Analyse matricielle

Lionel Magnis

1 Introduction et pré-requis

1.1 Notations

— $\mathbf{K}$ désigne le corps $\mathbf{R}$ ou $\mathbf{C}$. $\langle \cdot, \cdot \rangle$ est le produit hermitien (ou scalaire) usuel sur $\mathbf{K}^n$.

— Si A est une matrice à coefficients dans $\mathbf{K}$, A^* est sa conjuguée, (ou sa transposée $A^\top$ si $\mathbf{K} = \mathbf{R}$). I désigne la matrice identité, sa taille est donnée par le contexte.

— $TS_n(\mathbf{K})$ (respectivement $TI_n(\mathbf{K})$) est l'ensemble des matrices triangulaires supérieures (respectivement inférieures) de taille n.

— $TS_n^{++}(\mathbf{K})$ (respectivement $TI_n^{++}(\mathbf{K})$) désignent les matrices triangulaires dont les éléments diagonaux sont des réels strictement positifs. $TI_n^1(\mathbf{K})$ désigne les matrices triangulaires inférieures dont les éléments diagonaux valent tous 1.

— $U_n(\mathbf{K})$ désigne les matrices unitaires (ou orthogonales si $\mathbf{K} = \mathbf{R}$) de taille n, caractérisées par $O^*O = I$.

— $H_n^{++}(\mathbf{K})$ désigne l'ensemble des matrices hermitiennes (i.e. $A^* = A$) définies positives. Si $\mathbf{K} = \mathbf{R}$, ce sont les matrices symétriques définies positives, on le note aussi $S_n^{++}(\mathbf{R})$.

— $|x|_p$ désigne la norme p de $x \in \mathbf{K}^n$

— $\triangleq$ indique un objet défini par une égalité.

Proposition 1. *$TS_n^{++}(\mathbf{K})$, $TI_n^{++}(\mathbf{K})$ et $TI_n^1(\mathbf{K})$ sont des groupes pour la multiplication.*

1.2 Complexité

La complexité d'un calcul est le nombre d'opérations élémentaires qu'il met en œuvre. Pour simplifier, on ne considère que les multiplications et les divisions, qui sont plus longues que les additions.

Exemple 1. (fondamental) Un système linéaire triangulaire inversible

$$\begin{array}{ccccc} a_{1,1}x_1 & + \ldots + & a_{1,n}x_n & = & b_1 \\ & \ddots & \vdots & & \vdots \\ & & a_{n,n}x_n & = & b_n \end{array}$$

est résolu directement en

$$x_n = \frac{1}{a_{n,n}} b_n$$
$$x_{n-i} = \frac{1}{a_{n-i,n-i}} (b_i - \sum_{j=n-i+1}^{n} a_{i,j} x_j), \quad 1 \leq i \leq n-1$$

qui demande $\frac{n(n-1)}{2}$ additions, autant de multiplications et n divisions. Sa complexité est donc $n^2/2 + n/2$. En général, on considère que n est grand et on ne gardera que le terme dominant du développement asymptotique.

La résolution d'un système triangulaire est ainsi en $n^2/2$.

Exemple 2.

1. La multiplication d'un vecteur $x \in \mathbf{K}^n$ par une matrice $A \in M_n(\mathbf{K})$ est en n^2.
2. La multiplication (naïve) de deux matrices de $M_n(\mathbf{K})$ est en n^3.
3. Le produit scalaire de deux vecteurs de $\mathbf{K}^n$ est en ...
4. La projection d'un vecteur de $\mathbf{K}^n$ sur une droite est en ...
5. Une multiplication dans $\mathbf{C}$ vaut ... multiplications dans $\mathbf{R}$.

1.3 Pré-requis

A désigne une matrice de $M_n(\mathbf{K})$.

Définition 1. Le rayon spectral de A est défini par $\rho(A) \triangleq \max\{|\lambda|,\ \lambda \in \mathrm{sp}_{\mathbf{C}} A\}$.

Exemple 3.

$$\rho\left(\begin{pmatrix} 1 & i \\ 0 & 2 \end{pmatrix}\right) = 2, \quad \rho\left(\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}\right) = \dots$$

Définition 2. Considérons une norme $|\cdot|$ sur $\mathbf{K}^n$. La norme induite par $|\cdot|$ sur $M_n(\mathbf{K})$ est

$$\|A\| \triangleq \max_{|x|=1} |Ax| = \max_{x \neq 0} \frac{|Ax|}{|x|}$$

Elle vérifie les propriétés élémentaires suivantes.

1. $\forall x \in \mathbf{K}^n, \quad |Ax| \leq \|A\|\,|x|$
2. $\|I\| = 1$
3. $\forall A, B \in M_n(\mathbf{K}), \quad \|AB\| \leq \|A\|\,\|B\|$

Ce dernier point traduit le fait que $\|\cdot\|$ est une norme *matricielle*. Mais toutes les normes matricielles ne sont pas des normes induites.

Exercice 1. On définit la norme de Frobenius sur $M_n(\mathbf{K})$ par

$$\|A\|_F \triangleq \sqrt{\mathrm{Tr}\,(A^*A)} = \sqrt{\sum_{i,j} |a_{i,j}|^2}$$

Montrer que $\|\cdot\|_F$ est une norme matricielle mais n'est pas une norme induite.

Exemple 4. (*Quelques normes induites à connaître*) On note $\|\cdot\|_p$ la norme induite par la norme p. On a

$$\|A\|_1 = \max_j \sum_i |a_{i,j}|, \quad \|A\|_\infty = \max_i \sum_j |a_{i,j}| = \|A^*\|_1, \quad \|A\|_2 = \sqrt{\rho(A^*A)} = \|A^*\|_2$$

Exercice 2. (*Lien entre normes induites et rayon spectral*)

1. Montrer que $\rho(A^m) = \rho(A)^m$ et que $\rho(\alpha A) = |\alpha|\rho(A)$.
2. On suppose A diagonale. Montrer que, pour $1 \leq p \leq \infty$, $\|A\|_p = \rho(A)$.
3. On suppose que A est normale. Montrer que $\rho(A) = \|A\|_2$.
4. Montrer que ρ n'est pas une norme sur $M_n(\mathbf{K})$.
5. Montrer que pour toute norme induite $\|\cdot\|$ sur $M_n(\mathbf{C})$
$$\rho(A) \leq \|A\| \tag{1}$$
6. (*facultatif*) Montre que l'inéquation (1) est encore vraie pour une norme matricielle sur $M_n(\mathbf{C})$, ou même sur $M_n(\mathbf{R})$.

L'inéquation (1) montre que $\rho(A)$ minore l'ensemble $\{\|A\|,\ \|\cdot\| \text{ norme induite}\}$. Le théorème suivant montre que c'en est en fait la borne inférieure.

Théorème 1 (Householder)**.** *Pour toute matrice $A \in M_n(\mathbf{C})$ et tout $\varepsilon > 0$, il existe une norme induite $\|\cdot\|$ telle que*

$$\|A\| \leq \rho(A) + \varepsilon$$

Exercice 3.

1. Montrer qu'il existe une matrice D diagonale et $n-1$ matrices $M_1, \dots M_{n-1}$ telles que pour tout $\mu > 0$, il existe $Q \in GL_n(\mathbf{C})$ telle que
$$Q^{-1}AQ = D + \sum_{i=1}^{n} \mu^i M_i$$
2. Démontrer le Théorème 1.

Application 1. Montrer que la suite (A^m) converge vers 0 si et seulement si $\rho(A) < 1$, et que cette condition suffit pour que $I - A$ soit inversible.

Application 2. Montrer que pour toute norme induite $\|\cdot\|$

$$\rho(A) = \lim_{m\to\infty} \|A^m\|^{\frac{1}{m}}$$

On pourra considérer la matrice $\frac{A}{\rho(A)+\delta}$, pour $\delta > 0$.

2 Résolution de systèmes linéaires

On s'intéresse à la résolution d'un système

$$Ax = b \tag{2}$$

d'inconnue $x \in \mathbf{K}^n$ avec $A \in Gl_n(\mathbf{K})$ et $b \in \mathbf{K}^n$. Ce système admet une unique solution que l'on note $x^\#$. On a

$$x^\# = A^{-1}b$$

mais en pratique, on ne calcule jamais A^{-1}. On n'utilise pas non plus les formules de Cramer bien trop gourmandes en temps de calcul. On va plutôt se ramener à la résolution de systèmes triangulaires pour trouver la valeur exacte de $x^\#$ (méthodes directes, Section 2.2) ou bien une valeur approchée (méthodes itératives, Section 2.3). La résolution de systèmes linéaires apparait naturellement dans divers domaines de mathématiques appliquées. On en donne quelques exemples ci-dessous.

2.1 Exemples

2.1.1 Minimisation d'une fonction quadratique convexe

Soit $A \in S_n^{++}(\mathbf{R})$, $b \in \mathbf{R}^n$ et $c \in \mathbf{R}$. On considère la fonction f définie sur $\mathbf{R}^n$ par

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle + c$$

1. Montrer que f est convexe et que son gradient est

$$\nabla f(x) = Ax - b$$

2. Justifier que x minimise f si et seulement si $Ax = b$.

2.1.2 Moindres carrés [AK, Chapitre 7]

Soit $A \in M_{n,m}(\mathbf{R})$ et $b \in \mathbf{R}^n$. Considérons le système linéaire

$$Ax = b$$

d'inconnue $x \in \mathbf{R}^m$. Ce système n'a pas nécessairement de solution. On s'intéresse alors au problème suivant dit *aux moindres carrés*

$$(\mathcal{P}) : \text{Trouver } x_0 \in \mathbf{R}^m \text{ minimisant la fonction } x \mapsto |Ax - b|_2$$

1. En remarquant qu'on cherche en fait la projection orthogonale de b sur $\mathrm{Im}A$, montrer que $(\mathcal{P})$ admet au moins une solution. Montrer aussi que les solutions x de $(\mathcal{P})$ sont caractérisées par *l'équation normale associée*

$$A^\top Ax = A^\top b \tag{3}$$

2. Donner une CNS portant sur A pour que $(\mathcal{P})$ admette une unique solution.
3. $(\mathcal{P})$ équivaut à trouver x_0 minimisant la fonction $f : x \mapsto \frac{1}{2}|Ax-b|_2^2$. Montrer que

$$f(x) = \frac{1}{2}\left\langle A^\top Ax, x\right\rangle - \left\langle A^\top b, x\right\rangle + \frac{1}{2}\langle b, b\rangle$$

et retrouver ainsi l'équation normale.

Remarque 1. On a ainsi montré que pour $A \in M_{n,m}(\mathbf{R})$ quelconque, le système linéaire

$$A^\top Ax = A^\top b$$

admet toujours au moins une solution, résultat non-trivial a priori.

2.1.3 Solution approchée d'une équation différentielle

On considère l'équation du Laplacien en dimension 1 [Cia]

$$\begin{aligned} &u''(x) = -f(x), \quad 0 < x < 1 \\ &u(0) = 0, \quad u(1) = 0 \end{aligned}$$

qui modélise par exemple la position d'une corde dans un plan vertical (x, u), fixée en ses extrémités $x = 0, x = 1$ (on parle de *problème aux limites*) et soumise à une force verticale de densité continue f. C'est aussi une version stationnaire de l'équation de la chaleur avec terme source.

Ce problème admet une unique solution u, ce que l'on ne démontre pas ici, et que l'on suppose de classe $\mathcal{C}^4$. On ne cherche pas à la calculer explicitement mais on va en donner une expression approchée par la *méthode des différences finies*. On fixe $n \geq 1$ et on note

$$x_i \triangleq \frac{i}{n}, \quad 0 \leq i \leq n, \quad U \triangleq \begin{pmatrix} u(x_1) \\ \vdots \\ u(x_{n-1}) \end{pmatrix} \in \mathbf{R}^{n-1}, \quad b \triangleq \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_{n-1}) \end{pmatrix} \in \mathbf{R}^{n-1}$$

1. On note A la matrice symétrique

$$A \triangleq n^2 \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in M_{n-1}(\mathbf{R}) \tag{4}$$

Montrer que

$$AU = b + \delta b$$

avec $\delta b = \mathcal{O}(\frac{1}{n^2})$.

2. Montrer que les valeurs propres de A sont les

$$\lambda_k = 4n^2 \sin^2 \frac{\alpha_k}{2}, \quad \alpha_k = \frac{k\pi}{n}, \quad 1 \leq k \leq n-1$$

avec pour vecteur propre associé $v_k = (\sin j\alpha_k)_{1 \leq j \leq n-1}$. Pouvait-on voir autrement que $A \in S_{n-1}^{++}(\mathbf{R})$?

3. On note U_{approx} la solution du système linéaire $AX = b$. Montrer que

$$|U - U_{\text{approx}}|_2 = \mathcal{O}(\frac{1}{n^2})$$

On construit ainsi, en résolvant un système linéaire, une approximation de la solution u aux points x_i, avec une erreur d'approximation en $1/n^2$.

2.1.4 Solution approchée d'une équation aux dérivées partielles

On considère l'équation de la chaleur normalisée en dimension 1 [Cia]

$$\begin{aligned} &\frac{\partial u}{\partial t}(x,t) - \frac{\partial^2 u}{\partial x^2}(x,t) = f(x,t), \quad 0 < x < 1, \quad t > 0 \\ &u(0,t) = 0, \quad u(1,t) = 0 \\ &u(x,0) = u_0(x), \quad 0 \leq x \leq 1 \end{aligned}$$

qui modélise l'évolution dans le temps de la température $u(x,t)$ le long d'une barre maintenue à température 0 en ses extrémités $x = 0, x = 1$. f représente la densité de source de chaleur. u_0 représente la température initiale le long de la barre, supposée connue. On admet qu'il existe une unique solution est qu'elle est de classe $\mathcal{C}^4$. On ne sait en général pas la calculer. On fixe encore $n \geq 1$ et un pas de temps $\Delta_t > 0$. On définit les x_i comme à la section précédente et, pour tout entier $j \geq 0$

$$t_j \triangleq j\Delta_t, \quad U^j \triangleq \begin{pmatrix} u(x_1, t_j) \\ \vdots \\ u(x_{n-1}, t_j) \end{pmatrix} \in \mathbf{R}^{n-1}, \quad b^j \triangleq \begin{pmatrix} f(x_1, t_j) \\ \vdots \\ f(x_{n-1}, t_j) \end{pmatrix} \in \mathbf{R}^{n-1}$$

Pour bien distinguer le pas de temps et le pas d'espace, on note

$$\Delta_x \triangleq \frac{1}{n}$$

1. Montrer que

$$\frac{1}{\Delta_t} U^{j+1} = (\frac{1}{\Delta_t} I - A) U^j + b^j + \mathcal{O}(\Delta_t) + \mathcal{O}(\Delta_x^2)$$

où A est la matrice définie en (4). Ce schéma de discrétisation est dit *explicite*.

2. Montrer qu'on a aussi

$$(\frac{1}{\Delta_t}I + A)U^{j+1} = \frac{1}{\Delta_t}U^j + b^{j+1} + \mathcal{O}(\Delta_t) + \mathcal{O}(\Delta_x^2)$$

Ce schéma de discrétisation est dit *implicite*.

Dans les deux cas, on va calculer des approximations U^j_{approx} de U^j en négligeant les termes en $\mathcal{O}$, et en partant de $U^0 = (u_0(x_i))_{1\leq i\leq n-1}$ connu. Le but étant d'avoir

$$\lim_{\Delta_x,\Delta_t\to 0} |U^j - U^j_{\text{approx}}| = 0 \tag{5}$$

Le schéma implicite demande plus de calculs puisqu'il implique de résoudre un système linéaire

$$(\frac{1}{\Delta_t}I + A)X = \frac{1}{\Delta_t}U^j_{\text{approx}} + b^{j+1}$$

dont le second membre varie à chaque étape j. Néanmoins, il a de meilleures propriétés de *stabilité*. Sans rentrer dans le détail, (5) est vérifiée inconditionnellement pour le schéma implicite. Pour le schéma explicite, elle est vérifiée si Δ_x, Δ_t tendent vers 0 en respectant les conditions dites *CFL*

$$2\Delta_t \leq \Delta_x^2$$

Une étude avancée de la stabilité des schémas numériques pour l'équation de la chaleur et pour bien d'autres EDP est présentée dans [All, Chapitre 2].

2.1.5 Autres exemples

[Cia, Chapitre 3] présente d'autres exemples en particulier

— L'interpolation d'une fonction sur un segment par des splines cubiques.

— La méthode des éléments finis pour le problème du Laplacien en dimension 2.

2.2 Méthodes directes

2.2.1 Méthode d'élimination de Gauss

Le principe est de transformer (2) en un système triangulaire équivalent. La méthode s'appuie sur le résultat suivant.

Proposition 2. *Il existe $G \in GL_n(\mathbf{K})$ telle que $U \triangleq GA$ soit triangulaire supérieure.*

Il ne reste alors plus qu'à résoudre le système triangulaire $Ux = Gb$

Algorithmique 1. On peut construire G comme le produit d'au plus $n-1$ matrices triangulaires inférieures $T_1, \dots T_{n-1}$ et éventuellement de matrices de transposition s'il y a besoin de changer de *pivot*. En pratique, le terme Gb est mis à jour à chaque étape. On ne calcule pas explicitement G.

Exemple 5. Appliquer la méthode d'élimination de Gauss à

$$A = \begin{pmatrix} 0 & 2 & 3 \\ 1 & 4 & 1 \\ 2 & 1 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 1 \\ 8 \\ 5 \end{pmatrix}$$

Noter à chaque étape k : le $k-$ième élément diagonal q_k, la matrice de transposition P_k, le pivot p_k (si $q_k \neq 0$, $P_k = I$ et $p_k = q_k$), la matrice T_k.

Algorithmique 2. On retiendra que la méthode d'élimination de Gauss est en $n^3/3$.

Remarque 2. Les éléments diagonaux $u_{k,k}$ de U sont les pivots p_k.

Remarque 3. Si A est n'est pas inversible, la méthode d'élimination permet de rendre A triangulaire, mais U n'est pas inversible. Aussi, $Ux = Gb$ n'a pas nécessairement de solution.

Exercice 4. Montrer que

$$\det A = (-1)^p \prod_{k=1}^{n} u_{k,k}$$

où p est le nombre de matrices de transposition utilisées. La méthode d'élimination de Gauss permet ainsi un calcul de déterminant en $n^3/3$.

Algorithmique 3. Pour des raisons de stabilité numérique, il peut être utile de changer de pivot lorsque q_k est non nul mais petit. On peut prendre par exemple le plus grand (en module) dans la sous-colonne $[k : n, k]$ (*pivot partiel*) ou bien encore le plus grand dans la sous-matrice $[k : n, k : n]$ (*pivot total*). Un exemple concret est donné dans [Cia, page 78].

Remarque 4. La méthode de Gauss permet aussi d'échelonner des systèmes linéaires non-carrés. On trouve une bonne présentation des systèmes échelonnés dans [Esc, Chapitre 4].

2.2.2 Décomposition LU

Si q_k n'est jamais nul, il n'y a pas de matrice de transposition donc $G = T_{n-1} \dots T_1 \in TI_n(\mathbf{K})$ et en notant

$$L \triangleq G^{-1} = T_1^{-1} \dots T_{n-1}^{-1} \in TI_n(\mathbf{K})$$

on a

$$A = LU$$

La *décomposition LU* de A n'est donc rien d'autre que la mise en œuvre de la méthode de Gauss dans ce cas. Comme on va le voir, le fait que q_k n'est jamais nul se lit sur les mineurs principaux de A.

Théorème 2. *On suppose que les mineurs principaux de A sont tous non nuls. Il existe alors un unique couple de matrices (L, U) avec $L \in TI_n^1(\mathbf{K})$ et $U \in TS_n(\mathbf{K})$ inversible telles que $A = LU$.*

Remarque 5. Les notations L et U se rapportent aux termes anglais *lower* et *upper*.

Exercice 5. On va prouver le Théorème 2. On reprend la méthode d'élimination de Gauss.

1. (*Unicité*). Sous réserve d'existence, montrer que la décomposition LU est unique.
2. (*Existence*) On suppose qu'il existe k tel que $q_1, \dots, q_{k-1} \neq 0$ et $q_k = 0$. Montrer que le $k-$ième mineur principal de A est nul. Conclure.
3. Réciproquement, on suppose que A admet une telle décomposition. Montrer que les mineurs principaux de A sont tous non-nuls.

Algorithmique 4. On pourrait croire que la factorisation LU est plus coûteuse que la méthode de Gauss puisqu'elle demande de calculer en outre la matrice L. Il n'en est rien. La forme particulièrement simple des matrices T_i permet leur inversion par simple changement de signe de sa sous-colonne non nulle. On obtient alors L par concaténation de ces sous-colonnes (voir [Cia, page 83]). Pour s'en convaincre, on pourra chercher à calculer

$$\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix}^{-1} \quad \text{et} \quad \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 4 & 1 \end{pmatrix}$$

Ainsi, la factorisation LU est aussi en $n^3/3$.

On peut aussi calculer par récurrence les sous-matrices $[1:k, 1:k]$ de L et U, comme présenté dans [Ser, Chapitre 8]. Notons les L_k, U_k. On cherche *a priori* L_{k+1}, U_{k+1} sous la forme

$$L_{k+1} = \begin{pmatrix} L_k & 0 \\ l^* & 1 \end{pmatrix}, \quad U_{k+1} = \begin{pmatrix} U_k & u \\ 0 & z \end{pmatrix}$$

avec $u, l \in \mathbf{K}^k$, $z \in \mathbf{K}$. On trouve l et u en résolvant deux systèmes triangulaires de taille k, pour un coût en k^2. On a encore une complexité en $n^3/3$.

Remarque 6. Cette décomposition est utile lorsqu'on veut résoudre le système

$$Ax = b$$

pour plusieurs valeurs b. On résout alors deux systèmes triangulaires en cascade

$$Ly = b, \quad Ux = y$$

pour un coût en n^2. C'est le cas par exemple pour la résolution approchée d'EDP (Section 2.1.4) ou bien du calcul de A^{-1} présenté dans l'exercice qui suit.

Exercice 6. On suppose que les mineurs principaux de A sont tous non-nuls. Proposer une méthode de calcul de A^{-1} en $4n^3/3$. (*En fait, on peut réduire la complexité à* n^3).

Remarque 7. [Ser] présente un algorithme astucieux de multiplication et d'inversion de matrice en $\mathcal{O}(n^{\log 7/\log 2})$.

2.2.3 Factorisation de Cholesky [AK]

C'est une sorte de décomposition LU pour les matrices hermitiennes définies positives.

Théorème 3. *On suppose $A \in H_n^{++}(\mathbf{K})$. Il existe une unique $B \in TI_n^{++}(\mathbf{K})$ telle que*

$$A = BB^*$$

Exercice 7. On va prouver le Théorème 3.

1. Sous réserve d'existence, montrer l'unicité.
2. Justifier que A admet une décomposition LU et que les $u_{k,k}$ sont strictement positifs.
3. On note
$$D \triangleq \begin{pmatrix} \sqrt{u_{1,1}} & & \\ & \ddots & \\ & & \sqrt{u_{n,n}} \end{pmatrix}$$
Montrer que $B \triangleq LD$ convient.

Exemple 6. Etablir la factorisation de Cholesky de la matrice

$$A = \begin{pmatrix} 1 & 2 & 1 \\ 2 & 13 & 2 \\ 1 & 2 & 2 \end{pmatrix}$$

Algorithmique 5. On retiendra que Cholesky est deux fois plus rapide que LU (donc en $n^3/6$) tout en offrant le même avantage de ramener à une cascade de deux systèmes triangulaires

$$By = b, \quad B^*x = y$$

Aussi, on la préférera toujours à la décomposition LU lorsque $A \in H_n^{++}(\mathbf{K})$, ce qui est le cas dans beaucoup d'applications (Cf. Section 2.1).

Notons que cette factorisation permet un calcul du déterminant dans $H_n^{++}(\mathbf{K})$ en $n^3/6$.

Exercice 8. On revient sur le problème aux moindres carrés (Section 2.1.2). On suppose $n \geq m$ et A de rang m. Donner le coût de la résolution du problème à l'aide de l'équation normale.

Remarque 8. Il y a d'autres méthodes que l'équation normale pour résoudre un problème aux moindres carrés. On pourra consulter à ce sujet [AK, Chapitre 7].

2.3 Méthodes itératives

Ces méthodes ne permettent qu'un calcul approché de la solution $x^{\#}$, mais sont rapides comparées aux méthodes directes. Le principe de base est de décomposer A en

$$A = M - N$$

avec M inversible, de sorte que

$$Ax = b \quad \Leftrightarrow \quad Mx = Nx + b \quad \Leftrightarrow \quad x = M^{-1}(Nx + b)$$

La méthode itérative associée à la décomposition $A = M - N$ consiste alors à choisir $x^0 \in \mathbf{K}^n$ et à calculer les itérés x^k solution de

$$Mx^{k+1} = (Nx^k + b)$$

Ceci n'a d'intérêt que si M facilite le travail (typiquement M diagonale ou triangulaire). La pertinence de la décomposition se lit sur le rayon spectral de $M^{-1}N$.

Proposition 3. *(x^k) converge vers $x^{\#}$ quelque soit x^0 si et seulement si $\rho(M^{-1}N) < 1$. On dit alors que la méthode est convergente.*

Exercice 9. Prouver la Proposition 3. Montrer que la convergence est au moins linéaire et donner le taux de convergence.

Algorithmique 6. En pratique on ne connait pas $x^{\#}$ (on en cherche une approximation !) donc on ne sait pas quel terme x^k est satisfaisant. Un critère souvent utilisé est de s'arrêter lorsque

$$|Ax^k - b| \leq \epsilon$$

où ϵ est un seuil défini à l'avance. On note K le nombre d'itérations effectuées avant l'arrêt. En pratique, la complexité des méthodes itératives est en Kn^2. Elles sont donc intéressantes si $K \ll n$.

Remarque 9. C'est une méthode de point fixe pour la fonction $x \mapsto (M^{-1}Nx + b)$. L'inégalité $\rho(M^{-1}N) < 1$ équivaut au fait que la fonction est contractante.

2.3.1 Méthode de Jacobi

Dans cette méthode M est la diagonale de A, supposée inversible. Par exemple

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 13 & 2 \\ 1 & 1 & 2 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 13 & 0 \\ 0 & 0 & 2 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -2 & 0 \\ -2 & 0 & -2 \\ -1 & -1 & 0 \end{pmatrix}$$

Exercice 10. On suppose que A est à diagonale strictement dominante. Montrer que la méthode de Jacobi est convergente.

2.3.2 Méthode de Gauss-Seidel

Dans cette méthode M est la partie triangulaire inférieure de A, par exemple

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 13 & 2 \\ 1 & 1 & 2 \end{pmatrix}, \quad M = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 13 & 0 \\ 1 & 1 & 2 \end{pmatrix}, \quad N = \begin{pmatrix} 0 & -2 & 0 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix}$$

Proposition 4. *Si $A \in H_n^{++}(\mathbf{K})$, la méthode de Gauss-Seidel est convergente*

Exercice 11. On va prouver la Proposition 4.

1. Montrer que $M^* + N$ est diagonale et définie positive.
2. On note $|x|_A = \sqrt{\langle Ax, x\rangle}$ la norme associée à A sur $\mathbf{K}^n$. Montrer que, pour tout $x \neq 0$
$$\left|M^{-1}Nx\right|_A^2 < |x|_A^2$$
3. Conclure

Remarque 10. On vient en fait de montrer que, de manière générale, la méthode est convergente si $A \in H_n^{++}(\mathbf{K})$ et $M^* + N \in H_n^{++}(\mathbf{K})$.

Remarque 11. Un raffinement de la méthode de Gauss-Seidel consiste à pondérer la diagonale de A dans M par un facteur $\frac{1}{\omega}$. Par exemple pour $\omega = \frac{1}{3}$

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 13 & 2 \\ 1 & 1 & 2 \end{pmatrix}, \quad M = \begin{pmatrix} 3 & 0 & 0 \\ 2 & 39 & 0 \\ 1 & 1 & 6 \end{pmatrix}, \quad N = \begin{pmatrix} 2 & -2 & 0 \\ 0 & 26 & -2 \\ 0 & 0 & 4 \end{pmatrix}$$

C'est la *méthode de relaxation* [Ser, Chapitre 9], convergente si et seulement si $|\omega - 1| < 1$. L'idée est de choisir judicieusement ω pour rendre $\rho(M^{-1}N)$ petit, accélérant ainsi la convergence.

2.3.3 Méthode du gradient

Dans cette méthode $M = \frac{1}{\alpha}I$, où $\alpha \in \mathbf{R}^*$ est un paramète de réglage. Par exemple pour $\alpha = 0.1$

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 13 & 2 \\ 1 & 1 & 2 \end{pmatrix}, \quad M = \begin{pmatrix} 10 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 10 \end{pmatrix}, \quad N = \begin{pmatrix} 9 & -2 & 0 \\ -2 & -3 & -2 \\ -1 & -1 & 8 \end{pmatrix}$$

Exercice 12. On suppose A diagonalisable de valeurs propres $0 < \lambda_1 \leq \cdots \leq \lambda_n$

1. Montrer que la suite des itérés est
$$x^{k+1} = x^k - \alpha(Ax^k - b)$$
et justifier le nom de la méthode.

2. Montrer que la méthode du gradient converge si et seulement si $0 < \alpha < \dfrac{2}{\rho(A)}$.

3. Montrer que la valeur
$$\alpha = \frac{2}{\lambda_1 + \lambda_n}$$
minimise $\rho(M^{-1}N)$ qui vaut alors
$$\rho(M^{-1}N) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1}$$

Exercice 13. On suppose $A \in S_n^{++}(\mathbf{R})$. Une variante de la méthode du gradient consiste à faire varier α à chaque pas, en choisissant α_k qui minimise la fonction
$$\alpha \mapsto f(x_k - \alpha(Ax_k - b))$$
où $f : x \mapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$ est la fonction convexe sous-jacente. Montrer que ce minimum est atteint en un unique point
$$\alpha_k = \frac{\left|Ax^k - b\right|_2^2}{\langle A(Ax^k - b), Ax^k - b\rangle}$$
C'est la méthode *du gradient à pas optimal*.

Remarque 12. Ces deux méthodes sont délaissées au profit de l'algorithme beaucoup plus efficace du *gradient conjugué*, dans lequel les directions de descente sont itérativement orthonormées pour le produit scalaire induit par A sur $\mathbf{R}^n$. Voir [AK, Chapitre 9] pour un exposé détaillé.

2.4 Conditionnement

En pratique, la précision de A et b est limitée par la représentation des nombres en machine qui induit d'inévitables erreurs d'arrondi. Il se peut aussi que A et b ne soient pas connus parfaitement. C'est typiquement le cas s'ils représentent des mesures de quantité physique par des capteurs, nécessairement corrompues par du bruit, ou bien si on a fait une approximation par rapport à un autre système comme aux Sections 2.1.3 et 2.1.4. Ceci peut avoir des conséquences significatives sur la solution du système.

Exemple 7. (inspiré de [AK])
$$A = \begin{pmatrix} 8 & 6 & 4 & 1 \\ 1 & 4 & 5 & 1 \\ 8 & 4 & 1 & 1 \\ 1 & 4 & 3 & 6 \end{pmatrix}, \quad b = \begin{pmatrix} 8 \\ 10 \\ 2 \\ 6 \end{pmatrix} \quad \Rightarrow \quad x^{\#} = \begin{pmatrix} 0 \\ 0 \\ 2 \\ 0 \end{pmatrix}$$

En modifiant légèrement le second membre, on obtient une solution très différente

$$A = \begin{pmatrix} 8 & 6 & 4 & 1 \\ 1 & 4 & 5 & 1 \\ 8 & 4 & 1 & 1 \\ 1 & 4 & 3 & 6 \end{pmatrix}, \quad b = \begin{pmatrix} 8.05 \\ 10 \\ 2 \\ 6 \end{pmatrix} \quad \Rightarrow \quad x^{\#} = \begin{pmatrix} 2 \\ -5.225 \\ 5.5 \\ 1.4 \end{pmatrix}$$

On va voir que l'impact de ces erreurs d'arrondi sur la solution est liée au *conditionnement* de la matrice A.

Définition 3. Soit $|\cdot|$ une norme sur $\mathbf{K}^n$. Le conditionnement de A associé à la norme induite $\|\cdot\|$ est

$$\mathrm{cond}A = \|A\| \left\|A^{-1}\right\| \geq 1$$

Lorsqu'on considère la norme p, on note $\mathrm{cond}_p A$.

Proposition 5.

$$cond_2(A) = \sqrt{\frac{\lambda_{\max}(A^*A)}{\lambda_{\min}(A^*A)}}$$

Si A est normale

$$cond_2(A) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|}$$

Si $O \in U_n(\mathbf{K})$

$$cond_2(O) = 1, \quad cond_2(AO) = cond_2(OA) = cond_2(A)$$

Exercice 14.

1. Soit δb une perturbation du second membre et δx tel que $A(x^{\#} + \delta x) = b + \delta b$. Montrer que
$$\frac{|\delta x|}{|x^{\#}|} \leq \mathrm{cond}A \frac{|\delta b|}{|b|}$$
2. Montrer qu'il existe b et δb tels que l'égalité ait lieu. En ce sens, l'inégalité est optimale.
3. Soit δA une perturbation de la matrice et δx tel que $(A+\delta A)(x^{\#}+\delta x) = b$. Montrer que
$$\frac{|\delta x|}{|x^{\#} + \delta x|} \leq \mathrm{cond}A \frac{\|\delta A\|}{\|A\|}$$
et que cette inégalité est optimale pour un certain choix de b et δA.

Remarque 13. Ainsi, il est préférable que le conditionnement de A soit proche de 1. Ceci est d'autant plus vrai pour les méthodes itératives où les erreurs sont propagées et amplifiées à chaque itération. La technique du *préconditionnement* consiste à considérer un système linéaire équivalent

$$CAx = Cb$$

où $\mathrm{cond}(CA) < \mathrm{cond}A$. Il existe plusieurs méthodes pour choisir C. Certaines reposent sur des outils de factorisation de type LU. On pourra consulter à ce sujet [AK, Chapitre 5].

Remarque 14. Les inégalités de l'Exercice 14 sont certes optimales, mais elles sont en général très pessimistes (on dit aussi *conservatives*). Ainsi dans l'Exemple 7 on a

$$\frac{|\delta x|_2}{|x^\#|_2} \simeq 964 \frac{|\delta b|_2}{|b|_2}$$

ce qui fait déjà beaucoup, mais $\mathrm{cond}_2 A \simeq 3199$.

Exercice 15. [AK, page 95] On va analyser la matrice A du Laplacien discrétisé (Section 2.1.3) à la lumière de la Remarque 14.

1. Montrer que
$$\mathrm{cond}_2 A \underset{n\to\infty}{\sim} \frac{4n^2}{\pi^2}$$
de sorte que, quand n est grand, la matrice est très mal conditionnée.
2. En notant $x^\# = U_{\text{approx}}$, montrer que
$$\lim_{n\to\infty} \frac{1}{n} \left|x^\#\right|_2^2 = \int_0^1 u^2(x)dx, \quad \lim_{n\to\infty} \frac{1}{n} |b|_2^2 = \int_0^1 f^2(x)dx$$
3. Montrer enfin qu'il existe une constante $c > 0$ indépendante de n telle que
$$\frac{|\delta x|_2}{|x^\#|_2} \leq c \frac{|\delta b|_2}{|b|_2}$$

3 Recherche d'éléments propres

On cherche à déterminer des valeurs propres, et éventuellement des vecteurs propres, d'une matrice $A \in M_n(\mathbf{K})$. C'est un problème difficile, pour lequel on ne dispose pas de méthode exacte en général. On a des résultats qualitatifs qui permettent de localiser explicitement le spectre dans un domaine de $\mathbf{C}$ (Section 3.2), et des méthodes itératives qui permettent un calcul approché d'une ou plusieurs valeurs propres et de vecteurs propres (Section 3.3). On commence par donner quelques exemples de problèmes où apparait naturellement la recherche d'éléments propres.

3.1 Exemples

3.1.1 Racines d'un polynôme

Calculer les valeurs propres d'une matrice revient à chercher les racines de son polynôme caractéristique. Inversement, les racines d'un polynôme

$$P = X^n + \sum_{k=0}^{n-1} c_k X^k \in \mathbf{K}_n[X]$$

sont les valeurs propres de la matrice compagnon associée dans

$$\mathcal{C}(P) = \begin{pmatrix} 0 & \dots & 0 & -c_0 \\ 1 & \ddots & \vdots & \vdots \\ & \ddots & 0 & \vdots \\ & & 1 & -c_{n-1} \end{pmatrix} \in M_n(\mathbf{K})$$

Aussi, il est illusoire d'espérer des méthodes générales de calcul exact de valeur propre.

3.1.2 Stabilité d'une équation différentielle

On considère une équation différentielle linéaire à coefficients constants

$$\begin{aligned} X'(t) &= AX(t), \quad t \in \mathbf{R} \\ X(0) &= X_0 \end{aligned}$$

avec $A \in M_n(\mathbf{K})$. Le spectre de A renseigne sur la nature des solutions.

- si les valeurs propres de A sont à partie réelle < 0, $X(t)$ converge vers 0 (quand $t \to +\infty$) quel que soit X_0 . L'équation est dite *asymptotiquement stable*.
- si A admet une valeur propre à partie réelle > 0, $|X(t)|$ diverge vers l'infini pour presque toutes les conditions initiales X_0. L'équation est dite *instable*.

Remarque 15. Si l'une des valeurs propres est à partie réelle nulle, il faut étudier l'espace propre associé pour en savoir plus.

3.1.3 Analyse en composantes principales

On considère un nuage de m points de $\mathbf{R}^n$

$$X^k = (x_i^k)_{1 \le i \le n}, \quad 1 \le k \le m$$

Il n'est pas aisé de manipuler des données que l'on ne sait pas visualiser. Aussi, on cherche à représenter ce nuage de points dans un espace de petite dimension d (typiquement $d = 2$ ou 3) en perdant "peu" d'information. Quitte à centrer les X^k, on considère que

$$\sum_{k=1}^{m} X^k = 0$$

On note $\Gamma \in M_n(\mathbf{R})$ la matrice de covariance des X^k définie par

$$\Gamma_{i,j} = \frac{1}{m} \sum_{k=1}^{m} x_i^k x_j^k$$

Γ est symétrique, positive. Notons $\lambda_1 \ge \dots \ge \lambda_n \ge 0$ ses valeurs propres et $V_1, \dots V_n$ une base orthonormée de vecteurs propres associés. Les V_i représentent les axes principaux de

Γ (on dit aussi les *composantes principales*). λ_i est la variance (ou dispersion) du nuage de points dans la direction V_i.

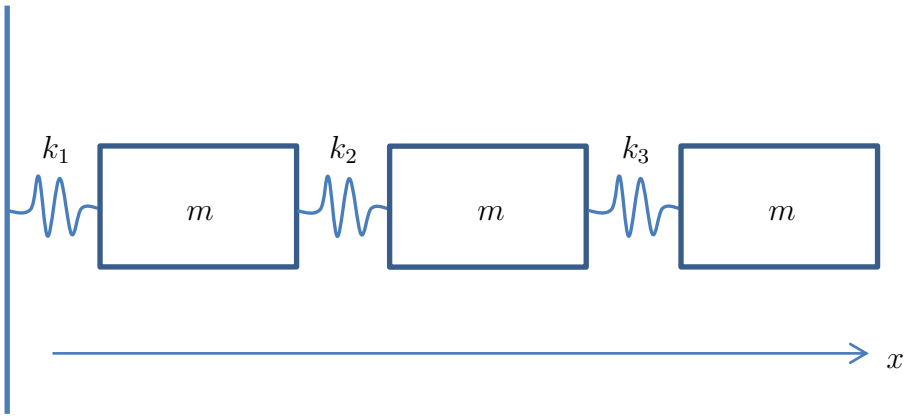
Si λ_i est petit, les points sont peu dispersés dans la direction V_i et on ne perd pas grand chose à considérer que le nuage est contenu dans le plan orthogonal à V_i. On sélectionne alors les d plus grandes valeurs propres $\lambda_1 \geq \cdots \geq \lambda_d$ et on projette le nuage de points dans l'espace engendré par $V_1, \ldots, V_d$:

$$Y^k = \sum_{i=1}^{d} \left\langle V_i, X^k \right\rangle V_i$$

On a ainsi "compressé" les X^k en un nuage de points Y^k en dimension d. Le taux de compression est défini par

$$\tau \triangleq \frac{\sum_{i=1}^{d} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

3.1.4 Vibrations d'un système mécanique [AK]



Considérons un système unidimensionnel de trois masses m reliées entre elles et à un support par des ressorts de constante de raideur k_1, k_2, k_3. On note x_1, x_2, x_3 le déplacement respectif des masses par rapport à leur position d'équilibre, $X \triangleq (x_1, x_2, x_3)$ satisfait l'équation différentielle

$$m\ddot{X} + KX = 0, \quad \text{avec } K = \begin{pmatrix} k_1 + k_2 & -k_2 & 0 \\ -k_2 & k_2 + k_3 & -k_3 \\ 0 & -k_3 & k_3 \end{pmatrix}$$

On cherche les pulsations propres du système c'est-à-dire les $\omega > 0$ telles qu'il existe une solution $X(t) = e^{i\omega t} X_0$. On a alors

$$KX_0 = m\omega^2 X_0$$

Ainsi, la recherche des pulsations propres est équivalente à la recherche des valeurs propres de K. Cette méthode est utilisée pour calculer les pulsations propres d'un immeuble représenté par n masses (plafonds) et ressorts (murs), en particulier pour prévoir son comportement en cas de séisme.

3.2 Etude qualitative du spectre

3.2.1 Domaine de Gershgörin [Ser]

Définition 4. Pour $1 \leq i \leq n$ on note D_i le disque fermé de centre $a_{i,i}$ et de rayon $\sum_{j \neq i} |a_{i,j}|$. On appelle *domaine de Gershgörin* de A

$$\mathcal{G}(A) \triangleq \bigcup_{i=1}^{n} D_i$$

Proposition 6. *(Théorème de Gershgörin). Le spectre de A est inclus dans son domaine de Gershgörin.*

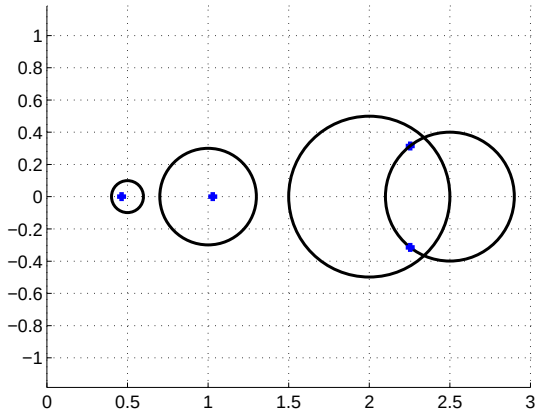
Exercice 16.

1. Montrer la Proposition 6.
2. En déduire qu'une matrice à diagonale strictement dominante est inversible.
3. Montrer aussi $\mathrm{sp}A \subset \mathcal{G}(A) \cap \mathcal{G}(A^\top)$.

Exemple 8.

On a représenté sur la figure ci-contre le domaine de Gershgörin (disques noirs) ainsi que les valeurs propres (croix bleues) de la matrice

$$A = \begin{pmatrix} 0.5 & 0.1 & 0 & 0 \\ 0.2 & 1 & 0.1 & 0 \\ 0 & 0.1 & 2 & -0.4 \\ 0 & 0 & 0.4 & 2.5 \end{pmatrix}$$



Sur cet exemple, les seuls disques qui contiennent deux valeurs propres sont ceux qui s'intersectent. Les autres en contiennent une et une seule. Ce phénomène, qui n'est pas anodin, est l'objet de l'exercice suivant.

Exercice 17. On note $\mathcal{E}$ une composante connexe de $\mathcal{G}(A)$ et p le nombre de disques D_i inclus dans $\mathcal{E}$. Ainsi, dans l'Exemple 8, $\mathcal{G}(A)$ a 3 composantes connexes

$$D_1 \ (p=1), \quad D_2 \ (p=1), \quad D_3 \cup D_4 \ (p=2)$$

Le but de l'exercice est de montrer que $\mathcal{E}$ contient exactement p valeurs propres de A (comptées avec leur ordre de multiplicité algébrique).

A cet effet, on note D la diagonale de A et, pour tout $0 \leq r \leq 1$

$$A_r \triangleq rA + (1-r)D$$

On note enfin $m(r)$ le nombre de valeurs propres de A_r dans $\mathcal{E}$. On cherche donc à montrer : $m(1) = p$.

1. Montrer que $m(0) = p$.
2. Montrer que $\mathcal{G}(A_r) \subset \mathcal{G}(A)$.
3. On suppose qu'il existe γ une arc simple $\mathcal{C}^1$ par morceaux orienté positivement séparant $\mathcal{E}$ (à l'intérieur) de $\mathcal{G}(A)\backslash\mathcal{E}$ (à l'extérieur). Montrer que
$$m(r) = \int_\gamma \frac{\chi_r'(z)}{\chi_r(z)} dz$$
où χ_r est le polynôme caractéristique de A_r.
4. En déduire que m est continue et conclure.
5. **Application.** On suppose que les disques D_i sont deux-à-deux disjoints. Montrer que A est diagonalisable.
6. Proposer une configuration où un tel arc γ n'existe pas. Comment peut-on conclure dans ce cas ?

Citons pour finir un résultat de [Ser] sur le domaine de Gershgörin d'une matrice A *irréductible* : si une valeur propre est sur la frontière de $\mathcal{G}(A)$, c'est un point d'intersection de tous les cercles.

3.2.2 Autres théorèmes

Il existe de nombreux résultats qualitatifs sur le spectre d'une matrice ou les racines d'un polynôme. En voici une liste non-exhaustive.

— théorème de Bauer-Fike [Cia, page 35], qui majore l'impact de la perturbation d'une matrice sur son spectre.

— suites de Sturm [Cia, page 122] qui donnent le nombre de racines d'un polynôme dans un intervalle donné.

— critère de Routh pour déterminer si les racines d'un polynôme sont à partie réelle < 0.

Citons enfin le

Théorème 4 (Gauss-Lucas)**.** *Les racines d'un polynôme dérivé P' sont dans l'enveloppe convexe des racines de P.*

Exercice 18. Démontrer le Théorème de Gauss-Lucas.

3.3 Méthodes itératives

3.3.1 Méthode de la puissance [AK]

Principe. On construit une suite (x^k) par récurrence
$$x^0 \in \mathbf{K}^n, \quad x^{k+1} = \frac{Ax^k}{|Ax^k|_2}$$

Exercice 19. On suppose que A est diagonalisable à valeurs propres réelles

$$|\lambda_1| \leq \cdots \leq |\lambda_{n-1}| < |\lambda_n|, \quad \text{avec} \quad \lambda_n > 0$$

On note $e_1, \ldots, e_n$ une base de vecteurs propres *normés* associés, et on décompose

$$x^0 = \sum_{i=1}^{n} \beta_i e_i$$

On suppose $\beta_n \neq 0$. Quitte à changer e_n en $\frac{\beta_n}{|\beta_n|} e_n$, on peut supposer $\beta_n > 0$.

1. Montrer que
$$x^k = e_n + \mathcal{O}(\tau^k), \quad \tau = \frac{|\lambda_{n-1}|}{|\lambda_n|}$$
Ainsi, x^k converge (au moins) linéairement vers un vecteur propre de A pour λ_n, au taux de convergence τ.
2. Montrer que
$$\left|Ax^k\right|_2 = \lambda_n + \mathcal{O}(\tau^k)$$
3. On suppose $e_1, \ldots e_n$ orthonormée. Montrer que
$$\left|Ax^k\right|_2 = \lambda_n + \mathcal{O}(\tau^{2k})$$
4. L'hypothèse $\beta_n \neq 0$ est-elle vraiment restrictive en pratique ?
5. Adapter la méthode si $\lambda_n < 0$.

Remarque 16. En pratique, on arrête les itérations lorsque

$$\left|x^{k+1} - x^k\right|_2 \leq \epsilon$$

où ϵ est un seuil fixé à l'avance.

3.3.2 Méthode de la puissance inverse [Ser]

Principe. On suppose A inversible avec $0 < |\lambda_1| < |\lambda_2|$ et on applique la méthode de la puissance à A^{-1}. Ceci permet un calcul approché de λ_1 et e_1 à un taux de convergence $\frac{|\lambda_1|}{|\lambda_2|}$.

Exercice 20.

1. On suppose qu'on connait une approximation μ d'une des valeurs propres λ_i de A, avec
$$|\lambda_j - \mu| > |\lambda_i - \mu| > 0, \quad \forall j \neq i$$
Montrer que la méthode de la puissance inverse appliquée à la matrice $A - \mu I$ permet un calcul approché de λ_i avec un taux de convergence
$$\frac{|\lambda_i - \mu|}{\min_{j \neq i} |\lambda_j - \mu|}$$
C'est en fait le principal intérêt de la méthode de la puissance inverse.
2. **Application.** On suppose qu'un des disques de Gershgörin D_i de A n'intersecte aucun autre. Proposer une méthode de calcul approché de l'unique valeur propre de A contenue dans D_i.

3.3.3 Méthode QR [Ser]

Théorème 5. *Soit $A \in GL_n(\mathbf{K})$. Il existe un unique couple de matrices (Q, R) avec $Q \in U_n(\mathbf{K})$ et $R \in TS_n^{++}(\mathbf{K})$ telles que*

$$A = QR$$

Ceci définit un homéomorphisme de $GL_n(\mathbf{K})$ sur $U_n(\mathbf{K}) \times TS_n^{++}(\mathbf{K})$.

Exercice 21. On va prouver le Théorème 5.

1. Sous réserve d'existence, montrer l'unicité
2. Montrer l'existence. *On pourra par exemple s'intéresser à la matrice A^*A.*
3. Montrer que la factorisation QR définit bien un homéomorphisme.

Remarque 17. C'est une version matricielle du procédé d'orthonormalisation de Gram-Schmidt appliqué aux colonnes de A.

Principe de la méthode. On construit une suite de matrices (A_k) par récurrence

$$A_1 = A, \quad A_{k+1} = R_k Q_k$$

où (Q_k, R_k) est la factorisation QR de A_k.

Théorème 6. *On suppose que A est diagonalisable*

$$A = Y^{-1} D Y$$

avec $D = diag(\lambda_1, \ldots, \lambda_n)$ et

$$|\lambda_1| > \cdots > |\lambda_n| > 0$$

On suppose en outre que Y admet une décomposition LU. Alors, la partie triangulaire inférieure de A_k converge vers D.

Exercice 22. (*difficile*) On va prouver le Théorème 6. Pour simplifier on suppose les λ_i réels strictement positifs.

1. On note
$$\mathcal{Q}_k \triangleq Q_1 \ldots Q_k, \quad \mathcal{R}_k \triangleq R_k \ldots R_1$$
Montrer que la factorisation QR de la $k-$ième puissance de A est
$$A^k = \mathcal{Q}_k \mathcal{R}_k$$
2. On note $Y = LU$ la décomposition de Y. Pour simplifier on suppose que la diagonale de U est strictement positive. On note encore
$$Y^{-1} = QR$$
la factorisation QR de Y^{-1}. Montrer que
$$\mathcal{Q}_k \mathcal{R}_k = QRD^k LU$$

3. Montrer que
$$D^k L D^{-k} = I + \mathcal{O}\left(\sigma^k\right)$$
où
$$\sigma = \max_j \frac{\lambda_{j+1}}{\lambda_j}$$
En déduire
$$\mathcal{Q}_k \mathcal{R}_k = Q B_k R D^k U$$
avec $B_k = I + o(1)$
4. On note $B_k = O_k T_k$ la factorisation QR de B_k. Montrer que
$$\mathcal{Q}_k = Q O_k, \quad \mathcal{R}_k = T_k R D^k U$$
5. Montrer enfin
$$(Q_k) \to I, \quad (R_k) \to R D R^{-1}$$
et conclure.

Remarque 18. Attention, dans l'exercice précédent la partie supérieure stricte de A_k converge. Ce n'est pas le cas en général. Ceci est du au fait qu'on a supposé les valeurs propres de A et les éléments diagonaux de U strictement positifs.

Algorithmique 7. En pratique, le taux de convergence σ peut être proche de 1, surtout pour des matrices de grande taille. Pour accélérer la convergence, on peut faire une approximation des valeurs propres par la méthode QR puis utiliser la puissance inverse comme présenté à l'Exercice 20 pour les dernières itérations. Ceci permet en outre de calculer des vecteurs propres.

Algorithmique 8.
1. Le calcul de la factorisation QR par le procédé de Gram-Schmidt est en n^3 mais cet algorithme a tendance à propager des erreurs d'arrondi. En pratique, on la calcule plutôt par l'algorithme de Householder [AK, Chapitre 7], qui consiste à multiplier A par $n-1$ matrices de symétrie orthogonale par rapport à des hyperplans bien choisis. Cette méthode est plus stable numériquement et plus rapide, en $2n^3/3$.
2. On pourrait penser à utiliser la factorisation $A = QR$ pour résoudre un système linéaire $Ax = b$. En effet, on est ramené au système triangulaire
$$Rx = Q^* b$$
Elle n'est jamais utilisée dans ce but car, même avec l'algorithme de Householder, la factorisation QR est deux fois plus lente que la méthode d'élimination de Gauss.
3. Dans le cadre du calcul de valeur propres, où on itère des factorisation QR, il est préférable de mettre A sous une forme réduite (dite *de Hessenberg*) pour un coût en $5n^3/3$. Le calcul de la factorisation à chaque étape est alors en $4n^2$. On pourra consulter à ce sujet [Ser, Chapitre 10].
4. On peut étendre la factorisation QR à des matrices non-carrées. C'est une des méthodes de résolution d'un problème aux moindres carrés présentées dans [AK, Chapitre 7].

3.3.4 Autres méthodes

Citons enfin deux méthodes qui s'appliquent à des matrices symétriques réelles [Cia, Chapitre 6].

— Jacobi, qui permet de calculer les valeurs propres et vecteurs propres

— Givens-Householder, qui s'appuie sur les suites de Sturm.

Références

[AK] G. Allaire and S. M. Kaber. *Algèbre linéaire numérique*. Ellipses, 2002.

[All] G. Allaire. *Analyse numérique et optimisation*. Editions de l'Ecole Polytechnique, 2005.

[Cia] P. G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation*. Masson, 1994.

[Esc] J.-P. Escofier. *Toute l'algèbre de la licence*. Dunod, 2006.

[Ser] D. Serre. *Les matrices*. Dunod, 2001.

Analyse matricielle : correction des exercices

Lionel Magnis

Exercice 7.

1. Soient $B_1, B_2 \in TI_n^{++}(\mathbf{K})$ telles que

$$A = B_1 B_1^* = B_2 B_2^*$$

On a

$$\left(B_1^{-1} B_2\right)^{-1} = \left(B_1^{-1} B_2\right)^*$$

Donc

$$B_1^{-1} B_2 \in TI_n^{++}(\mathbf{K}) \cap U_n(\mathbf{K}) = \{I\}$$

et $B_1 = B_2$.

2. $A \in H_n^{++}(\mathbf{K})$ donc par le critère de Sylvester, les mineurs principaux de A sont strictement positifs. On conclut immédiatement.

3. On a

$$A = LU = LD^2 D^{-2} U$$

et

$$A = A^* = \left(D^{-2}U\right)^* D^2 L^*$$

avec

$$\left(D^{-2}U\right)^* \in TI_n^1(\mathbf{K}), \quad D^2 L^* \in TS_n(\mathbf{K})$$

Par unicité de la décomposition LU de A, $D^{-2}U = L^*$ et finalement

$$A = LDDL^* = (LD)(LD)^*$$

avec $LD \in TI_n^{++}(\mathbf{K})$.

Exemple 6.

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & 13 & 2 \\ 1 & 2 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 1 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Exercice 8. L'équation normale est

$$A^\top Ax = A^\top b$$

Décomposons le temps de calcul.

- calcul (naïf) de $A^\top A$: nm^2
- calcul de $A^\top b$: nm
- factorisation de Cholesky de $A^\top A \in S_m^{++}(\mathbf{R})$: $m^3/6$
- résolution de deux systèmes triangulaires : m^2

Exercice 11.

1. On note D la diagonale de A. Comme A est hermitienne, on a

$$M = -N^* + D$$

De plus, A est définie positive donc ses coefficients diagonaux sont strictement positifs (pour le vérifier il suffit de calculer x^*Ax pour x vecteur de la base canonique). Bref

$$M^* + N = D^* = D \in H_n^{++}(\mathbf{K})$$

2.

$$\begin{aligned}
\left|M^{-1}Nx\right|_A^2 &= \left\langle AM^{-1}Nx, M^{-1}Nx\right\rangle \\
&= \left\langle AM^{-1}\left(M-A\right)x, M^{-1}\left(M-A\right)x\right\rangle \\
&= \langle Ax, x\rangle - \left\langle AM^{-1}Ax, x\right\rangle - \left\langle Ax, M^{-1}Ax\right\rangle + \left\langle AM^{-1}Ax, M^{-1}Ax\right\rangle \\
&= |x|_A^2 - \left(\left\langle M^{-1}Ax, Ax\right\rangle + \left\langle Ax, M^{-1}Ax\right\rangle - \left\langle AM^{-1}Ax, M^{-1}Ax\right\rangle\right)
\end{aligned}$$

Notons alors $y = M^{-1}Ax \neq 0$. On a

$$\begin{aligned}
\left\langle M^{-1}Ax, Ax\right\rangle + \left\langle Ax, M^{-1}Ax\right\rangle - \left\langle AM^{-1}Ax, M^{-1}Ax\right\rangle &= \langle y, My\rangle + \langle My, y\rangle - \langle Ay, y\rangle \\
&= \langle (M^* + M - A)\, y, y\rangle \\
&= \langle (M^* + N)\, y, y\rangle > 0
\end{aligned}$$

d'où la conclusion.

3. En passant au max sur $|x|_A = 1$ on obtient

$$\left\|M^{-1}N\right\|_A < 1$$

D'après l'Exercice 2, on a donc $\rho(A) < 1$ et par la Proposition 3, la méthode est convergente.

Exercice 13. Il s'agit de minimiser la fonction

$$\alpha \mapsto f(x - \alpha(Ax - b))$$

pour $x \neq A^{-1}b$ fixé. Pour simplifier on note $r = Ax - b \neq 0$. On a, pour $\alpha \in \mathbf{R}$.

$$\begin{aligned} f(x - \alpha r) &= \frac{1}{2} \langle A(x - \alpha r), x - \alpha r \rangle - \langle b, x - \alpha r \rangle \\ &= \frac{1}{2} \langle Ar, r \rangle \alpha^2 - |r|_2^2 \alpha + \langle Ax, x \rangle - \langle b, x \rangle \end{aligned}$$

où on a utilisé le fait que $A^\top = A$. Ce polynôme de degré 2 en α admet un minimum global en

$$\alpha = \frac{|r|_2^2}{\langle Ar, r \rangle}$$

Exercice 15.

1. A est symétrique donc normale. D'après la Proposition 5.

$$\mathrm{cond}_2(A) = \frac{|\lambda_{\max}(A)|}{|\lambda_{\min}(A)|} = \frac{\lambda_{n-1}}{\lambda_1} = \frac{\sin^2 \frac{n-1}{2n}\pi}{\sin^2 \frac{1}{2n}\pi} = \frac{1}{\tan^2 \frac{\pi}{2n}} \sim \frac{4n^2}{\pi^2}$$

2. On a

$$\begin{aligned} \frac{1}{n} |b|_2^2 &= \frac{1}{n} \sum_{i=1}^{n-1} f(\frac{i}{n})^2 \\ &= \frac{1}{n} \sum_{i=0}^{n-1} f(\frac{i}{n})^2 - \frac{f(0)^2}{n} \end{aligned}$$

On reconnait dans le terme $\sum$ la méthode des rectangles à gauche pour le calcul approché de

$$\int_0^1 f(t)^2 dt$$

La méthode converge en supposant par exemple f de classe $\mathcal{C}^1$. On a alors

$$\lim_{n\to\infty} \frac{1}{n} |b|_2^2 = \int_0^1 f(t)^2 dt$$

De même

$$\lim_{n\to\infty} \frac{1}{n} |U|_2^2 = \int_0^1 u(t)^2 dt$$

Par ailleurs

$$x^\# = U + \mathcal{O}(\frac{1}{n^2})$$

On en déduit

$$\lim_{n\to\infty} \frac{1}{n} \left|x^\#\right|_2^2 = \int_0^1 u(t)^2 dt$$

3. On suppose que f n'est pas la fonction nulle (sinon la solution de l'équation du Laplacien est triviale.) u n'est donc pas non plus la fonction nulle. On a déjà vu dans la Section 2.1.3

$$|\delta x|_2 \leq \frac{1}{\lambda_1} |\delta b|_2 = \frac{1}{4n^2 \sin^2 \frac{\pi}{2n}} |\delta b|_2$$

On a alors

$$\frac{|\delta x|_2}{|x^\#|_2} \leq c_n \frac{|\delta b|_2}{|b|_2}$$

avec

$$c_n = \frac{1}{4n^2 \sin^2 \frac{\pi}{2n}} \frac{|b|_2}{|x^\#|_2} \sim \frac{1}{\pi^2} \sqrt{\frac{\int_0^1 b^2(t)dt}{\int_0^1 u^2(t)dt}}$$

d'où le résultat.

Exercice 18. On considère

$$P = \lambda \prod_{i=1}^{p} (X - \lambda_i)^{\mu_i} \in \mathbf{C}[X]$$

Soit x une racine de P'. On veut montrer

$$x \in \mathrm{conv}(\lambda_1, \ldots, \lambda_p)$$

Si $P(x) = 0$, c'est fini. Sinon, on a aussi

$$0 = \frac{P'(x)}{P(x)} = \sum_{i=1}^{p} \frac{\mu_i}{x - \lambda_i} = \sum_{i=1}^{p} \frac{\mu_i(\bar{x} - \bar{\lambda}_i)}{|x - \lambda_i|^2}$$

En passant au conjugué il vient

$$\left(\sum_{i=1}^{p} \frac{\mu_i}{|x - \lambda_i|^2} \right) x = \sum_{i=1}^{p} \frac{\mu_i}{|x - \lambda_i|^2} \lambda_i$$

d'où le résultat.

Exercice 22.

1. On remarque déjà

$$A_{k+1} = Q_k^* Q_k R_k Q_k = Q_k^* A_k Q_k$$

et par une récurrence immédiate

$$A_{k+1} = \mathcal{Q}_k^* A \mathcal{Q}_k$$

d'où on déduit

$$A\mathcal{Q}_k = \mathcal{Q}_k A_{k+1}$$

Montrons alors par récurrence

$$A^k = \mathcal{Q}_k \mathcal{R}_k$$

Pour $k = 1$ c'est évident. Supposons le résultat vrai pour $k \geq 1$. On a

$$\begin{aligned} A^{k+1} &= AA^k \\ &= A\mathcal{Q}_k\mathcal{R}_k \\ &= \mathcal{Q}_k A_{k+1}\mathcal{R}_k \\ &= \mathcal{Q}_k Q_{k+1} R_{k+1}\mathcal{R}_k \\ &= \mathcal{Q}_{k+1}\mathcal{R}_{k+1} \end{aligned}$$

ce qui achève la récurrence.

2.

$$\begin{aligned} \mathcal{Q}_k\mathcal{R}_k &= A^k \\ &= Y^{-1}D^kY \\ &= QRD^kLU \end{aligned}$$

3. En conjuguant L par D^k, le coefficient d'indice (i, j) est multiplié par $(\lambda_i/\lambda_j)^k$. Ainsi

— les termes diagonaux restent égaux à 1

— les termes sur-diagonaux restent nuls

— les termes sous-diagonaux $(i > j)$ sont multipliés par un coefficient $\leq \sigma^k$

d'où le résultat. Notons $B_k = RD^kLD^{-k}R^{-1}$. On a alors

$$B_k = R\left(I + \mathcal{O}(\sigma^k)\right)R^{-1} = I + \mathcal{O}(\sigma^k)$$

et

$$\mathcal{Q}_k\mathcal{R}_k = QRD^kLD^{-k}R^{-1}RD^kU = QB_kRD^kU$$

4. On a

$$\mathcal{Q}_k\mathcal{R}_k = QO_kT_kRD^kU$$

avec

$$QO_k \in U_n(\mathbf{K}), \quad T_kRD^kU \in TS_n^{++}(\mathbf{K})$$

Par unicité de la factorisation QR, il vient

$$\mathcal{Q}_k = QO_k, \quad \mathcal{R}_k = T_kRD^kU$$

5. (B_k) converge vers I. La factorisation QR étant un homéomorphisme, (O_k) et (T_k) convergent également vers I. Ainsi

$$Q_k = \mathcal{Q}_{k-1}^* \mathcal{Q}_k = O_{k-1}^* Q^* Q O_k = O_{k-1}^* O_k \to I$$
$$R_k = \mathcal{R}_k \mathcal{R}_{k-1}^{-1} = T_k R D^k U U^{-1} D^{-(k-1)} R^{-1} T_{k-1}^{-1} \to RDR^{-1}$$

R étant triangulaire, la diagonale de RDR^{-1} est D. On trouve bien les valeurs propres de A, asymptotiquement.