Université de Tours
Département de Mathématiques

Contrôle Continu intermédaire

Analyse Numérique et Optimisation — MA2.4, 2019

Exercice : un peu de convexité

1. Soient f et g deux fonctions convexes de $\mathbb{R}$ dans $\mathbb{R}$, g étant croissante. Etablir que $g \circ f$ est convexe.
2. Soient f et g deux fonctions convexes de $\mathbb{R}$ dans $\mathbb{R}$, g étant affine. Etablir que $f \circ g$ est convexe.
3. En déduire que la fonction f_a définie sur $\mathbb{R}$ par $f_a(x) = |1 - ax|$ est convexe.
4. On se donne a et b deux réels tels que $0 < a < b$. Montrer que la fonction $f = \max(f_a, f_b)$ est convexe $[0, +\infty[$ et tracer son graphe sur $[0, +\infty[$.
5. Etablir que l'on a :
$$\min_{x \in [0,+\infty[} f(x) = f\left(\frac{2}{a+b}\right) = \frac{b-a}{a+b}.$$

Problème

Notations et définitions

Le cadre de tout ce problème est l'espace vectoriel $\mathbb{R}^N$. On notera $M_N(\mathbb{R})$ l'ensemble des matrices réelles à N lignes et N colonnes, $\mathcal{S}_N$ le sous-ensemble des matrices symétriques et $\mathcal{P}_N$ celui des matrices symétriques définies positives. Tout élément de $\mathbb{R}^N$ sera identifié à une matrice à une colonne.

Pour A élément de $\mathcal{P}_N$, on notera

$$\langle x, y \rangle_A = y^T A x = \langle x, Ay \rangle, \quad \forall x, y \in \mathbb{R}^N \tag{1}$$

le produit scalaire défini sur $\mathbb{R}^N$ à l'aide du produit canonique $\langle x, y \rangle$ (donc $\langle x, y \rangle = \langle x, y \rangle_I$, I désignant la matrice identité).

$$\|x\|_A = \sqrt{x^T A x}, \quad \forall x \in \mathbb{R}^N \tag{2}$$

la norme correspondante, et

$$|||M|||_A = \max_{x \in \mathbb{R}^N \setminus \{0\}} \frac{\|Mx\|_A}{\|x\|_A}, \quad \forall M \in M_N(\mathbb{R}) \tag{3}$$

la norme d'opérateur associée. On notera M_A^* l'adjoint de $M \in M_N(\mathbb{R})$ relativement au produit scalaire $\langle \cdot, \cdot \rangle_A$.
Pour $M \in M_N(\mathbb{R})$, on notera $\Lambda(M)$ l'ensemble de ses valeurs propres et $\rho(M)$ son rayon spectral défini par

$$\rho(M) = \max\{|\lambda|, \lambda \in \Lambda(M)\}. \tag{4}$$

Lorsque $\Lambda(M)$ est dans $\mathbb{R}$, on notera $\lambda_1(M)$ la plus petite et $\lambda_N(M)$ la plus grande des valeurs propres de M.
Pour M élément inversible de $M_N(\mathbb{R})$, on utilisera la notation M^{-T} pour $(M^T)^{-1} = (M^{-1})^T$.
Le conditionnement d'une matrice M inversible dans $M_N(\mathbb{R})$ est défini par

$$K(M) = |||M||| \cdot |||M^{-1}||| \tag{5}$$

avec $|||M||| = |||M|||_I$, I désignant la matrice identité.
L'objet du problème est d'étudier des méthodes optimales en terme du conditionnement de la matrice utilisée pour la résolution du système linéaire

$$Ax^* = b \tag{6}$$

où A et b sont donnés, respectivement dans $\mathcal{P}_N$ et $\mathbb{R}^N$, x^* étant l'unique solution de (6). On utilisera la fonctionnelle $J_{A,b}$ définie par

$$J_{A,b}(x) = \frac{1}{2} x^T A x - x^T b = \frac{1}{2}\langle x, Ax \rangle - \langle x, b \rangle, \quad \forall x \in \mathbb{R}^N. \tag{7}$$

Préliminaires

0. Notons $\|x\|_\infty = \max\{|x_i|, i \in [\![1,N]\!]\}$ où on a noté $x = (x_1, \ldots, x_N)$. Montrer que la norme $|||\cdot|||_\infty$ subordonnée à la norme $\|\cdot\|_\infty$ est donnée , si $A = (a_{ij})_{1\leq 1, j\leqslant N}$, par

$$|||A|||_\infty = \max_{i\in[\![1,N]\!]} \sum_{j=1}^{N} |a_{ij}|.$$

1. Etablir l'égalité

$$\Lambda(M_1M_2) = \Lambda(M_2M_1), \quad \forall M_1, M_2 \in M_N(\mathbb{R}) \tag{8}$$

et en déduire

$$\rho(M_1M_2) = \rho(M_2M_1), \quad \forall M_1, M_2 \in M_N(\mathbb{R}). \tag{9}$$

2. Montrer que tout élément S de $\mathcal{S}_N$ vérifie les relations

$$\lambda_1(S) = \min_{x\in\mathbb{R}^N\setminus\{0\}} \frac{x^TSx}{x^Tx}, \quad \lambda_N(S) = \max_{x\in\mathbb{R}^N\setminus\{0\}} \frac{x^TSx}{x^Tx}. \tag{10}$$

On pourra énoncer puis utiliser le théorème spectral pour la matrice S.

3. Montrer que pour tout $C \in \mathcal{P}_N$ et tout $S \in \mathcal{S}_N$, on a les relations

$$\lambda_1(C^{-1}S) = \lambda_1(L^{-1}SL^{-T}) = \min_{x\in\mathbb{R}^N\setminus\{0\}} \frac{x^TSx}{x^TCx}, \quad \lambda_N(C^{-1}S) = \max_{x\in\mathbb{R}^N\setminus\{0\}} \frac{x^TSx}{x^TCx}, \tag{11}$$

où on a utilisé la factorisation de Cholesky LL^T de C. On pourra poser $y = L^Tx$.

4. Dans le cadre de $\mathbb{R}^N$ muni du produit scalaire $\langle .,. \rangle_A$ (voir (1)),

   (a) Démontrer que, pour tout $M \in M_N(\mathbb{R})$, l'adjoint de M pour ce nouveau produit scalaire, M_A^*, vaut $A^{-1}M^TA$.

   (b) Etablir les relations

$$|||M|||_A^2 = \rho(M_A^*M) = |||M_A^*|||_A^2, \quad \forall M \in M_N(\mathbb{R}). \tag{12}$$

   (c) Montrer que, si M est autoadjoint pour $\langle .,. \rangle_A$ ($M_A^* = M$),

$$|||M|||_A = \rho(M), \tag{13}$$

   et retrouver ainsi que :

$$K(M) = \lambda_N(M)/\lambda_1(M), \quad \forall M \in \mathcal{P}_N. \tag{14}$$

5. On considère le problème modèle avec conditions de Dirichlet :

$$-u''(x) + \quad u(x) = f(x) \quad \forall x \in (0,1) \tag{15}$$

$$u(0) = 0 \qquad\qquad u(1) = 0 \tag{16}$$

On se donne un entier $N \geq 9$, on pose $h = 1N + 1$ et on définit la subdivision $(x_i = ih)_{1\leq i\leq N}$ de l'intervalle $[0,1]$.

   (a) Etablir que la mise en œuvre de la méthode des différences finies avec le schéma classique vu en cours, conduit à résoudre un système linéaire $A_N(h)U = b$, dont la matrice $A_N(h)$ est :

$$A_N(h) = \begin{pmatrix} 2+h^2 & -1 & 0 & \cdots & \cdots & 0 \\ -1 & 2+h^2 & -1 & 0 & \cdots & 0 \\ 0 & -1 & 2+h^2 & -1 & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \ddots & \vdots \\ 0 & & \ddots & \ddots & \ddots & -1 \\ 0 & \cdots & & 0 & -1 & 2+h^2 \end{pmatrix} \tag{17}$$

   (b) Etablir, grâce au théorème de Gerschgöring–Hadamard que l'on énoncera, que la matrice $A_N(h)$ est symétrique définie positive.

(c) Démontrer que $\Lambda(A_N(h)) = \{\lambda_k = h^2 + 4\sin^2(\frac{k\pi}{2(N+1)}), 1 \leq k \leq N\}$. On commencera par relier $\Lambda(A_N(h))$ à celui de la matrice $A_N(0)$, $\Lambda(A_N(0))$.

(d) En déduire que

$$K(A_N(h)) \simeq \frac{4}{\pi^2 + 1)h^2}. \tag{18}$$

6. Suivant les définitions (6) et (7),

(a) Etablir l'identité

$$J_{A,b}(x) = J_{A,b}(x^*) + \frac{1}{2}\|x - x^*\|_A^2, \quad \forall x \in \mathbb{R}^N \tag{19}$$

et en déduire que x^* est l'unique solution du problème de minimisation

$$\min\{J_{A,b}(x), x \in \mathbb{R}^N\}. \tag{20}$$

(b) B étant un élément inversible de $M_N(\mathbb{R})$, on effectue le changement de variable

$$\tilde{x} = B^T x. \tag{21}$$

Déterminer $\tilde{A}$ et $\tilde{b}$ définis par

$$J_{A,b}(x) = J_{\tilde{A},\tilde{b}}(\tilde{x}), \tag{22}$$

et montrer que

$$K(\tilde{A}) = \lambda_N(C^{-1}A)/\lambda_1(C^{-1}A) \tag{23}$$

où C est l'élément de $\mathcal{P}_N$ défini par

$$C = BB^T. \tag{24}$$

Dans la suite, le système de matrice $\tilde{A}$ et de second membre $\tilde{b}$ sera appelé système préconditionné, C étant dit précontitionneur, le but étant d'avoir $K(\tilde{A})$ plus petit que $K(A)$.

Partie 1

Dans le but de mieux comprendre ce que peut être un bon préconditionneur C pour A, on définit, compte tenu de (23), l'application σ de $\mathcal{P}_N \times \mathcal{P}_N$ dans $[1, \infty[$ par

$$\sigma(P_1, P_2) = \lambda_N(P_2^{-1}P_1)/\lambda_1(P_2^{-1}P_1). \tag{25}$$

6. (a) Montrer que σ vérifie les propriétés suivantes :

$$\sigma(P_1, P_2) = \sigma(P_2, P_1), \quad \forall P_1, P_2 \in \mathcal{P}_N \tag{26}$$

$$\sigma(\alpha_1 P_1, \alpha_2 P_2) = \sigma(P_1, P_2), \quad \forall P_1, P_2 \in \mathcal{P}_N, \forall \alpha_1, \alpha_2 \in \mathbb{R}_+^* \tag{27}$$

$$\sigma(P_1, P_2) \leqslant \sigma(P_1, P_3)\sigma(P_3, P_2), \quad \forall P_1, P_2, P_3 \in \mathcal{P}_N. \tag{28}$$

(b) Montrer que la relation définie dans $\mathcal{P}_N$ par

$$P_1 \mathcal{R} P_2 \quad \text{ssi} \quad \sigma(P_1, P_2) = 1 \tag{29}$$

est une relation d'équivalence qui s'exprime par $\exists \alpha \in \mathbb{R}_+^*$ tel que $P_2 = \alpha P_1$. On justifiera que toute matrice diagonalisable qui ne possède qu'une valeur propre α est égale à αI.

(c) Montrer qu'on peut définir pour l'ensemble quotient $\mathcal{P}_N/\mathcal{R}$ l'application $\dot{\sigma}$ par

$$\dot{\sigma}(\dot{P}_1, \dot{P}_2) = \sigma(Q_1, Q_2), \forall Q_1 \in \dot{P}_1, \forall Q_2 \in \dot{P}_2 \tag{30}$$

en notant $\dot{P}$ la classe d'équivalence de P, et une distante

$$d_\phi = \phi \circ \dot{\sigma} \tag{31}$$

avec $\phi : [1, \infty[\to \mathbb{R}_+$ vérifiant

$$\phi(t) = 0 \Leftrightarrow t = 1, \quad \phi(t_1 t_2) \leqslant \phi(t_1) + \phi(t_2).$$

Vérifier que d_1 qui correspond à la fonction $\phi_1(t) = \frac{t-1}{t+1}$ est une telle distance.

7. Pour la résolution itérative de (6), on considère la méthode associée à $P \in \mathcal{P}_N$, dont chaque pas est défini par
$$P(x_{k+1} - x_k) = b - Ax_k \tag{32}$$

   (a) Montrer que la matrice d'itération correspondante est $I - P^{-1}A$ et vérifie
$$|||I - P^{-1}A|||_A = \rho(I - P^{-1}A). \tag{33}$$

   (b) En utilisant la notation
$$\dot{\rho}(A, C) = \min\{|||I - P^{-1}A|||_A, P \in \dot{C}\}, \tag{34}$$
   montrer que
$$\dot{\rho}(A, C) = \min\{\rho(I - \alpha C^{-1}A), \alpha \in \mathbb{R}\} = \frac{\sigma(A, C) - 1}{\sigma(A, C) + 1}, \tag{35}$$
   On pourra utiliser le résultat de l'exercice et le réel :
$$\alpha = \frac{2}{\lambda_1(C^{-1}A) + \lambda_N(C^{-1}A)}.$$

   (c) Vérifier les relations
$$\dot{\rho}(A, C) = d_1(\dot{A}, \dot{C}), \tag{36}$$
$$K(\tilde{A}) = \frac{1 + d_1(\dot{A}, \dot{C})}{1 - d_1(\dot{A}, \dot{C})} \tag{37}$$
   qui montrent que l'étude de $\dot{\rho}(A, C)$ et celle de $K(\tilde{A})$ sont liées, et que l'intuition que le conditionnement qu'il en résulte est d'autant meilleur que "C est une bonne approximation de A" est justifiée rigoureusement par la distance d_1.

8. On considère ici, pour résoudre le système linéaire (6), une méthode itérative associée à M inversible non nécessairement dans $\mathcal{P}_N$:
$$M(x_{k+1} - x_k) = b - Ax_k, \tag{38}$$
   et on va construire une méthode associée à une matrice C de $\mathcal{P}_N$ pour l'utiliser comme préconditionneur (cf la question précédente). On note $U = I - M^{-1}A$ la matrice d'itération de la méthode (38).

   Cette méthode s'obtient par symétrisation de (38) : chaque pas consiste à effectuer un pas de la méthode (38) suivi d'un pas de la méthode obtenue en remplaçant M par M^T dans (38) (matrice d'itération notée V), c'est-à-dire que x_{k+1} est déduit de x_k par l'intermédiaire de $x_{k+1/2}$:
$$\begin{cases} M(x_{k+1/2} - x_k) = b - Ax_k \\ M^T(x_{k+1} - x_{k+1/2}) = b - Ax_{k+1/2} \end{cases} \tag{39}$$
   et on note W la matrice d'itération de cette méthode symétrisée.

   (a) Montrer que
$$V = U_A^*, \quad W = U_A^* U; \tag{40}$$
   en déduire que la méthode (40) converge si et seulement si $M + M^T - A$ est dans $\mathcal{P}_N$ et qu'elle peut alors s'écrire
$$C(x_{k+1} - x_k) = b - Ax_k, \tag{41}$$
   où la matrice
$$C = M(M + M^T - A)^{-1}M^T \tag{42}$$
   est élément de $\mathcal{P}_N$ et vérifie
$$\Lambda(C^{-1}A) \subset]0, 1]. \tag{43}$$

   (b) Etablir l'équivalence
$$|||I - M^{-1}A|||_A < 1 \Leftrightarrow M + M^T - A \in \mathcal{P}_N. \tag{44}$$

9. On va appliquer les résultats de la question précédente à la méthode classique dite de relaxation qui consiste à remplacer dans (38) M par

$$M(\omega) = \frac{1}{\omega}D - E, \quad \omega \in \mathbb{R}^*, \tag{45}$$

en utilisant les notations classiques pour la décomposition

$$A = D - E - E^T, \tag{46}$$

où

$$D_{ij} = \delta_{ij}A_{ij}, \quad E_{ij} = \begin{cases} -A_{ij} \text{ si } j < i \\ 0 \text{ si } j \geqslant i. \end{cases}$$

(a) Etablir l'équivalence

$$M(\omega) + M(\omega)^T - A \in \mathcal{P}_N \Leftrightarrow \omega \in]0, 2[\tag{47}$$

et montrer que pour tout $\omega \in]0, 2[$, la méthode symétrisée s'écrit sous la forme (41) en remplaçant C par

$$C(\omega) = \left(\frac{1}{\omega}D - E\right)\left(\frac{2-\omega}{\omega}D\right)^{-1}\left(\frac{1}{\omega}D - E\right)^T. \tag{48}$$

(b) Si on note $U(\omega)$ la matrice d'itération de la méthode de relaxation, établir l'équivalence

$$\|U(\omega)\|_A < 1 \Leftrightarrow \omega \in]0, 2[, \tag{49}$$

et en déduire la condition nécessaire et suffisante

$$\rho(U(\omega)) < 1 \Leftrightarrow \omega \in]0, 2[\tag{50}$$

après avoir vérifié que $\det(U(\omega)) = (1-\omega)^N$.

Partie 2

La suite consiste en une étude du préconditionneur $C(\omega)$, $\omega \in]0, 2[$, défini à partir de A par (48), connu sous le nom de préconditionneur S.S.O.R. (Symmetric Successive Over Relaxation), du fait qu'il a été obtenu par symétrisation de la méthode de relaxation dite S.O.R..

10. (a) En utilisant les résultats vus dans les préliminaires, la définition de σ ((25)) et (50), montrer que

$$\sigma(A, C(\omega)) \leqslant \max_{x \in \mathbb{R}^N \setminus \{0\}} \frac{x^T C(\omega) x}{x^T A x}. \tag{51}$$

(b) Vérifier l'identité

$$C(\omega) = \frac{1}{2-\omega}\left[A + \frac{(2-\omega)^2}{4\omega}D + \omega(ED^{-1}E^T - \frac{1}{4}D)\right] \tag{52}$$

et déduire la majoration

$$\sigma(A, C(\omega)) \leqslant \frac{1}{2-\omega}\left[1 + \frac{(2-\omega)^2}{4\omega}\mu(A) + \omega\delta(A)\right], \tag{53}$$

où on a utilisé les notations

$$\begin{cases} \mu(A) = \max\limits_{x \in \mathbb{R}^N \setminus \{0\}} \dfrac{x^T D x}{x^T A x}, \\ \delta(A) = \max\limits_{x \in \mathbb{R}^N \setminus \{0\}} \dfrac{x^T (ED^{-1}E^T) x - \frac{1}{4}x^T D x}{x^T A x} \end{cases} \tag{54}$$

11. (a) On suppose que $1 + 2\delta(A) > 0$. En étudiant la fonction φ définie sur $]0, 2[$ par

$$\varphi(t) = \frac{1}{2-t}[1 + \frac{(2-t)^2}{4t}\mu(A) + t\delta(A)],$$

montrer que la borne de (53) est minimale pour $\omega = \omega^*$, avec

$$\omega^* = \frac{2}{1+\sqrt{\frac{2(1+2\delta(A))}{\mu(A)}}} \tag{55}$$

et que

$$\sigma(A, C(\omega^*)) \leqslant \sqrt{\frac{1}{2}+\delta(A)}\sqrt{\mu(A)}+\frac{1}{2}. \tag{56}$$

(b) Montrer que, si la matrice A du système vérifie

$$\delta(A) \leqslant 0, \tag{57}$$

alors

$$\sigma(A, C(\omega_0^*)) \leqslant \sqrt{\frac{\mu(A)}{2}}+\frac{1}{2} \tag{58}$$

pour $\omega = \omega_0^*$ avec

$$\omega_0^* = \frac{2}{1+\sqrt{\frac{2}{\mu(A)}}}. \tag{59}$$

(c) Déduire que, dans ce cas, la matrice $\tilde{A}$ du système préconditionné par $C(\omega_0^*)$ vérifie

$$K(\tilde{A}) \leqslant \sqrt{\frac{K(A)}{2}}+\frac{1}{2}. \tag{60}$$

12. On introduit les notations

$$\widehat{A} = D^{-1/2}AD^{-1/2}, \widehat{E} = D^{-1/2}ED^{-1/2} \tag{61}$$

avec $(D^{1/2})_{ij} = \sqrt{D_{ij}})\delta_{ij}\sqrt{A_{ii}}$ et $D^{-1/2} = (D^{1/2})^{-1}$, d'où

$$\widehat{A} = I - \widehat{E} - \widehat{E}^T. \tag{62}$$

(a) Etablir l'identité

$$\delta(A) = \max_{y\in\mathbb{R}^N\setminus\{0\}} \frac{y^T\widehat{E}\widehat{E}^Ty - \frac{1}{4}y^Ty}{y^T\widehat{A}y}, \tag{63}$$

et déduire que la condition (57) est équivalente à la suivante

$$\rho(\widehat{E}\widehat{E}^T) \leqslant 1/4. \tag{64}$$

(b) Déduire l'implication

$$|||\widehat{E}|||_\infty \leqslant \frac{1}{2} \text{ et } |||\widehat{E}^T|||_\infty \leqslant \frac{1}{2} \Rightarrow \delta(A) \leqslant 0. \tag{65}$$

(c) Etablir la minoration

$$\delta(A) \geqslant \frac{\lambda_N(\widehat{A})-2}{4} \quad \text{puis que } 1+2\delta(A) > 0. \tag{66}$$

On pourra d'abord établir les relations suivantes :

$$\forall y \in \mathbb{R}^N \setminus \{0\}, \quad y^Ty.y^T\widehat{E}\widehat{E}^Ty \geqslant (y^T\widehat{E}y)^2 = \frac{1}{4}(y^T\widehat{A}y - y^Ty)^2.$$

Partie 3 Application au problème modèle (15)

Dans cette ultime partie, on quantifie le gain en terme de conditionnement que l'on fait en préconditionnant le système linéaire issu du problème (15), grâce au préconditionneur SSOR.

13. Etablir que le réel $\mu(A_N(h))$ défini par (54) vaut $\frac{2+h^2}{h^2+4\sin^2(\frac{\pi}{2(N+1)})}$. Donner un équivalent de $\mu(A_N(h))$ en terme de puissance de h.

14. (a) Etablir que si E_N et $\hat{E}_N$ sont les matrices définies dans la partie 2 (par (61)) pour $A_N(h)$, on a

$$|||\hat{E}_N|||_\infty < \frac{1}{2}, |||\hat{E}_N^T|||_\infty < \frac{1}{2}.$$

(b) En déduire que $\delta(A_N(h)) \leq 0$ puis que $\delta(A_N(h)) \geq -\frac{1}{2}\sin^2(\frac{\pi}{2(N+1)})$.

(c) Etablir que $\sin(x) \geq \frac{2}{\pi}x$ pour tout $x \in [0, \frac{\pi}{2}]$ et en déduire grâce à la partie 2 que

$$K(\tilde{A}_N(h)) \leq \frac{1}{h} + \frac{1}{2}.$$

15. Conclure.