Méthodes numériques : optimisation.
Préparation à l'examen de mai 2015

Amic Frouvelle

19 avril 2015

1 Laplacien discrétisé sur un carré

On considère le carré $\Omega =]0,1[\times]0,1[$, et on cherche à résoudre l'équation de Poisson avec condition nulle au bord sur ce carré (c'est une équation avec tout un tas d'applications en électrostatique, thermique, calcul de champ gravitationnel, etc.). Plus précisément, on se donne une fonction continue $f \in C(\overline{\Omega}, \mathbb{R})$, et on cherche une fonction $u \in C(\overline{\Omega}, \mathbb{R})$ telle que u soit C^2 sur Ω et vérifie l'équation de Poisson $\Delta u = f$ sur Ω, et $u = 0$ sur $\partial\Omega$:

$$\begin{cases} \partial^2_{xx} u(x,y) + \partial^2_{yy} u(x,y) = f(x,y) & \text{si } (x,y) \in]0,1[\times]0,1[\\ u(x,y) = 0 & \text{si } x = 0 \text{ ou } x = 1 \text{ ou } y = 0 \text{ ou } y = 1. \end{cases}$$

1. Montrer que si $g : \mathbb{R} \to \mathbb{R}$ est une fonction de classe C^4 avec $|g^{(4)}(x)| \leqslant C_4$ pour tout $x \in \mathbb{R}$, alors on a pour $x \in \mathbb{R}$ et $h \in \mathbb{R}$:

$$\left| g''(x) - \frac{g(x+h) - 2g(x) + g(x-h)}{h^2} \right| \leqslant \frac{C_4}{12} h^2. \tag{1}$$

On cherche à résoudre ce problème avec cette approximation de la dérivée seconde pour le calcul approché de $\partial^2_{xx} u$ et $\partial^2_{yy} u$. On prend $n \geqslant 1$, on pose $h = \frac{1}{n+1}$ et on approxime u aux points $(x_i, y_j) = (ih, jh)$ pour $0 \leqslant i, j \leqslant n+1$. Comme on sait déjà que u vaut zéro en de tels points si i ou j vaut 0 ou $n+1$ (situés sur le bord), on s'intéresse seulement aux valeurs u_{ij} approximant $u(x_i, y_j)$ pour $1 \leqslant i, j \leqslant n$.

2. Montrer que la résolution du système approché s'écrit

$$\Delta_1 U + U \Delta_1 + B = 0,$$

où $U = (u_{ij}) \in M_n(\mathbb{R})$ est l'inconnue, $B = (b_{ij}) \in M_n(\mathbb{R})$ avec $b_{ij} = h^2 f(x_i, y_j)$, et où Δ_1 est la matrice suivante de $M_n(\mathbb{R})$:

$$\Delta_1 = \begin{pmatrix} 2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2 & -1 & \ddots & & \vdots \\ 0 & -1 & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & -1 & 0 \\ \vdots & & \ddots & -1 & 2 & -1 \\ 0 & \cdots & \cdots & 0 & -1 & 2 \end{pmatrix}.$$

3. On munit $M_n(\mathbb{R})$ du produit scalaire $\langle U, V \rangle = \mathrm{Tr}(UV^T) = \sum_{1 \leqslant i,j \leqslant n} u_{ij} v_{ij}$. On pose A l'application linéaire de $M_n(\mathbb{R})$ dans lui-même qui à une matrice U associe $A(U) = \Delta_1 U + U \Delta_1$. On cherche donc à résoudre

$$A(U) + B = 0. \tag{2}$$

Montrer que l'application A est symétrique, au sens où l'on a $\langle A(U), V \rangle = \langle U, A(V) \rangle$ pour U, V dans $M_n(\mathbb{R})$.

4. Montrer que $-\sin(\theta-\varphi)+2\sin\theta-\sin(\theta+\varphi)=2\sin\theta\,(1-\cos\varphi)$ pour $\theta,\varphi\in\mathbb{R}$.
   En déduire que pour $1\leqslant k\leqslant n$, le vecteur $v_k=(\sin(\frac{jk\pi}{n+1}))_{1\leqslant j\leqslant n}\in\mathbb{R}^n$ vérifie $\Delta_1 v_k=\lambda_k v_k$, avec $\lambda_k=2(1-\cos(\frac{k\pi}{n+1}))$, (pour $1\leqslant k\leqslant n$). En déduire que Δ_1 est une matrice définie positive, et que les v_k sont orthogonaux deux à deux. Quel est le nombre de conditionnement de la matrice Δ_1 (donner un équivalent en fonction de n) ?
5. Pour $1\leqslant k,\ell\leqslant n$, on pose $V_{k,\ell}=v_k v_\ell^T$ (une matrice de $M_n(\mathbb{R})$). Montrer que les matrices $V_{k,\ell}$ sont orthogonales deux à deux et que l'on a $A(V_{k,l})=(\lambda_k+\lambda_\ell)V_{k,l}$.
6. En déduire que l'application A est symétrique définie positive de $M_n(\mathbb{R})$ dans lui-même, et que le problème (2) a une unique solution.
7. Pourquoi une résolution approchée du problème (2) est-elle satisfaisante ? Rappeler quel problème de minimisation est équivalent au problème (2). Pourquoi la méthode du gradient conjugué est adaptée à ce problème, et quel est approximativement le coût en nombre de multiplications et d'additions (entre réels) à chaque itération ? Mêmes questions pour la méthode de descente de gradient à pas fixe ou optimal.

On cherche à résoudre le problème avec une tolérance relative de l'ordre de $\frac{1}{n^2}$ (ceci provient en fait de l'estimation (1)).

8. Donner un équivalent du nombre de conditionnement de l'application A (cela correspond au nombre de conditionnement de la matrice correspondante si on se donnait une base $(e_1,\dots e_{n^2})$ de $M_n(\mathbb{R})$). En déduire une estimation de l'erreur en fonction du nombre d'itérations dans les deux cas de la méthode du gradient conjugué et de celle de descente de gradient à pas fixe ou optimal.
9. Quel est l'ordre de grandeur du nombre maximal d'itérations dans les deux cas pour atteindre la tolérance voulue ? En déduire que le coût total en nombre de multiplications est au plus de l'ordre de $n^3\ln n$ pour la méthode du gradient conjugué et de $n^4\ln n$ pour la méthode de gradient à pas fixe ou optimal. Quel est l'ordre de grandeur du nombre de réels à stocker en mémoire ?
10. * Quelle est la première difficulté à résoudre si on voulait résoudre le problème (2) de manière directe, comme un système linéaire, par exemple par pivot de Gauss ? On peut montrer que si on choisit convenablement la base de $M_n(\mathbb{R})$, alors la matrice correspondant à l'application A s'écrit par bloc (c'est une matrice $n^2\times n^2$, les blocs sont de taille $n\times n$)

$$\Delta_2=\begin{pmatrix} T & M & 0 & \cdots & \cdots & 0\\ M & T & M & \ddots & & \vdots\\ 0 & M & \ddots & \ddots & \ddots & \vdots\\ \vdots & \ddots & \ddots & \ddots & M & 0\\ \vdots & & \ddots & M & T & M\\ 0 & \cdots & \cdots & 0 & M & T \end{pmatrix},$$

   où $T=\Delta_1+2I_n$ et $M=-I_n$. On peut montrer que l'on peut faire des méthodes de décomposition du type du pivot de Gauss de manière efficace sans stocker tous les zéros. Mais les méthodes simples pour faire cela introduisent dans les matrices de décomposition beaucoup de facteurs non-nuls : ans la plupart des cas toutes les diagonales entre la diagonale principale de la grande matrice $n^2\times n^2$ et celles correspondant aux diagonales principales des matrices M dans la matrice par bloc ci-dessus sont non-nulles. Combien faut-il alors stocker de réels ? On montre également que dans ce cas, cela nécessite environ n^4 opérations. *Application numérique :* pour $n=10^3$, combien de place mémoire et de temps de calcul cela prendrait pour les différentes méthodes (on prendra 8 octets pour un réel, et 10^{10} opérations par seconde pour se donner une idée).

En fait il existe des méthodes directes un peu plus sophistiquées qui effectuent des permutations des lignes et des colonnes et qui nécessitent de stocker environ $n^2\ln n$ réels seulement, et qui nécessitent de l'ordre de n^3 opérations. Cependant la méthode du gradient conjuguée est bien plus flexible, et pourrait fonctionner même si les coefficients non-nuls de la grande matrice $n^2\times n^2$ sont répartis plus aléatoirement.

2 Taux de convergence de la méthode du gradient conjugué

On considère une matrice symétrique définie positive $A \in M_n(\mathbb{R})$ et un vecteur $b \in \mathbb{R}^n$. On cherche à approximer la solution x_* de l'équation $Ax + b = 0$ par la méthode du gradient conjugué. On suppose que n est grand et qu'on n'a pas les moyens d'effectuer n itérations de la méthode, donc on s'intéresse à l'estimation de l'erreur sur les premières itérations. On prend les notations $x_k, \alpha_k, r_k, p_k, \beta_k$ du cours.

1. Montrer que la norme $\|\cdot\|_A$ (définie par $\|x\|_A^2 = \langle x, Ax\rangle$) permet de mesurer l'erreur entre x et x_* au sens de la minimisation de la fonction $f(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle$: pour $x \in \mathbb{R}^n$, on a
$$f(x) - f(x_*) = \frac{1}{2}\|x - x_*\|_A^2.$$
2. On pose $e_k = x_k - x_*$. Montrer que $r_k = Ae_k$.
   En déduire que pour tout $k \geqslant 1$ $e_k - e_0 \in \mathrm{Vect}(Ae_0, A^2e_0, \ldots, A^ke_0)$, et qu'il existe donc un polynôme $P_k \in \mathbb{R}_k[X]$ tel que $P_k(0) = 1$ et $e_k = P_k(A)e_0$.
3. Montrer que e_k est l'unique minimiseur du problème suivant
$$\inf_{e \in e_0 + \mathrm{Vect}(Ae_0, A^2e_0, \ldots, A^ke_0)} \|e\|_A^2,$$
   et que donc P_k est l'unique minimiseur du problème
$$\inf_{P \in \mathbb{R}_k[X], P(0)=1} \|P(A)e_0\|_A^2.$$
   Indication : Raisonner par récurrence sur k et décomposer $e - e_0$ dans la base $(p_0, p_1, \ldots, p_{k-1})$.
4. On pose $(v_i)_{1\leqslant i\leqslant n}$ une base de vecteurs propres de A (associée aux valeurs propres $(\lambda_i)_{1\leqslant i\leqslant n}$). En décomposant e_0 dans cette base, si P est un polynôme, obtenir une expression de $P(A)e_k$, puis de $\|P(A)e_0\|_A$ faisant intervenir les $P(\lambda_i)$.
5. Montrer qu'on a alors, pour tout polynôme P de $\mathbb{R}_k[X]$ tel que $P(0) = 1$, en notant Λ l'ensemble des valeurs propres de A,
$$\|e_k\|_A^2 \leqslant \max_{\lambda\in\Lambda} |P(\lambda)|^2 \|e_0\|_A^2.$$
6. *Application :* on suppose que la matrice A a une seule « petite » valeur propre $\frac{1}{\kappa} > 0$ (avec $\kappa \gg 1$) et que toutes ses autres valeurs propres sont dans $[1-\rho, 1]$ avec $\rho \in]0, 1 - \frac{1}{\kappa}[$.
   (a) Quel est le nombre de conditionnement de la matrice A, dans le pire des cas ? À quel taux de convergence peut-on s'attendre d'après le cours ? Montrer qu'on a en fait, pour tout $k \geqslant 1$,
$$\|e_k\|_A \leqslant \kappa\rho^{k-1}\|e_0\|_A,$$
   et que donc pour certaines valeurs de κ et de ρ on a une convergence plus rapide. On pourra utiliser le polynôme $P(X) = (1 - \kappa X)(1 - X)^{k-1}$.
   (b) On suppose qu'on dispose du vecteur propre v_1 associé à la valeur propre $\frac{1}{\kappa}$, avec $\|v_1\| = 1$. On considère l'application linéaire $x \mapsto (\kappa - 1)(x \cdot v_1)v_1 + x$ de $\mathbb{R}^n$ dans $\mathbb{R}^n$, que l'on assimile à une matrice M^{-1}. Quel est le nombre de conditionnement de la matrice $M^{-1}A$? A-t-on intérêt à utiliser M^{-1} comme préconditionneur pour la méthode de gradient conjugué avec préconditionneur ?
7. * On veut démontrer le résultat donné dans le cours. On suppose donc que les valeurs propres de A sont dans $[L, \ell]$. On considère le polynôme de Tchebychev $T_k \in \mathbb{R}^k[X]$ donné par les relations suivantes
$$T_0 = 1, \quad T_1 = X, \quad T_{n+1} = 2XT_n - T_{n-1} \text{ pour } n \geqslant 1.$$
   (a) Montrer que pour $\theta \in \mathbb{R}$, $T_k(\cos\theta) = \cos k\theta$, et que donc $T_k([-1, 1]) \subset [-1, 1]$.

(b) En déduire que si on pose

$$P_k(x) = \frac{T_k(\frac{L+\ell-2x}{L-\ell})}{T_k(\frac{L+\ell}{L-\ell})},$$

alors on obtient $P_k(0) = 1$ et $|P_k(x)| \leqslant \left|T_k(\frac{L+\ell}{L-\ell})\right|^{-1}$ si $x \in [\ell, L]$.

(c) Montrer que pour $|x| > 1$, $T_k(x) = \frac{1}{2}[(x + \sqrt{x^2-1})^k + (x - \sqrt{x^2-1})^k]$.

(d) En déduire que l'on a donc

$$\|e_k\|_A \leqslant 2\left[\left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^k + \left(\frac{\sqrt{L}+\sqrt{\ell}}{\sqrt{L}-\sqrt{\ell}}\right)^k\right]^{-1} \|e_0\|_A \leqslant 2\left(\frac{\sqrt{L}-\sqrt{\ell}}{\sqrt{L}+\sqrt{\ell}}\right)^k \|e_0\|_A.$$

3 Méthode de moindres carrés non-linéaires (Gauss-Newton) [1]

Soient $r_i : \mathbb{R}^n \to \mathbb{R}$ pour $1 \leqslant i \leqslant k$ des fonctions de classe C^2. On notera $J(x) \in M_{k,n}(\mathbb{R})$ la matrice Jacobienne au point x de l'application $r : x \in \mathbb{R}^n \mapsto (r_i(x))_{1\leqslant i\leqslant k} \in \mathbb{R}^k$ (vu comme un vecteur colonne), c'est-à-dire que $J_{ij}(x) = \partial_{x_j} r_i(x)$. On cherche à minimiser $f(x) = \frac{1}{2}\sum_{i=1}^k r_i^2(x) = \frac{1}{2}\|r(x)\|^2$.

1. Calculer $\partial_{x_j} f(x)$ et en déduire une expression simple de $\nabla f(x)$ en fonction de $J(x)$ et de $r(x)$.
2. Calculer $\partial_{x_k}\partial_{x_j} f$ et en déduire une expression de la Hessienne de f au point x notée $H_f(x)$ en fonction de $J(x)$, des $r_i(x)$ et des Hessiennes des r_i au point x notées $H_{r_i}(x)$ pour $1 \leqslant i \leqslant k$.
3. Écrire alors l'équation vérifée par la direction de descente de la méthode de Newton.

On propose une méthode de descente où le choix de direction de descente se fait par la formule suivante : au point courant $x_k \in \mathbb{R}^n$, la direction d_k vérifie

$$J(x_k)^T J(x_k) d_k = -J(x_k)^T r(x_k)\,. \tag{3}$$

La suite $x_k \in \mathbb{R}^n$ est définie de manière habituelle à partir de $x_0 \in \mathbb{R}^n$ et par $x_{k+1} = x_k + \alpha_k d_k$, où le pas α_k est choisi à l'aide d'une méthode de recherche unidimensionnelle de telle sorte qu'il satisfasse la règle de Wolfe.

4. Pour quelle classe de fonctions r_i la formule (3) correspond-elle à la méthode de Newton ?

Soit $x_0 \in \mathbb{R}^n$. On suppose dans la suite que $S_0 = \{x \in \mathbb{R}^n \,|\, f(x) \leq f(x_0)\}$ est borné. On supposera aussi qu'il existe $\gamma > 0$ tel que pour $v \in \mathbb{R}^n$ et $x \in S_0$, on a $\|J(x)v\| \geq \gamma\|v\|$.

On rappelle le théorème de Zoutendijk :

Théorème. *On suppose que ∇f est L-Lipschitz sur S_0, que le pas α_k satisfait la règle de Wolfe, et que f est bornée inférieurement sur S_0.*

Si on note θ_k l'angle entre $-\nabla f(x_k)$ et d_k, de telle sorte que $\cos\theta_k = \frac{\langle -\nabla f(x_k), d_k\rangle}{\|\nabla f(x_k)\|\|d_k\|}$, alors on a

$$\sum_k \cos^2\theta_k \|\nabla f(x_k)\| < +\infty.$$

5. Montrer que les hypothèses du théorème de Zoutendijk sont satisfaites.
6. Minorer $|\cos\theta_k|$ et en déduire que $\lim_{k\to\infty} \nabla f(x_k) = 0$.
7. Que peut-on dire sur la convergence de l'algorithme si $J(x)$ n'est pas injective ?
8. Quel intérêt peut-il y avoir à ne pas considérer la direction de descente de Newton et privilégier la direction de descente définie par la formule (3) ?

1. Origine : examen 2014