert}{h^2}$	
1.00e-01	2.720547818529306e+00	2.265990070260848e-03	2.265990070260848e-01	
5.00e-02	2.718848184367850e+00	5.663559088047165e-04	2.265423635218866e-01	
1.00e-02	2.718304480882061e+00	2.265242301602299e-05	2.265242301602299e-01	
5.00e-03	2.718287491561000e+00	5.663101954933580e-06	2.265240781973432e-01	
1.00e-03	2.718282055003129e+00	2.265440843807198e-07	2.265440843807198e-01	
1.00e-04	2.718281866265215e+00	3.780617019444321e-08	3.780617019444321e+00	◀
1.00e-05	2.718287817060627e+00	5.988601581741193e-06	5.988601581741192e+04	
1.00e-06	2.718270053492233e+00	1.17749668[illegible]1722e-05	1.177496681181722e+07	
1.00e-08	0.000000000000000e+00	2.7182818[illegible]045e+00	2.718281828459045e+16	

h	r'''_h	$\lvert f'''(x_0) - r'''_h\rvert$	$\frac{\lvert f'''(x_0)-r'''_h\rvert}{h^2}$	
1.00e-01	2.721681521821839e+00	3.399693362793776e-03	3.399693362793776e-01	
5.00e-02	2.719131406571762e+00	8.495781127164648e-04	3.398312450865859e-01	
1.00e-02	2.718315809335791e+00	3.398087674622019e-05	3.398087674622019e-01	
5.00e-03	2.718290335934625e+00	8.507475580010748e-06	3.402990232004299e-01	
1.00e-03	2.718284264346948e+00	2.435887903384781e-06	2.435887903384781e+00	◀
1.00e-04	2.717381875072533e+00	8.999533865119425e-04	8.999533865119424e+04	
1.00e-05	8.881784197001250e-01	1.830103408758920e+00	1.830103408758920e+10	
1.00e-06	4.440892098500627e+02	4.413709280216036e+02	4.413709280216036e+14	
1.00e-08	1.776356839400250e+09	1.776356836681968e+09	1.776356836681968e+25	

Les valeurs exactes, à 16 chiffres significatifs, de $f'(1)$, $f''(1)$ et $f'''(1)$ sont toutes 2.718281828459045e+00.

Les conclusions sont les suivantes :

1) Avant que les erreurs d'arrondi ne prennent de l'importance, les valeurs des dernières colonnes sont presque constantes, i.e nous observons l'ordre de convergence en h^2.
2) Pour les petites valeurs de h les erreurs d'arrondi deviennent importantes vis-à-vis des erreurs de troncature. Il existe une valeur optimale de h, signalée par ◀, qui minimise l'erreur totale, i.e. qui fait un compromis entre l'erreur de troncature et l'erreur d'arrondi.
3) Ces h optimaux augmentent lorsque l'ordre de dérivation augmente. Des tableaux ci-dessus, nous tirons qu'ils valent approximativement $h_1 \simeq 10^{-5}$, $h_2 \simeq 10^{-4}$ et $h_3 \simeq 10^{-3}$. Les indices précisent l'ordre de dérivation.

   Une généralisation du raisonnement présenté au paragraphe 2.3 du cours donne les approximations des erreurs totales suivantes

$$E_1^h \simeq 2\eta\frac{|f(x_0)|}{h} + \frac{1}{24}\left|f^{(3)}(x_0)\right|h^2,$$
$$E_2^h \simeq 4\eta\frac{|f(x_0)|}{h^2} + \frac{1}{12}\left|f^{(4)}(x_0)\right|h^2 \quad \text{et}$$
$$E_3^h \simeq 8\eta\frac{|f(x_0)|}{h^3} + \frac{41}{320}\left|f^{(5)}(x_0)\right|h^2.$$

   Elles sont minimales pour les h suivants

$$h_1 = \left(24\eta\left|\frac{f(x_0)}{f^{(3)}(x_0)}\right|\right)^{\frac{1}{3}},$$
$$h_2 = \left(48\eta\left|\frac{f(x_0)}{f^{(4)}(x_0)}\right|\right)^{\frac{1}{4}} \quad \text{et}$$
$$h_3 = \left(\frac{3840}{41}\eta\left|\frac{f(x_0)}{f^{(5)}(x_0)}\right|\right)^{\frac{1}{5}}.$$

   Dans notre cas, $\eta = 10^{-16}$ et $f(x_0) = f^{(n)}(x_0)$, $\forall n \in \mathbb{N}$. Les h optimaux théoriques valent donc approximativement $h_1 \simeq 1.34\cdot 10^{-5}$, $h_2 \simeq 2.63\cdot 10^{-4}$ et $h_3 \simeq 1.56\cdot 10^{-3}$ et correspondent bien à nos résultats expérimentaux.
4) La dérivation numérique n'est pas très stable vis-à-vis des erreurs d'arrondi, i.e. nous perdons rapidement de la précision en augmentant l'ordre de dérivation.

Exercice 2

Point 2.1)

Nous avons

$$g(x) = f(x_0) + \frac{\Delta_h f(x_0)}{h}(x - x_0) + \frac{\Delta_h^2 f(x_0)}{2!h^2}(x - x_0)(x - x_1),$$

où $\Delta_h f(x_0) = f(x_0 + h) - f(x_0)$ et $\Delta_h^2 f(x_0) = f(x_0 + 2h) - 2f(x_0 + h) + f(x_0)$. Alors, en tenant compte des égalités $x_1 = x_0 + h$ et $x_2 = x_0 + 2h$, nous obtenons

$$\begin{aligned} g(x_0) &= f(x_0), \\ g(x_1) &= f(x_0) + \Delta_h f(x_0) = f(x_1) \quad \text{et} \\ g(x_2) &= f(x_0) + 2\Delta_h f(x_0) + \Delta_h^2 f(x_0) = f(x_2). \end{aligned}$$

Ceci montre que g est l'interpolant, de degré 2, de f passant par les points x_0, x_1, et x_2.

Considérons la fonction $r = f - g$. Nous observons que

1) r est (au moins) trois fois continûment dérivable et
2) $r(x_0) = r(x_1) = r(x_2) = 0$.

Par le théorème de Rolle, nous savons

a) $\exists \xi_0 \in [x_0, x_1] \quad t.q \quad r'(\xi_0) = 0 \quad i.e. \quad f'(\xi_0) = g'(\xi_0)$ et
b) $\exists \xi_1 \in [x_1, x_2] \quad t.q \quad r'(\xi_1) = 0 \quad i.e. \quad f'(\xi_1) = g'(\xi_1)$.

Point 2.2)

De 2.1), nous remarquons que $r'(\xi_0) = 0$, que $r'(\xi_1) = 0$ et que r' est une fonction (au moins) deux fois continûment dérivable. En utilisant encore le théorème de Rolle, mais sur r', nous trouvons que

$$\exists \eta \in [\xi_0, \xi_1] \quad t.q \quad r''(\eta) = 0.$$

Alors

$$r''(x) = \int_\eta^x r'''(s)ds = \int_\eta^x \{f'''(s) - g'''(s)\}ds = \int_\eta^x f'''(s)ds,$$

car g est un polynôme de degré deux.

Point 2.3)

Si $x \in [x_0, x_1]$, nous trouvons, en utilisant l'indication, que

$$\text{(A)} \qquad |r(x)| = \left|\int_{x_0}^x r'(s)ds\right| \leqslant \int_{x_0}^{x_1} |r'(s)|\, ds \leqslant h \max_{s\in[x_0,x_1]} |r'(s)| \quad \text{et}$$

$$\text{(B)} \qquad |r'(x)| = \left|\int_{\xi_0}^x r''(s)ds\right| \leqslant \int_{\xi_0}^{x_1} |r''(s)|\, ds \leqslant h \max_{s\in[x_0,x_1]} |r''(s)|.$$

De 2.2), nous savons que si $x \in [x_0, x_1]$ nous avons

$$|r''(x)| = \left|\int_\eta^x f'''(s)ds\right|.$$

Alors

$$|r''(x)| \leqslant \left\{\int_\eta^x |f'''(s)|\, ds\right\} \leqslant \int_{x_0}^{x_2} |f'''(s)|\, ds, \quad \text{et donc}$$

$$|r''(x)| \leqslant 2h \max_{s\in[x_0,x_2]} |f'''(s)|, \quad \text{car} \quad x_2 - x_0 = 2h.$$

Il découle de (B) que, pour $x \in [x_0, x_1]$, c[illegible] :

$$|r'(x)| \leqslant 2h^2 \max_{s\in[x_0,x_2]} |f'''(s)|.$$

Donc $\max_{x\in[x_0,x_1]} |r'(x)| \leqslant 2h^2 \max_{s\in[x_0,x_2]} |f'''(s)|$. Alors de (A), nous pouvons conclure que

$$\max_{x\in[x_0,x_2]} |r(x)| \leqslant 2h^3 \max_{t\in[x_0,x_2]} |f'''(t)|.$$

Idem pour l'intervalle $[x_1, x_2]$ en prenant ξ_1 à la place de ξ_0.

Comparaison avec la formule de Taylor :

Soit G le polynôme de Taylor, de degré 2, de la fonction f autour du point x_0 i.e.

$$G(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x_0)}{2}(x - x_0)^2.$$

Nous avons l'égalité suivante

$$f(x) = G(x) + \frac{f'''(\xi)}{3!}(x - x_0)^3, \quad \text{où} \quad \xi \in [x_0, x].$$

Comme $x_2 - x_0 = 2h$, nous obtenons l'estimation suivante, si $x \in [x_0, x_2]$,

$$|f(x) - G(x)| \leqslant \frac{4h^3}{3} \max_{\xi\in[x_0,x_2]} |f'''(\xi)|.$$

Nous pouvons en déduire que les deux p[illegible]ynômes, g et G, donnent une approximation de f qui est de l'ordre de h^3. Cependant, la formule de Taylor [illegible]ise les dérivées de f qui ne sont pas toujours aisées à obtenir. (Par exemple, lorsque la fonction f n'est don[illegible]e sous forme tabulée ou s'obtient par un calcul compliqué.)

Solution de la série 3

Exercice 1

La formule de Gauss-Legendre à 2 points, t_1 et t_2 qui sont les zéros du polynôme de Legendre P_2, est

$$J(g)=\omega_1 g(t_1)+\omega_2 g(t_2)=\int_{-1}^{1}\tilde{g}(t)dt,$$

où $\tilde{g}(t)=g(t_1)\varphi_1(t)+g(t_2)\varphi_2(t)$ avec $\varphi_1(t)=\dfrac{t-t_2}{t_1-t_2}$ et $\varphi_2(t)=\dfrac{t-t_1}{t_2-t_1}$.

Calcul des ω_i : en réécrivant la formule ci-dessus, nous obtenons

$$\omega_1 g(t_1)+\omega_2 g(t_2)=g(t_1)\int_{-1}^{1}\varphi_1(t)dt+g(t_2)\int_{-1}^{1}\varphi_2(t)dt.$$

Donc

$$\omega_1=\int_{-1}^{1}\varphi_1(t)dt=\frac{2t_2}{t_2-t_1}\quad\text{et}\quad\omega_2=\int_{-1}^{1}\varphi_2(t)dt=\frac{2t_1}{t_1-t_2}.$$

Calcul des t_i : le polynôme de Legendre de degré 2 est

$$P_2(t)=\frac{1}{2^2 2!}\frac{d^2}{dt^2}(t^2-1)^2=\frac{1}{2}(3t^2-1).$$

Ce polynôme a deux zéros, $t_1=\frac{-1}{\sqrt{3}}$ et $t_2=\frac{1}{\sqrt{3}}$, qui donnent les deux points de Gauss, et donc $\omega_1=\omega_2=1$.

Illustration : soit $I=\int_0^4 x^2e^x dx$.

En divisant l'intervalle $[0,4]$ en N parties égales, nous avons $h=\frac{4}{N}$ et $I=\displaystyle\int_0^4 x^2e^xdx=\sum_{j=0}^{N-1}\int_{jh}^{(j+1)h}x^2e^xdx$.

La transformation $t=\left(2\frac{x-jh}{h}-1\right)$ envoie l'intervalle $[jh,(j+1)h]$ sur $[-1,1]$. Nous utilisons la transformation inverse, i.e. $x=\frac{h}{2}(t+1)+jh$, comme changement de variable. Nous obtenons

$$\int_{jh}^{(j+1)h}x^2e^xdx=\int_{-1}^{1}\left(\frac{h}{2}(t+1)+jh\right)^2e^{jh}e^{\frac{h}{2}(t+1)}\frac{h}{2}dt=\int_{-1}^{1}G_j(t)dt,$$

où nous avons posé $G_j(t)=\left(\frac{h}{2}(t+1)+jh\right)^2e^{jh}e^{\frac{h}{2}(t+1)}\frac{h}{2}$. Nous pouvons donc écrire

$$I=\int_0^4x^2e^xdx=\sum_{j=0}^{N-1}\int_{-1}^{1}G_j(t)dt.$$

La formule de Gauss-Legendre à 2 points est $J(g)=g\left(\frac{-1}{\sqrt{3}}\right)+g\left(\frac{1}{\sqrt{3}}\right)$. Notons par I_h l'approximation de I en utilisant cette formule. Alors $I_h=\displaystyle\sum_{j=0}^{N-1}J(G_j)=\sum_{j=0}^{N-1}\left(G_j\left(\frac{-1}{\sqrt{3}}\right)+G_j\left(\frac{1}{\sqrt{3}}\right)\right)$.

Nous pouvons calculer I exactement en intégrant deux fois par parties. $I=\int_0^4x^2e^xdx=10e^4-2$. Sa valeur, à 10 chiffres significatifs, est 0.5439815003E+03.

N	h	I_h	$\lvert I-I_h\rvert$	$\frac{\lvert I-I_h\rvert}{h^4}$
10	$\frac{4}{10}$	0.543966776E+03	0.147242965E-01	0.575167832E+00
15	$\frac{4}{15}$	0.543978578E+03	0.292209611E-02	0.577855920E+00
20	$\frac{4}{20}$	0.543980574E+03	0.926107663E-03	0.578817289E+00
25	$\frac{4}{25}$	0.543981121E+03	0.379637394E-03	0.578817289E+00
30	$\frac{4}{30}$	0.543981317E+03	0.183166843E-03	0.579551340E+00

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Exercice 2

Soit p un polynôme de degré 2 défini sur $[-1,1]$.
Notons $\beta_1=p(t_1)=p(-\alpha)$, $\beta_2=p(t_2)=p(0)$, $\beta_3=p(t_3)=p(\alpha)$. Dans la base de Lagrange p s'écrit

$$p(t)=\sum_{i=1}^{3}\beta_i\varphi_i(t),$$

où

$$\varphi_1(t)=\frac{t}{\alpha}\cdot\frac{t-\alpha}{2\alpha},\quad\varphi_2(t)=\frac{t+\alpha}{\alpha}\cdot\frac{t-\alpha}{-\alpha}\quad\text{et}\quad\varphi_3(t)=\frac{t+\alpha}{2\alpha}\cdot\frac{t}{\alpha}.$$

Point 2.a)

$$J(p)=\int_{-1}^{1}p(t)dt=\int_{-1}^{1}\sum_{i=1}^{3}\beta_i\varphi_i(t)dt=\sum_{i=1}^{3}\beta_i\omega_i.$$

si nous définissons $\omega_i=\int_{-1}^{1}\varphi_i(t)dt$, avec $i=1,2,3$. Cela implique que

$$\omega_1=\int_{-1}^{1}\frac{t(t-\alpha)}{2\alpha^2}dt=\frac{1}{3\alpha^2},\quad\omega_2=\int_{-1}^{1}\frac{(t+\alpha)(t-\alpha)}{-\alpha^2}dt=\frac{-2}{3\alpha^2}+2\quad\text{et}\quad\omega_3=\int_{-1}^{1}\frac{(t+\alpha)t}{2\alpha^2}dt=\frac{1}{3\alpha^2}.$$

Avec ces ω_i, notre formule de quadrature intègre exactement les polynômes de degré 2.

Point 2.b)

Un polynôme quelconque p de degré 3 peut s'écrire $p(t)=at^3+q(t)$ où q est un polynôme de degré 2. Nous avons montré au point 2.a) que q est intégré exactement. Notre formule de quadrature est donc exacte, si elle intègre exactement le terme at^3. Ce qui est bien vrai, puisque nous avons la suite d'égalités suivantes :

$$J(at^3)=\sum_{i=1}^{3}at_i^3\omega_i=\frac{1}{3\alpha^2}a\alpha^3+\frac{1}{3\alpha^2}a(-\alpha)^3=0=\int_{-1}^{1}at^3dt,\quad\forall a\in\mathbb{R}$$

Point 2.c)

Un polynôme p quelconque de degré 5 peut s'écrire comme $p(t)=at^5+bt^4+q(t)$ où q est un polynôme de degré 3. Au point 2.b), nous avons montré que q est intégré exactement. Par le même raisonnement, nous voyons aussi que

$$J(at^5)=\sum_{i=1}^{3}at_i^5\omega_i=\frac{1}{3\alpha^2}a\alpha^5+\frac{1}{3\alpha^2}a(-\alpha)^5=0=\int_{-1}^{1}at^5dt,\quad\forall a\in\mathbb{R}.$$

Pour satisfaire $J(bt^4)=\int_{-1}^{1}bt^4dt$, il faut et il suffit que

$$\int_{-1}^{1}bt^4dt=\frac{2}{5}b=J(bt^4)=\frac{1}{3\alpha^2}b\alpha^4+\frac{1}{3\alpha^2}b(-\alpha)^4=\frac{2b}{3}\alpha^2,\quad\forall b\in\mathbb{R},$$

c'est-à-dire que $\alpha=\sqrt{\frac{3}{5}}$. Nous avons alors $\omega_1=\omega_3=\frac{5}{9}$ et $\omega_2=\frac{8}{9}$ et la formule d'intégration devient

$$J(g)=\frac{5}{9}g\left(-\sqrt{\frac{3}{5}}\right)+\frac{8}{9}g\left(0\right)+\frac{5}{9}g\left(\sqrt{\frac{3}{5}}\right).$$

Il est facile de vérifier que le polynôme de Legendre de degré 3, $P_3(t)$, s'annule en 0 et $\pm\sqrt{\frac{3}{5}}$. Ainsi, la formule ci-dessus est la formule de Gauss-Legendre à 3 points, que nous savons exacte pour des polynômes de degré 5.

ANALYSE NUMÉRIQUE
Prof. J. Rappaz

Le 4 mai 1995

Solution de la série 4

Exercice 1

Point a)

Résolution par élimination de Gauss

1ère étape : normalisation de la 1ère ligne.
Diviser la 1ère équation par le 1er pivot pour obtenir

$$x_1 - \frac{c}{a}x_2 + \frac{d}{a}x_N = \frac{b_1}{a}. \qquad (1')$$

et remplacer la 1ère équation par 1'.
2ème étape : élimination de x_1 dans la 2ème ligne.
Soustraire $(-c)$ fois l'équation 1' de la 2ème pour obtenir

$$\left(a - \frac{c^2}{a}\right)x_2 - cx_3 + \left(d + \frac{c}{a}d\right)x_N = b_2 + \frac{c}{a}b_1,$$

3ème étape : normalisation de la 2ème ligne.
Diviser cette dernière équation par le 2ème pivot, i.e. multiplier par $p = \left(a - \frac{c^2}{a}\right)^{-1}$, pour obtenir

$$x_2 - pcx_3 + p\left(d + \frac{c}{a}d\right)x_N = p\left(b_2 + \frac{c}{a}b_1\right) \iff x_2 - \frac{c}{a - \frac{c^2}{a}}x_3 + \frac{d + \frac{c}{a}d}{a - \frac{c^2}{a}}x_N = \frac{b_2 + \frac{c}{a}b_1}{a - \frac{c^2}{a}}. \qquad (2')$$

4ème étape : élimination de x_2 dans la 3ème ligne.
5ème étape : normalisation de la 3ème ligne.
etc

Nous arrivons finalement à un système d'équations de forme *flèche* et triangulaire supérieur, facile à résoudre.

Algorithme

Notons $\vec{c}$ le $(N-1)$-vecteur des coefficients sur-diagonaux et $\vec{d}$ le $(N-2)$-vecteur formé des coefficients facteurs des x_N dans les $(N-2)$ premières équations. Initialement, toutes les composantes c_i et d_i des vecteurs $\vec{c}$ et $\vec{d}$ valent c et d, respectivement. Les inconnues seront stockées à la place du vecteur $\vec{b}$.

```
p := 1/a;                          Initialisation de l'inverse du 1er pivot

Faire pour i = 1 à N-3
 |   c_i := c_i p;              }
 |   d_i := d_i p;              }  Normalisation de la i-ème ligne
 |   b_i := b_i p;              }
 |   p := 1/(a - c c_i);           Calcul de l'inverse du i-ème pivot
 |   d_{i+1} := d_{i+1} + c d_i; }  Elimination de la i-ème inconnue dans la (i+1)-ème ligne
 |   b_{i+1} := b_{i+1} + c b_i; }

c_{N-2} := c_{N-2} p;           }
d_{N-2} := d_{N-2} p;           }  Normalisation de la (N-2)-ème ligne
b_{N-2} := b_{N-2} p;           }
p := 1/(a - c c_{N-2});            Calcul de l'inverse du (N-2)-ème pivot
--> c_{N-1} := c_{N-1} - c d_{N-2}; }  Elimination de la (N-2)-ème inconnue dans la (N-1)-ème ligne
b_{N-1} := b_{N-1} + c b_{N-2};     }

c_{N-1} := c_{N-1} p;           }  Normalisation de la (N-1)-ème ligne
b_{N-1} := b_{N-1} p;           }
p := 1/(a - c c_{N-1});            Calcul de l'inverse du (N-1)-ème pivot
b_N := b_N + c b_{N-1};            Elimination de la (N-1)-ème inconnue dans la N-ème ligne

b_N := b_N p;                      Normalisation de la N-ème ligne et isolement de la N-ème inconnue

b_{N-1} := b_{N-1} + c_{N-1} b_N;  Isolement de la (N-1)-ème inconnue

Faire pour i = N-2 à 1 par pas de -1
 |   b_i := b_i + c_i b_{i+1} - d_i b_N;   Isolement de la i-ème inconnue
```

$$-c_i = -c_i p \Rightarrow c_i = c_i p$$

$$-c_{N-1} = -c_{N-1} + c d_{N-2} \Rightarrow c_{N-1} = c_{N-1} + c d_{N-2}$$

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

Point b)

Lorsque $a = 2$, $c = 1$, $d = 0$ et $b_j = 1$, $\forall j = 1, 2, \ldots, N$, il est facile de vérifier que les $x_j = \frac{1}{2}j(N+1-j)$, $\forall j = 1, 2, \ldots, N$, satisfont au système d'équations. Par exemple, pour la $k^{\text{ème}}$ ligne, nous avons

$$-x_{k-1} + 2x_k - x_{k+1} = -\frac{1}{2}(k-1)(N+1-k+1) + 2\frac{1}{2}k(N+1-k) - \frac{1}{2}(k+1)(N+1-k-1) = 1.$$

Point c)

Pour $N = 5$, nous avons les résultats suivant (avec 16 chiffres significatifs) :

j	Solution calculée x_j	Solution exacte x_j
1	$.2500000000000000 \cdot 10^1$	$.2500000000000000 \cdot 10^1$
2	$.3999999999999999 \cdot 10^1$	$.4000000000000000 \cdot 10^1$
3	$.4499999999999999 \cdot 10^1$	$.4500000000000000 \cdot 10^1$
4	$.4000000000000000 \cdot 10^1$	$.4000000000000000 \cdot 10^1$
5	$.2500000000000000 \cdot 10^1$	$.2500000000000000 \cdot 10^1$

Exercice 2

Point a)

Nous vérifions que le déterminant est différent de zéro et donc que le système est régulier.

Point b)

Le nombre de condition spectral d'une matrice régulière symétrique d'ordre 2 est

$$\chi(A) = \frac{max(|\lambda_1|, |\lambda_2|)}{min(|\lambda_1|, |\lambda_2|)},$$

où λ_1 et λ_2 sont les valeurs propres de A, donc les zéros du polynôme $det(A - \lambda I)$, i.e.

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{12} & a_{22} - \lambda \end{vmatrix} = 0 = \lambda^2 - \lambda(a_{22} + a_{11}) + a_{22}a_{11} - a_{12}^2.$$

Dans notre cas $a_{11} = 0.4218613, a_{12} = 0.6327917$ et $a_{22} = 0.9491883$. Ce qui donne $\chi(A) \simeq 0.3962570 \cdot 10^7$.

Point c)

Elimination de Gauss en retenant 7 chiffres significatifs lors de chaque opération :

$$\begin{cases} x_1 + 0.1499999 \cdot 10^1\, x_2 = 0.2499999 \cdot 10^1, \\ 0.6327917\, x_1 + 0.9491883\, x_2 = 0.1581980 \cdot 10^1. \end{cases}$$

Nous soustrayons 0.6327917 fois la première équation de la deuxième, ce qui donne le système

$$\begin{cases} x_1 + 0.1499999 \cdot 10^1\, x_2 = 0.2499999 \cdot 10^1, \\ 0.1400000 \cdot 10^{-5}\, x_2 = 0.2000000 \cdot 10^{-5}. \end{cases}$$

Nous obtenons $x_1 = 0.3571440$ et $x_2 = 0.1428571 \cdot 10^1$.

Conclusion : Le calcul avec 7 chiffres significatifs ne donne aucun chiffre significatif sur la solution, ce qui est cohérent avec $\chi(A)$ qui est de l'ordre de 10^7.

notc déf : $\chi(A) = \sqrt{\frac{max\ \lambda_i}{min\ \lambda_i}}$, préc. $= E(\log(\chi(A)))$

ANALYSE NUMÉRIQUE
Prof. J. Rappaz

Le 11 mai 1995

Solution de la série 5

Exercice 1

Point a)

Profils de L et U

Supposons que les matrices L et U aient les profils suivants :

$$L = \begin{pmatrix} a_1 & & & & & \\ e_1 & a_2 & & & & \\ & e_2 & \ddots & & & \\ \mbox{\Large 0} & & \ddots & \ddots & & \\ & & & e_{N-2} & a_{N-1} & \\ f_1 & f_2 & \dots & f_{N-2} & e_{N-1} & a_N \end{pmatrix} \quad \text{et } U = \begin{pmatrix} 1 & c_1 & & & & d_1 \\ & 1 & c_2 & & \mbox{\Large 0} & d_2 \\ & & \ddots & \ddots & & \vdots \\ & & & \ddots & c_{N-2} & d_{N-2} \\ & & & & 1 & c_{N-1} \\ & & & & & 1 \end{pmatrix}.$$

Nous remarquons en exécutant le produit LU que les régions sous-diagonales nulles de L, respectivement sur-diagonales de U, restent nulles dans la matrice LU. Puisque toutes les sous-matrices principales de A sont régulières par hypothèse, nous savons par le théorème du cours concernant la décomposition LU que les matrices L et U existent et sont uniques. Elles devront donc avoir les profils supposés ci-dessus.

Résolution

En multipliant L par la $1^{\text{ère}}$ colonne de U, nous obtenons directement la $1^{\text{ère}}$ colonne de A. Ainsi

$$a_1 = a, \qquad e_1 = e \qquad \text{et} \qquad f_1 = f.$$

En multipliant la $1^{\text{ère}}$ ligne de L par les $2^{\text{ème}}$ et $N^{\text{ème}}$ colonnes de U, nous obtenons respectivement les $2^{\text{ème}}$ et $N^{\text{ème}}$ termes de la $1^{\text{ère}}$ ligne de A. Ainsi, nous avons

$$a_1c_1 = c \qquad \text{et} \qquad a_1d_1 = d,$$

où, puisque a_1 est connu,

$$c_1 = \frac{c}{a_1} \qquad \text{et} \qquad d_1 = \frac{d}{a_1}.$$

Nous avons ainsi trouvé la $1^{\text{ère}}$ colonne de L et la $1^{\text{ère}}$ ligne de U, ou les $1^{\text{ères}}$ composantes de $\vec{a}$, $\vec{c}$, $\vec{d}$, $\vec{e}$ et $\vec{f}$.

En multipliant, les $2^{\text{ème}}$, $(2+1)^{\text{ème}}$ et $N^{\text{ème}}$ lignes de L par la $2^{\text{ème}}$ colonne de U, nous obtenons

$$e_{2-1}c_{2-1} + a_2 = a, \qquad e_2 = e \qquad \text{et} \qquad f_{2-1}c_{2-1} + f_2 = f.$$

Puisque les $(2-1)^{\text{èmes}}$ composantes de $\vec{c}$, $\vec{e}$ et $\vec{f}$ sont connues, nous pouvons déterminer que

$$a_2 = a - e_{2-1}c_{2-1}, \qquad e_2 = e \qquad \text{et} \qquad f_2 = f - f_{2-1}c_{2-1}.$$

En multipliant $2^{\text{ème}}$ ligne de L par les $(2+1)^{\text{ème}}$ et $N^{\text{ème}}$ colonnes de U, nous obtenons

$$a_2c_2 = c \qquad \text{et} \qquad e_{2-1}d_{2-1} + a_2d_2 = d.$$

Puisque nous connaissons, les $(2-1)^{\text{èmes}}$ composantes de $\vec{e}$ et $\vec{d}$, et la $2^{\text{ème}}$ de $\vec{a}$, nous pouvons déterminer que

$$c_2 = \frac{c}{a_2} \qquad \text{et} \qquad d_2 = \frac{d - e_{2-1}d_{2-1}}{a_2}.$$

Nous avons ainsi trouvé la $2^{\text{ème}}$ colonne de L et la $2^{\text{ème}}$ ligne de U, ou les $2^{\text{èmes}}$ composantes de $\vec{a}$, $\vec{c}$, $\vec{d}$, $\vec{e}$ et $\vec{f}$. Le même raisonnement peut s'appliquer pour calculer toutes les composantes encore inconnues, sauf les $(N-1)^{\text{èmes}}$ de $\vec{d}$, $\vec{c}$ et $\vec{e}$ et la $N^{\text{ème}}$ de $\vec{a}$.

Typeset by $\mathcal{A}_{\mathcal{M}}\mathcal{S}$-TEX

Ces dernières inconnues se déterminent facilement en utilisant toujours le même principe. En multipliant, les $(N-1)^{\text{ème}}$ et $N^{\text{ème}}$ lignes de L par la $(N-1)^{\text{ème}}$ colonne de U, nous obtenons

$$e_{N-2}c_{N-2} + a_{N-1} = a \qquad \text{et} \qquad f_{N-2}c_{N-2} + e_{N-1} = e,$$

qui implique que

$$a_{N-1} = a - e_{N-2}c_{N-2} \qquad \text{et} \qquad e_{N-1} = e - f_{N-2}c_{N-2}.$$

En multipliant la $(N-1)^{\text{ème}}$ ligne de L par la $N^{\text{ème}}$ colonne de U, nous obtenons

$$e_{N-2}d_{N-2} + a_{N-1}c_{N-1} = c, \qquad \text{ou que} \qquad c_{N-1} = \frac{c - e_{N-2}d_{N-2}}{a_{N-1}}.$$

Finalement, la multiplication de la $N^{\text{ème}}$ ligne de L avec la $N^{\text{ème}}$ colonne de U donne

$$\sum_{i=1}^{N-2} f_i d_i + e_{N-1}c_{N-1} + a_N = a, \qquad \text{ou que} \qquad a_N = a - \sum_{i=1}^{N-2} f_i d_i - e_{N-1}c_{N-1}.$$

Algorithme de décomposition LU :

$$\begin{array}{ll}
\left.\begin{array}{l} a_1 := a; \\ e_1 := e; \\ f_1 := f; \end{array}\right\} & \text{Calcul de la } 1^{\text{ère}} \text{ colonne de } L \\
\left.\begin{array}{l} c_1 := \frac{c}{a_1}; \\ d_1 := \frac{d}{a_1}; \end{array}\right\} & \text{Calcul de la } 1^{\text{ère}} \text{ ligne de } U \\
\text{Faire pour } i = 2 \text{ à } N-2 & \\
\quad \left.\begin{array}{|l} a_i := a - e_{i-1}c_{i-1}; \\ e_i := e; \\ f_i := f - f_{i-1}c_{i-1}; \end{array}\right\} & \text{Calcul de la } i^{\text{ème}} \text{ colonne de } L \\
\quad \left.\begin{array}{|l} c_i := \frac{c}{a_i}; \\ d_i := \frac{d - e_{i-1}d_{i-1}}{a_i}; \end{array}\right\} & \text{Calcul de la } i^{\text{ème}} \text{ ligne de } U \\
\left.\begin{array}{l} a_{N-1} := a - e_{N-2}c_{N-2}; \\ e_{N-1} := e - f_{N-2}c_{N-2}; \end{array}\right\} & \text{Calcul de la } (N-1)^{\text{ème}} \text{ colonne de } L \\
c_{N-1} := \frac{c - e_{N-2}d_{N-2}}{a_{N-1}}; & \text{Calcul de la } (N-1)^{\text{ème}} \text{ ligne de } U \\
\left.\begin{array}{l} a_N := a - e_{N-1}c_{N-1}; \\ \text{Faire pour } i = 1 \text{ à } N-2 \\ \quad | \quad a_N := a_N - f_i d_i; \end{array}\right\} & \text{Calcul de } a_N
\end{array}$$

Point b)

Après avoir trouvé la décomposition LU, il reste à résoudre successivement deux systèmes triangulaires, premièrement le système triangulaire inférieur $L\vec{y} = \vec{b}$, deuxièmement le système supérieur $U\vec{x} = \vec{y}$. Ces types de systèmes ont été vu dans la série 4, exercice 1. Nous savons qu'ils donnent les algorithmes suivants, en stockant les vecteurs intermédiaires $\vec{x}$ et $\vec{y}$ dans le vecteur initial $\vec{b}$:

$$\begin{array}{ll}
\left.\begin{array}{l} b_1 := \frac{b_1}{a_1}; \\ \text{Faire pour } i = 2 \text{ à } N-1 \\ \quad | \quad b_i := b_i - e_{i-1}b_{i-1}; \\ \quad | \quad b_i := b_i/a_i; \\ b_N := b_N - e_{N-1}b_{N-1}; \\ \text{Faire pour } i = 1 \text{ à } N-2 \\ \quad | \quad b_N := b_N - f_i b_i; \\ b_N := \frac{b_N}{a_N}; \end{array}\right\} & \text{Algorithme de résolution } L\vec{y} = \vec{b} \\
\left.\begin{array}{l} b_{N-1} := b_{N-1} - c_{N-1}b_N; \\ \text{Faire pour } i = N-2 \text{ à } 1 \text{ par pas de -1} \\ \quad | \quad b_i := b_i - c_i b_{i+1} - d_i b_N; \end{array}\right\} & \text{Algorithme de résolution } U\vec{x} = \vec{y}
\end{array}$$

Point c)

Cas particulier : série 4, exercice 1b.

Pour $N = 5$, nous avons les résultats suivants (avec 16 chiffres significatifs), lorsque $a = 2$, $c = e = -1$, $d = f = 0$ et $b_i = 1$, $\forall i = 1, 2, \dots N$:

j	Solution calculée x_j	Solution exacte x_j
1	$.2500000000000000 \cdot 10^1$	$.2500000000000000 \cdot 10^1$
2	$.3999999999999999 \cdot 10^1$	$.4000000000000000 \cdot 10^1$
3	$.4499999999999999 \cdot 10^1$	$.4500000000000000 \cdot 10^1$
4	$.4000000000000000 \cdot 10^1$	$.4000000000000000 \cdot 10^1$
5	$.2500000000000000 \cdot 10^1$	$.2500000000000000 \cdot 10^1$

Exercice 2

Point a)

De façon évidente, A est symétrique. Vérifions qu'elle est bien définie positive. Pour le faire, nous distinguons deux cas.

- Si $c = 0$, nous avons $\vec{x}^T A\vec{x} = a\|\vec{x}\|^2 \geqslant 0$, $\forall \vec{x}$ et $\vec{x}^T A\vec{x} = 0 \Rightarrow \vec{x} = 0$.
- Si $c \neq 0$:

$$\begin{aligned}\vec{x}^T A\vec{x} &= ax_1^2 + 2cx_1x_2 + ax_2^2 + 2cx_2x_3 + ax_3^2 + \cdots + ax_{N-1}^2 + 2cx_{N-1}x_N + ax_N^2 \\ &\geqslant 2|c|\,x_1^2 + 2cx_1x_2 + 2|c|\,x_2^2 + 2cx_2x_3 + 2|c|\,x_3^2 + \cdots + 2|c|\,x_{N-1}^2 + 2cx_{N-1}x_N + 2|c|\,x_N^2 \\ &\geqslant |c|\left\{x_1^2 + (x_1 \pm x_2)^2 + (x_2 \pm x_3)^2 + \cdots + (x_{N-1} \pm x_N)^2 + x_N^2\right\}, \\ &\qquad \text{le signe est } + \text{ si } c > 0, - \text{ si } c < 0.\end{aligned}$$

Ainsi, $\vec{x}^T A\vec{x} \geqslant 0$.
Si $\vec{x}^T A\vec{x} = 0$, alors $x_1 = x_N = 0$ et $(x_{j+1} \pm x_j)^2 = 0$, $\forall j = 1, 2, \dots, (N-1)$, donc $\vec{x} = 0$.

Point b)

Puisque A est symétrique définie positive, le théorème du cours concernant les matrices de bande dit que la matrice L sera, comme A, une matrice de bande de demi-largeur 2. Donc, $A = LL^T$ s'écrira

$$\begin{pmatrix} a & c & & & & \\ c & a & c & & & \\ & c & \ddots & \ddots & & \\ & & \ddots & \ddots & c & \\ & & & c & a & c \\ & & & & c & a \end{pmatrix} = \begin{pmatrix} a_1 & & & & & \\ c_1 & a_2 & & & & \\ & c_2 & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & c_{N-2} & a_{N-1} & \\ & & & & c_{N-1} & a_N \end{pmatrix} \begin{pmatrix} a_1 & c_1 & & & & \\ & a_2 & c_2 & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & c_{N-2} & \\ & & & & a_{N-1} & c_{N-1} \\ & & & & & a_N \end{pmatrix}$$

$$= \begin{pmatrix} a_1^2 & * & & & & \\ a_1c_1 & c_1^2 + a_2^2 & * & & & \\ & a_2c_2 & \ddots & * & & \\ & & \ddots & \ddots & * & \\ & & & c_{N-2}a_{N-2} & c_{N-2}^2 + a_{N-1}^2 & * \\ & & & & c_{N-1}a_{N-1} & c_{N-1}^2 + a_N^2 \end{pmatrix}.$$

Remarquons qu'il n'est pas nécessaire d'utiliser les termes sur-diagonaux du produit LL^T - nous les avons remplacés par des $*$ - et rappelons que les termes diagonaux de L sont positifs. En égalant terme à terme A et LL^T :

- nous trouvons que $a_1 = \sqrt{a}$,
- ensuite, en connaissant le coefficient diagonal d'une ligne, nous calculons que le coefficient sous-diagonal de la ligne suivante vaut $c_i = \frac{c}{a_i}$,
- puis, en connaissant le coefficient sous-diagonal, nous trouvons que le coefficient diagonal de la même ligne vaut $a_{i+1} = \sqrt{a - c_i^2}$.

Algorithme de Cholesky :

$a_1 := \sqrt{a}$;	Calcul du coefficient a_1
Faire pour $i = 1$ à $N - 1$	
| $c_i := \frac{c}{a_i}$;	Calcul du coefficient sous-diagonal de la $(i+1)^{\text{ème}}$ ligne
| $a_{i+1} := \sqrt{a - c_i^2}$;	Calcul du coefficient diagonal de la $(i+1)^{\text{ème}}$ ligne

ANALYSE NUMÉRIQUE
Prof. J. Rappaz

Le 1 juin 1995

Solution de la série 7

Exercice 1

Point a)

En remplaçant A, x, x^3 et b par leur valeur et f par le vecteur formé des composantes f_1, f_2, f_3, la définition $f(x) = Ax + x^3 + b$ s'écrit

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \end{pmatrix} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} + \begin{pmatrix} x_1^3 \\ x_2^3 \\ x_3^3 \end{pmatrix} + \begin{pmatrix} -4 \\ 6 \\ -13 \end{pmatrix} = \begin{pmatrix} 2\ x_1 -1\ x_2 +0\ x_3 + x_1^3 - 4 \\ -1\ x_1 +2\ x_2 -1\ x_3 + x_2^3 + 6 \\ 0\ x_1 -1\ x_2 +2\ x_3 + x_3^3 - 13 \end{pmatrix}$$

Chercher les zéros de $f(x)$ est équivalent à chercher x_1, x_2, x_3 tel que

$$\begin{pmatrix} 2\ x_1 -1\ x_2 +0\ x_3 + x_1^3 - 4 \\ -1\ x_1 +2\ x_2 -1\ x_3 + x_2^3 + 6 \\ 0\ x_1 -1\ x_2 +2\ x_3 + x_3^3 - 13 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

que nous pouvons aussi écrire

$$\begin{pmatrix} 2\ x_1 -1\ x_2 +0\ x_3 + x_1^3 \\ -1\ x_1 +2\ x_2 -1\ x_3 + x_2^3 \\ 0\ x_1 -1\ x_2 +2\ x_3 + x_3^3 \end{pmatrix} = \begin{pmatrix} 4 \\ -6 \\ 13 \end{pmatrix}.$$

Ce dernier système d'équations est le système (P_1).

Point b)

$$Df(y) \overset{\text{déf.}}{=} \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(y) & \frac{\partial f_1}{\partial x_2}(y) & \frac{\partial f_1}{\partial x_3}(y) \\ \frac{\partial f_2}{\partial x_1}(y) & \frac{\partial f_2}{\partial x_2}(y) & \frac{\partial f_2}{\partial x_3}(y) \\ \frac{\partial f_3}{\partial x_1}(y) & \frac{\partial f_3}{\partial x_2}(y) & \frac{\partial f_3}{\partial x_3}(y) \end{pmatrix} \overset{\substack{\text{déf. de } f_1, f_2, f_3 \\ \text{du point a)}}}{=} \begin{pmatrix} 2+3y_1^2 & -1 & 0 \\ -1 & 2+3y_2^2 & -1 \\ 0 & -1 & 2+3y_3^2 \end{pmatrix}.$$

Remarque : la matrice $Df(y)$ est symétrique définie positive.

Point c)

La méthode de Newton-corde donne la formule de récurrence suivante (cf. cours)

$$x^{n+1} = x^n - Df(x^0)^{-1} f(x^n), \quad n = 0, 1, 2 \ldots$$

Algorithme de Newton-corde

Calcul préliminaire

$x_1 := 2;\ x_2 := 2;\ x_3 := 2;$	Initialisation de x^0 dans le vecteur $\vec{x}$
Faire pour $i = 1$ à 3 $\quad a_i := 2 + 3x_i^2;$	Calcul et stockage de la diagonale de la matrice $Df(x^0)$
Faire pour $i = 1$ à 2 $\quad c_i := -1;$	Stockage de la sous-diagonale de la matrice $Df(x^0)$
$a_1 := \sqrt{a_1};$ Faire pour $i = 1$ à 2 $\quad c_i := c_i/a_i;$ $\quad a_{i+1} := \sqrt{a_{i+1} - c_i^2};$	Algorithme de Cholesky de la matrice symétrique définie positive $Df(x^0)$ (cf. série 5, exercice 2)

Récurrence

Faire $\quad f_1 := 2x_1 - x_2 + x_1^3 - 4;$ $\quad f_2 := -x_1 + 2x_2 - x_3 + x_2^3 + 6;$ $\quad f_3 := -x_2 + 2x_3 + x_3^3 - 13;$	Calcul de $f(x^n)$
$\quad f_1 := f_1/a_1;$ $\quad$Faire pour $i = 2$ à 3 $\qquad f_i := f_i - c_{i-1}f_{i-1};$ $\qquad f_i := f_i/a_i;$	Algorithme de résolution $L\vec{z} = \vec{f}$ (stockage de $\vec{z}$ à la place de $\vec{f}$)
$\quad f_3 := f_3/a_3;$ $\quad$Faire pour $i = 2$ à 1 par pas de -1 $\qquad f_i := f_i - c_i f_{i+1};$ $\qquad f_i := f_i/a_i;$	Algorithme de résolution $L^T\vec{y} = \vec{f}$ (stockage de $\vec{y}$ à la place de $\vec{f}$)
$\quad$Faire pour $i = 1$ à 3 $\qquad x_i := x_i - f_i;$	Calcul de x^{n+1} (stockage de x^{n+1} à la place de $\vec{x}$)
tant que la condition d'arrêt n'est pas satisfaite.	

Remarque : ici, nous avons pris comme condition d'arrêt $\sum_{i=1}^{3} f_i^2 < 10^{-13}$: stop.

Point d)

Un programme en C, correspondant à cet algorithme, est :

```
#include <stdio.h>
#include <math.h>

void main(void)
{
 int i, n=0;
 double corr, a[3], c[2], f[3], x[3]={2., 2., 2.};

 for(i=0; i<3; i++)
   a[i] = 2.+3.*x[i]*x[i];
 for(i=0; i<2; i++)
   c[i] = -1.;
 a[0] = sqrt(a[0]);
 for(i=0; i<2; i++)
  {
   c[i]/= a[i];
   a[i+1] = sqrt(a[i+1]-c[i]*c[i]);
  }
 do
  {
   n++;
   f[0] = 2.*x[0]- x[1] +pow(x[0],3.)  -4.;
   f[1] = -x[0]+2.*x[1]- x[2]+pow(x[1],3.)  +6.;
   f[2] = -x[1]+2.*x[2]+pow(x[2],3.)-13.;
   f[0]/= a[0];
   for(i=1; i<3; i++)
    {
     f[i]-= c[i-1]*f[i-1];
     f[i]/= a[i];
    }
   f[2]/= a[2];
   for(i=1; 0<=i; i--)
    {
     f[i]-= c[i]*f[i+1];
```

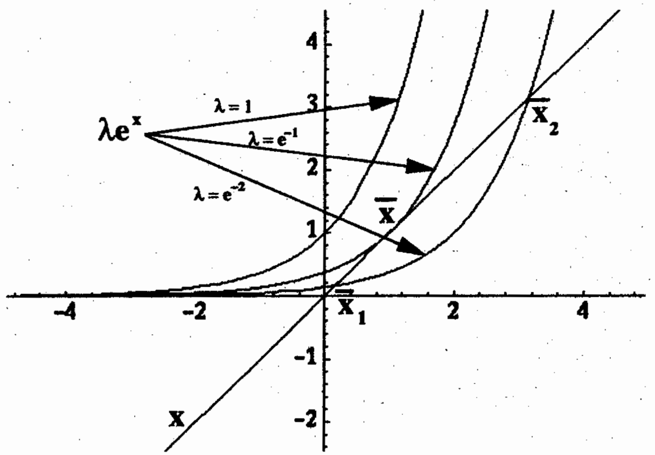
```
        f[i]/= a[i];
      }
     for(i=0; i<3; i++)
        x[i]-= f[i];
     for(i=0, corr=0; i<3; corr+= f[i]*f[i++]);
   }
   while(corr<1.e-13);

   printf("Solution après %d pas :  \n", n);
   for (i=0; i<3; i++)
     printf("x[%d] = %lf \n", i, x[i]);
}
```

Résultats :

- Solution après 1 pas : $x = \begin{pmatrix} 1.498159 \\ 0.974227 \\ 2.141016 \end{pmatrix}$.
- Solution après 45 pas : $x = \begin{pmatrix} 1.000000 \\ -1.000000 \\ 2.000000 \end{pmatrix}$.
- Solution exacte : $x = \begin{pmatrix} 1. \\ -1. \\ 2. \end{pmatrix}$.

<u>**Exercice 2**</u>

Point a)



Le graphique montre que

- si $\lambda > \frac{1}{e}$ - ici, $\lambda = 1$ -, le problème (P_2) n'a pas de solution;
- si $\lambda = \frac{1}{e}$, les deux graphes $y = x$ et $y = \lambda e^x = e^{x-1}$ sont tangents en $x = 1$; l'équation admet $\overline{x} = 1$ comme seule solution;
- si $\lambda < \frac{1}{e}$, - ici, $\lambda = \frac{1}{e^2}$ -, le problème (P_2) a deux solutions positives, $0 < \overline{x}_1 < 1 < \overline{x}_2$.

Dans la suite, $0 < \lambda < \frac{1}{e}$.

Point b)

(i) La suite $x_{n+1} = \lambda e^{x_n}$, $n = 0, 1, 2, \ldots$, est la suite correspondant au problème de point fixe $x = g(x)$, pour $g(x) \equiv \lambda e^x$, selon la définition du paragraphe 7.2 du cours.

La dérivée de $g(x)$ est $g'(x) = \lambda e^x$, donc, $|g'(x)| < 1$, si $x \in]0, 1[$.

Puisque $\overline{x}_1$ est le point fixe appartenant à l'intervalle $x \in]0, 1[$, le théorème du paragraphe 7.2 du cours affirme qu'il existe un $\epsilon > 0$ tel que si x_0 satisfait $|\overline{x}_1 - x_0| \leqslant \epsilon$, la suite donnée par $x_{n+1} = \lambda e^{x_n}$, $n = 0, 1, 2, \ldots$, converge vers $\overline{x}_1$ lorsque n tend vers l'infini.

(ii) Remarquons que pour $x > 0$, nous avons

$$x = \lambda e^x \iff x = \ln x - \ln \lambda.$$

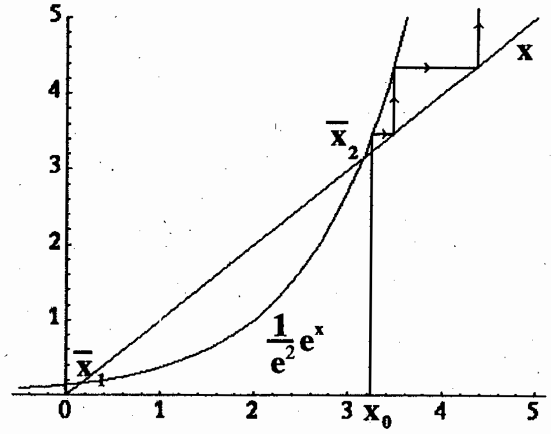
La suite $x_{n+1} = \ln x_n - \ln \lambda$, $n = 0, 1, 2, \ldots$, correspond au problème de point fixe $x = g(x)$, pour $g(x) \equiv \ln x - \ln \lambda$.

La dérivée de $g(x)$ est $g'(x) = \frac{1}{x}$, et $|g'(x)| < 1$ lorsque $x > 1$.

Le point fixe supérieur à 1 est $\overline{x}_2$. De même qu'au point (i), il existe un $\epsilon > 0$ tel que si x_0 satisfait $|\overline{x}_2 - x_0| \leqslant \epsilon$, la suite donnée par $x_{n+1} = \ln x - \ln \lambda$, $n = 0, 1, 2, \ldots$, converge vers $\overline{x}_2$ lorsque n tend vers l'infini.

Remarque : Si nous prenons l'algorithme (i) pour calculer $\overline{x}_2$, ou l'algorithme (ii) pour calculer $\overline{x}_1$, la condition $|g'(x)| < 1$ n'est plus satisfaite et le théorème ne s'applique plus.

Comme nous le voyons sur le graphique ci-dessous, l'algorithme (i) diverge si $x_0 > \overline{x}_2$, même si x_0 est "très proche" de x_2.



ANALYSE NUMÉRIQUE Le 8 juin 1995
Prof. J. Rappaz

Solution de la série 8

Exercice 1

Définissons $f(x,t) = e^{-x^2}$, indépendant de t. Notre problème de Cauchy se note :

trouver une fonction $u : t \in [0,\infty[\mapsto u(t) \in \mathbb{R}$ qui satisfait

$$(P_1') \quad \begin{cases} \dot u(t) = f(u(t),t), & t>0, \\ u(0) = 0. \end{cases}$$

Point a)

Puisque le maximum de la fonction $\left|se^{-s^2}\right|$ vaut $\frac{e^{-\frac12}}{\sqrt2}$ – maximum atteint en $s = \pm\frac{1}{\sqrt2}$ –, nous avons, pour x, $y \in \mathbb{R}$ quelconques,

$$|f(y,t) - f(x,t)| = \left|e^{-y^2} - e^{-x^2}\right| = \left|\int_x^y -2se^{-s^2}ds\right| \leqslant \sqrt2 e^{-\frac12}\left|\int_y^x ds\right| = \sqrt2 e^{-\frac12}|y-x|.$$

Comme $f(x,t)$ est, en plus, continue sur $\mathbb{R}\times\mathbb{R}^+$, les hypothèses du théorème de Cauchy-Lipschitz sont remplies et le problème (P_1'), et donc (P_1), a une solution unique.

Point b)

Si u^n est l'approximation de $u(t)$ par le schéma de Heun, on a

$$\begin{cases} p^1 = f(u^n, t_n), \\ p^2 = f(u^n + h_n p^1, t_{n+1}), \\ u^{n+1} = u^n + h_n\left(\frac12 p^1 + \frac12 p^2\right), \end{cases}$$

pour $n = 0, 1, 2, \ldots$, avec $u^0 = u(t_0) = 0$ et $h_n = h$. Le schéma devient donc

$$\begin{cases} p^1 = e^{-(u^n)^2}, \\ p^2 = e^{-(u^n + hp^1)^2}, \\ u^{n+1} = u^n + \frac{h}{2}(p^1 + p^2), \end{cases}$$

pour $n = 0, 1, 2, \ldots$, avec $u^0 = 0$.

Point c)

Le premier pas de ce schéma donne

$$\begin{cases} p^1 = e^{-(u^0)^2} = 1, \\ p^2 = e^{-(u^0+hp^1)^2} = e^{-h^2}, \\ u^1 = u^0 + \frac{h}{2}(p^1+p^2) = \frac{h}{2}\left(1 + e^{-h^2}\right). \end{cases}$$

Point d)

Le schéma d'Euler progressif est (cf. cours §8.3a) $v^{n+1} = v^n + h_n f(v^n, t_n)$, pour $n = 0, 1, 2, \ldots$, avec $v^0 = u(t_0)$. Dans notre cas, $v^0 = u(0) = 0$, donc $v^1 = v^0 + hf(v^0, t_0) = v^0 + he^{-(v^0)^2} = h$.

La différence $\left|u^1 - v^1\right|$ vaut

$$\left|u^1 - v^1\right| = \left|\frac{h}{2}\left(1 + e^{-h^2}\right) - h\right| = \left|\frac{h}{2}\left(1 - e^{-h^2}\right)\right| = \frac{h}{2}\left(1 - e^{-h^2}\right).$$

$$\left|-\frac{h}{2}\left(1 - e^{-h^2}\right)\right|$$
$$= \left|\frac{h}{2}\left(1 - e^{-h^2}\right)\right|$$

trouvé comment ?

Comme le maximum de la fonction $\frac{1-e^{-x^2}}{x^2}$ vaut 1 – maximum atteint en $x = 0$ –,

$$1 \geqslant \frac{1 - e^{-h^2}}{h^2} = \frac{\frac{h}{2}\left(1 - e^{-h^2}\right)}{\frac{h^3}{2}} = \frac{\left|u^1 - v^1\right|}{\frac{h^3}{2}}.$$

Finalement, nous avons que

$$\left|u^1 - v^1\right| \leqslant \frac{h^3}{2}.$$

Exercice 2

Point a)

Le méthode de Runge-Kutta classique, pour un système différentiel noté vectoriellement

$$\begin{cases} \dot{\vec u}(t) = \vec f(\vec u(t), t), & t > 0, \\ \vec u(0) = 0, \end{cases}$$

est (cf. cours §8.5 et §8.6)

$$\begin{cases} \vec p^{\,1} = \vec f(\vec u^{\,n}, t_n), \\ \vec p^{\,2} = \vec f\left(\vec u^{\,n} + \frac{h_n}{2}\vec p^{\,1}, t_n + \frac{h_n}{2}\right), \\ \vec p^{\,3} = \vec f\left(\vec u^{\,n} + \frac{h_n}{2}\vec p^{\,2}, t_n + \frac{h_n}{2}\right), \\ \vec p^{\,4} = \vec f\left(\vec u^{\,n} + h_n\vec p^{\,3}, t_{n+1}\right), \\ \vec u^{\,n+1} = \vec u^{\,n} + \frac{h_n}{6}\left(\vec p^{\,1} + 2\vec p^{\,2} + 2\vec p^{\,3} + \vec p^{\,4}\right), \end{cases}$$

pour $n = 0, 1, 2, \ldots$, avec $\vec u^{\,0} = \vec u(0)$. Remarquons que nous avons noté les indices des vecteurs $\vec p$ en exposant contrairement à ce qui fait dans le cours, pour garder la notation avec indices inférieurs pour les composante des vecteurs.

Dans notre cas, la fonction $\vec f$ est définie par

$$\vec f(\vec v) = \begin{pmatrix} \sin v_2 \\ \cos v_1 \end{pmatrix},$$

qui ne dépend pas explicitement du temps, les pas de temps sont tous égaux, $h_n = h$, pour $n = 0, 1, 2, \ldots$ e la condition initiale vaut

$$\vec u^{\,0} = \vec u(0) = \begin{pmatrix} u_1(0) \\ u_2(0) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Ainsi le schéma ci-dessus devient

$$\begin{cases} \vec p^{\,1} = \vec f(\vec u^{\,n}), \\ \vec p^{\,2} = \vec f\left(\vec u^{\,n} + \frac{h_n}{2}\vec p^{\,1}\right), \\ \vec p^{\,3} = \vec f\left(\vec u^{\,n} + \frac{h_n}{2}\vec p^{\,2}\right), \\ \vec p^{\,4} = \vec f\left(\vec u^{\,n} + h_n\vec p^{\,3}\right), \\ \vec u^{\,n+1} = \vec u^{\,n} + \frac{h_n}{6}\left(\vec p^{\,1} + 2\vec p^{\,2} + 2\vec p^{\,3} + \vec p^{\,4}\right), \end{cases}$$

pour $n = 0, 1, 2, \ldots$, avec $\vec u^{\,0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Remarque : plus concrètement, en remplaçant $\vec{f}(\vec{v})$ par sa définition et en explicitant selon les composantes des vecteurs, nous obtenons le schéma

$$\left\{\begin{array}{l}
\begin{pmatrix} p_1^1 \\ p_2^1 \end{pmatrix} = \begin{pmatrix} \sin(u_2^n) \\ \cos(u_1^n) \end{pmatrix}, \\
\begin{pmatrix} p_1^2 \\ p_2^2 \end{pmatrix} = \begin{pmatrix} \sin\left(u_2^n + \frac{h}{2}p_2^1\right) \\ \cos\left(u_1^n + \frac{h}{2}p_1^1\right) \end{pmatrix}, \\
\begin{pmatrix} p_1^3 \\ p_2^3 \end{pmatrix} = \begin{pmatrix} \sin\left(u_2^n + \frac{h}{2}p_2^2\right) \\ \cos\left(u_1^n + \frac{h}{2}p_1^2\right) \end{pmatrix}, \\
\begin{pmatrix} p_1^4 \\ p_2^4 \end{pmatrix} = \begin{pmatrix} \sin\left(u_2^n + hp_2^3\right) \\ \cos\left(u_1^n + hp_1^3\right) \end{pmatrix}, \\
\begin{pmatrix} u_1^{n+1} \\ u_2^{n+1} \end{pmatrix} = \begin{pmatrix} u_1^n \\ u_2^n \end{pmatrix} + \frac{h}{6}\begin{pmatrix} p_1^1 + 2p_1^2 + 2p_1^3 + p_1^4 \\ p_2^1 + 2p_2^2 + 2p_2^3 + p_2^4 \end{pmatrix},
\end{array}\right.$$

pour $n = 0, 1, 2, \ldots$, avec $\vec{u}^{\,0} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$.

Point b)

Un programme, en C++, qui calcule une approximation de $u_1(T)$ et $u_2(T)$, avec N pas égaux entre 0 et T, est

```
#include <iostream.h>
#include <math.h>

class vct { public:  double c[2];
                     friend vct operator* (const double, const vct &);
                     vct operator+ (const vct &);
                     vct& operator+= (const vct &);
                     vct f(const vct &); };

vct operator* (const double x, const vct &in)
 {vct out = {x*in.c[0], x*in.c[1]}; return out;}
vct vct::operator+ (const vct &in)
 {vct out = {c[0]+in.c[0], c[1]+in.c[1]}; return out;}
vct& vct::operator+= (const vct &in)
 {c[0]+= in.c[0]; c[1]+= in.c[1]; return *this;}
vct f(const vct &in)
 {vct out = {sin(in.c[1]), cos(in.c[0])}; return out;}

void main(void)
{
 int n, N;  double h, h2, h6, T;  vct u={0., 0.}, p1, p2, p3, p4;

 cout << "Entrer la valeur à atteindre, T : "; cin >> T;
 cout << "et le nombre de pas pour y arriver, N : "; cin >> N;
 h = T/N;  h2 = h/2.;  h6 = h/6.;

 for(n=0; n<N; n++)
  {
   p1 = f(u);
   p2 = f(u+h2*p1);
   p3 = f(u+h2*p2);
   p4 = f(u+h*p3);
   u += h6*(p1+2.*p2+2.*p3+p4);
  }
 cout << "u(" << T << ")=(" << u.c[0] << "," << u.c[1] << ")\n";
}
```

Résultats

Pour $T = 1.$, nous trouvons les approximations de $\vec{u}(T)$ suivantes, en fonction du nombre N de subdivisions de T ou de la largeur des pas h :

$$\begin{array}{lcl}
\text{avec } N = 1 & : & \vec{u}(1) \simeq \vec{u}^{\,1} = \begin{pmatrix} 0.45723 \\ 0.971679 \end{pmatrix}, \\
\text{avec } N = 5 & : & \vec{u}(1) \simeq \vec{u}^{\,5} = \begin{pmatrix} 0.45722 \\ 0.978159 \end{pmatrix}, \\
\text{avec } N = 10 & : & \vec{u}(1) \simeq \vec{u}^{\,10} = \begin{pmatrix} 0.457221 \\ 0.978167 \end{pmatrix}.
\end{array}$$

ANALYSE NUMÉRIQUE
Prof. J. Rappaz

Le 15 juin 1995

Solution de la série 9

Exercice 1

Point a)

Soit v une fonction continue qui a les deux propriétés suivantes

i) sa première dérivée est continue par morceaux sur $[0,1]$

ii) $v(0) = v(1) = 0$.

Multiplions l'équation différentielle par v et intégrons entre 0 et 1. Nous obtenons

$$\int_0^1 -v(x)\frac{d}{dx}\left\{\frac{1}{1+x}\cdot\frac{du}{dx}(x)\right\}dx = \int_0^1 f(x)v(x)dx.$$

L'intégration par parties du premier terme donne

$$\int_0^1 -v(x)\frac{d}{dx}\left\{\frac{1}{1+x}\cdot\frac{du}{dx}(x)\right\}dx = \int_0^1 \frac{dv}{dx}\left(\frac{1}{1+x}\cdot\frac{du}{dx}(x)\right)dx,$$

parce que le terme de bord est nul grâce à la propriété (ii). Notons $v'(x) = \frac{dv}{dx}$ et $u'(x) = \frac{du}{dx}$, alors nous avons

$$\int_0^1 \frac{1}{1+x}\cdot u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx.$$

Soit V l'ensemble de toutes les fonctions continues, de dérivée première continue par morceaux, et qui s'annulent en 0 et 1. Le problème : chercher une fonction $u \in V$ qui satisfait

$$\text{(PV)} \qquad \int_0^1 \frac{1}{1+x}\cdot u'(x)v'(x)dx = \int_0^1 f(x)v(x)dx, \qquad \forall v \in V,$$

est la formulation variationnelle du problème (P).

Point b)

Soit $\{\varphi_i\}_{i=1}^N$ les fonctions de base de type "éléments finis" comme dans le cours, et soit V_h le sous-espace vectoriel de V engendré par ces fonctions, c'est-à-dire les φ_i, $i = 1,2,\ldots,N$, forment une base de V_h. Une approximation de (PV) est : trouver une fonction $u_h \in V_h$ t.q.

$$(P_hV) \qquad \int_0^1 \frac{1}{1+x}\cdot u_h'(x)v_h'(x)dx = \int_0^1 f(x)v_h(x)dx, \qquad \forall v_h \in V_h.$$

Comme cette équation est valable pour toute fonction v_h dans V_h, il est équivalent à dire qu'elle est valable si $v_h = \varphi_j \in V_h$, $\forall j = 1,2,\ldots,N$. (P_hV) peut s'écrire

$$\int_0^1 \frac{1}{1+x}\cdot u_h'(x)\varphi_j'(x)dx = \int_0^1 f(x)\varphi_j(x)dx, \qquad \forall j = 1,2,\ldots,N.$$

La fonction u_h est un élément de V_h, i.e. elle s'écrit $u_h(x) = \sum_{i=1}^N u_i\varphi_i(x)$, où $u_1, u_2, \ldots, u_N$ sont N nombres réels. Trouver u_h revient à déterminer ces N coefficients. En substituant l'expression de u_h dans l'équation ci-dessus, nous obtenons

$$(P_\Sigma V) \qquad \sum_{i=1}^N u_i\left\{\int_0^1 \frac{1}{1+x}\cdot\varphi_i'(x)\varphi_j'(x)dx\right\} = \int_0^1 f(x)\varphi_j(x)dx, \qquad \forall j = 1,2,\ldots,N.$$

Soit A la $N\times N$ matrice de coefficients $a_{ji} = \int_0^1 \frac{1}{1+x}\cdot\varphi_i'(x)\varphi_j'(x)dx$, $\vec{f}$ le N-vecteur dont la $j^{\text{ième}}$ composante est $f_j = \int_0^1 f(x)\varphi_j(x)dx$ et $\vec{u}$ le N-vecteur de composantes $u_1, u_2, \ldots, u_N$, alors l'équation $(P_\Sigma V)$ s'écrit $A\vec{u} = \vec{f}$.

Calcul de A

Clairement $a_{ij} = a_{ji}$, i.e. A est une matrice symétrique et nous ne calculerons que les cas où $i \geqslant j$. Rappelons que $x_j = jh$, $j = 0, 1, \ldots, N+1$, $h = \frac{1}{N+1}$ et

$$\varphi_i(x) = \begin{cases} \frac{1}{h}(x - x_{i-1}), & \text{si } x \in [(i-1)h, ih], \\ -\frac{1}{h}(x - x_{i+1}), & \text{si } x \in [ih, (i+1)h], \\ 0, & \text{ailleurs.} \end{cases}$$

La dérivée de φ_i est

$$\varphi_i'(x) = \begin{cases} \frac{1}{h}, & \text{si } x \in [(i-1)h, ih], \\ -\frac{1}{h}, & \text{si } x \in [ih, (i+1)h], \\ 0, & \text{ailleurs.} \end{cases}$$

$(j+1)h$ arctg(x) $j\cdot h$

Nous distinguons 3 cas :

a) si $|i-j| > 1$, alors $a_{ji} = 0$,

b) si $|i-j| = 1$, i.e. $i = j+1$, alors

$$a_{ji} = a_{j,j+1} = -\frac{1}{h^2}\int_{jh}^{(j+1)h} \frac{1}{1+x}dx = -\frac{1}{h^2}\ln\left(\frac{1+(j+1)h}{1+jh}\right)$$

car le produit $\varphi_i'\varphi_j' = 0$ en dehors de l'intervalle $[x_j, x_{j+1}]$,

c) si $i = j$, alors

$$a_{ii} = \frac{1}{h^2}\int_{(i-1)h}^{(i+1)h} \frac{1}{1+x}dx = \frac{1}{h^2}\ln\left(\frac{1+(i+1)h}{1+(i-1)h}\right)$$

La matrice A est symétrique et tridiagonale.

Calcul du second membre

On veut évaluer le second membre en utilisant la méthode des trapèzes. Cette méthode approxime l'intégrale $\int_{x_0}^{x_{N+1}} g(x)dx$ par $L_hg \overset{\text{déf.}}{=} \frac{h}{2}\left\{g(x_0) + 2\sum_{k=1}^N g(x_k) + g(x_{N+1})\right\}$.

Dans notre cas $g(x) = f(x)\varphi_j(x)$, car $f_j = \int_0^1 f(x)\varphi_j(x)dx$. Notons par $\tilde{f}_j = L_h(f\varphi_j)$. Alors

$$\tilde{f}_j = \frac{h}{2}\left\{f(x_0)\varphi_j(x_0) + 2\sum_{k=1}^N f(x_k)\varphi_j(x_k) + f(x_{N+1})\varphi_j(x_{N+1})\right\} = hf(x_j), \quad j = 1,\ldots,N,$$

car $\varphi_j(x_k) = 1$, si $j = k$, et $\varphi_j(x_k) = 0$ sinon.

Point c)

Lorsque $f(x) = x^2$, nous avons $\tilde{f}_j = hx_j^2 = j^2h^3$, $\forall j = 1, 2, \ldots, N$.

Dans ce cas la solution exacte de l'équation différentielle se calcule facilement. En intégrant une fois l'équation du problème (P), nous obtenons $\frac{1}{1+x}\cdot u'(x) = -\frac{1}{3}x^3 + K_1$, où K_1 est une constante d'intégration. En multipliant par $(1+x)$, puis en intégrant de nouveau, nous avons finalement $u(x) = -\frac{1}{15}x^5 - \frac{1}{12}x^4 + \frac{K_1}{2}x^2 + K_1x + K_2$, où K_2 est une deuxième constante d'intégration. Les conditions aux bords sont satisfaites si $K_2 = 0$ et $K_1 = \frac{1}{10}$.

Le calcul en double précision, donne

x	Solution exacte	Solution calculée avec N=7	Solution calculée avec N=3	Solution calculée avec N=1
0.00000000e+00	0.00000000e+00	0.00000000e+00	0.00000000e+00	0.00000000e+00
1.25000000e-01	1.32588704e-02	1.30385046e-02		
2.50000000e-01	2.77343750e-02	2.73246635e-02	2.60958673e-02	
3.75000000e-01	4.23889160e-02	4.18365106e-02		
5.00000000e-01	5.52083333e-02	5.45758625e-02	5.26783631e-02	4.50842200e-02
6.25000000e-01	6.29577637e-02	6.23240675e-02		
7.50000000e-01	6.09375000e-02	6.03977950e-02	5.87783168e-02	
8.75000000e-01	4.27388509e-02	4.24048497e-02		
1.00000000e+00	0.00000000e+00	0.00000000e+00	0.00000000e+00	0.00000000e+00

Remarque

Le passage entre u, la solution exacte, et $\vec{u}$, le vecteur calculé, se fait à l'aide de 2 approximations :

i) *approcher u par une méthode d'éléments finis faites avec des fonctions linéaires par morceaux; c'est l'étape $u_h = \sum_{i=1}^{N} u_i \varphi_i(x)$. L'erreur $e \equiv u - u_h$ est de l'ordre $O(h^2)$.*

ii) *utiliser les formules de quadrature, pour effectuer les intégrations, dans le calcul de la matrice A et du vecteur $\vec{f}$. Dans notre cas les éléments de A sont calculés exactement et les éléments de $\vec{f}$ sont calculés à l'aide de la formule des trapèzes.*

Il est possible de montrer que l'intégration numérique proposée dans ii) ne donne pas lieu à des perturbations plus grandes que celles obtenues en n'utilisant pas d'intégration numérique.

ANALYSE NUMÉRIQUE
Prof. J. Rappaz

27 septembre 1993

PROPÉDEUTIQUE

Problème 1 (3 points)

?

On considère le système différentiel du premier ordre :

$$\begin{cases} \dot{u}_1(t) & = & \sin\big(u_2(t)\big) \\ \dot{u}_2(t) & = & \cos\big(u_1(t)\big) \end{cases} \qquad t > 0$$

avec la condition initiale $u_1(0) = u_2(0) = 1$. Si h est un nombre positif donné, on pose $t_j = jh$ avec $j = 0, 1, 2, \ldots$

√ **a)** En supposant que le vecteur $\vec{u}^n = (u_1^n; u_2^n)$ est une approximation de $\vec{u}(t_n) = \big(u_1(t_n); u_2(t_n)\big)$, écrire le schéma d'*Euler rétrograde* qui permettra de calculer $\vec{u}^{n+1}$ à partir de $\vec{u}^n$. Quelles sont les équations qu'il faudra résoudre pour calculer u_1^{n+1} et u_2^{n+1}?

√ **b)** Faire un seul pas de la méthode de Newton pour calculer numériquement le vecteur $\vec{u}^1$ à partir de $\vec{u}^0$ en prenant $h = 0.1$.

√ Problème 2 (2 points)

Exprimer la base de Lagrange des polynômes de degré 4 associée aux points -2, -1, 0, 1 et 2. Utiliser cette base pour trouver un polynôme p de degré 4 qui satisfait :

$$p(-2) = 3, \ p(-1) = 0, \ p(0) = -3, \ p(1) = 12, \ p(2) = 87.$$

Expliciter le polynôme obtenu.

$f(x_0-h) = f(x_0) - f'(x_0)h - \frac{1}{2}f''(\xi)h^2 \quad \xi \in [x_0, x_0+h]$

$f(x_0) - f(x_0-h) = f'(x_0)h + \frac{1}{2}f''(\xi)h^2$

$\left|f'(x_0) - \frac{f(x_0)-f(x_0-h)}{h}\right| = \frac{1}{2}f''(\xi)h^2$

tourner, svp

EXAMEN D'ANALYSE NUMÉRIQUE — le 13 juillet 1993

Prof. J. Rappaz — Génie Civil, Mécanique, Physique

Début 8h15, Fin 10h00.

SÉRIE B

Problème 1 (3 points). BON EXERCICE.

Soit $f : x \in [0,1] \mapsto f(x) \in \mathbb{R}$ une fonction continue donnée; on se propose de chercher $u : x \in [0,1] \mapsto u(x) \in \mathbb{R}$ qui satisfait

$$(P) \quad \begin{cases} -u''(x) + \cos u(x) = f(x) & si \quad x \in (0,1), \\ u(0) = u(1) = 0. \end{cases}$$

a) *Discrétiser le problème (P) par la méthode des différences finies en posant $h = \frac{1}{N+1}$, où N est un entier positif donné, $x_j = jh$ avec $j = 0, 1, 2, \ldots, N+1$, et u_j est une approximation de $u(x_j)$ i.e. $u_j \simeq u(x_j)$, $1 \le j \le N$. Ecrire explicitement le problème non-linéaire obtenu pour déterminer les valeurs $(u_j)_{1 \le j \le N}$; on appelera ce problème (P_h).*

b) *On considère le cas particulier où $N = 2$ et $f(x) = 0.5$ $\forall x \in [0,1]$. Ecrire explicitement le problème non-linéaire (P_h) et, pour le résoudre numériquement, faire un seul pas de la méthode de Newton en prenant pour valeur de départ pour u_1 et u_2 la valeur 1.*

Problème 2 (2 points).

exercice série 8

Soit $f : (x,t) \in \mathbb{R} \times \mathbb{R}^+ \mapsto f(x,t) \in \mathbb{R}$ une fonction continue donnée et considérons le problème de Cauchy

$$\begin{cases} \dot{u}(t) = f(u(t), t) & si \quad t > 0, \\ u(0) = u_0, \end{cases}$$

où $\dot{u} = \frac{du}{dt}$ et u_0 est un nombre réel donné.

Si h est un paramètre positif donné, si $t_j = jh$ avec $j = 0, 1, 2, \ldots$, on suppose avoir obtenu, pour $n \ge 2$, des approximations $u^0, u^1, \ldots, u^n$ de $u(t_0), u(t_1), \ldots, u(t_n)$ respectivement. Pour calculer u^{n+1} on procède ainsi

i) *on calcule le polynôme $\varphi(t)$ de degré 1 tel que*

$$\varphi(t_{n-1}) = f(u^{n-1}, t_{n-1}), \ \varphi(t_n) = f(u^n, t_n);$$

ii) *on pose $u^{n+1} = u^n + \int_{t_n}^{t_{n+1}} \varphi(t)dt$.*

a) *Justifier cette méthode.*

b) *Expliciter cette méthode, c'est-à-dire expliciter u^{n+1} en fonction de u^n et u^{n-1}.*

tournez s.v.p.

ANALYSE NUMÉRIQUE
Prof. J. Rappaz

le 29 mars 1993
Génie Civil, Mécanique, Physique

Durée d'examen 1h45.

Exercice 1 (5 points). *On considère la $N \times N$ matrice*

$$A = \begin{pmatrix} 8 & 3 & & & & & \\ 1 & 8 & 3 & & & 0 & \\ & 1 & 8 & 3 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & 0 & & & 1 & 8 & 3 \\ & & & & & 1 & 8 \end{pmatrix}.$$

(Plus explicitement $a_{ii} = 8$ si $1 \leq i \leq N$, $a_{i,i+1} = 3$ si $1 \leq i \leq N-1$, $a_{i+1,i} = 1$ si $1 \leq i \leq N-1$ et $a_{ij} = 0$ si $1 \leq i,j \leq N$ avec $|i-j| > 1$).

1) *Démontrer que toutes les sous-matrices principales de A sont régulières et en déduire qu'il existe une unique décomposition $A = LU$ où L est une matrice triangulaire inférieure et U est une matrice triangulaire supérieure avec des 1 dans sa diagonale.*
(Indication: pour démontrer que A est régulière, on montre que $\vec{x}^T A\vec{x} = 0$ implique $\vec{x} = 0$).
2) *Montrer en les calculant que les matrices L et U de la décomposition ci-dessus prennent la forme:*

$$L = \begin{pmatrix} l_1 & & & & \\ m_1 & l_2 & & 0 & \\ & m_2 & l_3 & & \\ & & \ddots & \ddots & \\ 0 & & & m_{N-1} & l_N \end{pmatrix} \quad et \quad U = \begin{pmatrix} 1 & u_1 & & & \\ & 1 & u_2 & & 0 \\ & & 1 & u_3 & \\ & & & \ddots & \ddots \\ & 0 & & & u_{N-1} \\ & & & & 1 \end{pmatrix}.$$

Déterminer de façon récursive les valeurs l_j, $1 \leq j \leq N$; m_j et u_j, $1 \leq j \leq N-1$.
3) *Si B est une $N \times N$ matrice de coefficients $(b_{ij})_{1\leq i,j\leq N}$ données, écrire un algorithme qui permet d'obtenir $A^{-1}B$ en n'utilisant que le tableau B et les valeurs $l_1, \ldots, l_N, m_1, \ldots, m_{N-1}, u_1, \ldots, u_{N-1}$.*

GC - MEC - PHYS. 26.09.1989

Examen Propédeutique II.
Analyse numérique.

CHIMIE 2

Ex. 1 (6 points).

Soient $f(t) = t + 0.1\,t^2$, $g(\alpha,\beta,t) = \mathrm{sh}\,(\alpha t - \beta)$. Déterminer α et $\beta \in \mathbb{R}$ tels que $f(1) = g(\alpha,\beta,1)$ et $f(-1) = g(\alpha,\beta,-1)$.

pas besoin !

a) Expliciter pour α et β un système non linéaire de deux équations à deux inconnues.

b) Résoudre ce système en effectuant deux pas de la méthode de Newton et en partant de l'approximation initiale $(\alpha_0,\beta_0) = (1,0)$. Les systèmes linéaires seront résolus par la méthode canonique (précision des calculs : 4 chiffres après la virgule).

calculatrice ?

Ex. 2 (4 points).

Soient $a \in \mathbb{R}$, $b \in \mathbb{R}$, $h > 0$, $k > 0$,

$$f : \mathbb{R}^2 \longrightarrow \mathbb{R}\,,\quad I = \int_a^{a+h}\int_b^{b+k} f(x,y)\,dxdy\,.$$

a) Déterminer, en s'inspirant de la formule de Simpson, une approximation K de I de la forme $K = \sum_{i=0}^{2}\sum_{j=0}^{2} a_{ij}\, f(x_i,y_j)$ où a_{ij} ne dépend que de h et k et où $x_0 = a$, $x_1 = a + \frac{h}{2}$, $x_2 = a + h$, $y_0 = b$, $y_1 = b + \frac{k}{2}$, $y_2 = b + k$.

b) Déterminer le plus grand entier n tel que $K = I$ $\forall f \in \mathcal{F}_n$ où $\mathcal{F}_n = \left\{ f : \mathbb{R}^2 \longrightarrow \mathbb{R} \mid f(x,y) = \sum_{\substack{i=0\\j=0}}^{n} a_{ij}x^iy^j \right\}$; donner une justification.

27 mai 1997
Informatique
Physique
Physique-faculté

TEST ECRIT (SERIE A)
(tiré d'un examen propédeutique)
Prof. J. Rappaz
3 exercices à faire en maximum 1h

Exercice 1. (4 points)

Soit a, c, d trois nombres réels donnés. On définit la $N \times N$ matrice A par

$$A = \begin{pmatrix} a & & 0 & & c & d \\ & a & & & c & d \\ & & \ddots & & \vdots & \vdots \\ & 0 & & \ddots & \vdots & \vdots \\ & & & \ddots & c & d \\ c & c & \dots & \dots & c & a & d \\ d & d & \dots & \dots & d & d & a \end{pmatrix}$$

a) On suppose que les nombres a, c et d sont donnés de telle sorte que A soit une matrice symétrique définie positive. Vérifier que la décomposition de Cholesky LL^T de A donne lieu à une matrice L de la forme

$$L = \begin{pmatrix} l_1 & & & & & & \\ & l_2 & & & 0 & & \\ & & \ddots & & & & \\ & 0 & & \ddots & & & \\ & & & & \ddots & & \\ m_1 & m_2 & \dots & \dots & m_{N-2} & l_{N-1} & \\ n_1 & n_2 & \dots & \dots & n_{N-2} & n_{N-1} & l_N \end{pmatrix}$$

Donner explicitement les valeurs l_j, $1 \leqslant j \leqslant N$; m_j, $1 \leqslant j \leqslant N-2$ et n_j, $1 \leqslant j \leqslant N-1$ en fonction de a, c et d.

b) Soit $\vec{b}$ un N-vecteur donné. Ecrire un algorithme qui permet de résoudre le système $A\vec{x} = \vec{b}$ en utilisant (a).

Exercice 2. (4.5 points)

Soit deux nombres α et β donnés tels que $-1 \leqslant \alpha < 0$ et $0 < \beta \leqslant 1$. On pose $t_1 = \alpha$, $t_2 = 0$, $t_3 = \beta$ et nous sommes intéressés à trouver trois nombres $\omega_1, \omega_2, \omega_3$ qui définirons la formule de quadrature

$$J(g) = \sum_{j=1}^{3} \omega_j g(t_j),$$

où g est une fonction continue donnée sur $[-1, 1]$.

a) Trouver ω_1, ω_2, ω_3 en fonction de α et β tels que $J(p) = \int_{-1}^{1} p(t)dt$, pour tout polynôme p de degré 2.

b) Montrer que si $J(p) = \int_{-1}^{1} p(t)dt$, pour tout polynôme p de degré 3, alors on a nécessairement $\alpha = -\beta$.

c) Existe-t-il α et β tels que $J(p) = \int_{-1}^{1} p(t)dt$, pour tout polynôme p de degré 4 ? Si oui, que valent α et β ?

TOURNER LA PAGE, S.V.P.

Exercice 3. (1.5 points)

On considère le problème de Cauchy :
trouver une fonction $u : t \in [0, \infty[\to u(t) \in \mathbb{R}$ qui satisfait

$$(P) \qquad \begin{cases} \dot{u}(t) = e^{-\left(u(t)\right)^2} \cdot \sin t, & \text{si } t > 0, \\ u(0) = 0. \end{cases}$$

Si h est un nombre positif donné et si $t_n = nh$ avec $n = 0, 1, 2, \ldots,$ écrire un algorithme qui permet de calculer numériquement, pour $n = 0, 1, 2, \ldots, N$, les approximations u^n de $u(t_n)$ par le schéma d'Euler progressif.

27 septembre 1994

EXAMEN D'ANALYSE NUMERIQUE — Physique

Prof. J. Rappaz

4 exercices à faire entre 8h15 et 10h00

Exercice 1. (3 points)

Soit deux points t_1, t_2 donnés tels que $0 \leqslant t_1 < t_2$. Si $f(t)$ est une fonction continue donnée sur l'intervalle $[0, \infty[$, on veut définir une formule de quadrature de la forme $J(f) = \omega_1 f(t_1) + \omega_2 f(t_2)$ pour approcher numériquement la quantité $\int_0^\infty e^{-t} f(t)dt$.

a) Trouver les poids ω_1, ω_2 en fonction de t_1, t_2 tels que $J(p) = \int_0^\infty e^{-t} p(t)dt$, pour tout polynôme p de degré 1.

b) Montrer que si $t_1 = 2 - \sqrt{2}$ et $t_2 = 2 + \sqrt{2}$, alors la formule de quadrature est exacte pour les polynômes de degré 3.

Exercice 2. (2 points) ⊗ pb. classique : revient chaque année

On considère le problème de Cauchy :
trouver une fonction $u : t \in [0,1] \mapsto u(t) \in \mathbf{R}$ qui satisfait

$$(P) \qquad \begin{cases} \dot{u}(t) = \cos u(t) & si \;\; t > 0\,, \\ u(0) = 0. \end{cases}$$

a) Si h est un nombre positif donné et si $t_n = nh$ avec $n = 0, 1, 2, \ldots$, écrire le schéma des trapèzes (ou Crank-Nicolson) pour calculer numériquement les approximations u^n de $u(t_n)$ où $u(t)$ est solution de (P).

b) Pour calculer u^{n+1} à partir de u^n dans le schéma ci-dessus, on utilise la méthode de Newton. Ecrire explicitement cet algorithme.

c) Faire un seul pas de l'algorithme donné en (b) pour calculer u^1 à partir de u^0.

Exercice 3. (3 points) ⊗

On considère la $N \times N$ matrice A donnée par

$$A = \begin{pmatrix} a & & & & & c \\ & a & & 0 & & c \\ & & \ddots & & & \vdots \\ & & & \ddots & & \vdots \\ & 0 & & & \ddots & c \\ & & & & a & c \\ c & c & \ldots & \ldots & c & a \end{pmatrix}$$

où a et c sont des nombres positifs donnés tels que A puisse être considérée comme une matrice symétrique définie positive.

Vérifier que la décomposition de Choleski LL^T de A donne lieu à une matrice de la forme

$$L = \begin{pmatrix} l_1 & & & & & \\ & l_2 & & & 0 & \\ & & l_3 & & & \\ & & & \ddots & & \\ & 0 & & & \ddots & \\ & & & & l_{N-1} & \\ m_1 & m_2 & m_3 & \ldots & \ldots & m_{N-1} & l_N \end{pmatrix}$$

et écrire un algorithme qui permet de calculer le vecteur $\vec{l}$ de composantes $l_1, l_2, \ldots, l_N$, et le vecteur $\vec{m}$ de composantes $m_1, m_2, \ldots, m_{N-1}$.

(L'algorithme ne devra **pas utiliser de tableau** à 2 indices !)

Exercice 4. (2 points)

L'évolution de la masse d'un isotope radioactif au cours du temps est gouvernée par la relation $m(t) = ae^{-bt}$ où a et b sont des nombres réels à déterminer.

Après avoir fait des mesures, on obtient les valeurs suivantes pour (t, m) : $(1, m_1)$, $(2, m_2)$ et $(3, m_3)$, où m_i, $i = 1, 2, 3$ sont des nombres positifs correspondant aux masses mesurées.

a) Comment utiliser la méthode des moindres carrés pour déterminer des valeurs raisonnables de a et b en fonction de m_1, m_2, m_3 ?
Indication : prendre le logarithme de la relation !

b) Trouvez-vous qu'il y a un rapport entre la solution obtenue au point (a) et la solution du problème :

$$\text{Trouver } a \text{ et } b \text{ tels que } \sum_{i=1}^{3} \left(ae^{-bi} - m_i\right)^2 = \min \, ?$$

Justifier votre réponse.

Problème 3 (5 points).

Soit A la $N \times N$ matrice symétrique définie positive donnée par

$$A = \begin{pmatrix} 8 & 3 & 1 & & & & & & \\ 3 & 8 & 3 & 1 & & & & 0 & \\ 1 & 3 & 8 & 3 & 1 & & & & \\ & 1 & 3 & 8 & 3 & 1 & & & \\ & & 1 & 3 & 8 & 3 & 1 & & \\ & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots \\ & 0 & & & & \ddots & \ddots & \ddots & 1 \\ & & & & & & \ddots & \ddots & 3 \\ & & & & & & 1 & 3 & 8 \end{pmatrix}.$$

(Plus explicitement $a_{ii} = 8$ si $1 \leq i \leq N$; $a_{i,i+1} = a_{i+1,i} = 3$ si $1 \leq i \leq N-1$; $a_{i,i+2} = a_{i+2,i} = 1$ si $1 \leq i \leq N-2$; $a_{ij} = 0$ si $|i-j| \geq 3$.)

a) *Si on considère la décomposition de Cholesky $A = LL^T$, montrer en calculant L que cette dernière prend la forme :*

$$L = \begin{pmatrix} l_1 & & & & & & \\ m_1 & l_2 & & & & 0 & \\ n_1 & m_2 & l_3 & & & & \\ & n_2 & m_3 & l_4 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & 0 & & & \ddots & \ddots & \ddots \\ & & & & n_{N-2} & m_{N-1} & l_N \end{pmatrix}.$$

On ne demande pas de démontrer que A est définie positive.

b) *Ecrire un algorithme qui aura pour sortie les valeurs $l_j, 1 \leq j \leq N; m_j, 1 \leq j \leq (N-1)$ et $n_j, 1 \leq j \leq (N-2)$. (Cet algorithme n'utilisera pas d'autres tableaux que $\vec{l} = (l_1, l_2, \ldots, l_N), \vec{m} = (m_1, m_2, \ldots, m_{N-1})$ et $\vec{n} = (n_1, n_2, \ldots, n_{N-2})$.)*