Compte-rendu en *Méthodes numériques de base*

Résultats sur l'identification de conductivité

Aurélien PEPIN, Antonin KLOPP-TOSSER

2 mai 2017

1 Méthode des différences finies

QUESTION 1. L'écriture sous forme matricielle du θ-schéma *(10)* exprime l'itération $k+1$ du vecteur U en fonction de l'itération k. On isole donc, dans le θ-schéma, les termes en $u_i^{(k+1)}$.

$$
\begin{aligned}
& u_i^{(k+1)} - \mu\theta \left(C_{i+1/2}\, u_{i+1}^{(k+1)} - (C_{i+1/2} + C_{i-1/2})\, u_i^{(k+1)} + C_{i-1/2}\, u_{i-1}^{(k+1)} \right) \\
& = \mu(1-\theta) \left(C_{i+1/2}\, u_{i+1}^{k} - (C_{i+1/2} + C_{i-1/2})\, u_i^{k} + C_{i-1/2}\, u_{i-1}^{k} \right) + u_i^k \\
\\
\iff & \left(1 + \theta\mu(C_{i+1/2} + C_{i-1/2})\right)\, u_i^{(k+1)} - \theta\mu(C_{i+1/2}\, u_{i+1}^{(k+1)}) - \theta\mu(C_{i-1/2}\, u_{i-1}^{(k+1)}) \\
& = \left(1 + (\theta-1)\mu(C_{i+1/2} + C_{i-1/2})\right)\, u_i^{k} - (\theta-1)\mu\, C_{i+1/2}\, u_{i+1}^{k} - (\theta-1)\mu\, C_{i-1/2}\, u_{i-1}^{k}
\end{aligned}
$$

Sachant que $U^{(k)}$ est la matrice-colonne des $u_i^{(k)}$ $\forall i \in [\![1; n]\!]$, on identifie les termes un à un :

$$
\overbrace{\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}}^{I} + \theta\mu \overbrace{\begin{pmatrix} C_{1+1/2} + C_{1-1/2} & -C_{1+1/2} & & 0 \\ -C_{2-1/2} & \ddots & \ddots & \\ & \ddots & \ddots & -C_{(n-1)+1/2} \\ 0 & & -C_{n-1/2} & C_{n+1/2} + C_{n-1/2} \end{pmatrix}}^{A} \cdot \overbrace{\begin{pmatrix} u \\ u \\ u \\ u \end{pmatrix}}^{U}
$$

Question2 :
Soit $x = (x_1, x_2, ..., x_n)$
$x^T Ax = \sum_{i=1}^{n} (C_{i+1/2} + C_{i-1/2}) x_i^2 - 2 \sum_{i=1}^{n-1} C_{i+1/2} x_i x_{i+1}$
$= \sum_{i=1}^{n} C_{i+1/2} x_i^2 + \sum_{j=0}^{n-1} C_{j+1/2} x_{j+1}^2 + -2 \sum_{i=1}^{n-1} C_{i+1/2} x_i x_{i+1}$ On pose j = i-1
$= \sum_{i=1}^{n-1} C_{i+1/2} (x_i - x_{i+1})^2 + C_{1/2} x_{1/2}^2 + C_{n+1/2} x_n^2 > 0$
Comme $x^T Ax > 0 \forall x \in \mathbb{R}^n$ alors la matrice A est symétrique définie positive.

2 Factorisation de Cholesky dans le cas tridiagonal

3 Problème stationnaire

Question 6.
D'après l'équation (9),
$\frac{\partial}{\partial x}[C(x)\frac{\partial u}{\partial x}](x_i,t_k) \approx \frac{C_{i+1/2}u_{i+1}^{(k)}-(C_{i+1/2}+C_{i-1/2})u_i^{(k)}+C_{i-1/2}u_{i-1}^{(k)}}{\delta_x^2}$
Si on écrit ce système, $\forall i$ sous forme matricielle, on obtient :
$\frac{\partial}{\partial x}[C(x)\frac{\partial u}{\partial x}] = \frac{1}{\delta^2}(-Au + \overset{U}{\begin{pmatrix} u_0 C_{1/2} \\ 0 \\ 0 \\ 0 \end{pmatrix}})$ car $u(-l) = u_0$ avec $u = (u_1^{(k)}, ..., u_n^{(k)})$ et A la matrice trouvée à la question 1.
Donc A u = B, avec $B = (b_0, ..., b_n)$ où, $b_0 = C_{1/2}u_0 * \delta_x^2$ et $b_i = 0\ \forall i \in [2, n]$

Ce système admet une solution unique car A est définie et donc inversible.

Question 7.
On résout l'équation :

$$\begin{cases} \frac{\partial}{\partial x}[C(x)\frac{\partial u}{\partial x}] = 0 \\ u(-l) = u_0 \\ u(l) = 0 \end{cases}$$

$\frac{\partial}{\partial x}[C(x)\frac{\partial u}{\partial x}] = 0\ -\frac{1}{l}e^{-\frac{x}{l}}\frac{\partial u}{\partial x} + e^{-\frac{x}{l}}\frac{\partial^2 u}{\partial x^2} = 0\ \frac{\partial^2 u}{\partial x^2} - \frac{1}{l}\frac{\partial u}{\partial x} = 0$
$u(x) = \alpha + \beta e^{\frac{x}{l}}$
Calcul des constantes :

$$\begin{cases} u(-l) = \alpha + \beta e^{-1} = u_0 = 1 \\ u(l) = \alpha + \beta e = 0 \end{cases}$$

$$\begin{cases} \beta = \frac{1}{e^{-1}-e} \\ \alpha = -\frac{e}{e^{-1}-e} \end{cases}$$

Donc : $u(x) = -\frac{e}{e^{-1}-e} + \frac{1}{e^{-1}-e}e^{\frac{x}{l}}$

4 Évolution d'une donnée stationnaire

Question 8.
Question 9.
$MU^{(k+1)} = NU^{(k)} + B$ Cette méthode de décomposition converge uniquement si $\rho(M^{-1}N) < 1$, ce qui est le cas ici(question 8).
Cette décomposition converge vers $Ax = b$, avec A = M - N.
$M - N = I + \theta\mu A - I - (\theta - 1)\mu A = \mu A$
$Ax = B/\mu$ On retrouve ici les matrices A et B de la question (6).

5 Étude du problème inverse

Annexes

6 Sensibilisation à l'arithmétique machine

Exercice 1

Avec $x = 1^{30}$ et $y = 1^{-8}$, les résultats attendus sont $z = w = 1$.
Or, ces calculs sous `Scilab` donnent bien $w = 1$ mais aussi $z = 0$. Étudions la commande :

```
--> x + y
```

Le résultat obtenu est erroné, il vaut x. En effet, le résultat de l'addition requiert **38 chiffres significatifs** ce qui est supérieur à la précision arbitraire permise par la commande `format(20)`. Seuls les premiers chiffres sont pris en compte, c'est une conséquence de la limite de la taille de la mantisse avec la norme IEEE 754 utilisée dans `Scilab`.

Même s'ils devraient mathématiquement être égaux à 1, l'ordre des opérations dans z et dans w altère le résultat. Le fait de mettre des parenthèses va indiquer à la machine quelle opération faire en premier.

— Dans le cas de z, $(y + x) = x$ à cause du manque de précision et donc $(y + x) - x = 0$;
— Dans le cas de w, $(x - x) = 0$ donc $\frac{y+(x-x)}{y} = 1$.

Exercice 2

La figure 1 ci-dessous présente en bleu la valeur attendue du calcul de $f(x)$ et en vert la valeur calculée par `Scilab` sur l'intervalle $[0, 4]$. Si $f(x) = x$ pour tout x de l'intervalle, on voit que ce n'est pas ce qu'on obtient par le calcul. En effet :

— sur l'intervalle $[0, 1]$, $f(x) = 0$;
— sur l'intervalle $[1, 4]$, $f(x) = 1$.

FIGURE 1 – Graphique de l'exercice 2

Méthode. Nous avons calculé y et $f(x)$ grâce à une boucle de 128 itérations. Sur l'intervalle $[0,1]$, le calcul de y donne un résultat très proche de 1 mais strictement inférieur. Ceci est dû à la précision machine et aux algorithmes de calcul de la racine carrée qui ne fournissent pas une valeur exacte mais une valeur approchée par des polynômes.
Le problème est le même sur l'intervalle $[1,4]$ où les écarts de précision introduits par la racine carrée sont amplifiés par la mise au carré. Le calcul de y donne un résultat égal à 1 alors qu'il devrait être strictement supérieur.
Ainsi, sur l'intervalle $[0,1]$, $y^{256} = 0$, alors que sur l'intervalle $[1,4]$, $y = 1$ donc $y^{256} = 1$.

Exercice 3

1. Grâce à une intégration par parties, on établit les égalités suivantes :

$$\begin{aligned} I_n &= \int_0^1 x^n e^x dx \\ &= [x^n e^x]_0^1 - \int_0^1 n x^{n-1} e^x dx \\ &= e - n \int_0^1 x^{n-1} e^x dx \\ &= e - n I_{n-1} \end{aligned}$$

Ce qui conduit à une suite définie par :

$$\begin{cases} I_n &= e - n I_{n-1} \\ I_0 &= e - 1 \end{cases}$$

Évaluation. On implémente la récurrence dans `Scilab` grâce à une fonction récursive. Le résultat de l'appel de la fonction `integrale1(20)` vaut environ `-129.264`. La fonction $x^n e^x$ est positive entre 0 et 1 quel que soit n. L'intégrale d'une fonction positive étant aussi positive, on en déduit que ce résultat est faux.
La valeur de I_n devient en effet très petite au fur et à mesure que n croît. La précision de calcul limitée fausse les résultats, il faut donc choisir une autre méthode.

2. D'après le développement en série de e^x, on sait que :

$$e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!} \implies I_n = \int_0^1 x^n e^x dx = \int_0^1 \sum_{n=0}^{+\infty} \frac{x^n}{n!} dx$$

La suite $u_n = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$ est positive et croissante. D'après le **théorème de convergence monotone**, il est possible d'intervertir la somme et l'intégrale :

$$\begin{aligned}\int_0^1 \sum_{n=0}^{+\infty} \frac{x^n}{n!} dx &= \sum_{n=0}^{+\infty} \int_0^1 \frac{x^n}{n!} dx \\ &= \sum_{n=0}^{+\infty} \frac{1}{n!} \left[\frac{x^{21+n}}{21+n}\right]_0^1 \\ &= \sum_{n=0}^{+\infty} \frac{1}{n!(21+n)}\end{aligned}$$

Évaluation. Le terme général de la série ci-dessus tend rapidement vers 0. On obtiendra donc une bonne approximation de l'intégrale sans calculer énormément de termes. Soit `N` le nombre de termes calculés. Pour `N` = 10, l'appel à `integrale2(N)` vaut environ `0.1238`.

3. Le résultat de la fonction `integrale2` est plus crédible que celui de la fonction `integrale1`. Dans le calcul par récurrence, les erreurs de précision se propagent au fur et à mesure des appels et faussent beaucoup le résultat. Dans le calcul itératif, les erreurs de précision sont minimes et n'ont plus d'influence quand le terme tend vers zéro.

   La fonction `integrate` de `Scilab` permet de vérifier que le calcul itératif est correct.

Exercice 4

En subdivisant l'intervalle d'intégration en `N` points, on peut obtenir une approximation de l'intégrale I_n par la méthode des rectangles *à gauche* ou *à droite*. On note I^G l'approximation par la gauche et I^D l'approximation par la droite.

- Pour `N = 100` points, $I^G = 0.1106$ et $I^D = 0.1379$
- Pour `N = 1000` points, $I^G = 0.1224$ et $I^D = 0.1251$
- Pour `N = 10000` points, $I^G = 0.1236$ et $I^D = 0.1239$
- Pour `N = 100000` points, $I^G = 0.1237$ et $I^D = 0.1238$

Á gauche comme à droite, les valeurs de la méthode des rectangles tendent bien vers `0.1238`.

FIGURE 2 – Graphique de l'exercice 4

7 Étude du phénomène de Gibbs

Exercice 5

Pour calculer la série de Fourier de la fonction $f(x)$, on détermine ses coefficients de Fourier.

> **Note**. Comme la fonction f est impaire, les a_n **sont nuls**. En effet, la fonction $cos(x)$ est paire et le produit d'une fonction paire et d'une fonction impaire reste impair. Puisque l'intervalle sur lequel on intègre est centré en zéro, la partie gauche (< 0) et la partie droite (> 0) de l'intégrale s'annulent.

On procède alors au calcul des $b_n(f)$.

$$
\begin{aligned}
b_n(f) &= \frac{2}{T}\int_{-\frac{1}{2}}^{\frac{1}{2}} f(t)sin(2\pi\frac{n}{t}t)dt \\
&= 2\int_{-\frac{1}{2}}^{\frac{1}{2}} f(t)sin(2\pi nt)dt \text{ car } T=1 \\
&= -2\int_{-\frac{1}{2}}^{0} sin(2\pi nt)dt + 2\int_{0}^{\frac{1}{2}} sin(2\pi nt)dt \\
&= -2\left[\frac{-cos(2\pi nt)}{2\pi n}\right]_{-\frac{1}{2}}^{0} + 2\left[-\frac{cos(2\pi nt)}{2\pi n}\right]_{\frac{1}{2}}^{0} \\
&= \frac{1}{\pi n}(1-cos(\pi n)) - \frac{1}{\pi n}(cos(\pi n)-1) \\
&= \frac{1}{\pi n}(1-(-1)^n)) - \frac{1}{\pi n}((-1)^n-1) \\
&= \frac{2}{\pi n}(1-(-1)^n))
\end{aligned}
$$

Grâce à ce résultat, la série de Fourier de la fonction f est :

$$
\begin{aligned}
f(t) &= \sum_{n=0}^{+\infty} \frac{2}{\pi n}(1-(-1)^n)sin(2\pi nt) \\
&= \sum_{n=0}^{+\infty} \frac{4}{\pi(2n+1)}sin(2(2n+1)\pi t) \text{ car } \forall n \in 2\mathbb{Z}, f(t)=0
\end{aligned}
$$

La fonction $f(x)$ est une fonction affine et dérivable par morceaux. Elle possède trois points de discontinuité : $\{-\frac{1}{2}, 0, \frac{1}{2}\}$. Sa série de Fourier est une fonction continue qui ne peut pas effectuer de « sauts », elle devient alors invalide en ces points. Ainsi, $f(0) = -1$, $f(-\frac{1}{2}) = -1$ et $f(\frac{1}{2}) = 1$ mais la série de Fourier passe par 0 pour ces trois points.

Exercice 6

La figure 3 illustre la représentation en série de Fourier *tronquée* de la fonction $f(x)$ définie dans l'exercice 5. La courbe bleue montre le résultat de la somme des `10` premiers `Ntermes` tandis que la courbe verte est générée à partir de `70 Ntermes`. Plus le nombre de `Ntermes` croît, plus la série de Fourier approche la fonction $f(x)$.
Quel que soit le nombre de termes calculé, il se produit toutefois aux points de discontinuité du signal de fortes oscillations issues du **phénomène de Gibbs**.

FIGURE 3 – Graphique de l'exercice 6

Le phénomène de Gibbs donne l'intuition que la convergence de la série de Fourier quand $n \to \infty$ ne peut pas être uniforme sur $[-\frac{1}{2}; \frac{1}{2}]$. Plus formellement, le *théorème de Dirichlet* permet de conclure que la série de Fourier de la fonction f (dérivable par morceaux) :

- converge simplement vers la fonction f ;
- converge uniformément vers f sur tout $[a, b]$ qui ne contient pas de point de discontinuité ;
- converge vers $\frac{f(t^+)+f(t^-)}{2}$ pour tout point t de discontinuité.

8 Théorème de Gershgorin

Exercice 7

1. On appelle disque de Gershgorin, le disque D_k défini par :

$$D_k = \{z \in \mathbb{C} : |z - akk| \leq \Lambda_k = \sum_{j=1, j\neq k}^{N} |a_{kj}|\}$$

Soit la matrice A, carrée de taille N, telle que $A = (a_{ij})_{\substack{1\leq i\leq N \\ 1\leq j\leq N}}$.

Soit λ un vecteur propre de A et $v = (v_1, v_2, ..., v_N)$ le vecteur propre associé.
D'après l'égalité $(A - \lambda I)v = 0$, on a :

$$\forall i \in [\![1, N]\!], (a_{ii} - \lambda)v_i + \sum_{\substack{j=1 \\ i\neq j}}^{N} a_{ij}v_j = 0 \text{ d'où on déduit que } |a_{ii} - \lambda v_i| = \sum_{\substack{j=1 \\ i\neq j}}^{N} |a_{ij}v_j|$$

On choisit i tel que $|v_i| = sup_{k\in[\![1,N]\!]}|v_k|$. On peut former le quotient $\frac{v_j}{v_i}$ car v est un vecteur propre, donc il est non nul. On a alors $v_i = \underset{k\in[\![1,N]\!]}{sup} |v_k| \neq 0$.

$$\begin{aligned} |a_{ii} - \lambda| &= |\sum_{\substack{j=1 \\ i\neq j}}^{N} a_{ij}\frac{v_j}{v_i}| \\ &\leq \sum_{\substack{j=1 \\ i\neq j}}^{N} |a_{ij}\frac{v_j}{v_i}| \\ &\leq \sum_{\substack{j=1 \\ i\neq j}}^{N} |a_{ij}| \text{ étant donné que } \frac{v_j}{v_i} \leq 1 \end{aligned}$$

À partir de cette inégalité, on vérifie que $\exists i \in [\![1, N]\!], \lambda \in D_i$ soit $\lambda \in \bigcup_{k=1}^{N} D_k$.

2. Dans `Scilab`, on représente un disque de Gershgorin D_k dans le plan complexe où :
   — Le centre (x_O, y_O) du disque vaut $(\Re(a_{kk}), \Im(a_{kk}))$;
   — Le rayon r du disque vaut Λ_k.
   La fonction `spec` permet d'obtenir les valeurs propres de la matrice A. La question suivante montre un exemple de génération des disques grâce à `Scilab`.

3. Ce graphique montre que les valeurs propres sont bien dans l'union des disques de A.

FIGURE 4 – Application du théorème de Gershgorin sur la matrice A

4. Soit A une matrice **strictement dominante** telle que $A = (a_{ij})_{\substack{1\leq i\leq N \\ 1\leq j\leq N}}$.

Comme A est strictement dominante, on a : $\forall i \sum_{k \neq i} |a_{ik}| < |a_{ii}|$.

Soit λ une valeur propre de A. D'après la question 1, $\lambda \in \bigcup_{k=1}^{N} D_k$. Alors :

$$\exists k : |\lambda - a_{kk}| \leq \Lambda_k = \sum_{j=1, j \neq k}^{N} |a_{kj}|$$

$$|\lambda - a_{kk}| \leq \Lambda_k = \sum_{j=1, j \neq k}^{N} |a_{kj}| < |a_{kk}|$$

$$|\lambda - a_{kk}| < |a_{kk}|$$

On en déduit donc que $\lambda \neq 0$. Puisque toutes les valeurs propres de la matrice A sont différentes de 0, la matrice A est **inversible**.