Cours ANN 203 – Examen écrit, jeudi 2 mai 2024

NOTE: Les matrices et les vecteurs considérés dans tous les exercices sont supposés à valeurs réelles. La notation $\|x\|$ désigne la norme 2 (euclidienne) du vecteur x . $\text{span}(a_1, \dots, a_k)$ désigne l'espace engendré par la famille $a_1, \dots, a_k$. E24-0

Exercice E24-1 Pour $\varepsilon > 0$, on considère la matrice

$$A = \begin{bmatrix} 1 & 1 \\ 0 & \varepsilon \end{bmatrix}$$

Déterminer, en fonction du paramètre ε , les valeurs singulières de A et son conditionnement $\kappa_2(A)$. Commenter l'effet de ε sur le résultat.

Solution: on a $A = USV^T$, où U et V sont les matrices contenant (par colonne) les vecteurs propres de $A^T A = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon^2 \end{bmatrix}$ et $AA^T = \begin{bmatrix} 1 & 1 \\ 1 & 1+\varepsilon^2 \end{bmatrix}$ et S est la matrice diagonale contenant les valeurs propres (identiques) de $A^T A$ ou AA^T ; on trouve

$$S = \text{diag}(\lambda_1, \lambda_2) = \frac{1}{2} \text{diag}(2 + \varepsilon^2 + \sqrt{2 + \varepsilon^4}, 2 + \varepsilon^2 - \sqrt{2 + \varepsilon^4})$$

Les valeurs singulières de S sont $\sigma_1 = \sqrt{\lambda_1}$, $\sigma_2 = \sqrt{\lambda_2}$. Son nombre de conditionnement est

$$\kappa_2(A) = \sigma_1/\sigma_2 = \frac{2 + \varepsilon^2 + \sqrt{2 + \varepsilon^4}}{2 + \varepsilon^2 - \sqrt{2 + \varepsilon^4}} = \frac{(2 + \varepsilon^2 + \sqrt{2 + \varepsilon^4})^2}{\varepsilon^4 - \varepsilon^2} = \frac{(2 + \sqrt{2})^2}{\varepsilon^2} \left(1 + O(\varepsilon^2)\right)$$

La matrice A est donc d'autant plus mal conditionnée que ε est petit.

Exercice E24-2 (orthogonalisation de Gram-Schmidt) Soit $A \in \mathbb{R}^{m \times n}$ avec $m \geq n$, supposée de rang maximal. Les colonnes de $A = [a_1, \dots, a_n]$ définissent une famille de n vecteurs $a_j \in \mathbb{R}^m$. Il est fréquemment utile ou important de trouver une famille orthonormée de vecteurs $q_1, \dots, q_n \in \mathbb{R}^m$ telle que $\text{span}(a_1, \dots, a_n) = \text{span}(q_1, \dots, q_n)$. Le procédé de Gram-Schmidt (et ses diverses extensions), objet de cet exercice, consiste à construire séquentiellement les vecteurs q_i de sorte que $\text{span}(q_1) = \text{span}(a_1)$, $\text{span}(q_1, q_2) = \text{span}(a_1, a_2)$, $\text{span}(q_1, q_2, q_3) = \text{span}(a_1, a_2, a_3) \dots$

(a) Définir q_1 .

(b) Soit $2 \leq k \leq n$. On suppose $(q_1, \dots, q_{k-1})$ construite. Montrer qu'on peut construire q_k tel que

$$a_k = b_{k1}q_1 + \dots + b_{kk}q_k, \quad q_i^T q_k = 0 \quad (i < k), \quad \|q_k\| = 1 \quad (*)$$

q_k est-il unique? Montrer que $\text{span}(q_1, \dots, q_k) = \text{span}(a_1, \dots, a_k)$ pour tout k .

(c) Montrer que le procédé séquentiel ci-dessus, une fois complet ($k = n$ atteint), revient à écrire

$$A = QB$$

avec $Q = [q_1, \dots, q_n]$. Que peut-on dire de la matrice B ?

(d) Quelles sont les conséquences de ce qui précède sur l'existence (ou non) et l'unicité (ou non) de la factorisation QR de A ? Énoncer une règle simple sur le processus d'orthonormalisation permettant de rendre unique cette factorisation?

Solution:

(a) Le choix naturel est $q_1 = a_1/\|a_1\|$ (son opposé convient aussi).

(b) Multipliant (*) à gauche par q_i^T ($i < k$), on trouve

$$b_{ik} = q_i^T a_k, \quad 1 \leq i \leq k-1.$$

Le coefficient b_{kk} s'obtient alors en réécrivant (*) comme

$$b_{kk}q_k = a_k - b_{k1}q_1 - \dots - b_{k(k-1)}q_{k-1}$$

et en imposant la condition de normalisation $\|q_k\| = 1$; son signe peut donc être choisi librement, ce qui donne deux signes opposés possibles pour q_k .

On procède ensuite par récurrence: supposons que $\text{span}(q_1, \dots, q_{k-1}) = \text{span}(a_1, \dots, a_{k-1})$ pour $k \geq 2$. Cela est clairement vrai pour $k = 2$. Soit alors $u \in \text{span}(a_1, \dots, a_k)$, que l'on peut écrire comme $u = v + ta_k$ avec $v \in \text{span}(a_1, \dots, a_{k-1})$. La relation $(\star)$ entraîne que $a_k \in \text{span}(q_1, \dots, q_k)$, tandis que $v \in \text{span}(q_1, \dots, q_{k-1})$ par hypothèse de récurrence. Par conséquent, $u \in \text{span}(a_1, \dots, a_k) \implies u \in \text{span}(q_1, \dots, q_k)$, et donc $\text{span}(a_1, \dots, a_k) \subset \text{span}(q_1, \dots, q_k)$. Les deux familles de vecteurs a_i et q_i étant linéairement indépendantes et de même cardinal pour tout k , on doit alors avoir $\text{span}(q_1, \dots, q_k) = \text{span}(a_1, \dots, a_k)$ pour tout k .

(c) Les relations $(\star)$ définissant a_k pour tout k peuvent être écrites comme

$$A = QB, \quad B = \begin{bmatrix} b_{11} & b_{21} & b_{31} & \dots & b_{k1} \\ 0 & b_{22} & b_{32} & \dots & b_{k2} \\ 0 & & \ddots & & \vdots \\ \vdots & & & \ddots & b_{k(k-1)} \\ 0 & \dots & \dots & \dots & b_{kk} \end{bmatrix},$$

et B est donc triangulaire supérieure.

(d) $A = QB$ constitue une factorisation QR de A (Q étant unitaire et B triangulaire supérieure). La condition de normalisation $\|q_k\| = 1$ permet pour chaque k de choisir librement le signe de b_{kk} , et il y a donc 2^n factorisations $A = QB$ possibles. Les coefficients b_{kk} étant sur la diagonale de B , un choix naturel est de les prendre tous positifs.

Exercice E24-3 (convergence des itérations de Rayleigh) *Les itérations de Rayleigh (algorithme 5.3, rappelé ci-dessous), qui sont une version améliorée des itérations inverses, convergent rapidement vers une valeur propre et un vecteur propre d'une matrice $A \in \mathbb{R}^{n \times n}$ symétrique définie positive (de valeurs propres $\lambda_1, \dots, \lambda_n$ et vecteurs propres orthonormés associés $q_1, \dots, q_n$). L'objet de cet exercice est de quantifier cette convergence.*

Algorithm Rayleigh quotient iteration

```

1:  $A \in \mathbb{K}^{n \times n}$  Hermitian (input),  $x^{(0)} \in \mathbb{K}^n$  with  $\|x_0\| = 1$  (initialization)
2:  $\lambda^{(0)} = (x^{(0)})^H A x^{(0)}$  (Initialize Rayleigh quotient)
3: for  $k = 1, 2, \dots$  do
4:   solve  $(A - \lambda^{(k-1)}I)x^{(k)} = x^{(k-1)}$  (apply  $(A - \lambda^{(k-1)}I)^{-1}$  to  $x^{(k-1)}$ )
5:    $x^{(k)} = x^{(k)} / \|x^{(k)}\|$  (next normalized iterate)
6:    $\lambda^{(k)} = (x^{(k)})^H A x^{(k)}$  (Rayleigh quotient)
7:   Stop if convergence, set  $\lambda = \lambda^{(k)}$ ,  $q = x^{(k)}$ 
8: end for

```

On examine l'effet d'une itération générique k . Pour simplifier les notations, on utilise $x = x^{(k-1)} = x_1 q_1 + \dots + x_n q_n$ et $y = x^{(k)} = y_1 q_1 + \dots + y_n q_n$ pour les vecteurs (unitaires) $x^{(k-1)}$ et $x^{(k)}$ en début et fin d'itération, et note $\lambda = \lambda^{(k-1)}$, $\mu = \lambda^{(k)}$ les quotients de Rayleigh en début et fin d'itération. On considère la situation où les itérations convergent vers la valeur propre λ_m , supposée simple (multiplicité 1); le numéro m de cette limite (qui est donc fixé pour tout l'exercice) dépend du choix du vecteur initial $x^{(0)}$. On suppose que $x_m = 1 - \varepsilon$ ($0 \leq \varepsilon \ll 1$), ce qui exprime que x est déjà assez proche de q_m au début de l'itération. On cherche le comportement en fonction de ε des autres quantités intervenant dans l'itération, afin de déterminer la vitesse de la convergence de la suite $(x^{(k)}, \lambda^{(k)})_{k \geq 0}$ vers (q_m, λ_m) .

(a) Montrer que $x_i^2 = O(\varepsilon)$ pour tout $i \neq m$.

(b) Calculer λ (étape 2). Montrer que $\lambda - \lambda_m = O(\varepsilon)$, et que $\lambda - \lambda_i = O(1)$ pour tout $i \neq m$.

(c) Montrer que

$$\sum_{i=1}^n \frac{x_i^2}{(\lambda_i - \lambda)^2} = \frac{x_m^2}{(\lambda_m - \lambda)^2} [1 + O(\varepsilon^3)]$$

- (d) Exprimer y (étapes 4, 5) en fonction de x , λ et λ_i .
- (e) Calculer μ (étape 6), et déterminer l'ordre en ε de $\mu - \lambda_m$. Que conclure sur la vitesse de convergence de l'approximation de λ_m par les itérations de Rayleigh?
- (f) Donner le comportement asymptotique en ε de y_m . Que conclure sur la vitesse de convergence de l'approximation de q_m par les itérations de Rayleigh?

Solution:

- (a) Le caractère unitaire de x et l'hypothèse sur x_m donnent

$$1 = x_1^2 \dots + x_{m-1}^2 + (1 - \varepsilon)^2 + x_{m+1}^2 \dots + x_n^2 \implies 0 = x_1^2 \dots + x_{m-1}^2 + \varepsilon^2 - 2\varepsilon + x_{m+1}^2 \dots + x_n^2,$$

ce qui entraîne $x_i^2 = O(\varepsilon)$ pour tout $i \neq m$.

- (b) Le quotient de Rayleigh initial vérifie

$$\lambda - \lambda_m = \lambda_1 x_1^2 + \dots + \lambda_m (x_m^2 - 1) + \dots + \lambda_n x_n^2 = \lambda_m (\varepsilon^2 - 2\varepsilon)^2 + O(\varepsilon),$$

tandis que, pour $i \neq m$

$$\lambda - \lambda_i = \lambda_1 x_1^2 + \dots + \lambda_i (x_i^2 - 1) + \lambda_m x_m^2 + \dots + \lambda_n x_n^2 = \lambda_i + \lambda_m + O(\varepsilon) = O(1)$$

- (c) Compte tenu de ce qui précède, on a

$$\sum_{i=1}^n \frac{x_i^2}{(\lambda_i - \lambda)^2} = \frac{x_m^2}{(\lambda_m - \lambda)^2} \left(1 + \sum_{i \neq m} \frac{x_i^2}{(\lambda_i - \lambda)^2} \frac{(\lambda_m - \lambda)^2}{x_m^2} \right) = \frac{x_m^2}{(\lambda_m - \lambda)^2} [1 + O(\varepsilon^3)]$$

- (d) Avec les notations adoptées, on a $(A - \lambda I)y = x$, soit

$$(\lambda_i - \lambda)y_i = x_i \quad \text{et} \quad y_1^2 + \dots + y_n^2 = \sum_{i=1}^n \frac{x_i^2}{(\lambda_i - \lambda)^2} = \frac{x_m^2}{(\lambda_m - \lambda)^2} [1 + O(\varepsilon^3)]$$

- (e) Le quotient de Rayleigh donnant $\lambda^{(k)} =: \mu$ est

$$\mu = \frac{\lambda_1 y_1^2 + \dots + \lambda_n y_n^2}{y_1^2 + \dots + y_n^2} = \frac{\sum_{i=1}^n \frac{\lambda_i x_i^2}{(\lambda_i - \lambda)^2}}{\sum_{i=1}^n \frac{x_i^2}{(\lambda_i - \lambda)^2}},$$

et on obtient donc

$$\mu - \lambda_m = \frac{\sum_{i=1}^n \frac{\lambda_i - \lambda_m}{(\lambda_i - \lambda)^2} x_i^2}{\sum_{i=1}^n \frac{x_i^2}{(\lambda_i - \lambda)^2}} = \frac{\sum_{i=1}^n \frac{\lambda_i - \lambda_m}{(\lambda_i - \lambda)^2} x_i^2}{\frac{x_m^2}{(\lambda_m - \lambda)^2} [1 + O(\varepsilon^3)]} = \frac{\sum_{i \neq m} \frac{\lambda_i - \lambda_m}{(\lambda_i - \lambda)^2} x_i^2}{\frac{x_m^2}{(\lambda_m - \lambda)^2} [1 + O(\varepsilon^3)]}$$

Dans la formule ci-dessus, on a $x_i^2 = O(\varepsilon)$ ($i \neq m$) et $\lambda_m - \lambda = O(\varepsilon)$, les autres facteurs étant $O(1)$, ce qui permet de déduire

$$\mu - \lambda_m = \frac{O(\varepsilon)}{O(\varepsilon^{-2})[1 + O(\varepsilon^3)]} = O(\varepsilon^3),$$

ce qui montre la convergence cubique des itérés de quotient de Rayleigh vers λ_m .

- (f) Pour la composante y_m du vecteur y unitaire, on obtient

$$y_m = \frac{x_m}{\lambda_m - \lambda} \left(\sum_{i=1}^n \frac{x_i^2}{(\lambda_i - \lambda)^2} \right)^{-1/2} = \frac{x_m}{\lambda_m - \lambda} \left(\frac{x_m^2}{(\lambda_m - \lambda)^2} \right)^{-1/2} [1 + O(\varepsilon^3)]^{-1/2} \\ = \text{sign}(x_m) [1 + O(\varepsilon^3)],$$

ce qui montre de même la convergence cubique de l'approximation de q_m par les itérations de Rayleigh (puisque les autres composantes y_i sont, par voie de conséquence, $O(\varepsilon^3)$).

Exercice E24-4 (convergence de la méthode de plus grande pente) On a vu (cours) que, pour une matrice $A \in \mathbb{R}^{n \times n}$ symétrique définie positive, résoudre $Ax = b$ ($b \in \mathbb{R}^n$ donné) est équivalent à chercher x qui minimise $J(x) := \frac{1}{2}x^T Ax - x^T b$. Cet exercice porte sur la convergence de la méthode de la plus grande pente (steepest descent method), qui fut la première méthode proposée pour la minimisation de J . On note x^* la solution de $Ax = b$ (qui minimise donc aussi J) et, pour $x \in \mathbb{R}^n$ quelconque, on pose $d(x) = x^* - x$.

(a) Montrer que $J(x) = \frac{1}{2}\|d(x)\|_A^2$ (à une constante additive près, sans importance et mise à zéro dans la suite), où $\|z\|_A^2 := z^T Az$ définit la norme en énergie associée à A .

(b) L'itération k de plus grande pente partant de l'itéré actuel $x = x^{(k-1)}$ consiste à chercher l'itéré suivant $y = x^{(k)}$ de la forme $y = x + \alpha r$ avec $r := b - Ax$, $\alpha > 0$ puis à choisir α selon $\alpha^* = \arg \min_{\alpha} J(x + \alpha r)$; l'itéré suivant ainsi obtenu est donc donné par $y = x + \alpha^* r$.

Montrer que $r = Ad(x)$. Expliquer en quoi r est la direction de plus grande pente. Calculer $J(y)$ et montrer que $J(y) < J(x)$.

(c) A l'aide des étapes précédentes, exprimer $\|d(y)\|_A / \|d(x)\|_A$ en fonction de A et r .

(d) Toute matrice A symétrique définie positive vérifie l'inégalité (admise) de Kantorovich

$$\frac{(z^T Az)(z^T A^{-1}z)}{(z^T z)^2} \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1 \lambda_n} \quad \text{pour tout } z \in \mathbb{R}^n, z \neq 0,$$

où λ_1 et λ_n sont les valeurs propres extrêmes de A . En appliquant cette inégalité, montrer que

$$\frac{\|d(y)\|_A}{\|d(x)\|_A} \leq \frac{\lambda_1 - \lambda_n}{\lambda_1 + \lambda_n}$$

et exprimer le majorant en termes du conditionnement $\kappa_2(A)$ de A . Quelles sont les conséquences de ce résultat? Comparer en particulier avec la majoration correspondante vérifiée par l'algorithme du gradient conjugué appliqué au même problème.

Solution:

(a) Posant $x = x^* - d$, on trouve

$$J(x) = \frac{1}{2}(x^* - d)^T A(x^* - d) - (x^* - d)^T b = \frac{1}{2}\|d\|_A^2 + J(x^*) + d^T(b - Ax^*) = \frac{1}{2}\|d\|_A^2 + J(x^*),$$

ce qui prouve (a) avec la constante donnée par $J(x^*)$.

(b) On a facilement $r = b - Ax = (b - Ax^*) + Ad = Ad$. Le vecteur r est la direction de plus grande pente car $r = -\nabla J(x)$. On obtient ensuite

$$\begin{aligned} J(y) &= \frac{1}{2}(x + \alpha r)^T A(x + \alpha r) - (x + \alpha r)^T b = J(x) + \frac{1}{2}\alpha^2 r^T A r - \alpha r^T b + \alpha r^T A x \\ &= J(x) + \frac{1}{2}\alpha^2 r^T A r - \alpha r^T r. \end{aligned}$$

La valeur α^* de α produisant la valeur minimale de $J(y)$ est donnée par

$$\alpha r^T A r - r^T r = 0 \quad \implies \quad \alpha^* = \frac{r^T r}{r^T A r}$$

Pour cette valeur de α , on obtient

$$J(y) = J(x) - \frac{1}{2} \frac{(r^T r)^2}{r^T A r} < J(x)$$

(c) On trouve

$$d(y) = x^* - y = x^* - x - \alpha^* r = d(x) - \alpha^* r$$

et

$$\|d(y)\|_A^2 = \|d(x)\|_A^2 + (\alpha^*)^2 r^T A r - 2\alpha^* r^T A d(x) = \|d(x)\|_A^2 - \frac{(r^T r)^2}{r^T A r}$$

Par conséquent, on obtient (avec $\|d(x)\|_A^2 = d(x)^T A d(x) = r^T A^{-1} r$)

$$\frac{\|d(y)\|_A^2}{\|d(x)\|_A^2} = 1 - \frac{(r^T r)^2}{(r^T A r)(r^T A^{-1} r)}.$$

(d) L'inégalité de Kantorovich donne alors

$$\frac{\|d(y)\|_A^2}{\|d(x)\|_A^2} \leq 1 - \frac{4\lambda_1\lambda_n}{(\lambda_1 + \lambda_n)^2} = \frac{(\lambda_1 - \lambda_n)^2}{(\lambda_1 + \lambda_n)^2}.$$

Par convention, $\lambda_1 \geq \lambda_n$. Puisque $\kappa_2(A) = \lambda_1/\lambda_n \geq 1$ (conditionnement de A pour la norme spectrale), on trouve

$$\frac{\|d(y)\|_A}{\|d(x)\|_A} \leq \frac{\kappa_2(A) - 1}{\kappa_2(A) + 1},$$

ce qui permet d'estimer le taux de convergence de la méthode de plus grande pente vers la solution x^* (pour laquelle $d(x^*) = 0$), et en particulier de comparer ce taux à celui de la méthode du gradient conjugué, pour laquelle

$$\frac{\|d(y)\|_A}{\|d(x)\|_A} \leq \frac{\sqrt{\kappa_2(A)} - 1}{\sqrt{\kappa_2(A)} + 1},$$