UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Section de mathématiques

Analyse Numérique
Corrigé Série 15

5 mars 2020

1. ($\star$, tout l'exercice)(**Régularisation de Tikhonov**) Considérons le problème de trouver $\boldsymbol{x}^* \in \mathbb{R}^n$ tel que
$$\|A\boldsymbol{x}^* - \boldsymbol{b}\|_2^2 + \mu\|\boldsymbol{x}^*\|_2^2 = \min_{x\in\mathbb{R}^n} \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \mu\|\boldsymbol{x}\|_2^2,$$
avec $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \geq n$, et $\mu \geq 0$. Ce problème est appelé *régularisation de Tikhonov.*

(a) **(0.25 points)** Réécrire ce problème comme la minimisation d'une fonction $f(x) = \|\widetilde{A}\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|_2^2$ où la matrice $\widetilde{A}$ et le vecteur $\widetilde{\boldsymbol{b}}$ sont à préciser. Donner aussi la taille de $\widetilde{A}$ et $\widetilde{\boldsymbol{b}}$.

Sol.: *On pose* $\widetilde{A} = \begin{pmatrix} A \\ \sqrt{\mu}I_n \end{pmatrix} \in \mathbb{R}^{(m+n)\times n}$ *et* $\widetilde{\boldsymbol{b}} = \begin{pmatrix} \boldsymbol{b} \\ 0_n \end{pmatrix} \in \mathbb{R}^{m+n}$, *alors*
$$f(x) = \|\widetilde{A}\boldsymbol{x} - \widetilde{\boldsymbol{b}}\|_2^2 = \|A\boldsymbol{x} - \boldsymbol{b}\|_2^2 + \mu\|\boldsymbol{x}\|_2^2.$$

(b) **(0.25 points)** Écrire les équations normales associées uniquement en fonction de A, $\boldsymbol{b}$ et μ.

Sol.: *Les équations normales sont* $\widetilde{A}^\top\widetilde{A}\boldsymbol{x}^* = \widetilde{A}^\top\widetilde{\boldsymbol{b}}$. *Or* $\widetilde{A}^\top\widetilde{A} = A^\top A + \mu I_n$ *et* $\widetilde{A}^\top\widetilde{\boldsymbol{b}} = A^\top\boldsymbol{b}$, *donc on peut réécrire les équations normales comme*
$$(A^\top A + \mu I_n)\boldsymbol{x}^* = A^\top\boldsymbol{b}.$$

(c) **(0.25 points)** Pour quels μ les équations normales ont-elles une unique solution $\boldsymbol{x}^*$? Donner dans ce cas une formule explicite pour $\boldsymbol{x}^*$.

Sol.: *Comme la matrice* $A^\top A$ *est symétrique semi-définie positive, la matrice* $B = A^\top A + \mu I_n$ *est elle aussi symétrique semi-définie positive. De plus,* $\mathrm{Sp}(B) = \mathrm{Sp}(A^\top A) + \mu$, *donc si* $\mu > 0$, B *est définie positive, et donc inversible. Les équations normales ont donc une unique solution dès que* $\mu > 0$ *et elle est donnée par*
$$\boldsymbol{x}^* = (A^\top A + \mu I_n)^{-1}A^\top\boldsymbol{b}.$$

(d) **(0.25 points)** En fonction du rang de A, quelle peut être l'utilité de considérer la régularisation de Tikhonov plutôt que le problème habituel ($\mu = 0$) ?

Sol.: *Si* $\mathrm{rang}(A) = n$, *la méthode des moindres carrés peut être appliquée comme vue en cours. Par contre, si* $\mathrm{rang}(A) < n$, $A^\top A$ *n'est pas définie positive et on perd l'unicité dans les équations normales. En modifiant le problème avec* $\mu > 0$, *on obtient une nouvelle matrice* $\widetilde{A}$ *proche de* A *si* μ *est petit, et de rang maximal. Ainsi la régularisation de Tikhonov permet d'appliquer la méthode des moindres carrés à un problème proche dans le cas où* A *n'est pas de rang maximal.*

2. (**Factorisation QR avec Gram–Schmidt classique et modifiée**) En cours (Section 6.2 dans le polycopié), on a vu les méthodes de Gram–Schmidt classique (GSC) et modifiée (GSM) pour calculer la factorisation QR d'une matrice $A \in \mathbb{R}^{m\times n}$. Par commodité, on redonne ici les deux algorithmes.

Algorithme 1 (Factorisation QR – Gram–Schmidt classique (pas stable)).
Soit $A \in \mathbb{R}^{m\times n}$ où $m \geq n$ et $\mathrm{rang}(A) = n$.

for $j = 1, \dots, n$ **do**
  $\tilde{\boldsymbol{q}}_j \leftarrow \boldsymbol{a}_j$
  for $i = 1, \dots, j-1$ **do**
    $r_{ij} \leftarrow \boldsymbol{a}_j^\top \boldsymbol{q}_i$
    $\tilde{\boldsymbol{q}}_j \leftarrow \tilde{\boldsymbol{q}}_j - r_{ij}\boldsymbol{q}_i$
  end for
  $r_{jj} \leftarrow \|\tilde{\boldsymbol{q}}_j\|_2$
  $\boldsymbol{q}_j \leftarrow \tilde{\boldsymbol{q}}_j / r_{jj}$
end for

Algorithme 2 (Factorisation QR – Gram–Schmidt modifiée (stable)).
Soit $A \in \mathbb{R}^{m\times n}$ où $m \geq n$ et $\operatorname{rang}(A) = n$.

for $i = 1, \ldots, n$ **do**
  $\tilde{\boldsymbol{q}}_i \leftarrow \boldsymbol{a}_i$
end for
for $i = 1, \ldots, n$ **do**
  $r_{ii} \leftarrow \|\tilde{\boldsymbol{q}}_i\|_2$
  $\boldsymbol{q}_i \leftarrow \tilde{\boldsymbol{q}}_i / r_{ii}$
  for $j = i+1, \ldots, n$ **do**
    $r_{ij} \leftarrow \tilde{\boldsymbol{q}}_j^\top \boldsymbol{q}_i$
    $\tilde{\boldsymbol{q}}_j \leftarrow \tilde{\boldsymbol{q}}_j - r_{ij}\boldsymbol{q}_i$
  end for
end for

(a) Considérer la matrice

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

où ϵ est un petit nombre tel que le résultat de l'opération $1 + \epsilon^2$ en précision finie est toujours arrondi à 1, c.-à-d., $\mathrm{fl}(1+\epsilon^2) = 1$. En supposant l'absence d'autres erreurs d'arrondi, calculer les matrices Q produites par les méthodes GSC et GSM, et évaluer leur orthonormalité en norme $\|\cdot\|_2$, i.e. estimer $\|Q^\top Q - I\|_2$.

Indication : *Il n'est pas nécessaire de faire les calculs à la main. Utiliser un outil numérique (par exemple, MATLAB, Maple, Mathematica, Wolfram Alpha, ...).*

Sol.: *En appliquant la méthode classique (Algorithme 1) on obtient :*

$$Q_\mathrm{C} = \begin{bmatrix} 1 & 0 & 0 \\ \epsilon & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ 0 & \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix}, \qquad Q_\mathrm{C}^\top Q_\mathrm{C} = \begin{bmatrix} 1 & -\frac{\epsilon}{\sqrt{2}} & -\frac{\epsilon}{\sqrt{2}} \\ -\frac{\epsilon}{\sqrt{2}} & 1 & \frac{1}{2} \\ -\frac{\epsilon}{\sqrt{2}} & \frac{1}{2} & 1 \end{bmatrix}, \qquad \|Q_\mathrm{C}^\top Q_\mathrm{C} - I\|_2 \approx \frac{1}{2}.$$

La méthode modifiée (Algorithme 2) nous donne :

$$Q_\mathrm{M} = \begin{bmatrix} 1 & 0 & 0 \\ \epsilon & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{6}} \\ 0 & 0 & \sqrt{\frac{2}{3}} \end{bmatrix}, \qquad Q_\mathrm{M}^\top Q_\mathrm{M} = \begin{bmatrix} 1 & -\frac{\epsilon}{\sqrt{2}} & -\frac{\epsilon}{\sqrt{6}} \\ -\frac{\epsilon}{\sqrt{2}} & 1 & 0 \\ -\frac{\epsilon}{\sqrt{6}} & 0 & 1 \end{bmatrix}, \qquad \|Q_\mathrm{M}^\top Q_\mathrm{M} - I\|_2 \approx \sqrt{\frac{2}{3}}\,\epsilon.$$

Il est évident que l'algorithme GSC n'est pas stable : numériquement, il produit des vecteurs $\boldsymbol{q}_j$ qui sont très loin d'être orthogonaux, car $\|Q_\mathrm{C}^\top Q_\mathrm{C} - I\|_2 \approx \frac{1}{2} \gg \epsilon$. En revanche, l'algorithme GSM donne des vecteurs $\boldsymbol{q}_j$ qui sont beaucoup plus précis, c.-à-d., orthogonaux. C'est l'algorithme que l'on utilise en pratique.

(b) La méthode GSC et la méthode GSM appliquées à la matrice $A \in \mathbb{R}^{m\times n}$ $(m \geq n)$ calculent une factorisation $A = QR$, $Q \in \mathbb{R}^{m\times n}$.

i. En usant de projections orthogonales P_j bien choisies, réécrire les deux méthodes avec ces projections. Puis, montrer que la méthode GSC (Algorithme 1) correspond aux opérations

$$\tilde{\boldsymbol{q}}_j \leftarrow (I - P_{j-1} - P_{j-2} - \cdots - P_1)\,\boldsymbol{a}_j, \qquad j = 1, \ldots, n,$$

et que la méthode GSM (Algorithme 2) correspond à

$$\tilde{\boldsymbol{q}}_j \leftarrow (I - P_{j-1})\,(I - P_{j-2}) \cdots (I - P_1)\,\boldsymbol{a}_j, \qquad j = 1, \ldots, n,$$

Rappel : *La matrice* $P = \boldsymbol{v}\boldsymbol{v}^\top \in \mathbb{R}^{n\times n}$*, avec* $\boldsymbol{v} \in \mathbb{R}^n$*,* $\|\boldsymbol{v}\| = 1$*, représente une projection orthogonale sur* $\mathrm{Vect}(\boldsymbol{v})$*, c.-à-d., le sous-espace vectoriel engendré par* $\boldsymbol{v}$*.*

Sol.: *On définit la projection orthogonale* $P_i = \boldsymbol{q}_i\boldsymbol{q}_i^\top$*, où* $\boldsymbol{q}_i$ *dénote la i-ème colonne de* Q*. Par la méthode GSC (Algorithme 1), on a*

$$\tilde{\boldsymbol{q}}_j = \boldsymbol{a}_j - \sum_{i=1}^{j-1} (\boldsymbol{a}_j^\top \boldsymbol{q}_i)\,\boldsymbol{q}_i = \boldsymbol{a}_j - \sum_{i=1}^{j-1} (\boldsymbol{q}_i \boldsymbol{q}_i^\top)\,\boldsymbol{a}_j = \left(I - \sum_{i=1}^{j-1} P_i\right) \boldsymbol{a}_j, \qquad j = 1, \ldots, n.$$

ii. En utilisant la formulation avec les projections orthogonales, montrer que, en arithmétique exacte, les méthode GSC et GSM sont équivalentes.

Sol.: *Il faut montrer que*

$$(I - P_{j-1})\,(I - P_{j-2}) \cdots (I - P_1) = I - P_{j-1} - P_{j-2} - \cdots - P_1.$$

En usant du fait que $P_iP_j = 0$ *si* $i \neq j$*, on montre aisément le résultat par récurrence.*

3. ($\star$, questions a-b-c)(**Les transformations de Givens pour la décomposition QR**)

(a) **(0.25 points)** Étant donné un vecteur $\begin{pmatrix} a \\ b \end{pmatrix}$, trouver toutes les matrices de la forme

$$A(\theta) = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix},$$

telles que le vecteur $y = A(\theta)x$ vérifie $y_2 = 0$ (seconde composante nulle). Interpréter graphiquement le résultat dans le cas où $a = 0$ et dans le cas général.

Sol.: *Supposons le vecteur* $\begin{pmatrix} a \\ b \end{pmatrix}$ *non nul. Un calcul donne*

$$y = A(\theta)x = \begin{pmatrix} a\cos\theta - b\sin\theta \\ a\sin\theta + b\cos\theta \end{pmatrix}$$

Ainsi, $y_2 = a\sin\theta + b\cos\theta$ *et on a* $y_2 = 0$ *si* $\theta = -\arctan(b/a) + k\pi$ *avec* $k \in \mathbb{N}$ *et la convention* $\theta = \pi/2 + k\pi$ *si* $a = 0$*. Ceci donne*

$$A(\theta) = \pm\frac{1}{\sqrt{a^2 + b^2}} \begin{pmatrix} a & b \\ -b & a \end{pmatrix}.$$

La matrice $A(\theta)$ *ci-dessus est une rotation d'angle* θ *dans le plan. L'angle* θ *ainsi défini correspond ici à tous les angles orientés entre* x *et le vecteur* $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$*. Si* $a = 0$*, on a* $\theta = \pm\frac{\pi}{2}$ *(à* $2k\pi$ *près). C'est logique, car on doit faire tourner de* $\pm\frac{\pi}{2}$ *un vecteur sur l'axe des ordonnées pour l'amener sur l'axe des abscisses.*

(b) **(0.25 points)** Répétez la question a) en trouvant les transformations A telles que le vecteur $y = A(\theta)x$ vérifie $y_1 = y_2$. Interpréter graphiquement le résultat dans le cas où $a = 0$ et dans le cas général.

Sol.: *Les mêmes calculs que dans la question précédente donnent la condition* $(a - b)\cos\theta = (a + b)\sin\theta$*, c'est à dire* $\theta = \arctan((a - b)/(a + b)) + k\pi$ *avec la convention* $\theta = \pi/2$ *si* $a + b = 0$*. La matrice* $A(\theta)$ *est donc*

$$A(\theta) = \pm\frac{1}{\sqrt{2}\sqrt{a^2 + b^2}} \begin{pmatrix} a + b & b - a \\ a - b & a + b \end{pmatrix}.$$

On a encore une rotation, mais au lieu d'envoyer le vecteur sur l'axe des abscisses, on l'envoie sur la droite $y = x$*. Si* $a = 0$*, on a* $\theta = -\frac{\pi}{4} + k\pi$*, c'est normal car l'axe des ordonnées fait un angle de* $-\frac{\pi}{4}$ *avec la droite* $y = x$*.*

(c) **(0.5 points)** Étant donné un vecteur $x \in \mathbb{R}^n$, trouver une matrice orthogonale Q telle qu'il existe $\alpha \in \mathbb{R}$ avec

$$Qx = \alpha e_1 = \begin{pmatrix} \alpha \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Indication : considérer des matrices de la forme

$$A_{n_1,n_2,n_3}(a,b) = \begin{pmatrix} I_{n_1} & 0 & 0 & 0 & 0 \\ 0 & \frac{a}{\sqrt{a^2+b^2}} & 0 & \frac{b}{\sqrt{a^2+b^2}} & 0 \\ 0 & 0 & I_{n_2} & 0 & 0 \\ 0 & -\frac{b}{\sqrt{a^2+b^2}} & 0 & \frac{a}{\sqrt{a^2+b^2}} & 0 \\ 0 & 0 & 0 & 0 & I_{n_3} \end{pmatrix} \tag{1}$$

où $I_{n_1}, I_{n_2}, I_{n_3}$ désignent les matrices identités de taille $n_1 \times n_1, n_2 \times n_2, n_3 \times n_3$, respectivement.

Sol.: *On considère d'abord la matrice $A_{0,0,n-2}(x_1, x_2)$ qui d'après la question précédente permet d'annuler la seconde composante de x, les composantes 3 à n restant inchangées :*

$$x^{(1)} = A_{0,0,n-2}(x_1, x_2)x = \begin{pmatrix} \alpha_2 \\ 0 \\ x_3 \\ \vdots \\ x_n \end{pmatrix}.$$

On multiplie ensuite successivement par les matrices $A_{0,i,n-i-2}(\alpha_i, x_{i+2}), i = 1, \ldots, n-2$, choisies de manière analogue pour annuler la composante en ligne $i+2$ (les composantes des lignes autres que 1 et $i+2$ restant inchangées). Au final, on obtient, avec $\alpha_i = \sqrt{x_1^2 + \ldots + x_i^2}$,

$$\underbrace{A_{0,n-2,0}(\alpha_n, x_n) \cdots A_{0,1,n-3}(\alpha_2, x_3) A_{0,0,n-2}(x_1, x_2)}_{Q} x = e_1.$$

La matrice Q ainsi obtenue est une matrice orthogonale comme produit de matrices orthogonales.

(d) En déduire un algorithme de de décomposition QR utilisant les transformations précédentes (transformations de Givens), au lieu des réflexions de Householder qui seront vues en cours. Combien de transformations de la forme (1) l'algorithme QR obtenu nécessite t-il ?

Sol.: *On reprend le même algorithme, en remplaçant les réflexions de Householder par la transformations ainsi définies à la question précédente. Il y a autant de matrices de forme (1) que de zéros à faire apparaître sous la diagonale dans l'algorithme QR, c'est-à-dire $n(n-1)/2$.*