Dernière édition : 23 mars 2020

Notes prises par Téofil Adamski

ANALYSE NUMÉRIQUE

(ANU)

Benjamin Boutin

1A maths 2019, ENS de Rennes

Chapitre 1

Résolution des systèmes linéaires

1.1 Exemple important : matrice du laplacien unidimensionnel

Soit Ω un ouvert de $\mathbb{R}^n$. De nombreux problèmes font intervenir l'opérateur laplacien Δ définie par

$$\forall u \in \mathscr{C}^2(\Omega), \quad \Delta u \coloneqq \sum_{j=1}^{d} \partial_{jj} u.$$

Par exemples, en physique, les équations de la chaleur et de propagation s'écrivent respectivement

$$\partial_x u = \kappa\, \Delta u \quad \text{et} \quad \partial_{tt} u - c^2\, \Delta u = 0.$$

▷ Exemple. Soit $f \in \mathscr{C}^\infty([0,1])$. On cherche à résoudre le problème

$$\begin{cases} -u''(x) = f(x), & 0 < x < 1, \\ u(0) = u(1) = 0. \end{cases} \tag{P}$$

Si $f = 0$, alors l'unique solution est $u = 0$. Sinon on peut trouver une formule par intégration successive.

1.1.1 Discrétisation aux différences finies

Soit $n \geqslant 2$. On pose $h \coloneqq (n+1)^{-1}$ et $x_i \coloneqq ih$ pour $i \in [\![0, n+1]\!]$. On obtient alors n points équiréparties dans $[0,1]$. Supposons qu'il existe une unique solution $u \in \mathscr{C}^4(\mathbb{R})$ au problème (P). Pour tout $i \in [\![1, n]\!]$, par la formule de Taylor en x_i, on obtient que

$$u''(x_i) = -f(x_i) = \frac{1}{h}\left(\frac{u(x_{i+1}) - u(x_i)}{h} - \frac{u(x_i) - u(x_{i-1})}{h}\right) + \frac{h^2}{12} u^{(4)}(x_i + \theta h), \quad \theta \in [-1, 1].$$

On simplifie le problème en dimension finie sous la forme du problème linéaire suivant.

Problème. On veut trouver un vecteur $U \coloneqq (u_1, \dots, u_n) \in \mathbb{R}^n$ tel que

$$\forall i \in [\![1, n]\!], \quad \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = -f(x_i)$$

où on a posé $u_0 = u_{n+1} = 0$. Ici, on a supprimé le reste de Taylor. Matriciellement, on doit résoudre le système

$$(n+1)^2 A_n U = F \quad \text{avec} \quad A_n \coloneqq \begin{pmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{pmatrix} \in \mathscr{M}_n(\mathbb{R}) \quad \text{et} \quad F \coloneqq (f(x_i))_{1 \leqslant i \leqslant}.$$

1.1.2 Étude spectrale de la matrice du laplacien

Afin d'approcher le problème (P), on veut résoudre l'équation $(n+1)^2 A_n U = F$ d'inconnue $U \in \mathbb{R}^n$ où le vecteur $F \in \mathbb{R}^n$ est donné par la fonction f.

PROPOSITION 1.1. La matrice A_n est symétrique définie positive.

Preuve Soit $U \coloneqq (u_1, \dots, u_n) \in \mathbb{R}^n$. On pose $u_0 = u_{n+1} = 0$. Alors

$$\begin{aligned}\langle A_n U, U\rangle &= \sum_{i=1}^{n} u_i(-u_{i-1} + 2u_i - u_{i+1}) \\ &= \sum_{i=2}^{n}(u_i - u_{i-1})u_{i-1} + \sum_{i=1}^{n}(u_i - u_{i-1})u_i + u_n^2 \\ &= \sum_{i=2}^{n}(u_i - u_{i-1})^2 + u_1^2 + u_n^2 \geqslant 0.\end{aligned}$$

De plus, si $\langle A_n U, U\rangle = 0$, alors $U = 0$ par récurrence. Donc A est définie positive. □

CONSÉQUENCE. On a $\sigma(A_n) \subset]0, +\infty[$. Comme $\rho(A_n) \leqslant \|A_n\|_\infty \leqslant 4$, on a $\sigma(A_n) \subset]0, 4]$. On peut montrer que la matrice $A_n - 4I_n$ est symétrique définie négative, donc $\sigma(A) \in]0, 4[$

PROPOSITION 1.2. On a

$$\sigma(A_n) = \left\{ 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \;\middle|\; 1 \leqslant k \leqslant n \right\}.$$

Preuve Cherchons $U \coloneqq (u_1, \dots, u_n) \in \mathbb{R}^n$ et $\lambda \in]0, 1[$ tels que $U \neq 0$ et $A_n U = \lambda U$. On résout

$$\forall i \in [\![1, n]\!], \quad -u_{i-1} + 2u_i - u_{i+1} = \lambda u_i.$$

La suite $(u_1, \dots, u_n)$ est donc solution d'une récurrence linéaire d'ordre 2. Le polynôme caractéristique de cette récurrence s'écrit $X^2 - (2 - \lambda)X + 1$ et il admet deux racines

$$r \coloneqq \frac{1}{2}(2 - \lambda) + i\sqrt{\lambda(4 - \lambda)} \quad \text{et} \quad \overline{r}.$$

Ainsi pour tout $j \in [\![1, n]\!]$, on a $u_j = \alpha r^j + \beta \overline{r}^j$ avec $\alpha, \beta \in \mathbb{C}$. Or $|r|^2 = r\overline{r} = 1$, donc il existe $\theta \in \mathbb{R}$ tel que $r = e^{i\theta}$. Les deux conditions $u_0 = 0$ et u_{n+1} permette d'obtenir α et β. En l'occurrence, on a

$$\alpha + \beta = 0 \quad \text{et} \quad \alpha e^{i(n+1)\theta} - \beta e^{-i(n+1)\theta} = 0.$$

Si $e^{i(n+1)\theta} - e^{-i(n+1)\theta} \neq 0$, alors $\alpha = 0 = \beta$, donc $U = 0$, donc $\lambda \notin \sigma(A_n)$. On suppose que $e^{i(n+1)\theta} - \beta e^{-i(n+1)\theta} = 0$. On obtient alors une droite vectorielle de solutions de la forme $u_j = \alpha(r^j - r^{-j})$ avec $\alpha \in \mathbb{C}$ pour tout $j \in [\![1, n]\!]$. Il suffit de trouver $\lambda \in]0, 4[$ tel que $r = e^{i\theta}$ vérifie $e^{i(n+1)\theta} = e^{i(n+1)\theta}$, *i. e.*

$$\theta = \frac{k\pi}{n+1}, \quad k \in \mathbb{Z}.$$

Pour résoudre $\lambda \in]0, 1[$ tel que $r = e^{ik\pi/(n+1)}$ avec $k \in \mathbb{Z}$, on observe que

$$\operatorname{Re} r = \cos\left(\frac{k\pi}{n+1}\right) = \frac{1}{2}(2 - \lambda),$$

donc

$$\lambda = 2 - 2\cos\left(\frac{k\pi}{n+1}\right) = 4\sin^2\left(\frac{k\pi}{2(n+1)}\right).$$

Parmi ces valeurs, il s'y trouve exactement n valeurs deux à deux distincts. On a donc montré que

$$\sigma(A_n) \subset \left\{ 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \;\middle|\; k \in \mathbb{Z} \right\} = \left\{ 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \;\middle|\; 1 \leqslant k \leqslant n \right\}.$$

L'inclusion inverse est donnée par la synthèse de la preuve. □

On peut donc identifier explicitement sa norme 2 qui vaut

$$\|A_n\|_2 = \rho(A_n) = 2\sin^2\left(\frac{n\pi}{2(n+1)}\right)$$

et $\|A_n^{-1}\|_2 = (\|A_n\|_2)^{-1}$ et on peut en déduire $\operatorname{cond}_2(A) = \|A_n\|_2 \|A_n^{-1}\|_2$

DÉFINITION 1.3. Un vecteur $v \in \mathbb{R}^n$ est dit *positif* si chacune de ses composantes sont positifs et on note $v \geqslant 0$. Une matrice $A \in \mathscr{M}_n(\mathbb{R})$ est dite positive si chacun de ses coefficients est positives et on note $A \geqslant 0$. De plus, on dit que A est *monotone* si elle est inversible d'inverse positif.

PROPOSITION 1.4. Soit $A \in \mathscr{M}_n(\mathbb{R})$. Alors A est monotone si et seulement si

$$\forall v \in \mathbb{R}^n, \quad Av \geqslant 0 \quad \Longrightarrow \quad v \geqslant 0.$$

Preuve On suppose que A est monotone. Soit $v \in \mathbb{R}^n$ tel que $Av \geqslant 0$. Alors $v = A^{-1}(Av) \geqslant 0$. Réciproquement, on suppose que $Av \geqslant 0 \Rightarrow v \geqslant 0$ pour tout $v \in \mathbb{R}^n$. Montrons que $\operatorname{Ker} A = \{0\}$. Soit $v \in \operatorname{Ker} A$. Alors $Av = 0$, donc $Av \geqslant 0$ et $A(-v) \geqslant 0$, donc $v \geqslant 0$ et $-v \geqslant 0$, donc $v = 0$. On en déduit que A est inversible. De plus, si e_i désigne le i-ième vecteur de la base canonique de $\mathbb{R}^n$, alors $e_i \geqslant 0$, donc $A^{-1}e_i \geqslant 0$. D'où $A^{-1} \geqslant 0$. D'où la monotonie de A. □

PROPOSITION 1.5. La matrice A_n est monotone.

Preuve Soit $v \coloneqq (v_1, \dots, v_n) \in \mathbb{R}^n$ tel que $A_n v \geqslant 0$. Montrons que $v \geqslant 0$. En posant $v_0 = v_{n+1} = 0$, on a

$$\forall i \in [\![1, n]\!], \quad -v_{i-1} + 2vi - v_{i+1} \geqslant 0.$$

La première équation donne $2v_1 - v_2 \geqslant 0$. Soit $k \in [\![1, n]\!]$ tel que $v_k = \min_{i \in [\![1,n]\!]} v_i$. Montrons que $v_k \geqslant 0$.

- Si $k = 1$, alors $2v_1 - v_2 \geqslant 0$, donc $v_1 = v_k \geqslant v_2 - v_1 \geqslant 0$.
- Si $k = n$, alors $2v_n - v_{n-1} \geqslant 0$, donc $v_n = v_k \geqslant v_{n-1} - v_n \geqslant 0$.
- Si $k \in [\![2, n-1]\!]$, alors la k-ième équation donne $(Av)_k = -v_{k+1} + 2v_k - v_{k-1} \geqslant 0$ de sorte que

$$\underbrace{(-v_{k+1} + v_k)}_{\leqslant 0} + \underbrace{(v_k - v_{k-1})}_{\leqslant 0} \geqslant 0,$$

donc $v_k = v_{k-1} = v_{k+1}$.

Par récurrence, on peut montrer que $v_1 = \cdots = v_k = \cdots = v_n$ et donc, pas le premier cas, on a $v_k \geqslant 0$. □

CONSÉQUENCE. Pour $n \geqslant 3$, on peut montrer que

$$\|A_n^{-1}\|_\infty = \begin{cases} (n+1)^2/8 & \text{si } n \text{ est impair,} \\ n(n+2)/8 & \text{sinon.} \end{cases}$$

Preuve La fonction

$$\varphi \colon \left| \begin{array}{l} [0,1] \longrightarrow \mathbb{R}, \\ \quad x \longmapsto \frac{1}{2}x(1-x) \end{array} \right.$$

est solution du problème $-\varphi'' = 1$ sur $]0,1[$ et $\varphi(0) = \varphi(1) = 0$. Soit $n \geqslant 3$. On pose $\phi \coloneqq (\phi_1, \dots, \phi_n) \in \mathbb{R}^n$ tel que

$$\forall i \in [\![1, n]\!], \quad \phi_i \coloneqq \varphi\Big(\frac{i}{n+1}\Big) = \frac{1}{2(n+1)^2} i(n+1-1).$$

En posant $\mathbb{1} \coloneqq (1, \dots, 1) \in \mathbb{R}^n$, on a $(n+1)^2 A_n \phi = \mathbb{1}$. Soit $F \in \mathbb{R}^n$. On a $\|F\|_\infty \mathbb{1} - F \geqslant 0$ et $F + \|F\|_\infty \mathbb{1} \geqslant 0$. Or $(n+1)^2 A_n$ est monotone, donc

$$0 \leqslant ((n+1)^2 A_n)^{-1}(\|F\|_\infty \mathbb{1} - F) \quad \text{et} \quad 0 \leqslant ((n+1)^2 A_n)^{-1}(F + \|F\|_\infty \mathbb{1}).$$

Cependant, on a $((n+1)^2 A_n)^{-1}\mathbb{1} = \phi$, donc

$$-\|F\|_\infty \phi \leqslant U \coloneqq ((n+1)^2 A_n)^{-1} F \leqslant \|F\|_\infty \phi,$$

donc $\|U\|_\infty \leqslant \|F\|_\infty \|\phi\|_\infty$. Finalement, on a $\|A_n^{-1}\|_\infty \leqslant (n+1)^2 \|\phi\|_\infty$. En considérant $F = \mathbb{1}$, on obtient que

$$\|A_n^{-1}\|_\infty = (n+1)^2 \|\phi\|_\infty .$$

En calculant $\|\phi\|_\infty$, on obtient le résultat voulue. □

CONVERGENCE. On considère $f \in \mathscr{C}^2([0,1])$. Soit $u \in \mathscr{C}^4([0,1])$ l'unique solution au problème (P). Soit $n \geqslant 1$. On considère

$$U^{\text{ex}} = \Big(u\Big(\frac{i}{n+1}\Big)\Big)_{1 \leqslant i \leqslant n} \quad \text{et} \quad F = \Big(f\Big(\frac{i}{n+1}\Big)\Big)_{1 \leqslant i \leqslant n}.$$

En particulier, il existe une erreur de consistance $\varepsilon \coloneqq (\varepsilon_1, \dots, \varepsilon_n) \in \mathbb{R}^n$ telle que

$$(n+1)^2 A_n U^{\text{ex}} = F + \varepsilon.$$

Par un développement de TAYLOR, il existe $(\theta_1, \dots, \theta_n) \in [-1, 1]^n$ tel que

$$\varepsilon_i = \frac{1}{(n+1)^2} \frac{1}{12} u^{(4)}\Big(\frac{i}{n+1} + \frac{\theta_i}{n+1}\Big), \qquad 1 \leqslant i \leqslant n.$$

Alors

$$\|\varepsilon\|_\infty \leqslant \frac{1}{12(n+1)^2} \max_{x \in [0,1]} |f''(x)| .$$

On note $U^{(n)} \in \mathbb{R}^n$ une solution du problème approché, *i. e.* vérifiant $(n+1)^2 A_n U^{(n)} = F$. La suite $(U^{(n)})_{n\geqslant 1}$ converge-t-elle uniformément vers U^{ex} ? Pour tout $n \geqslant 1$, on a

$$(n+1)^2 A_n (U^{\text{ex}} - U^{(n)}) = F + \varepsilon - F = \varepsilon,$$

donc

$$\|U^{\text{ex}} - U^{(n)}\|_\infty \leqslant (n+1)^{-2} \|A_n^{-1}\|_\infty \|\varepsilon\|_\infty \sim n^{-2} \frac{n^2}{8} \frac{n^{-2}}{12} \|f''\|_\infty = C n^{-1} \|f''\|_\infty \longrightarrow 0$$

avec $C \coloneqq 1/96 \in \mathbb{R}^*$.

1.2 Méthodes directes

1.2.1 Remarques préliminaires

Soient $A \in \mathrm{GL}_n(\mathbb{K})$ et $b \in \mathbb{K}^n$. Pour résoudre le système $Ax + b$, on peut le faire par la formule de Cramer et le calcul des déterminants de taille n. La complexité de ce calcul est en $O(nn!)$. Sur un supercalculateur à 200 petaflops, pour $n = 50$, il faut 2×10^{41} années. En réalité, on peut trouver des algorithmes en $O(n^3)$. Pour $n = 10\,000$, une machine classique suffit pour exécuter l'algorithme en quelques secondes.

Le cas particulier des matrices triangulaires mène à deux algorithmes : un algorithme de descente pour les matrices triangulaires inférieures en $O(n^2)$ et un de remonté pour les matrices triangulaires supérieures toujours avec la même complexité.

1.2.2 Factorisation LU

Pour résoudre le système $Ax = b$, les opérations du pivot de Gauss peuvent être stockées dans une factorisation LU de A. De même, inverser A revient à résoudre n systèmes linéaires $Ax = e_i$ où e_i est le i-ième vecteur de la base canonique de $\mathbb{K}^n$.

Définition 1.6. Soit $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{M}_n(\mathbb{K})$. Les *mineurs principaux dominants* de A sont les déterminants $\det(a_{i,j})_{1\leqslant i,j\leqslant k}$ pour $k \in [\![1, n]\!]$.

◇ Remarque. Toute matrice dont les mineurs principaux dominants sont non nuls est inversible. La réciproque est fausse en considérant

$$A \coloneqq \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Tout matrice symétrique définie positive à ses mineurs principaux dominants tous non nuls : c'est une conséquence du théorème spectral.

Lemme 1.7. Soit $A \in \mathscr{M}_n(\mathbb{K})$ dont les mineurs principaux dominants sont non nuls. Alors pour tout $\lambda \in \mathbb{K}$ et tous $i, j \in [\![1, n]\!]$ tels que $i < j$, la matrice $T(\lambda, i, j)A$ a aussi ses mineurs principaux dominants non nuls où

$$T(\lambda, i, j) \coloneqq I_n + \lambda E_{i,j}.$$

Preuve Soit $k \in [\![1, n]\!]$. On note

$$T(\lambda, i, j) = \begin{pmatrix} T_k & 0 \\ \tilde{T} & T_{n-k} \end{pmatrix} \quad \text{et} \quad A = \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix} \quad \text{avec} \quad T_k, A_k \in \mathscr{M}_k(\mathbb{K}).$$

Alors

$$T(\lambda, i, j)A = \begin{pmatrix} T_k A_k & * \\ * & * \end{pmatrix},$$

donc le mineur principal dominant de taille k est $\det(T_k A_k)$. Or la matrice T_k s'écrit

$$T_k = \begin{cases} I_k & \text{si } k \geqslant i \text{ et } k \geqslant j, \\ I_k + \lambda E_{i,j} & \text{sinon,} \end{cases}$$

donc elle est de déterminant 1. On en déduit que $\det(T_k A_k) = \det A_k$ □

Théorème 1.8. Soit $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{M}_n(\mathbb{K})$ dont les mineurs principaux dominants sont non nuls. Alors il existe un unique couple $(L, U) \in \mathrm{T}_n^\ell(\mathbb{K}) \times \mathrm{T}_n^{\mathrm{u}}(\mathbb{K})$ tel que $A = LU$ et la matrice L contienne que des uns sur sa diagonale.

Preuve • *Unicité.* Soient (L_1, U_1) et (L_2, U_2) deux tels couples. Alors $L_2^{-1}L_1 = U_2U_1^{-1}$ est à la fois triangulaire inférieure et supérieures. Comme les coefficients diagonaux de L_1 et L_2 ne sont que des uns, on a $L_2^{-1}L_1 = I_n$, donc $L_1 = L_2$ puis $U_1 = U_2$.

• *Existence.* Soit $k \in [\![1, n-1]\!]$. Supposons déterminées des matrices T_1, …, T_{k-1} chacune produit de transvections inférieures telles que

$$T_{k-1}\cdots T_1 A = \begin{pmatrix} a_{1,1}^{(k)} & & & & \\ & \ddots & & & (a_{i,j}^{(k)}) \\ & & a_{k-1,k-1}^{(k)} & & \\ & & & a_{k,k}^{(k)} & \\ & (0) & & \vdots & \\ & & & a_{n,k}^{(k)} & \end{pmatrix}.$$

Par le lemme, pour tout $j \in [\![1, k]\!]$, on a $a_{j,j}^{(k)} \neq 0$. En particulier, le pivot $a_{k,k}^{(k)}$ est non nul. Pour tout $i > k$, on peut opérer

$$L_i \longleftarrow L_i - x_{i,k}^{(k)} \quad \text{avec} \quad x_{i,k}^{(k)} \coloneqq \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.$$

On pose

$$T_k \coloneqq \prod_{i=k+1}^{n} T(-x_{i,k}^{(k)}, i, k).$$

Alors

$$T_k\cdots T_1 A = \begin{pmatrix} a_{1,1}^{(k)} & & & & \\ & \ddots & & & (a_{i,j}^{(k)}) \\ & & a_{k,k}^{(k)} & & \\ & & & & \\ & (0) & & (a_{i,j}^{(k+1)}) & \end{pmatrix}.$$

On procède alors par récurrence sur k pour obtenir des matrices T_i.

Obtenons une décomposition LU. On pose $U \coloneqq T_{n-1}\cdots T_1 A \in \mathrm{T}_n^{\mathrm{u}}(\mathbb{K})$ et $L \coloneqq (T_{n-1}\cdots T_1)^{-1} \in \mathrm{T}_n^{\ell}(\mathbb{K})$. Le couple (L, U) vérifie alors les conditions recherchées. □

Conséquence. Pour résoudre un système $Ax = b$, on calcul la décomposition LU de $A = LU$ et on résout les systèmes $Ly = b$ et $Ux = y$ en $O(n^2)$ opérations. On a stocké les opérations du pivot du second membre b dans la matrice L.

Sur le choix du pivot. Si l'un des mineurs principaux dominants est nul, disons le k-ième étant le plus petit nul, l'algorithme s'arrête à la k-ième étape car $a_{k,k}^{(k)} = 0$. Mais si A est inversible, on a $a_{i,k}^{(k)} \neq 0$ pour $i > k+1$ (sans quoi on aurait $\operatorname{rg} A \leqslant n-1$) et on pourrait permuter les lignes k et i en utilisant $a_{i,k}^{(k)}$ comme pivot. Cela donne une factorisation de A sous la forme $A = PLU$.

Si un pivot est très petit, une difficulté numérique se pose sur l'arithmétique approchée. L'erreur de calcul sur $a_{i,k}^{(k)}$ est multipliée par $|(a_{k,k}^{(k)})^{-1}| \gg 1$. On peut alors permuter pour limiter cet effet :

– on permute avec la ligne i telle que $\max_{i \geqslant k} |a_{i,k}^{(k)}|$ est réalisé sur la diagonale. ;
– on permute lignes et colonnes pour utiliser cette valeur comme pivot et on obtient une factorisation $A = PLUQ$.

1.2.3 Décomposition de Cholesky

Théorème 1.9. Soit $A \in \mathscr{S}_n^{++}(\mathbb{R})$. Alors il existe une unique matrice $B \in \mathrm{T}_n^{\ell}(\mathbb{R})$ telle que $A = B\,{}^{\mathrm{t}}B$ et les coefficients diagonaux de B sont strictement positifs.

Preuve La matrice A admet une unique factorisation $A = LU$ où les coefficients diagonaux de L valent 1. On note $U = (u_{i,j})_{1 \leqslant i,j \leqslant n}$. Alors on peut écrire $U = D\tilde{U}$ avec $D \coloneqq \operatorname{diag}(u_{i,i})_{1 \leqslant i \leqslant n}$ et $\tilde{U} \in \mathrm{T}_n^{\mathrm{u}}(\mathbb{R})$ dont les coefficients diagonaux valent 1. Or $LD\tilde{U} = A = {}^{\mathrm{t}}A = {}^{\mathrm{t}}U\,{}^{\mathrm{t}}L = {}^{\mathrm{t}}\tilde{U}\,{}^{\mathrm{t}}D\,{}^{\mathrm{t}}L$. Par unicité, on en déduit que $L = {}^{\mathrm{t}}\tilde{U}$. D'où $A = LD\,{}^{\mathrm{t}}L$. Pour tout $x \neq 0$, on a

$$0 < \langle Ax, x\rangle = \langle D\,{}^{\mathrm{t}}Lx, {}^{\mathrm{t}}Lx\rangle,$$

donc D est symétrique définie positive. On en déduit que $u_{i,i} > 0$ pour tout $i \in [\![1, n]\!]$. On pose

$$B \coloneqq L \operatorname{diag}(\sqrt{u_{i,i}})_{1 \leqslant i \leqslant n} \in \mathrm{T}_n^\ell(\mathbb{R}).$$

Par ce qui précède, la matrice B vérifient les hypothèses. □

1.3 Méthodes itératives

1.3.1 Généralité

Soient $A \in \mathrm{GL}_n(\mathbb{K})$ et $b \in \mathbb{K}^n$. On veut résoudre le système $Ax = b$. On choisit $M, N \in \mathscr{M}_n(\mathbb{K})$ telles que $A = M - N$ et M est inversible et facile à inverser. On définit alors pour un vecteur donnée $x_0 \in \mathbb{K}^n$ la suite $(x_k)_{k\in\mathbb{N}}$ vérifiant

$$Mx_{k+1} = Nx_k + b, \quad k \in \mathbb{N}.$$

Définition 1.10. On dit qu'une méthode itérative est convergente si, pour tout $x_0 \in \mathbb{K}^n$, la suite $(x_k)_{k\in\mathbb{N}}$ converge.

◇ Remarque. La matrice A étant inversible, la seule limite possible est la solution $x = A^{-1}b$.

Définition 1.11. On appelle
- *résidu* à l'étape $k \in \mathbb{N}$ le vecteur $r_k \coloneqq b - Ax_k$;
- *erreur* à l'étape $k \in \mathbb{N}$ le vecteur $e_k \coloneqq x - x_k$

Alors $x_k \to x$ si et seulement si $e_k \to 0$ si et seulement si $r_k \to 0$.

◇ Remarque. Pour tout $k \in \mathbb{N}$, on a $e_k = A^{-1}r_k$, donc $\|e_k\| \leqslant |||A^{-1}||| \, \|r_k\|$.

Théorème 1.12. La méthode itérative pour $A = M - N$ est convergente si et seulement si $\rho(M^{-1}N) < 1$.

Preuve Pour tout $k \in \mathbb{N}$, on a $x_{k+1} = M^{-1}Nx_k + M^{-1}b$ et $e_{k+1} = M^{-1}Ne_k$. On en déduit que $e_k = (M^{-1}N)^k e_0$ pour tout $k \in \mathbb{N}$. On souhaite avoir que $e_n \to 0$ pour tout $e_0 \in \mathbb{K}^n$, *i. e.* $(M^{-1}N)^k \to 0$, *i. e.* $\rho(M^{-1}N) < 1$. □

◇ Remarque. La vitesse de convergence est au plus géométrique et est liée à la quantité $\rho(M^{-1}N)$. En effet, soit $\varepsilon > 0$. Il existe une norme $\|\ \|$ sur $\mathbb{K}^n$ telle que $\left|||M^{-1}N|||\right| \leqslant \rho(M^{-1}N) = \varepsilon$, donc

$$\forall k \in \mathbb{N}, \quad \|e_k\| \leqslant |||(M^{-1}N)^k||| \, \|e_0\| \leqslant (\rho(M^{-1}N) + \varepsilon)^k \, \|e_0\| \, .$$

1.3.2 Méthodes usuelles

(i) Méthode de Richardson

Soit $\alpha \in \mathbb{R}^*$. On pose $M \coloneqq \alpha^{-1}I_n$ et $N \coloneqq \alpha^{-1}I_n - A$ de sorte que $M - N = A$. La suite récurrente est alors définie par

$$x_{k+1} = x_k + \alpha(b - Ax_k), \quad k \in \mathbb{N}.$$

De plus, on a $M^{-1}N = I_n - \alpha A$ et $\sigma(M^{-1}N) = \{1 - \alpha\lambda \mid \lambda \in \sigma(A)\}$. Supposons que $A \in \mathscr{S}_n^{++}(\mathbb{R})$. On note $0 < \lambda_1 \leqslant \cdots \leqslant \lambda_n$ ses valeurs propres. En faisant un dessin, on remarque que

$$\rho(M^{-1}N) = \begin{cases} |1 - \alpha\lambda_n| & \text{si } \alpha \notin [0, \alpha_*], \\ |1 - \alpha\lambda_1| & \text{sinon} \end{cases} \quad \text{avec} \quad \alpha_* \coloneqq \frac{2}{\lambda_n + \lambda_n}.$$

En particulier, le rayon spectral de $M^{-1}N$ est strictement inférieur à 1 si et seulement si $\alpha \in]0, 2/\rho(A)[$. De plus, il est minimal lorsque $\alpha = \alpha_*$ et il vaut alors

$$\rho(M^{-1}N) = \frac{\lambda_n - \lambda_1}{\lambda_n + \lambda_1} = \frac{\operatorname{cond}_2 A - 1}{\operatorname{cond}_2 A + 1}.$$

(ii) Méthode de Jacob

On pose $M \in \mathscr{M}_n(\mathbb{R})$ la diagonale D de A et on suppose que D est inversible. La matrice d'itération s'écrit alors $M^{-1}N = I_n - D^{-1}A$. En pratique, l'algorithme « en place » prend la forme suivante. Soit $x_0^{(0)} \in \mathbb{R}^n$. Pour $k \in \mathbb{N}$ et $i \in [\![1, n]\!]$, on pose

$$x_i^{(k+1)} \coloneqq \frac{1}{a_{i,i}}\Big(b_i - \sum_{j\neq i} a_{i,j}x_j^{(k)}\Big).$$

(iii) Méthode de Gauss-Seidel

On pose $D \in \mathscr{M}_n(\mathbb{R})$ la diagonale de A, $E \coloneqq (-a_{i,j}\mathbb{1}_{i>j})_{1\leqslant i,j\leqslant n}$ et $F \coloneqq (-a_{i,j}\mathbb{1}_{i<j})_{1\leqslant i,j\leqslant n}$. On pose alors $M \coloneqq D - E$ et $N \coloneqq F$. Pour $k \in \mathbb{N}$ et $i \in [\![1, n]\!]$, on pose

$$x_i^{(k+1)} \coloneqq \frac{1}{a_{i,i}}\Big(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j}x_j^{(k)}\Big).$$

(iv) Méthode de relaxation (SOR)

Soit $\omega > 0$. On pose $M = \omega^{-1}D - E$ et $N \coloneqq (1-\omega)\omega^{-1}D + F$. Alors $M - N = D - E - F$.

1.3.3 Critères explicites de convergences

(i) Matrices à diagonale dominante

Définition 1.13. Une matrice $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{M}_n(\mathbb{C})$ est à diagonale strictement dominante si

$$\forall i \in [\![1, n]\!], \quad |a_{i,i}| > \sum_{j\neq i} |a_{i,j}|.$$

◇ Remarque. Toute telle matrice a une diagonale inversible et les algorithmes précédents peuvent s'appliquer.

Proposition 1.14. Soit $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{M}_n(\mathbb{C})$ à diagonale strictement dominante. Alors A est inversible.

Preuve Soit $x \coloneqq (x_1, \dots, x_n) \in \operatorname{Ker} A$. Alors pour tout $i \in [\![1, n]\!]$, on a

$$\sum_{j=1}^{n} a_{i,j}x_j = 0, \quad \text{donc} \quad a_{i,i}x_i = -\sum_{j\neq i} a_{i,j}x_j,$$

donc

$$|a_{i,i}|\,|x_i| \leqslant \Big(\sum_{j\neq i} |a_{i,j}|\Big)\,\|x\|_\infty.$$

Or il existe $i \in [\![1, n]\!]$ tel que $|x_i| = \|x\|_\infty$, donc

$$\underbrace{\Big(\sum_{j\neq i} |a_{i,j}| - |a_{i,i}|\Big)}_{>0}\,\|x\|_\infty \geqslant 0,$$

donc $x = 0$. On en déduit que A est inversible. □

Théorème 1.15. Soit $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{M}_n(\mathbb{C})$ à diagonale strictement dominante. Alors les méthodes de Jacobi, Gauss-Seidel et de relation pour $\omega \in]0, 1]$ convergent.

Preuve Montrons le résultat pour la méthode de Jacobi. On a $M^{-1}N = I_n - D^{-1}A$. Soit $\lambda \in \sigma(M^{-1}N)$ telle que $|\lambda| = \rho(M^{-1}N)$. On note $x \in \mathbb{C}^n$ un vecteur propre associé. Alors $Nx = \lambda Mx$, donc $(D - A)x = \lambda Dx$. On considère $i \in [\![1, n]\!]$ telle que $|x_i| = \|x\|_\infty \neq 0$. Alors

$$-\sum_{j\neq i} a_{i,j}x_j = \lambda a_{i,i}x_i, \quad \text{donc} \quad \lambda = -\frac{1}{a_{i,i}x_i}\sum_{j\neq i} a_{i,j}x_j.$$

L'inégalité triangulaire donne alors

$$|\lambda| \leqslant \frac{1}{|a_{i,j}|\,\|x\|_\infty}\sum_{j\neq i} |a_{i,j}|\,|x_j| \leqslant \frac{1}{|a_{i,j}|}\sum_{j\neq i} |a_{i,j}| < 1.$$

Cela montre que $\rho(M^{-1}N) < 1$ ce qui conclut.. □

(ii) Matrices hermitiennes définies positives

Théorème 1.16. Soit $A \in \mathscr{H}_n^{++}(\mathbb{C})$. On suppose qu'il existe $M \in \mathrm{GL}_n(\mathbb{C})$ telle que $A = M - N$. Alors $M^* + N$ est hermitienne. De plus, si $M^* + N$ est hermitienne définie positive, alors $\rho(M^{-1}N) < 1$.

Preuve La matrice $M^* + N$ est clairement hermitienne. On suppose qu'elle est hermitienne définie positive. Il suffit de trouver une norme subordonnée $|||\ |||$ sur $\mathscr{M}_n(\mathbb{C})$ telle que $|||M^{-1}N||| < 1$. La produit scalaire induit par la matrice A

$$(x, y) \longmapsto \langle x, y\rangle_A \coloneqq \langle Ax, y\rangle$$

induit une norme et, par conséquent, une norme subordonnée $|||\ |||_A$. Soit $x \in \mathbb{C}^n$ tel que $\|x\|_A = 1$. Alors

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

On pose $y \coloneqq M^{-1}Ax$. Alors

$$\begin{aligned}\|M^{-1}Nx\|_A^2 &= \langle Ax - Ay, x - y\rangle \\ &= \langle Ax, x\rangle - \langle Ay, x\rangle - \langle Ax, y\rangle + \langle Ay, y\rangle \\ &= 1 - \langle y, Ax\rangle - \langle Ny, y\rangle \\ &= 1 - \langle y, My\rangle - \langle Ny, y\rangle \\ &= 1 - \langle M^*y, y\rangle - \langle Ny, y\rangle \\ &= 1 - \langle (M^* + N)y, y\rangle.\end{aligned}$$

Or comme M et A sont inversibles et $x \neq 0$, on a $y \neq 0$. Donc $\|M^{-1}Nx\|_A^2 < 1$ et ceci pour tout $x \in \mathbb{C}^n$ tel que $\|x\|_A = 1$, donc $|||M^{-1}N||| < 1$. □

1.4 Résolution de systèmes linéaires au sens des moindres carrées

Soient $n, p \in \mathbb{N}^*$, $A \in \mathscr{M}_{n,p}(\mathbb{K})$ et $b \in \mathbb{K}^n$. On cherche un vecteur $x \in \mathbb{K}^p$ tel que $Ax = b$. On dit que le système linéaire $Ax = b$ est

- *carré* si $n = p$;
- *surdéterminé* si $n > p$;
- *sousdéterminé* si $n < p$;
- de *rang maximal* si $\operatorname{rg} A = \min(n, p)$;
- *incompatible* si $b \notin \operatorname{Im} A$;
- *compatible* si $b \in \operatorname{Im} A$.

Si le système est compatible, alors il existe des solutions : si $\operatorname{rg} A = p$, il existe une unique solution et, si $\operatorname{rg} A < p$, il existe une infinité de solutions. En particulier, s'il est surdéterminé, de rang maximal et compatible, alors il existe une unique solution. S'il est sous-déterminé de rang maximal, alors il est compatible et il existe une infinité de solutions.

Question. Que se passe-t-il si le système est incompatible ?

Définition 1.17. Résoudre le système $Ax = b$ au sens des moindres carrés signifie trouver une vecteur $x \in \mathbb{K}^p$ minimisant $\|Ax - b\|_2^2$

1.4.1 Existence et unicité

Théorème 1.18. L'ensemble des solutions du système $Ax = b$ au sens des moindres carrés est

$$\{x \in \mathbb{K}^p \mid Ax = p(b)\}$$

où l'application p est la projection orthogonale sur $\operatorname{Im} A$.

Preuve Comme $p(b) \in \operatorname{Im} A$, la caractérisation du projeté donne

$$\forall w \in \operatorname{Im} A, \quad \langle p(b) - b, w\rangle = 0.$$

Soit $v \in \operatorname{Im} A$. Le théorème de Pythagore donne alors

$$\|v - b\|^2 = \|v - p(b)\|^2 + \|p(b) - b\|^2 \geqslant \|p(b) - b\|^2$$

où l'égalité est vraie si et seulement si $v = p(b)$. De plus, il existe $x_0 \in \mathbb{K}^p$ tel que $Ax_0 = p(b)$. L'unicité est garantie si et seulement si $\operatorname{Ker} A = \{0\}$. □

1.4.2 Équation normale

LEMME 1.19. Toute solution du système $Ax = b$ au sens des moindres carrés est solution du système $A^*Ax = A^*b$ et réciproquement.

Preuve Soient $x, y \in \mathbb{K}^p$. Alors

$$\begin{aligned}\|Ay - b\|^2 = \|Ay - Ax + Ax - b\|^2 &= \|Ay - Ax\|^2 + \|Ax - b\|^2 + 2\langle Ay - Ax, Ax - b\rangle \\ &= \|Ay - Ax\|^2 + \|Ax - b\|^2 + 2\langle y - x, A^*Ax - A^*b\rangle.\end{aligned}$$

Si le vecteur x est solution de l'équation $A^*Ax = A^*b$, alors $\|Ay - b\|^2 \geqslant \|Ax - b\|^2$ pour tout $y \in \mathbb{R}^p$, donc il est solution au sens des moindres carrés. Réciproquement, on suppose que le vecteur x est solution au sens des moindres carrés. On écrit $y = x + tz$ avec $t \in \mathbb{R}$ et $z \in \mathbb{K}^p$. Or

$$\|Ay - b\|^2 = \|Ax - b\|^2 + 2t\langle z, A^*Ax - A^*b\rangle + t^2\|Az\|^2.$$

Si $t > 0$, alors $2\langle z, A^*Ax - A^*b\rangle + t\|Az\|^2 \geqslant 0$. En laissant tendre t vers 0, on a $\langle z, A^*Ax - A^*b\rangle \geqslant 0$. De même, on considérant le cas $t < 0$, on montre que $\langle z, A^*Ax - A^*b\rangle = 0$. Ceci est vrai pour tout $z \in \mathbb{K}^p$. On en déduit alors que $A^*Ax = A^*b$. □

◇ REMARQUE. L'équation normale $A^*Ax = A^*b$ fait intervenir une matrice A^*A qui est hermitienne et positive. De plus, son noyau est égal à celui de A. Ainsi le système carré $AA^*x = b$ est inversible et il existe une unique solution pour les moindres carrés.

1.4.3 Exemple de la régression linéaire

Soit $(x_i, y_i)_{1\leqslant i\leqslant n}$ une famille de points de $\mathbb{R}^2$. On cherche un couple $(a, b) \in \mathbb{R}^2$ tel que

$$\forall i \in [\![1, n]\!], \quad ax_i + b = y_i.$$

Matriciellement, le problème s'écrit sous la forme $A(a, b) = b$ avec

$$A \coloneqq \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \quad \text{et} \quad b \coloneqq \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

L'équation normale fait intervenir la matrice et le vecteur

$$A^*A = \begin{pmatrix} \sum_{i=1}^n x_i^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n 1 \end{pmatrix} \quad \text{et} \quad A^*b = \begin{pmatrix} \sum_{i=1}^n x_iy_i \\ \sum_{i=1}^n y_i \end{pmatrix}.$$

De plus, on a $\operatorname{Ker} A = \operatorname{Ker} A^*A = \{0\}$ dès que $n \geqslant 2$ et qu'au moins deux des réels x_i sont distincts. Après résolution de l'équation normale, on obtient que

$$\begin{cases} a = \dfrac{n\sum_{i=1}^n x_iy_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \dfrac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}, \\ b = \overline{y} - a\overline{x} \end{cases} \quad \text{avec} \quad \overline{x} \coloneqq \frac{1}{n}\sum_{i=1}^n x_i^2 \quad \text{et} \quad \overline{y} \coloneqq \frac{1}{n}\sum_{i=1}^n y_i^2.$$

1.4.4 Factorisation QR par les matrices de HOUSEHOLDER

PRINCIPE. On veut utiliser des symétries orthogonales pour produire une factorisation QR.

◇ REMARQUE. Si $Q \in \mathscr{H}_n(\mathbb{C}) \cap \mathrm{U}_n(\mathbb{C})$, alors $Q^{-1} = Q^* = Q$

DÉFINITION 1.20. On appelle *matrice de HOUSEHOLDER* associée à un vecteur $v \in \mathbb{R}^n$ la matrice

$$H(v) \coloneqq \begin{cases} I_n & \text{si } v = 0, \\ I_n - 2\dfrac{v\,{}^t v}{{}^t v v} & \text{sinon.} \end{cases}$$

PROPOSITION 1.21. Soit $v \in \mathbb{R}^n$. Alors

1. la matrice $H(v)$ est symétrique et orthogonale ;
2. si $v \neq 0$, alors $H(v)$ est la matrice de la symétrie orthogonale sur $v^\perp$ parallèlement à $\operatorname{Vect} v$;
3. pour tout $e \in \mathbb{R}^n$ tel que $e^*e = 1$, on a $H(v + \|v\|e)v = -\|v\|e$ et $H(v - \|v\|e)v = \|v\|e$.

NOUVEL ALGORITHME D'ÉLIMINATION. Soit $A \in \mathscr{M}_{n,p}(\mathbb{R})$. On note $a^{(1)} \in \mathbb{R}^n$ la première colonne de A et $e_1 \in \mathbb{R}^n$ le premier vecteur de la base canonique de $\mathbb{R}^n$. On pose $H^{(1)} \coloneqq H(a^{(1)} - \|a^{(1)}\|a_1)$. Alors

$$H^{(1)}a^{(1)} = \|a^{(1)}\|e_1 \quad \text{et} \quad H^{(1)}A = \begin{pmatrix} \|a^{(1)}\| & \\ \vdots & * \\ 0 & \end{pmatrix}$$

où cette dernière matrice est du même rang que A. Supposons, à l'étape $k \in [\![1, n]\!]$, qu'on a

$$H^{(k-1)} \cdots H^{(1)}A = \begin{pmatrix} T & & * \\ & | & \\ (0) & a^{(k)} & * \\ & | & \end{pmatrix}$$

où T est triangulaire supérieure et $a^{(k)} \in \mathbb{R}^{n-k+1}$ est non nul si $\operatorname{rg} A > k$. On pose alors

$$H^{(k)} = \begin{pmatrix} I_{k-1} & 0 \\ 0 & H(a^{(k)} - \|a^{(k)}\|(1, 0, \dots, 0)) \end{pmatrix}$$

de sorte que

$$H^{(k)} \cdots H^{(1)}A = \begin{pmatrix} T & & * \\ & \|a^{(k+1)}\| & \\ & 0 & \\ (0) & \vdots & * \\ & 0 & \end{pmatrix}.$$

◇ REMARQUES. – Si $n = p$ et A est inversible, on obtient $A = QR$ avec $R \in \mathrm{T}_n^{\mathrm{u}}(\mathbb{R})$ et $Q = H^{(1)} \cdots H^{(n)} \in \mathrm{O}_n(\mathbb{R})$. On a également $R \in \mathrm{GL}_n(\mathbb{R})$.

– Si $n > p$ et $\operatorname{rg} A = p$, on obtient $A = QR$ avec $Q \in \mathrm{O}_n(\mathbb{R})$ et $R \in \mathscr{M}_{n,p}(\mathbb{R})$. La matrice R peut être mise sous la forme

$$R = \begin{pmatrix} R_1 \\ 0_{n-p} \end{pmatrix}$$

avec $R_1 \in \mathrm{T}_p^{\mathrm{u}}(\mathbb{R}) \cap \mathrm{GL}_p(\mathbb{R})$ et la matrice Q peut se décomposer sous la forme $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$. De cette manière, on a $A = Q_1R_1$. De plus, pour tout $x \in \mathbb{R}^p$, on a

$$\begin{aligned} \|Ax - b\|^2 &= \|QRx - b\|^2 \\ &= \|Rx - Q^*b\|^2 \\ &= \|R_1x - Q_1^*b\|^2 + \|Q_2^*b\|^2. \end{aligned}$$

On trouve donc la solution au sens des moindre carrées qui est $x = R_1^{-1}Q_1^*b$ et le résidé est $\|Ax - b\| = \|Q_2^*b\|$.

– Si $n > p$ et $r \coloneqq \operatorname{rg} A < p$, on obtient $Q = QR$ où la matrice R se met sous la forme

$$R = \begin{pmatrix} R_1 & R_2 \\ 0 & 0 \end{pmatrix}$$

avec $R_1 \in \mathrm{T}_r^{\mathrm{u}}(\mathbb{R})$ et $R_2 \in \mathscr{M}_{r,p-r}(\mathbb{R})$ et la matrice Q se met sous la forme $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$ avec $Q_1 \in \mathscr{M}_{n,r}(\mathbb{R})$. Alors, pour tout $x \coloneqq (x_1, x_2) \in \mathbb{R}^r \times \mathbb{R}^{p-r}$, on a

$$\|Rx - Q^*b\|^2 = \|R_1x_1 + R_2x_2 - Q_1^*b\|^2 + \|Q_2^*b\|^2.$$

Cette quantité est minimale pour $x_1 = R_1^{-1}Q_1^*b - R_1^{-1}R_2x_22$, *i. e.* pour

$$x = \begin{pmatrix} R_1^{-1}Q_1^*b \\ 0 \end{pmatrix} + \underbrace{\begin{pmatrix} -R_1^{-1}R_2x_2 \\ x_2 \end{pmatrix}}_{\in \operatorname{Ker} A} \quad \text{avec} \quad x_2 \in \mathbb{R}^{p-r}.$$

Un défaut d'unicité apparaît par la présence du vecteur $x_2 \in \mathbb{R}^{p-r}$ en paramètre qui est traité en choisissant la solution de norme minimal

1.4.5 Décomposition en valeurs singulières

DÉFINITION 1.22. Soit $A \in \mathscr{M}_{n,p}(\mathbb{C})$. Les *valeurs singulières* sont les réels positives $\sigma_i = \sqrt{\lambda_i}$ où les λ_i sont les valeurs propres réelles non nulles de A^*A.

◇ REMARQUE. Comme les matrices A et A^*A ont le même noyau, elles ont le même rang, donc la matrice A^*A admet exactement $r \coloneqq \operatorname{rg} A$ valeurs singulières comptées avec multiplicité.

THÉORÈME 1.23. Soit $A \in \mathscr{M}_{n,p}(\mathbb{C})$ de rang $r \leqslant \min(n,p)$. Alors A admet exactement r valeurs singulières comptées avec multiplicité et il existe $U \in \mathrm{U}_n(\mathbb{C})$ et $V \in \mathrm{U}_p(\mathbb{C})$ telles que $A = U\Sigma V^*$ avec

$$\Sigma \coloneqq \mathrm{diag}(\sigma_1, \dots, \sigma_r, 0) \in \mathscr{M}_{n,p}(\mathbb{R})$$

où les réels σ_i sont les valeurs singulières de A.

DÉFINITION 1.24. On appelle *pseudo-inverse* d'une matrice $A \in \mathscr{M}_{p,n}(\mathbb{C})$ la matrice

$$A^\dagger \coloneqq V\Sigma^\dagger U^* \quad \text{avec} \quad \Sigma^\dagger \coloneqq \mathrm{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0) \in \mathscr{M}_{p,n}(\mathbb{R})$$

en reprenant les notations du théorème précédent.

◇ REMARQUE. La pseudo-inverse ne dépend pas du choix de la décomposition en valeurs singulières.

APPLICATION AUX MOINDRES CARRÉS. On écrit $A = U\Sigma V^*$ la décomposition en valeurs singulières de A. Soit $x \in \mathbb{C}^p$. Alors $\|Ax - b\| = \|U^*Ax - U^*b\| = \|\Sigma V^*x - U^*b\|$. On pose $c \coloneqq U^*b \in \mathbb{C}^n$ et $y \coloneqq V^*x \in \mathbb{C}^p$. Alors

$$\|\Sigma y - c\|^2 = \sum_{i=1}^{r} |\sigma_i y_i - c_i|^2 + \sum_{i=r+1}^{n} |c_i|^2$$

et ce terme est minimal si $y_i = \sigma_i^{-1} c_i$ pour tout $i \in [\![1, r]\!]$, *i. e.* $y = \Sigma^\dagger c + y_0$ avec $y_0 \coloneqq (0, \dots, 0, y_{r+1}, \dots, y_p) \in \mathbb{C}^p$. D'où $x = A^\dagger b + V y_0$. Parmi ces solutions, le vecteur x est de norme minimale si et seulement si le vecteur y l'est avec

$$\|y\|^2 = \sum_{i=1}^{r} |\sigma_i^{-1} c_i|^2 + \sum_{i=r+1}^{n} |c_i|^2,$$

i. e. si et seulement si $x = A^\dagger b$.

1.5 Méthodes variationnelles

1.5.1 Principe

Soient $A \in \mathscr{S}_n(\mathbb{R})$ et $b \in \mathbb{R}^n$. On pose $f \colon x \in \mathbb{R}^n \longmapsto \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle$. La fonction f est de classe $\mathscr{C}^\infty$. Le gradient de f s'écrit $\nabla f(x) = Ax - b$ pour tout $x \in \mathbb{R}^n$.

PROPOSITION 1.25. 1. Si $A \in \mathscr{S}_n^{++}(\mathbb{R})$, alors f admet un unique minimum sur $\mathbb{R}^n$ qui est solution de $Ax = b$.
2. Si $A \in \mathscr{S}_n^{+}(\mathbb{R}) \setminus \mathscr{S}_n^{++}(\mathbb{R})$ et $b \in \mathrm{Im}\, A$, alors f atteint son minimum sur $x_0 + \mathrm{Ker}\, A$ avec $Ax_0 = b$.
3. Si A n'est pas positive ou $b \notin \mathrm{Im}\, A$, alors $\inf_{x\in\mathbb{R}^n} f(x) = -\infty$.

Preuve 1. Soit $x \in \mathbb{R}^n$ tel que $Ax = b$. Alors pour $h \in \mathbb{R}^n$, on a $f(x+h) = f(x) + \frac{1}{2}\langle Ah, h\rangle$ avec $\langle Ah, h\rangle \geqslant 0$. Donc le vecteur $x = A^{-1}b$ est l'unique minimum de f sur $\mathbb{R}^n$. Réciproquement, soit $x \in \mathbb{R}^n$ tel que $Ax \neq b$. Soit $\alpha \in \mathbb{R}$. On pose $h \coloneqq \alpha(Ax - b) \in \mathbb{R}^n$. Alors

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

Pour $\alpha > 0$ assez proche de 0, on a $f(x+h) < f(x)$, donc x n'est pas un minimum. □

1.5.2 Algorithme du gradient à pas fixe

Soit $\alpha \in \mathbb{R}$. On considère une suite $(x_j)_{j\in\mathbb{N}}$ de $\mathbb{R}^n$ définie par

$$\forall j \in \mathbb{N}, \quad x_{j+1} = x_j - \alpha\nabla f(x_j).$$

On a $\rho(I_n - \alpha A) \Leftrightarrow \alpha \in]0, 2/\rho(A)[$. Ainsi la suite converge et cette méthode est optimale si $\rho(I_n - \alpha A)$ est minimal, *i. e.* si $\alpha = 2/(\lambda_{\min} + \lambda_{\max})$.

1.5.3 Algorithme du gradient à pas optimal

On considère une suite $(x_j)_{j\in\mathbb{N}}$ de $\mathbb{R}^n$ définie par

$$\forall j \in \mathbb{N}, \quad x_{j+1} = x_j - \alpha_j\nabla f(x_j) \quad \text{avec} \quad \alpha_j \coloneqq \mathrm{argmin}_{\alpha\in\mathbb{R}}\, f(x_j - \alpha\nabla f(x_j)).$$

◇ Remarque. On suppose que $A \in \mathscr{S}_n^{++}(\mathbb{R})$. Soit $j \in [\![1, n]\!]$. Alors pour tout $\alpha \in \mathbb{R}$, on a

$$g(\alpha) \coloneqq f(x_j - \alpha \nabla f(x_j)) = f(x) - \alpha \|\nabla f(x_j)\|^2 + \frac{\alpha^2}{2} \langle A \nabla f(x_j), \nabla f(x_j) \rangle.$$

Le minimum de g est atteint pour $g'(\alpha) = 0$. On en déduit que

$$\alpha_j = \frac{\|\nabla f(x_j)\|^2}{\langle A \nabla f(x_j), \nabla f(x_j) \rangle}.$$

Proposition 1.26. On note $\kappa \coloneqq \operatorname{cond}_2 A$. Soit $x \in \mathbb{C}^n$ une solution de $Ax = b$. Pour tout $j \in \mathbb{N}$, on a

$$\|x_{j+1} - x\|_A \leqslant \frac{\kappa - 1}{\kappa + 1} \|x_j - x\|_A.$$

Preuve Soit $j \in \mathbb{N}$. On note $e_j \coloneqq x_j - x$. Alors

$$\begin{aligned} \|e_{j+1}\|_A &= \min_{\alpha \in \mathbb{R}} \|x_j - \alpha(Ax_j - b) - x\|_A \\ &\leqslant \|e_{j+1}^{\mathrm{R}}\|_A \quad \text{avec} \quad e_{j+1}^{\mathrm{R}} \coloneqq x_j - \alpha_{\mathrm{R}}(Ax_j - b) - x \end{aligned}$$

où $\alpha_{\mathrm{R}} \in \mathbb{R}$ est la constante optimale dans la méthode itérative de Richardson. On a

$$\begin{aligned} \|e_{j+1}^{\mathrm{R}}\|_A = \|A^{1/2} e_{j+1}^{\mathrm{R}}\|_2 &= \|A^{1/2} R A^{-1/2} A^{1/2} e_j\|_A \quad \text{avec} \quad R \coloneqq I_n - \alpha_{\mathrm{R}} A \\ &\leqslant \|A^{1/2} R A^{-1/2}\|_2 \|A^{1/2} e_j\|_2 \\ &= \rho(A^{1/2} R A^{-1/2}) \|e_j\|_A \end{aligned}$$

avec $\rho(A^{1/2} R A^{-1/2}) = \rho(R) = (\kappa - 1)/(\kappa + 1)$. □

1.5.4 Espace de Krylov

Définition 1.27. Soient $r \in \mathbb{R}^n$, $j \in \mathbb{N}^*$ et $A \in \mathscr{M}_n(\mathbb{R})$. On appelle *espace de Krylov* de A associé à r et d'ordre j l'espace

$$K_j(A, r) \coloneqq \operatorname{Vect}(r, Ar, \dots, A^{j-1} r).$$

On note $K_0 = \{0\}$.

Proposition 1.28. Soient $r \in \mathbb{R}^n$ et $A \in \mathscr{M}_n(\mathbb{R})$. Alors la suite $(K_j)_{j \geqslant 0}$ vérifie

$$K_0 \subsetneq K_1 \subsetneq \cdots \subsetneq K_\ell = K_{\ell+1} = \cdots$$

pour un certain entier $\ell \leqslant n$. En particulier, pour tout $j \geqslant 0$, on a

$$\dim K_j = \begin{cases} j & \text{si } j \leqslant \ell, \\ \ell & \text{sinon.} \end{cases}$$

◇ Remarque. Pour les méthodes du gradient à pas fixe et optimal, la suite $(x_j)_{j \geqslant 0}$ vérifie

$$b - Ax_j \in K_{j+1}(A, r_0) \quad \text{et} \quad x_j \in x_0 + K_j(A, r_0)$$

pour tout $j \geqslant 0$.

Question. Peut-on modifier les algorithmes précédents de sorte que $x_j = x = A^{-1} b$ pour $j \geqslant 0$ assez grand avec $x_j \in x_0 + K_j(A, r_0)$, *i. e.* il existe $p \in \mathbb{R}[X]$ tel que $x - x_0 = p(A) r_0$?

Proposition 1.29. Soient $A \in \mathrm{GL}_n(\mathbb{R})$ et $x_0 \in \mathbb{R}^n$. Alors $A^{-1} b \in x_0 + K_\ell(A, r_0)$ avec $r_0 \coloneqq b - Ax_0$.

1.5.5 Algorithme du gradient conjugué

Idée. On veut construire récursivement les projetés A-orthogonaux du vecteur $x = A^{-1} b$ sur les sous-espaces affines $x_0 + K_j(A, r_0)$ pour $j \geqslant 0$ à partir d'un choix $x_0 \in \mathbb{R}^n$.

Soit $x_0 \in \mathbb{R}^n$. On pose $r_0 \coloneqq b - Ax_0$. On considère $p_0 \coloneqq r_0$ une direction de descente. Alors la famille (p_0) est une base A-orthogonale de $K_1(A, r_0)$. L'objectif est de construire une base $(p_0, \dots, p_{\ell-1})$ A-orthogonale de $K_\ell(A, r_0)$ en drapeau. Soit $j \in \mathbb{N}$. Connaissant x_j, on cherche $x_{j+1} = x_j + \alpha_j p_j$ pour un réel $\alpha_j \in \mathbb{R}$ bien choisi. On sait que $x_j = p_{E_j}(x_0)$ avec $E_j = x_0 + K_j(A, r_0)$. Pour $i \leqslant j - 1$, on remarque que $\langle x_{j+1} - x, p_i \rangle_A = 0$ et

$$\langle x_{j+1} - x, p_j \rangle_A = \alpha_j \|p_j\|_A^2 + \langle x_j - x, p_0 \rangle_A = \alpha_j \|p_j\|_A^2 + \langle Ax_j - b, p_j \rangle_2.$$

On pose alors

$$\alpha_j \coloneqq \frac{\langle r_j, p_j \rangle}{\|p_j\|_A^2} \quad \text{avec} \quad r_j \coloneqq b - Ax_j.$$

Comment trouver la suite $(p_j)_{j\in\mathbb{N}}$? On utilise le procédé de Gram-Schmidt pour déduire p_j de la base $(p_0, \dots, p_{j-1})$ de K_j et de r_j qui complète cette base en une base de K_{j+1}. Or pour $i \leqslant j-2$, on a $\langle r_j, p_i\rangle_A = \langle r_j, Ap_i\rangle_A = 0$ car $Ap_i \in K_{j-1}$ et $r_j \perp_A K_j$. Donc on pose $p_i \coloneqq r_j - \beta_{j-1} rj - 1$ avec

$$\beta_{j-1} = \frac{\langle r_j, p_{j-1}\rangle_A}{\|p_{j-1}\|_A^2}.$$

Chapitre 2

Approximation spectrale

2.1 Méthodes numériques

2.1.1 Méthode de la puissance

Idée. Soit $A \in \mathscr{M}_n(\mathbb{C})$. Pour trouver la valeur propre $\lambda \in \mathbb{C}$ tel que $|\lambda| = \rho(A)$, on s'appuie sur l'observation vague $A^k x \simeq \lambda^k x$ pour tout $k \in \mathbb{N}^*$ et tout $x \in \mathbb{C}^n$.

Algorithme. Soit $(x_k)_{k\geqslant 0} \in \mathbb{C}^{\mathbb{N}}$ telle que

$$\forall k \geqslant 0, \quad x_{k+1} = \frac{Ax_k}{\|Ax_k\|_2}.$$

Pour $k \geqslant 0$, on pose $\nu_k \coloneqq \langle x_k, Ax_k\rangle$.

Proposition 2.1. On note $\lambda_1, \ldots, \lambda_d$ les valeurs propres de A telles que $|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_d|$. On note

$$\pi_A = \prod_{i=1}^{d} (X - \lambda_i)^{n_i}.$$

On suppose que $x_0 \in \mathbb{C}^n \setminus \bigoplus_{i=2}^{d} \operatorname{Ker}(A - \lambda_i I_n)^{n_i}$. Alors les suites $(x_k)_{k\geqslant 0}$ et $(\nu_k)_{k\geqslant 0}$ sont bien définies et

$$\lim_{k\to+\infty} \nu_k = \lambda_1 \quad \text{et} \quad \lim_{k\to+\infty} \left(\frac{\overline{\lambda_1}}{|\lambda_1|}\right)^k x_k \in \operatorname{Ker}(A - \lambda_1 I_n).$$

Méthode de la puissance inverse avec translation. Étant donnée $A \in \mathscr{M}_n(\mathbb{C})$, la méthode précédente se généralise de sorte à identifier d'autres valeurs propres de A et vecteurs propres associés, par exemple la valeur propre de plus petit module de A. Pour ce faire, il suffit d'appliquer l'algorithme à la matrice A^{-1} dans le cas où A est inversible, ce qui revient en pratique à résoudre à chaque itération un système de la forme $Ax = b$. En considérant des translations dans le plan complexe, on peut encore déterminer la valeur propre de A la plus proche d'un complexe $\mu \in \mathbb{C}$ donné, pourvu que les hypothèses du théorème s'appliquent à $(A - \mu I_n)^{-1}$. En particulier, cela suppose que $\mu \notin \sigma(A)$.

2.1.2 Forme de Hessenberg

Le théorème de Schur garantit l'existence d'une base orthonormée dans laquelle un endomorphisme de $\mathbb{C}^n$ admet une matrice triangulaire supérieure. En pratique, la construction d'une telle base n'est pas accessible par une méthode directe (pas plus que la diagonalisation ne l'est). On peut néanmoins s'approcher dans un premier temps de la forme triangulaire par une mise sous forme Hessenberg. Ce pré-calcul permet ensuite d'accélérer le coût de calcul des méthodes d'approximation, en particulier dans le cas symétrique ou hermitien pour lequel la matrice de Hessenberg se trouve être tridiagonale (symétrique ou hermitienne respectivement).

Définition 2.2. Une matrice $(a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{M}_n(\mathbb{C})$ est dite de Hessenberg si

$$\forall i, j \in [\![1, n]\!], \quad i - j \geqslant 2 \quad \Longrightarrow \quad a_{i,j} = 0.$$

Contrairement au cas de la trigonalisation ou diagonalisation complète, on dispose d'un algorithme « élémentaire » et relativement peu coûteux qui permet ici de construire la matrice U. Il s'appuie sur les matrices de symétries orthogonales de Householder et coûte de l'ordre de $5n^3/3 + O(n^2)$ multiplications et $4n^3/3 + O(n^2)$ additions. Dans le cas d'une matrice hermitienne, le coût de calcul est légèrement réduit par le fait des symétries.

Proposition 2.3. 1. Soit $A \in \mathscr{M}_n(\mathbb{C})$. Alors il existe $U \in \mathrm{U}_n(\mathbb{C})$ telle que U^*AU est de Hessenberg.
2. Soit $A \in \mathscr{M}_R(\mathbb{C})$. Alors il existe $O \in \mathrm{O}_n(\mathbb{R})$ telle que ${}^tOAO$ est de Hessenberg.
3. Si A est hermitienne, alors toute forme de Hessenberg U^*AU est tridiagonale.

2.1.3 Entrelacement des racines

Soit $A \in \mathscr{H}_n(\mathbb{C})$. On vient de voir qu'il est possible de déterminer une matrice unitairement semblable à cette matrice qui soit hermitienne et tridiagonale. On peut alors localiser ses valeurs propres (réelles) par la proposition suivante.

PROPOSITION 2.4. Soit $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n} \in \mathscr{H}_n(\mathbb{C})$ tridiagonale telle que $a_{j+1,j} \neq 0$ pour tout $j \in [\![1, n-1]\!]$. Soit $k \in [\![1, n]\!]$. On note $A_k \coloneqq (a_{i,j})_{n-k+1\leqslant i,j\leqslant n} \in \mathscr{M}_k(\mathbb{C})$ le bloc diagonal inférieur de A de taille k. Alors les valeurs propres de A_k sont réelles et simples et elles séparent strictement celles de A_{k+1}.

Preuve On procède par récurrence sur k. Pour $k \in [\![2, n-1]\!]$, en notant P_k le polynôme caractéristique de A_k, on obtient la relation de récurrence

$$P_{k+1} = (a_{n-k,n-k} - X)P_k - |a_{n-k+1,n-k}|^2 P_{k-1}.$$

Par convention, on pose $P_0 \coloneqq 1$. Alors une récurrence assure que les polynômes $P_k \in \mathbb{C}_k[X]$ sont à coefficients réels puisque $a_{n-k,n-k} = \langle Ae_{n-k}, e_{n-k}\rangle \in \mathbb{R}$ où le vecteur $e_{n-k} \in \mathbb{C}^n$ est le $(n-k)$-ième vecteur de la base canonique de $\mathbb{C}^n$. De plus, on a $P_k(t) \longrightarrow (\mp 1)^k \infty$ quand $t \to \pm\infty$.

L'hypothèse de récurrence assure que les racines de P_k sont réelles simples et, en les notant $\alpha_1^{(k)} < \cdots < \alpha_k^{(k)}$, elles s'entrelacent, *i. e.* elles sont telles que

$$\alpha_1^{(k)} < \alpha_1^{(k-1)} < \cdots < \alpha_{k-1}^{(k-1)} < \alpha_k^{(k)}.$$

Ainsi, avec le comportement à l'infini de P_{k-1}, on trouve le signe de P_{k-1} en chacune des racines de P_k. En examinant la valeur de P_{k+1} en chaque racine de P_k, on observe que

$$P_{k-1}(\alpha_m^{(k)})P_{k+1}(\alpha_m^{(k)}) < 0, \quad \forall m \in [\![1, k]\!].$$

Ces changements de signes garantissent l'existence d'au moins $k-1$ racines de P_{k+1} dans l'intervalle $[\alpha_1^{(k)}, \alpha_k^{(k)}]$ et son signe à l'infini donne deux racines supplémentaires. De ce fait, les au plus $k+1$ racines de P_{k+1} sont toutes réelles, simples et clairement entrelacées entre celle de P_k ce qui conclut la récurrence. □

La suite de polynômes $(P_k)_{1\leqslant k\leqslant n}$, construite dans la preuve précédente, constitue ce qu'on appelle une suite de STURM et les propriétés d'entrelacement de leurs racines permettent de dénombrer le nombre de racines inférieures à un réel x simplement en comptant le nombre de changements de signes dans la suite $(P_k(x))_{1\leqslant k\leqslant n}$.

2.1.4 Méthode QR

La méthode QR est d'une stupéfiante simplicité. Étant donnée une matrice $A \in \mathscr{M}_n(\mathbb{C})$, on calcule la décomposition QR de A, notée $A = QR$ avec Q orthogonale réelle et r triangulaire supérieure à coefficients diagonaux positifs. On remplace ensuite A par le produit RQ et on réitère le procédé.

Sous de bonnes hypothèses sur la matrice A, la suite $(A_k)_{k\in\mathbb{N}}$ ainsi construite converge alors vers une matrice triangulaire supérieure semblable à A dont les coefficients diagonaux sont alors les valeurs propres de A ordonnées en module décroissant. L'algorithme peut être interprété comme une réduction asymptotique de SCHUR.

On admet le théorème suivant qui donne un algorithme permettant de trouver très rapidement les valeurs propres de A. Même s'il ne fait pas de place à une vraie convergence, il permet d'approximer le spectre.

THÉORÈME 2.5. Soit $A \in \mathrm{GL}_n(\mathbb{C})$ une matrice dont les n valeurs propres λ_i vérifient $|\lambda_1| > \cdots > |\lambda_n| > 0$. Alors A est diagonalisable et il existe $P \in \mathrm{GL}_n(\mathbb{C})$ telle que $A = P^{-1}\operatorname{diag}(\lambda_1, \ldots, \lambda_n)P$. On suppose que P admet une factorisation LU. Alors

1. la suite des parties triangulaires inférieures strictes de A_k converge vers 0 ;
2. la suite des diagonales de A_k converge vers $\operatorname{diag}(\lambda_1, \ldots, \lambda_n)$;
3. la suite des parties triangulaires supérieures strictes de A_k est bornée.

De plus, la convergence est au plus géométrique de raison $\max_{1\leqslant i\leqslant n-1} |\lambda_{i+1}/\lambda_i|$.

2.2 Résultats théoriques

2.2.1 Localisation du spectre

THÉORÈME 2.6 *(HADAMARD).* Soit $A \coloneqq (a_{i,j})_{1\leqslant i,j\leqslant n}$ telle que

$$\forall i \in [\![1,n]\!], \quad |a_{i,i}| > \sum_{j\neq i} |a_{i,j}|\,.$$

Alors A est inversible.

Preuve Ce théorème a déjà été montré. □

THÉORÈME 2.7 *(GERSHGÖRIN).* Soit $A \in \mathscr{M}_n(\mathbb{C})$. Pour $i \in [\![1,n]\!]$, on note

$$D_i \coloneqq \{z \in \mathbb{C} \mid |z - a_{i,i}| \leqslant \sum_{j\neq i} |a_{i,j}|\}.$$

Alors

$$\sigma(A) \subset \bigcup_{i=1}^{n} D_i.$$

Preuve Soit $\lambda \in \sigma(A)$. La matrice $A - \lambda I_n$ n'est pas inversible, donc elle n'est pas à diagonale strictement dominante, donc il existe $i \in [\![1,n]\!]$ tel que $\lambda_i \in D_i$ ce qui montre l'inclusion. □

THÉORÈME 2.8 *(GERSHGÖRIN 2).* Dans chaque composante connexe de $\bigcup_{i=1}^{n} D_i$, on compte autant de valeurs propres de A (comptées avec leur multiplicité algébrique) que de disques de GERSHGÖRIN.

Preuve C'est un résultat déduit d'un principe de continuité des racines d'un polynôme en fonction de ses coefficients (ou de continuité des valeurs propres en fonction des coefficients de la matrice). On abordera plus loin ce résultat utilisant le théorème de ROUCHÉ en analyse complexe.

Pour $t \in [0,1]$, on pose $B(t) \coloneqq D + t(A - D) \in \mathscr{M}_n(\mathbb{C})$ où on a noté D la diagonale de A. On remarque que $B(1) = A$ et $\sigma(B(0)) = \{a_{i,i} \mid 1 \leqslant i \leqslant n\}$. On admet le théorème de relèvement des racines : il existe n fonctions continues $\lambda_i \colon [0,1] \to \mathbb{C}$ telles que $\lambda_i(0) = a_{i,i}$ pour tout $i \in [\![1,n]\!]$ et $\{\lambda_i(1) \mid 1 \leqslant i \leqslant n\} = \sigma(A)$. Soit $t \in [0,1]$. Le théorème de GERSHGÖRIN appliqué à $B(t)$ donne

$$\sigma(B(t)) = \{\lambda_i(t) \mid 1 \leqslant i \leqslant n\} \subset K(t) \coloneqq \bigcup_{i=1}^{n} D_i(t) \quad \text{avec} \quad D_i(t) \coloneqq \{z \in \mathbb{C} \mid |z - a_{i,i}| \leqslant t\sum_{j\neq i} |a_{i,j}|\}.$$

Soit $I \subset [\![1,n]\!]$ telle que $M \coloneqq \bigcup_{i\in I} D_i(1)$ soit une composante connexe de $K \coloneqq K(1)$. On peut écrire $K = M \sqcup N$ où M et N sont deux fermés disjoints. Comme les applications $t \longmapsto D_i(t)$ sont croissante, pour tout $t \in [0,1]$, l'ensemble $\bigcup_{i\in I} D_i(t)$ est un sous-ensemble de M totalement disjoint de N.

Soit $i \in I$. L'application continue λ_i est à valeurs dans $K = M \sqcup N$ et vérifie $\lambda_i(0) \in M$, donc elle est à valeurs dans M. En particulier, l'ensemble $\{\lambda_i(1) \mid i \in I\}$ est contenu dans M et contient $\sharp I$ valeurs propres de A ce qui termine la preuve. □

COROLLAIRE 2.9. 1. Si un disque est isolé, alors il contient exactement une valeur propre.
2. Si les disques de GERSHGÖRIN sont tous disjoints deux à deux, alors chacun contient exactement une valeur propre.

2.2.2 Continuité des valeurs propres

PROPOSITION 2.10. Soient $A \in \mathscr{M}_n(\mathbb{C})$ diagonalisable et $\lambda \in \mathbb{C}$ une valeur propre simple de A. Alors il existe un voisinage ouvert V de A dans $\mathscr{M}_n(\mathbb{C})$ et une fonction $\mu \in \mathscr{C}^\infty(V, \mathbb{C})$ tels que $\mu(A) = \lambda$ et

$$\forall B \in V, \quad \mu(B) \in \sigma(B).$$

Preuve Appliquons le théorème des fonctions implicites à la fonction

$$\psi \colon \left| \begin{array}{l} \mathscr{M}_n(\mathbb{C}) \times \mathbb{C} \times \mathbb{C}^n \longrightarrow \mathbb{C} \times \mathbb{C}^n, \\ \qquad (B, \mu, y) \longmapsto (\|y\|_2^2 - 1, By - \mu y). \end{array} \right.$$

Les zéros de cette application sont les triplets (B, μ, y) pour lesquels μ est une valeur propre de B et Y est un vecteur propre associé. Soit $(A, \lambda, x) \in \mathscr{M}_n(\mathbb{C}) \times \mathbb{C} \times \mathbb{C}^n$ tel que $\psi(A, \lambda, x) = (0, 0)$. Pour tout $(\mu, y) \in \mathbb{C} \times \mathbb{C}^n$, on a

$$\begin{aligned} \psi(A, \lambda + \mu, x + y) &= (\|x + y\|^2 - 1, A(x + y) - (\lambda + \mu)(x + y)) \\ &= \psi(A, \lambda, x) + (2x^* y, Ay - \mu x - \lambda y) + o(\mu, y). \end{aligned}$$

Il suffit alors de montrer que la dérivée partielle

$$f \coloneqq \frac{\partial \psi}{\partial(\mu, y)}(A, \lambda, x) \colon \left| \begin{array}{l} \mathbb{C} \times \mathbb{C}^n \longrightarrow \mathbb{C} \times \mathbb{C}^n, \\ \quad (\mu, y) \longmapsto (2x^* y, Ay - \mu x - \lambda y). \end{array} \right.$$

est inversible. Soit $(\mu, y) \in \operatorname{Ker} f$. Alors $x^* y = 0$ et $Ay - \mu x - \lambda y = 0$. En appliquant $A - \lambda I_n$ à la seconde égalité, on obtient que $(A - \lambda I_n)^2 y = \mu(A - \lambda I_n)x = 0$, donc $y \in \operatorname{Ker}(A - \lambda I_n)^2 = \operatorname{Ker}(A - \lambda I_n) = \operatorname{Vect}\{x\}$ puisque A est diagonalisable et λ est simple. Alors il existe $\alpha \in \mathbb{C}$ tel que $y = \alpha x$. Comme $x^* y = 0$, on a $\alpha \|x\|_2^2 = 0$, donc $\alpha = 0$, donc $Y = 0$ et $X = 0$. On en déduit que l'application f est injective, donc qu'elle est inversible. Le théorème des fonctions implicites assure l'existence d'un voisinage ouvert V de A dans $\mathscr{M}_n(\mathbb{C})$, d'un voisinage ouvert J de λ dans $\mathbb{C}$, d'un voisinage ouvert O de x dans $\mathbb{C}^n$ et d'une application $\phi \in \mathscr{C}^\infty(V, J \times O)$ telle que

$$\forall (B, \mu, y) \in V \times J \times O, \quad \psi(B, \mu, y) = 0 \quad \Longleftrightarrow \quad (\mu, y) = \phi(B).$$

On prend alors la projection sur la première coordonnées de ϕ. □

◇ Remarque. Le résultat précédent est seulement local et tombe en défaut dès qu'on approche une matrice comportant par exemple un bloc de Jordan non-trivial, le contre-exemple suivant illustre ce fait explicitement. Pour $x \in \mathbb{R}$, on pose

$$A(x) \coloneqq \begin{pmatrix} 0 & 1 \\ x & 0 \end{pmatrix}.$$

On a $\sigma(A(x)) = \{\pm\sqrt{x}\}$ si $x \geqslant 0$ et $\sigma(A(x)) = \left\{\pm i\sqrt{-x}\right\}$ si $x < 0$. Le spectre est alors continûment paramétrable au voisinage de $x = 0$, mais pas de manière $\mathscr{C}^1$.

Dans la suite de cette partie, nous proposons une approche plus globale qui permet de démontrer le résultat de continuité que l'on peut déjà observer sur le contre-exemple précédent. Bien sûr, en contrepartie on n'obtiendra pas la régularité plus précise des valeurs propres comme précédemment.

Lemme 2.11. Soit $n \in \mathbb{N}$. On munit de $\mathbb{C}_n[X]$ de la norme

$$P \coloneqq \sum_{j=0}^{n} p_j X^j \longmapsto \|P\| \coloneqq \max_{0 \leqslant j \leqslant n} |p_j|.$$

Soient $P \in \mathbb{C}_n[X]$ non constant et $x \in \mathbb{C}$ une de ses racines. On note $\eta > 0$ la distance de x à l'ensemble des autres racines de P. Soit $\rho \in]0, \eta[$. Alors il existe $\delta > 0$ tel que tout polynôme $Q \in \mathbb{C}_n[X]$ vérifiant $\|Q - P\| \leqslant \delta$ compte exactement μ racines dans $D \coloneqq \overline{\mathrm{B}}(x, \rho) \subset \mathbb{C}$, comptées avec multiplicité.

Preuve Notons γ le lacet orienté paramétrant la frontière ∂D. Comme P ne s'annule pas sur γ qui est un domaine bornée, il existe $\alpha, \beta > 0$ tels que

$$\forall z \in \gamma, \quad |P(z)| \geqslant \alpha \quad \text{et} \quad \sum_{j=0}^{n} |z|^j \leqslant \beta.$$

Soit $\delta \in]0, \alpha/\beta[$. Soit $Q \in \mathbb{C}_n[X]$ tel que $\|Q - P\| \leqslant \delta$. Alors pour tout $z \in \gamma$, on a

$$|P(z) - Q(z)| \leqslant \|P - Q\| \, \beta < a \leqslant |P(z)|.$$

Le théorème de Rouché assure alors que le nombre de zéros de P et Q dans D sont égaux. □

ne difficulté subsiste pour en arriver à la continuité des valeurs propres, qui concerne la topologie retenue pour les ensemble spectraux, afin de tenir compte des éventuelles multiplicités. Pour ce faire, on définit l'application $\tilde{\sigma}$ qui à une matrice $A \in \mathscr{M}_n(\mathbb{C})$ associe le n-uplet $\tilde{\sigma}(A) \in \mathbb{C}^n$ de ses valeurs propres, chacune étant répétée avec sa multiplicité algébrique. Afin de supprimer l'arbitraire, on se place dans le quotient $\mathbb{C}^n/\!\sim$ de $\mathbb{C}^n$ par la relation d'équivalence $\sim$ induite par les permutations d'indices. Pour $a \in \mathbb{C}^n$, on a $\mathscr{C}\ell(a)$ la classe d'équivalence de a.

Le résultat de continuité ne concerne pas à proprement parler l'application de spectre σ, mais plutôt celle qui à A associe la classe d'équivalence $\mathscr{C}\ell(\tilde{\sigma}(A))$ qui tient compte des multiplicités algébriques éventuelles. Il reste à préciser la topologie pour laquelle on est en mesure de quantifier le résultat. On munit $\mathbb{C}^n/\!\sim$ de la distance d_{H} définie par

$$d_{\mathrm{H}}(a, b) \coloneqq \min_{\sigma, \tau \in \mathfrak{S}_n} \max_{1 \leqslant j \leqslant n} |a_{\sigma(i)} - b_{\tau(j)}|, \quad a \coloneqq (a_1, \ldots, a_n) \in \mathbb{C}^n, \; b \coloneqq (b_1, \ldots, b_n) \in \mathbb{C}^n.$$

On considère une norme sur $\mathbb{C}^n$ et on munit $\mathscr{M}_n(\mathbb{C})$ de la norme subordonnée.

Théorème 2.12. L'application $\mathscr{C}\ell(\tilde{\sigma}(\cdot)) \colon \mathscr{M}_n(\mathbb{C}) \longrightarrow (\mathbb{C}^n/\!\sim, d_{\mathrm{H}})$ est continue.

Preuve L'application $A \in \mathscr{M}_n(\mathbb{C}) \longmapsto \chi_A \in \mathbb{C}_n[X]$ est clairement continue : on peut munir $\mathbb{C}_n[X]$ de la topologie induite par la distance $(P,Q) \longmapsto \|P-Q\|$ introduite précédemment. Soit $P \in \mathbb{C}_n[X]$. On note

$$\eta := \min\{|x-y| \mid P(x)=P(y)=0,\ x \neq y\}$$

la plus petite distance entre ses racines. Soit $\varepsilon > 0$ tel que $\varepsilon < \eta/3$. D'après le lemme précédent, pour toute racine $x \in \mathbb{C}$ de P, il existe $\delta_x > 0$ tel que, pour tout polynôme $Q \in \mathbb{C}_n[X]$ vérifiant $\|Q-P\| \leqslant \delta_x$, la boule $\overline{\mathrm{B}}(x,\varepsilon)$ contienne autant de zéros de P que de zéros de Q. On pose alors $\delta := \min_x \delta_x$ de sorte que, pour tout polynôme $P \in \mathbb{C}_n[X]$ vérifiant $\|P-Q\| < \delta$, chacun des boules $\overline{\mathrm{B}}(x,\varepsilon)$ contienne autant de zéros de Q que de zéros de P. Soit $Q \in \mathbb{C}_n[X]$ tel que $\|P-Q\| < \delta$. Par ailleurs, en comptant les racines ainsi constituée, le polynôme Q n'admet pas de racines en dehors des boules $\overline{\mathrm{B}}(x,\varepsilon)$. Par conséquent, en notant $Z_P \in \mathbb{C}^n$ le n-uplet des zéros de P et $Z_Q \in \mathbb{C}^n$ celui de Q, on obtient la majoration $d_{\mathrm{H}}(\mathscr{C}\ell(Z_P),\mathscr{C}\ell(Z_Q)) \leqslant \varepsilon$ ce qui montre la continuité. □

2.2.3 Conditionnement du problème aux valeurs propres

Le théorème suivant permet de préciser, au moins dans le cas d'une matrice diagonalisable, la manière donc le spectre dépend de cette matrice. On a vu, dans les contre-exemples précédents, que, sans cette hypothèse, on perd généralement le caractère lipschitizien. Cette étude est primordiale dans les applications, dans la mesure où les objets ne sont connus qu'approximativement, ce malgré quoi il faut être en mesure de certifier la qualité du résultat (souvenez-vous du pont de Tacoma !).

THÉORÈME 2.13. On munit $\mathbb{C}^n$ d'un norme et $\mathscr{M}_n(\mathbb{C})$ de la norme subordonnée associée. Soient $A, E \in \mathscr{M}_n(\mathbb{C})$. On suppose que A est diagonalisable, *i. e.* il existe $P \in \mathrm{GL}_n(\mathbb{C})$ telle que $D := P^{-1}AP$ soit diagonale. Alors

$$\forall \mu \in \sigma(A+E), \quad \mathrm{d}(\mu,\sigma(A)) \leqslant (\operatorname{cond} P)\|E\|.$$

Preuve Soit $\mu \in \sigma(A+E)$. Le résultat est immédiat si $\mu \in \sigma(A)$. Supposons que $\mu \notin \sigma(A)$. Alors

$$A+E-\mu I_n = P(D-\mu I_n)[I_n + (D-\mu I_n)^{-1}P^{-1}EP]P^{-1}$$

ce qui montre que -1 est une valeur propre de $C := (D-\mu I_n)^{-1}P^{-1}EP$. On en déduit que

$$1 \leqslant \rho(C) \leqslant \|(D-\mu I_n)^{-1}\|\,\|E\|\operatorname{cond} P.$$

Comme la matrice $D-\mu I_n$ est diagonale, on a

$$\|(D-\mu I_n)^{-1}\| = [\min_{\lambda\in\sigma(A)} |\lambda-\mu|]^{-1}$$

ce qui permet de conclure. □

◇ REMARQUE. Avec les notations précédemment introduites, on a

$$d_{\mathrm{H}}(\mathscr{C}\ell(\tilde\sigma(A+E)),\mathscr{C}\ell(\tilde\sigma(A))) \leqslant (\operatorname{cond} P)\|E\|.$$

En réalité, comme on l'a vu pour la méthode de la puissance, il se pose également la question de connaître la précision du spectre d'une matrice, ayant obtenu au préalable une approximation d'une solution (λ, x) de l'équation aux valeurs propres $Ax = \lambda x$.

PROPOSITION 2.14. Soient $A \in \mathscr{H}_n(\mathbb{C})$ et (λ, x) un couple propre de A. Soit $(\tilde\lambda, \tilde x)$ l'approximation de (λ, x). On pose $\tilde r := A\tilde x - \tilde\lambda \tilde x$. Alors

$$\mathrm{d}(\tilde\lambda,\sigma(A)) \leqslant \frac{\|\tilde r\|_2}{\|\tilde x\|_2}.$$

Preuve Soit $(u_1,\dots,u_n)$ une base orthonormée constituée de vecteurs propres u_i de A associés aux valeurs propres λ_i. On écrit $\tilde x = \sum_{i=1}^n \alpha_i u_i$. Alors

$$\tilde r = \sum_{i=1}^n \alpha_i(\lambda_i - \tilde\lambda)u_i,$$

donc

$$\|\tilde r\|_2 = \sum_{i=1}^n |\alpha_i|^2|\lambda_i-\tilde\lambda|^2 \quad \text{et} \quad \|\tilde x\|_2 = \sum_{i=1}^n |\alpha_i|^2.$$

On en déduit que

$$\frac{\|\tilde r\|_2}{\|\tilde x\|_2} \geqslant \min_{1\leqslant i\leqslant n} |\lambda_i-\tilde\lambda|^2 \geqslant \mathrm{d}(\tilde\lambda,\sigma(A)).$$ □

◇ REMARQUE. On admet le résultat plus général suivant. Soit $A \in \mathscr{M}_n(\mathbb{C})$ une matrice diagonalisable. Il existe $P \in \mathrm{GL}_n(\mathbb{C})$ telle que $P^{-1}AP$ soit diagonale. Alors pour tout $\varepsilon > 0$ tel que $\|\tilde r\|_2 \leqslant \varepsilon\|\tilde x\|_2$, on a

$$\mathrm{d}(\tilde\lambda,\sigma(A)) \leqslant \varepsilon \operatorname{cond} P.$$