Analyse numérique

Thomas Harbreteau

13 avril 2020

Notes du cours de Benjamin Boutin.
Université de Rennes 1, année 2019/2020.

Sommaire

Chapitre 1

Résolution de systèmes linéaires

1 Rappels et notations

1.1 Rappels

Soit $A \in \mathcal{M}_n(\mathbf{K})$ et $\lambda \in \sigma(A)$.

- Polynôme caractéristique : $\chi_A = \det(A - XI_n)$.
- Multiplicité algébrique de λ : multiplicité en tant que racine du polynôme caractéristique.
- Multiplicité géométrique de λ : dimension du sous-espace propre associé à λ.
- Valeur propre défective : multiplicité géométriques et algébriques différentes.

1.2 Notations

- $T_n^u(\mathbf{K})$ l'ensemble des matrices triangulaires supérieures.
- $U_n(\mathbf{K})$ l'ensemble des matrices unitaires, i. e. dont les colonnes forment une base orthonormée pour le produit scalaire hermitien (analogue aux matrices orthogonales réelles). Caractérisation matricielle : $U^*U = UU^* = I_n$.

2 Réduction en dimension finie

Théorème 1.2.1 (Réduction de Jordan) *Soit $A \in \mathcal{M}_n(\mathbf{K})$, supposée trigonalisable sur $\mathbf{K}$,*

$$\exists P \in \mathcal{GL}_n(\mathbf{K}), \quad P^{-1}AP = \begin{pmatrix} J_{k_1}(\lambda_1) & 0 & \dots & 0 \\ 0 & J_{k_2}(\lambda_2) & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & J_{k_m}(\lambda_m) \end{pmatrix},$$

où

- *les $(\lambda_i)_{1\leq i\leq m}$ non-nécéssairement distincts deux à deux,*
- *pour tout $i \in \{1, \dots, m\}$, $J_{k_i}(\lambda_i)$ désigne un bloc de Jordan de taille $k_i \geq 1$:*

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \dots & \dots & 0 & \lambda \end{pmatrix}.$$

2.1 Réduction en base orthonormée

Proposition 1.2.2 (Produit scalaire hermitien) $\forall x, y, z \in \mathbf{C}^*, \forall \lambda, \mu \in \mathbf{C}, \forall A \in \mathcal{M}_n(\mathbf{C})$,

- $\langle x, y + z\rangle = \langle x, y\rangle + \langle x, z\rangle$,
- $\langle x, y\rangle = \overline{\langle y, x\rangle}$,
- $\langle \lambda x, \mu y\rangle = \bar{\lambda}\mu\langle x, y\rangle$,
- $\langle Ax, y\rangle = \langle x, A^*x\rangle$, *où* $A^* = \bar{A}^T$ *(définition de l'adjoint),*
- $|\langle x, y\rangle| \leq \|x\|_2\|y\|_2$.

Proposition 1.2.3 (Projection orthogonale) *On considère F un sous-espace de $\mathbf{K}^n$ et $(e_1, \dots, e_n)$ une base orthonormée de F. Alors la projection orthogonale sur F parallèlement à $F^\perp$, notée p_F, est obtenue comme :*

$$\forall x \in \mathbf{K}^n, \quad p_F(x) = \sum_{i=1}^{m} \langle e_i, x \rangle = \left(\sum_{i=1}^{m} e_i e_i^* \right) x.$$

PREUVE.

$$\sum_{i=1}^{m} \langle e_i, x \rangle ei = \sum_{i=1}^{m} (e_i^* x) e_i = \sum_{i=1}^{m} e_i (e_i^* x) = \left[\sum_{i=1}^{m} (e_i e_i^*) \right] x.$$

□

Proposition 1.2.4 (Orthonormalisation de Gram-Schmidt) *Soit $(v_i)_{1 \le i \le n}$ une base de $\mathbf{K}^n$. On construit une base orthonormée $(e_i)_{1 \le i \le n}$ de $\mathbf{K}^n$ par projections successives sur les directions précédemment construites :*

$$u_1 := v_1, \quad e_1 := u_1 / \|u\|,$$

et pour tout $k \in \{2, \dots, n\}$,

$$u_k := v_k - p_{\mathrm{Vect}(e_1, \dots, e_{k-1})}(v_k) = v_k - \sum_{i=1}^{k-1} \frac{\langle u_i, v_k \rangle}{\langle u_i, u_i \rangle} u_i, \quad e_k := \frac{u_k}{\|u_k\|}.$$

Proposition 1.2.5 (Factorisation QR) *Soit $A \in \mathcal{GL}_n(\mathbf{C})$, alors il existe $Q \in U_n(\mathbf{C})$ et $R \in T_n^u(\mathbf{C})$ telles que $A = QR$. Cette décomposition est de plus unique si on suppose les termes diagonaux de R réels positifs.*
PREUVE.

$$\forall k \in \{1, \dots, n\}, \quad v_k = u_k + \sum_{i=1}^{k-1} \frac{\langle u_i, v_k \rangle}{\langle u_i, u_i \rangle} u_i = \|u_n\| e_k + \sum_{i=1}^{k-1} \langle e_i, v_k \rangle e_i.$$

Ainsi,

$$A = \begin{pmatrix} \vdots & & \vdots \\ v_1 & \dots & v_n \\ \vdots & & \vdots \end{pmatrix} = \begin{pmatrix} \vdots & & \vdots \\ e_1 & \dots & e_n \\ \vdots & & \vdots \end{pmatrix} \begin{pmatrix} \|u_1\| & & (\langle e_i, v_j \rangle) \\ & \ddots & \\ (0) & & \|u_n\| \end{pmatrix}.$$

□

Théorème 1.2.6 (Schur) *Soit $A \in \mathcal{M}_n(\mathbf{C})$, alors il existe $U \in U_n(\mathbf{C})$ et $T \in T_n^u(\mathbf{C})$ telles que $U^* A U = T$.*
PREUVE. P admet une factorisation QR, notée $P = QR$, où $Q \in U_n(\mathbf{C})$ et $R \in T_n^u(\mathbf{C})$. Alors

$$A = QRT(QR)^{-1} = QRTR^{-1}Q^{-1} = Q(\underbrace{RTR^{-1}}_{\in T_n^u(\mathbf{C})})Q^*.$$

□

2.2 Réduction des matrices normales

Définition 1.2.7 (Matrice normale) *Une matrice $A \in \mathcal{M}_n(\mathbf{C})$ est dite normale si $AA^* = A^*A$.*

Exemple 1.2.8 $S_n(\mathbf{R})$; $H_n(\mathbf{C})$ hermitiennes complexes $(A^* = A)$; $O_n(\mathbf{R})$; $U_n(\mathbf{C})$.

Lemme 1.2.9 *Toute matrice triangulaire (supérieure ou inférieure) et normale est diagonale.*
PREUVE. Soit $A \in T_n^u(\mathbf{C})$.

$$\forall i \in \{1, \dots, n\}, \quad \begin{cases} (A^*A)_{i,i} &= \displaystyle\sum_{j=1}^{i} \bar{A}_{j,i} A_{j,i} &= \displaystyle\sum_{j=1}^{i} |a_{i,j}^2| \\ (AA^*)_{i,i} &= |a_{i,i}|^2. \end{cases}$$

□

Théorème 1.2.10 *Une matrice $A \in \mathcal{M}_n(\mathbf{C})$ est normale si et seulement si elle est diagonalisable en base orthonormée.*

PREUVE. ($\Longleftarrow$) : Supposons A diagonalisable en base orthonormée. On note $A = UDU^*$, où $U \in U_n(\mathbf{C})$ et D diagonale. Alors

$$\begin{cases} A^*A &= (UDU^*)^*(UDU^*) = UD^*(U^*U)DU^* = UD^*DU^* \\ AA^* &= UDD^*U^*. \end{cases}$$

Ainsi,

$$AA^* = A^*A \iff DD^* = D^*D \iff \forall i \in \{1, \dots, n\}, \quad D_{i,i}\bar{D}_{i,i} = \bar{D}_{i,i}D_{i,i}.$$

($\Longrightarrow$) : Soit $A \in \mathcal{M}_n(\mathbf{C})$ normale, $AA^* = A^*A$. Par théorème de Schur, il existe $U \in U_n(\mathbf{C})$ et $T \in T_n^u(\mathbf{C})$ telles que $A = UTU^*$. Du fait de l'égalité $AA^* = A^*A$, on trouve $TT^* = T^*T$ après calculs. Ainsi, T est normale et triangulaire supérieure, donc est diagonale par le lemme précédent. □

Proposition 1.2.11 *Les valeurs propres d'une matrice normale sont*

- *réelles si $A \in S_n(\mathbf{R})$, ou plus généralement si $A \in H_n(\mathbf{C})$,*
- *réelles positives si $A \in S_n^+(\mathbf{R})$ ou si $A \in H_n^+(\mathbf{R})$,*
- *réelles strictement positives si $A \in S_n^{++}(\mathbf{R})$ ou si $A \in H_n^{++}(\mathbf{R})$,*
- *imaginaires pures si A est anti-symétrique réelle ou anti-hermitienne,*
- *complexes de module 1 si $A \in O_n(\mathbf{R})$ ou plus généralement si $A \in U_n(\mathbf{C})$.*

PREUVE.

- Valeurs propres de $A \in U_n(\mathbf{C})$? Soit A normale, alors $A = UDU^*$ avec U unitaire et D vérifie $D^*D = DD^* = I_n$ (car $A^*A = AA^* = I_n$). On déduit de cette dernière égalité que

$$\forall i \in \{1, \dots, n\}, \quad |D_{i,i}|^2 = 1,$$

  donc $\sigma(A) = \sigma(D) = \{z \in \mathbf{C} \mid |z| = 1\}$.

□

3 Propriétés spectrales des matrices hermitiennes

Définition 1.3.1 (Quotient de Rayleigh) *Soit $A \in H_n(\mathbf{C})$, on définit l'application*

$$R_A : \begin{array}{ccc} (\mathbf{C}^n)^* & \longrightarrow & \mathbf{R} \\ x & \longmapsto & \dfrac{\langle Ax, x\rangle}{\langle x, x\rangle} \end{array}.$$

PREUVE. Si $A \in H_n(\mathbf{C})$,

$$\forall x \in \mathbf{C}^n, \quad \langle Ax, x\rangle = (Ax)^*x = x^*(A^*x) = \langle x, A^*x\rangle = \langle x, Ax\rangle,$$

car $A = A^*$. Par caractère hermitien du produit scalaire, $\langle Ax, x\rangle = \overline{\langle x, Ax\rangle}$, donc ici, $\langle x, Ax\rangle = \overline{\langle x, Ax\rangle} \in \mathbf{R}$. Ainsi, $R_A(x) \in \mathbf{R}$. □

Théorème 1.3.2 *Soit $A \in H_n(\mathbf{C})$ de valeurs propres (non nécéssairement positives) $\lambda_{\min} = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n = \lambda_{\max}$, alors*

$$\max_{x \neq 0} R_A(x) = \lambda_{\max}, \quad \text{et} \quad \min_{x \neq 0} R_A(x) = \lambda_{\min}.$$

PREUVE. Soient $x \in \mathbf{C}^n \setminus \{0\}$ et $z \in \mathbf{C} \setminus \{0\}$,

$$R_A(zx) = \frac{\langle Azx, zx\rangle}{\langle zx, zx\rangle} = \frac{\bar{z}z\langle Ax, x\rangle}{\bar{z}z\langle x, x\rangle} = R_A(x).$$

Étudier les extremums de R_A sur $\mathbf{C}^n \setminus \{0\}$ se ramène à les étudier sur $\mathcal{S} := \{x \in \mathbf{C}^n \mid \|x\|_2 = 1\}$.

$\mathcal{S}$ étant compacte et R_A continue sur $\mathcal{S}$, R_A est bornée et atteint ses bornes.

A étant hermitienne, elle est diagonalisable en base orthonormée. Notons $A = UDU^*$, avec $U \in U_n(\mathbf{C})$ et D diagonale réelle (car A est hermitienne), $D = \mathrm{diag}(\lambda_1, \dots, \lambda_n)$ avec $\lambda_1 \leq \cdots \leq \lambda_n$. Alors

$$R_A(x) = \frac{\langle UDU^*x, x\rangle}{UU^*x, x\rangle} = \frac{\langle DU^*x, U^*x\rangle}{\langle U^*x, U^*x\rangle} = R_D(U^*x).$$

L'application

$$\sigma : \begin{array}{ccc} \mathcal{S} & \longrightarrow & \mathcal{S} \\ x & \longmapsto & U^*x \end{array}$$

définit une bijection sur $\mathcal{S}$. Les extremums de R_A sur $\mathcal{S}$ sont ceux de R_D. Soit $y \in \mathcal{S}$,

$$R_D(y) = \frac{\langle Dy, y\rangle}{\langle y, y\rangle} = \langle Dy, y\rangle = y^* D^* y = y^* Dy = \sum_{i=1}^{n} \bar{y}_i \lambda_i y_i = \sum_{i=1}^{n} \lambda_i |y_i|^2.$$

Comme y est de norme 1,

$$\lambda_{\min} \underbrace{\sum_{i=1}^{n} |y_i|^2}_{=1} \leq R_D(y) \leq \lambda_{\max} \underbrace{\sum_{i=1}^{n} |y_i|^2}_{=1}.$$

Les égalités sont atteintes à gauche pour $y = (1, 0, \ldots, 0)^T$ et à droite pour $y = (0, \ldots, 0, 1)^T$. □

Remarque 1.3.3 On peut bien permuter les vecteurs colonnes de D. Soit P une matrice de permutation, $\tilde{D} = PDP^{-1} = PDP^*$, alors

$$A = UDU^* = UP^*\tilde{D}PU^* = (\underbrace{UP^*}_{\text{unitaire}})\tilde{D}(UP^*)^*.$$

4 Normes subordonnées et rayon spectral

4.1 Définitions

Définition 1.4.1 (Normes usuelles) *On considère $x \in \mathbf{C}^n$ et $A \in \mathcal{M}_n(\mathbf{C})$.*

$$\|x\|_1 = \sum_{j=1}^{n} |x_i|, \quad \|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |A_{i,j}|.$$

$$\|x\|_\infty = \max_{1\leq j\leq n} |x_j|, \quad \|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |A_{i,j}|.$$

4.2 Rayon spectral

Définition 1.4.2 (Rayon spectral) *Soit $A \in \mathcal{M}_n(\mathbf{K})$. On appelle rayon spectral de A la quantité*

$$\rho(A) = \max\{|\lambda| \mid \lambda \in \sigma(A)\}.$$

Lemme 1.4.3 *Si A est normale, alors $\|A\|_2 = \rho(A)$. Plus généralement, pour toute matrice $A \in \mathcal{M}_n(\mathbf{C})$, on a $\|A\|_2 = \sqrt{\rho(A^*A)}$.*

Théorème 1.4.4 *Soit $A \in \mathcal{M}_n(\mathbf{K})$.*

1. Pour toute norme subordonnée $\|\cdot\|$, on a $\rho(A) \leq \|A\|$.

2. Réciproquement, pour tout $\varepsilon > 0$, il existe une norme subordonnée $\|\cdot\|_\varepsilon$ telle que $\|A\|_\varepsilon \leq \rho(A) + \varepsilon$.

Ainsi,

$$\rho(A) = \inf_{\|\cdot\| \text{ subordonnée}} \|A\|.$$

Preuve.

1. Soit $\lambda \in \sigma(A)$ telle que $\rho(A) = |\lambda|$, et $x \in \mathbf{C}^n \setminus \{0\}$, $Ax = \lambda x$.

$$\|Ax\| = \|\lambda x\| = |\lambda|\|x\| = \rho(A)\|x\|,$$

et

$$\frac{\|Ax\|}{\|x\|} = \rho(A) = \max_{y\in\mathbf{C}^n\setminus\{0\}} \frac{\|Ay\|}{\|y\|} = \|A\|_{\mathcal{M}_n(\mathbf{C})}.$$

2. Soit $A \in \mathcal{M}_n(\mathbf{C})$. Par théorème de Schur, il existe $U \in U_n(\mathbf{C})$ et $T \in T_n^u(\mathbf{C})$ telles que $A = UTU^*$. Alors $\rho(A) = \rho(T)$, mais a-t-on une relation entre $\|A\|$ et $\|T\|$, pour une norme quelconque ? On calcule

$$\|T_1\| = \max_{1\leq j\leq n} \left(|\lambda_j| + \underbrace{\sum_{i=1}^{j-1} |T_{i,j}|}_{\text{à « éliminer »}} \right).$$

Soit $\eta > 0$, on note

$$D(\eta) := \begin{pmatrix} \eta & & & (0) \\ & \eta^2 & & \\ & & \ddots & \\ (0) & & & \eta^n \end{pmatrix},$$

et on définit

$$T(\eta) := \begin{pmatrix} \lambda_1 & & (\eta^{-i+j} T_{i,j}) \\ & \ddots & \\ (0) & & \lambda_n \end{pmatrix}.$$

On remarque que $T(\eta) \to_{(\eta \to 0)} \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, donc

$$\|T(\eta)\|_1 = \max_j \left(|\lambda_j| + \eta \sum_{i=1}^{j-1} \eta^{j-1-i} |T_{i,j}| \right) \xrightarrow[\eta \to 0]{} \max_j |\lambda_j| = \rho(A).$$

Soient $\varepsilon > 0$ et η suffisamment petit de sorte que $\|T(\eta)\|_1 \leq \rho(A) + \varepsilon$. De plus, on a

$$A = (UD(\eta)) T(\eta) (D(\eta)^{-1} U^*).$$

On considère sur $\mathbf{C}^n$ la norme

$$\forall x \in \mathbf{C}^n, \quad \|x\|_\varepsilon = \|D(\eta)^{-1} U^* x\|_1.$$

Ainsi,

$$\begin{aligned} \|Ax\|_\varepsilon &= \|D(\eta)^{-1} U^* A x\|_1 \\ &= \| \underbrace{(D(\eta)^{-1} U^*)(UD(\eta))}_{=I_n} T(\eta) (D(\eta)^{-1} U^*) x\|_1 \\ &= \|T(\eta) D(\eta)^{-1} U^* x\|_1 \\ &\leq \|T(\eta)\|_1 \|D(\eta) U^* x\|_1 \\ &\leq (\rho(A) + \varepsilon) \|x\|_1. \end{aligned}$$

Par conséquent,

$$\rho(A) = \inf_{\|\cdot\| \text{ subordonnée}} \|A\|.$$

□

4.3 Utilisation du rayon spectral

Théorème 1.4.5 *Soit $A \in \mathcal{M}_n(\mathbf{K})$, il y a équivalence entre les propositions suivantes.*

1. $\lim_{(k \to +\infty)} A^k = 0$.
2. *Il existe une norme subordonnée $\|\cdot\|$ telle que $\|A\| < 1$.*
3. $\rho(A) < 1$.

Preuve. 2) $\implies$ 1) : $A^k \to 0 \iff \|A^k\| \to 0$, mais $\|A^k\| \leq \|A\|^k \to 0$.

2) $\iff$ 3) : Conséquence du théorème précédent.

1) $\implies$ 3) : Par contraposée, supposons que $\rho(A) \geq 1$. Alors il existe $\lambda \in \mathbf{C}$ et $x \in \mathbf{C}^n$ tels que $|\lambda| \geq 1$, $x \neq 0$ et $Ax = \lambda x$. Par récurrence, pour tout $k \in \mathbf{N}$, $A^k x = \lambda^k x$, donc

$$\|A^k x\| = |\lambda^k| \|x\| \geq \|x\| \neq 0.$$

Ainsi,

$$\|A^k\| = \sup_{y \neq 0} \frac{\|A^k y\|}{\|y\|} \geq 1,$$

donc $(A^k)_k$ ne peut tendre vers 0. □

Corollaire 1.4.6 *Soit $A \in \mathcal{M}_n(\mathbf{K})$, pour toute norme subordonnée $\|\cdot\|$, on a*

$$\lim_{k \to +\infty} \|A^k\|^{1/k} = \rho(A).$$

PREUVE. Soient $\lambda \in \mathbf{C}$ et $x \in \mathbf{C}^n$ tels que $x \neq 0$, $Ax = \lambda x$ et $|\lambda| = \rho(A)$. Alors

$$\|A^k x\| = |\lambda|^k \|x\| = \rho(A)^k \|x\|,$$

donc $\|A^k\| \geq \rho(A)^k$, d'où $\|A^k\|^{1/k} \geq \rho(A)$, et ce pour tout $k \in \mathbf{N}$.

Soit $\varepsilon > 0$, On souhaiterait obtenir pour k assez grand $\|A^k\|^{1/k} \leq \rho(A) + \varepsilon$. Posons

$$A_\varepsilon = \frac{A}{\rho(A) + \varepsilon}.$$

On a alors $\rho(A_\varepsilon) = \rho(A)/(\rho(A) + \varepsilon) < 1$, donc $A_\varepsilon^k \to_{(k \to +\infty)} 0$. À partir d'un certain rang k_0, $\|A_\varepsilon^k\| \leq 1$, or

$$\|A_\varepsilon^k\| = \left(\frac{1}{\rho(A) + \varepsilon}\right)^k \|A\|^k.$$

Par conséquent, pour $k \geq k_0$, $\|A^k\| \leq (\rho(A) + \varepsilon)^k$, d'où $\|A^k\|^{1/k} \leq \rho(A) + \varepsilon$. □

Théorème 1.4.7 *Soit $A \in \mathcal{M}_n(\mathbf{K})$, il y a équivalence entre les propositions suivantes.*

1. *La suite $(A^k)_{k \geq 0}$ est bornée.*
2. *$\rho(A) \leq 1$ et les valeurs propres de A de module 1 sont toutes semi-simples.*

PREUVE. Idée essentielle : T la réduite de Jordan de A, de blocs de Jordan les $J_{k_j}(\lambda_j)$, alors

$$(A^k)_k \text{ bornée} \iff \forall j \in \{1, \ldots, n\}, \quad (J_{k_j}(\lambda_j)^k)_k \text{ est bornée.}$$

or

$$J_{k_j}(\lambda_j)^k = \begin{pmatrix} \lambda_j^k & k\lambda_j^{k-1} & \ldots & \binom{k}{i}\lambda_j^{k-i} \\ & \ddots & \ddots & \vdots \\ & & \ddots & k\lambda_j^{k-1} \\ & & & \lambda_j^k \end{pmatrix}.$$

Ainsi,

- si $|\lambda_j| < 1$, $J_{k_j}(\lambda_j)^k \to 0$,
- si $|\lambda_j| > 1$, $|\lambda_j^k| \to +\infty$, donc $(J_{k_j}(\lambda_j)^k)$ n'est pas bornée.

□

Exemple 1.4.8

- Soit
$$A := \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 5 \\ 0 & 0 & 1/2 \end{pmatrix},$$
alors $\rho(A) = 1 > 1$, donc $(A^k)_k$ n'est pas bornée.
- Soit
$$B := \begin{pmatrix} \mathrm{i} & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix},$$
alors $\rho(B) = 1$ et $\sigma(B) = \{1, \mathrm{i}\}$. La valeur propre 1 est défective (associée à un bloc de Jordan de taille $2 > 1$), donc $(B^k)_k$ n'est pas bornée.
- Soit
$$C := \begin{pmatrix} e^{\mathrm{i}\pi} & 0 & 0 \\ 0 & e^{\mathrm{i}\pi/6} & 0 \\ 0 & 0 & e^{\mathrm{i}\pi/6} \end{pmatrix}.$$
Alors $\rho(C) = 1$ et $(C^k)_k$ est bornée.

5 Conditionnement

On se place dans $\mathbf{K}^n$ muni d'une norme vectorielle $\|\cdot\|$ et $\mathcal{M}_n(\mathbf{K})$ muni de la norme induite $\|\cdot\|$.

Définition 1.5.1 (Conditionnement) *Étant donnée $A \in \mathcal{GL}_n(\mathbf{K})$, on appelle conditionnement de A dans la norme $\|\cdot\|$ la quantité*

$$\mathrm{cond}(A) = \|A\|\|A^{-1}\|.$$

Proposition 1.5.2
1. $\mathrm{cond}(A) \geq 1$.
2. $\forall \alpha \in \mathbf{K}^*, \quad \mathrm{cond}(\alpha A) = \mathrm{cond}(A)$.
3. $\forall A, B \in \mathcal{GL}_n(\mathbf{K}), \quad \mathrm{cond}(AB) \leq \mathrm{cond}(A)\,\mathrm{cond}(B)$.

Proposition 1.5.3 (Conditionnement en norme 2)
1. *Soit* $A \in \mathcal{GL}_n(\mathbf{K})$,
$$\mathrm{cond}_2(A) = \sqrt{\frac{\max\{|\lambda| \mid \lambda \in \sigma(A^*A)\}}{\min\{|\lambda| \mid \lambda \in \sigma(A^*A)\}}}.$$
2. *Si de plus* A *est normale, alors*
$$\mathrm{cond}_2(A) = \frac{\max\{|\lambda| \mid \lambda \in \sigma(A)\}}{\min\{|\lambda| \mid \lambda \in \sigma(A)\}}.$$
3. *En particulier, si* $A \in U_n(\mathbf{C})$, *ou si* $A \in O_n(\mathbf{R})$, *alors* $\mathrm{cond}_2(A) = 1$.

PREUVE. Si A est normale, on sait que $\rho(A) = \|A\|_2 = \max\{|\lambda| \mid \lambda \in \sigma(A)\}$. De plus, si $A \in \mathcal{GL}_n(\mathbf{C})$,
$$\rho(A^{-1}) = \max\{|\lambda^{-1}| \mid \lambda \in \sigma(A)\} = \min\{|\lambda| \mid \lambda \in \sigma(A)\}^{-1}.$$
Mais A^{-1} est normale ($(AA^*)^{-1} = (A^*A)^{-1}$), donc $\rho(A^{-1}) = \|A^{-1}\|_2$. Ainsi,
$$\mathrm{cond}_2(A) = \frac{\max|\lambda|}{\min|\lambda|}.$$
□

Théorème 1.5.4 *Soient* $A \in \mathcal{GL}_n(\mathbf{K})$, $b \in \mathbf{K}^n$ *et* $x \in \mathbf{K}^n$ *tels que* $Ax = b$.
1. *Soient* $\delta x \in \mathbf{K}^n$ *et* $\delta b \in \mathbf{K}^n$ *tels que* $A(x + \delta x) = b + \delta b$, *alors*
$$\frac{\|\delta x\|}{\|x\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}.$$
2. *Soient* $\delta x \in \mathbf{K}^n$ *et* $\delta A \in \mathcal{M}_n(\mathbf{K})$ *tels que* $A + \delta A \in \mathcal{GL}_n(\mathbf{K})$ *et* $(A + \delta A)(x + \delta x) = b$, *alors*
$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \mathrm{cond}(A)\frac{\|\delta A\|}{\|A\|}.$$

PREUVE. Soient $A \in \mathcal{GL}_n(\mathbf{C})$, $b \in \mathbf{C}^n$ et $x \in \mathbf{C}^n$ tels que $Ax = b$. Soit $\delta b \in \mathbf{C}^n$ une perturbation. Notons $y := x + \delta x$ l'unique solution de $Ay = b + \delta b$. Donc si $Ax + A\delta x = b + \delta b$, alors $\delta x = A^{-1}\delta b$. Par conséquent,
$$\|\delta x\| = \|A^{-1}\delta b\| \leq \|A^{-1}\|\|\delta b\|,$$
donc
$$\|b\| = \|Ax\| \leq \|A\|\|x\|,$$
d'où
$$\frac{\|\delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\delta b\|}{\|b\|}.$$
□

Exemple 1.5.5 Probèmes mal conditionnés :
- Interpolation polynomiale : Soient $(x_i)_{1\leq i\leq n}$ et $(y_i)_{1\leq i\leq n}$ des complexes. On suppose les $(x_i)_i$ deux à deux distincts. On cherche le polynôme $P \in \mathbf{C}[X]$ de degré minimal vérifaint pour tout $i \in \{1, \ldots, n\}$, $P(x_i) = y_i$.
  Il existe un unique $P \in \mathbf{C}_{n-1}[X]$ solution. On peut le chercher dans la base canonique de $\mathbf{C}_{n-1}[X]$,
$$P = \sum_{j=0}^{n-1} a_{j+1}X^j.$$
  Le vecteur $A = (a_1, \ldots, a_n)^T \in \mathbf{C}^n$ est alors solution de
$$VA = \begin{pmatrix} y_1 \\ \vdots \\ y_1 \end{pmatrix}, \quad \text{avec } V = (x_i^{j-1})_{1\leq i,j\leq n} \text{ la matrice de Vandermonde (inversible).}$$
  $\mathrm{cond}(V)$ est très grand pour n assez grand. Conséquence (en présence d'une arithmétique approchée) : la solution obtenue numériquement est loin d'être bonne. Sur le test, $\mathrm{cond}(V) \simeq 10^{24}$, avec une erreur machine de 10^{-16}. On peut donc s'attendre à $\|\delta x\|/\|x\| \simeq 10^8$.
  Pour contourner cette difficulté, on reformule le problème au niveau algébrique. Cela revient à changer de base dans la formulation d'origine. Par exemple :

— Base de Lagrange associée aux points $(x_i)_{1\leq i\leq n}$,

$$L_i = \prod_{\substack{j=1\\ j\neq i}}^{n} \frac{X - x_j}{x_i - x_j}, \quad P = \sum_{i=1}^{n} y_i L_i.$$

Dans ce cas, la formulation « algèbre linéaire »revient à inverser la matrice identité, de conditionnement 1.

— Base de Newton :

$$P = b_0 + b_1(X - x_1) + b_2(X - x_1)(X - x_2) + \cdots + b_{n-1}(X - x_1)\dots(X - x_{n-1}).$$

L'algorithme des différences divisées (mémo Scilab) permet de trouver les $(b_i)_{1\leq i\leq n-1}$ (de type « taux d'accroissement »).

6 Exemple important : matrice du laplacien 1d

De nombreux problèmes sur des applications des mathématiques font intervenir l'opérateur laplacien

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \quad \text{en dimension 3,}$$

comme

- l'équation de la chaleur : $\dfrac{\partial u}{\partial t} = K\Delta u$,
- l'équation des ondes : $\dfrac{\partial^2 u}{\partial t^2} - c^2\Delta u = 0$.

Exemple 1.6.1 $f \in \mathcal{C}^0([0,1])$ est donnée, on considère le système

$$(\mathcal{S}) : \left\{ \begin{array}{ccccccc} -u''(x) & = & f(x) & , & x \in]0,1[\\ u(0) & = & u(1) & = & 0 \end{array} \right. .$$

Si $f = 0$, l'unique $u \in \mathcal{C}^2([0,1])$ est $u = 0$. Sinon, on peut trouver une formule par intégrations successives (en dimension plus grande, cette méthode échoue).

6.1 Discrétisation aux différences finies

Soit $n \in \mathbf{N}$, $n \geq 1$, on pose $h := 1/(n+1)$ ainsi que $x_i := ih$ pour $0 \leq i \leq n+1$,

$$0 = x_0 \leq x_1 \leq \cdots \leq x_{n+1} = 1.$$

Supposons qu'il existe une unique solution $u \in \mathcal{C}^4([0,1])$ au problème précédent. Par formules de Taylor,

$$\forall i \in \{1,\dots,n\}, \quad u''(x_i) = -f(x_i) = \frac{\dfrac{u(x_{i+1}) - u(x_i)}{h} - \dfrac{u(x_i) - u(x_{i-1})}{h}}{h} + \frac{h^2}{12}u^{(4)}(x_i + \theta_i h), \quad \theta_i \in [-1,1].$$

On simplifie le problème (en réalité on l'approche) en dimension finie, sous forme linéaire en le problème suivant.

Trouver $U \in \mathbf{R}^n$ tel que, ayant posé $U_0 = U_{n+1} = 0$ (conditions aux limites), on ait

$$\forall i \in \{1,\dots,n\}, \quad \frac{U_{i+1} - 2U_i + U_{i-1}}{h^2} = -f(x_i),$$

i. e. $AU = F$, où $F = (f(x_i))_i$ et

$$A = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & (0) \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ (0) & & -1 & 2 \end{pmatrix}.$$

Définition 1.6.2 (Laplacien) *La matrice*

$$A_n = \frac{1}{h^2}\begin{pmatrix} 2 & -1 & & (0) \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ (0) & & -1 & 2 \end{pmatrix}.$$

est égale à l'opposé de la matrice du laplacien.

Le problème à résoudre est $(n+1)^2 A_n U = F$, avec $U \in \mathbf{R}^n$ et $F \in \mathbf{R}^n$ donnée, pour approcher $(\mathcal{S})$.

6.2 Propriétés hermitiennes de A_n

Proposition 1.6.3 *La matrice A_n est symétrique définie positive.*

PREUVE. Soit $U \in \mathbf{R}^n$, on impose $U_0 = U_{n+1} = 0$. Par sommation d'Abel,

$$\langle A_n U, U\rangle = \sum_{i=1}^{n} U_i \underbrace{(-U_{i-1} + 2U_i - U_{i+1})}_{=-(U_{i-1}+U_i)+(U_i-U_{i-1})} = \sum_{i=2}^{n}(U_i - U_{i-1})U_{i-1} + \sum_{i=1}^{n}(U_i - U_{i-1})U_i + U_n^2 = \sum_{i=2}^{n}(U_i - U_{i-1})^2 + U_1^2 + U_n^2.$$

Si $\langle A_N U, U\rangle = 0$, alors $U_1 = U_n = 0$ et pour tout $n \in \{2, \dots, n\}$, $U_i - U_{i-1} = 0$, donc $U = 0$. Par conséquent, $\sigma(A_n) \subset \mathbf{R}_+^*$. En utilisant le rayon spectral, $\rho(A_n) \leq \|A_n\|_\infty \leq 4$ (égalité si $n \geq 3$). Ainsi, $\sigma(A_n) \subset]0, 4]$. □

Remarque 1.6.4 On peut montrer que $A_n - 4I_n$ est symétrique défini négative, et donc $\sigma(A_n) \subset]0, 4]$.

Proposition 1.6.5 (Spectre de A_n)

$$\sigma(A_n) = \left\{ 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \mid 1 \leq k \leq n \right\}.$$

PREUVE. 1ère méthode : Voir TD, par coïncidence algébrique (avec le problème continue).

2ème méthode : Cherchons $U \in \mathbf{R}^n$ et $\lambda \in]0, 4]$ tels que $U \neq 0$ et $A_n U = \lambda U$. On résout

$$\forall i \in \{1, \dots, n\}, \quad -U_{i-1} + 2U_i - U_{i+1} - U_{i+1} = \lambda U_i,$$

avec $U_0 = U_{n+1} = 0$. $(U_i)_{1 \leq i \leq n}$ est solution d'une récurrence linéaire d'ordre 2. Le polynôme caractéristique de cette récurrence est $X^2 - (2 - \lambda)X + 1 = 0$, de racines

$$r := \frac{1}{2}[(2 - \lambda) + \mathrm{i}\sqrt{\lambda(4 - \lambda)}], \quad \text{et} \quad \bar{r}.$$

Ainsi, pour tout j, $U_j = \alpha r^j + \beta \bar{r}^j$, avec $\alpha, \beta \in \mathbf{C}$. Comme $|r|^2 = r\bar{r} = 1$, on a $r = e^{\mathrm{i}\theta}$, avec $\theta \in \mathbf{R}$. Les conditions $U_0 = U_{n+1} = 0$ permettent d'obtenir α et β. En l'occurence, en (α, β),

$$\left\{ \begin{array}{ccccc} U_0 & = & \alpha + \beta & = & 0 \\ U_{n+1} & = & \alpha r^{n+1} + \beta \bar{r}^{n+1} & = & \alpha[e^{\mathrm{i}(n+1)\theta} - e^{-\mathrm{i}(n+1)\theta}]. \end{array} \right.$$

Si la différence d'exponentielles est non nulle, alors $\alpha = \beta = 0$, et $U = 0$, et $\lambda \notin \sigma(A_n)$. Si elle est nulle, on obtient une droite vectorielle de solutions de la forme $U_j = \alpha(r^j - \bar{r}^j)$. Il suffit de trouver $\lambda \in]0, 4[$ tel que $r = e^{\mathrm{i}\theta}$ vérifie

$$e^{\mathrm{i}(n+1)\theta} = e^{-\mathrm{i}(n+1)\theta},$$

ce qui est équivalent à $\theta = k\pi/(n+1)$, où $k \in \mathbf{Z}$. Pour résoudre $\lambda \in]0, 4[$ tel que $r(\lambda) = e^{\mathrm{i}k\pi/(n+1)}$, où $k \in \mathbf{Z}$, on observe que

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

d'où

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

Parmis ces valeurs pour $k \in \mathbf{Z}$, il y a exactement n valeur deux à deux distinctes. Ainsi,

$$\sigma(A_n) \subset \left\{ 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \mid k \in \mathbf{Z} \right\} = \left\{ 4\sin^2\left(\frac{k\pi}{2(n+1)}\right) \mid 1 \leq k \leq n \right\}.$$

L'inclusion réciproque est dans la synthèse de la preuve. □

On peut ainsi identifier explicitement

$$\|A_n\|_2 = \rho(A_n) = 4\sin^2\left(\frac{n\pi}{2(n+1)}\right),$$

et également

$$\|A_n^{-1}\|_2 = \rho(A_n^{-1}) = \left[4\sin^2\left(\frac{n\pi}{2(n+1)}\right)\right]^{-1}.$$

De plus, $\mathrm{cond}_2(A_n) = \|A_n\|_2 \|A_n^{-1}\|_2$.

6.3 Monotonie de la matrice A_n

Définition 1.6.6 (Positivité et monotonie)
- *Un vecteur $v \in \mathbf{R}^n$ est dit positif si ses composantes le sont. On note alors $v \geq 0$.*
- *Une matrice $A \in \mathcal{M}_n(\mathbf{R})$ est dite positive, noté $A \geq 0$, si ses coefficients le sont.*
- *Une matrice $A \in \mathcal{M}_n(\mathbf{R})$ est dite monotone si $A \in \mathcal{GL}_n(\mathbf{R})$ et $A^{-1} \geq 0$.*

Remarque 1.6.7 Il n'y a aucun lien entre la positivité des matrices symétriques et cette positivité ci. Par exemple, la matrice

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

est positive mais de spectre $\{-1, 3\}$.

Proposition 1.6.8 *Soit $A \in \mathcal{M}_n(\mathbf{R})$. Il y a équivalence entre*
- *A est monotone,*
- $\forall v \in \mathbf{R}^n, \quad Av \geq 0 \implies v \geq 0$.

PREUVE. ($\Longrightarrow$) : On suppose $A \in \mathcal{GL}_n(\mathbf{R})$ et $A^{-1} \geq 0$. Soit $v \in \mathbf{R}^n$ tel que $Av \geq 0$, alors $v = A^{-1}(Av)$, d'où

$$\forall j \in \{1, \dots, n\}, \quad v_j = \sum_{i=1}^{n} \underbrace{(A^{-1})_{j,i}}_{\geq 0} \underbrace{(Av)_i}_{\geq 0} \geq 0,$$

donc $v \geq 0$.

($\Longleftarrow$) : Soit $v \in \mathbf{R}^n$ tel que $Av = 0$, alors $Av \geq 0$, donc $v \geq 0$, et $A(-v) \geq 0$, donc $v \leq 0$. Donc $v = 0$, et $\mathrm{Ker}(A) = \{0\}$ et $A \in \mathcal{GL}_n(\mathbf{R})$. De plus, on pose e_i le $i^{\text{ème}}$ vecteur de la base canonique de $\mathbf{R}^n$. $e_i \geq 0$ donc $A^{-1}e_i \geq 0$ par la propriété. Ce vecteur étant la $i^{\text{ème}}$ colonne de A^{-1}, on a bien $A^{-1} \geq 0$. □

Si A est monotone et $x, y \in \mathbf{R}^n$ sont tels que $A(y - x) \geq 0$, alors $y - x \geq 0$. On obtient un principe de comparaison des solutions de $Ax = f$.

Proposition 1.6.9 *Pour tout $n \in \mathbf{N}^*$, A_n est monotone.*

PREUVE. Soit $v \in \mathbf{R}^n$ tel que $A_n v \geq 0$. Montrons que $v \geq 0$. Les coefficients de v vérifient

$$\forall i \in \{1, \dots, n\}, \quad -v_{i-1} + 2v_i - v_{i+1} \geq 0,$$

après avoir posé $v_0 = v_{n+1} = 0$. La première équation est $2v_1 - v_2 \geq 0$. Posons $k \in \{1, \dots, n\}$ tel que $v_k = \min_{1 \leq i \leq n} v_i$. Il suffit de montrer que $v_k \geq 0$.

Si $k = 1$, alors

$$2v_1 - v_2 \geq 0 \implies v_1 = v_k \underbrace{\geq}_{\text{équation}} v_2 - v_1 \underbrace{\geq}_{\text{définition de } k} 0.$$

Si $k = n$, alors

$$2v_n - v_{n-1} \geq 0 \implies v_n = v_k \geq v_{n-1} - v_n \geq 0.$$

Si $k \in \{2, \dots, n-1\}$, alors la $k^{\text{ème}}$ équation $(Av_k)_k = -v_{k+1} + 2v_k - v_{k-1} \geq 0$, de sorte que

$$\underbrace{(-v_{k+1} + v_k)}_{\leq 0} + \underbrace{(v_k - v_{k-1})}_{\leq 0} \geq 0.$$

Ainsi, $v_k = v_{k-1} = v_{k+1}$. On peut, par récurrence sur l'indice, montrer que les v_i sont tous égaux □

Conséquence : Pour $n \geq 3$, on peut montrer que $\|A_n\|_\infty$ vaut $(n+1)^2/8$ si n est impair et $n(n+2)/8$ sinon.

PREUVE. Observons que la fonction

$$\varphi : \begin{array}{ccc} [0,1] & \longrightarrow & \mathbf{R} \\ x & \longmapsto & \frac{1}{2}x(1-x) \end{array}$$

est solution de $-\varphi'' = 1$ sur $]0, 1[$, et que $\varphi(0) = \varphi(1) = 0$. Posons pour $n \geq 3$, $\phi \in \mathbf{R}^n$ tel que

$$\phi_i = \varphi\left(\frac{i}{n+1}\right) = \frac{1}{2(n+1)^2} i(n+1-i).$$

On peut calculer

$$(n+1)^2 A_n \phi = \frac{1}{2}[2i(n+1-i) - (i-1)(n+1-i+1) - (i+1)(n+1-i-1)]_{1 \leq i \leq n} = (1)_{1 \leq i \leq n}.$$

Soit $F \in \mathbf{R}^n$ tel que

$$\forall i \in \{1, \dots, n\}, \quad \|F\|_\infty \leq F_i \leq \|F\|_\infty,$$

alors

$$0 \leq v_1 := \|F\|_\infty \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} - F, \quad \text{et} \quad v_2 := F + \|F\|_\infty \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \geq 0.$$

Or $(n+1)^2 A_n$ est monotone, donc

$$(n+1)^2 A_n v \geq 0 \implies v \geq 0 \quad \text{ou} \quad v \geq 0 \implies ((n+1)^2 A_n)^{-1} V \geq 0.$$

On en déduit que

$$\begin{cases} 0 & \leq & ((n+1)^2 A_n)^{-1} v_1 \\ 0 & \leq & ((n+1)^2 A_n)^{-1} v_2 \end{cases},$$

mais

$$((n+1)^2 A_n)^{-1} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \phi,$$

d'où, en comparant composante par composante,

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

Ainsi,

$$\|U\|_\infty \leq \|F\|_\infty \|\phi\|_\infty.$$

On peut interpréter cela comme

$$\|A_n^{-1}\| = \sup_{F \neq 0} \frac{A_n^{-1} F\|_\infty}{\|F\|_\infty} = \sup_{F \neq 0} \frac{(n+1)^2 \|((n+1)^2 A_n)^{-1} F\|_\infty}{\|F\|_\infty} \leq (n+1)^2 \|\phi\|_\infty.$$

Il y a égalité pour $F = (1, \dots, 1)^T$. Reste à calculer $\|\phi\|_\infty$. □

6.4 Convergence de la méthode (en norme uniforme)

Soit $f \in \mathcal{C}^2([0,1])$ et u l'unique solution du système

$$(\mathcal{S}) : \begin{cases} -u'' & = & f \\ u(0) & = & u(1) = 0 \end{cases}$$

En fait, $u \in \mathcal{C}^4([0,1])$. Pour $n \in \mathbf{N}^*$, on considère les objets suivants.

- $U^{ex} \in \mathbf{R}^n$ le vecteur de composantes $U_i^{ex} = u(i/(n+1))$, pour $1 \leq i \leq n$. En particulier, il vérifie $(n+1)^2 A_n U^{ex} = F + \varepsilon$, avec $F = (f(i/(n+1)))_{1 \leq i \leq n}$, et $\varepsilon \in \mathbf{R}^n$, appelé *erreur de coïncidence*. Par développement de Taylor,

$$\forall i \in \{1, \dots, N\}, \exists \theta_i \in [-1, 1], \quad \varepsilon_i = \frac{1}{(n+1)^2} \frac{1}{12} u^{(4)} \left(\frac{i}{n+1} + \frac{\theta_i}{n+1} \right).$$

  Donc

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

- $U^{(n)} \in \mathbf{R}^n$ la solution du problème approché $(n+1)^2 A_n U^{(n)} = F$.

Question : $\|U^{ex} - U^{(n)}\|_\infty \to_{(n \to +\infty)} 0$? On a

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

donc

$$\|U^{ex} - U^{(n)}\|_\infty \leq (n+1)^{-2} \underbrace{\|A_n^{-1}\|_\infty}_{\sim n^2/8} \underbrace{\|\varepsilon\|_\infty}_{\sim n^{-2} \|f^{(2)}\|_\infty / 12}.$$

Ainsi,

$$\|U^{ex} - U^{(n)}\|_\infty \leq C n^{-2} \|f^{(2)}\|_\infty \xrightarrow[n \to \infty]{} 0,$$

pour $C \in \mathbf{R}^n$ une constante indépendante de n et de f.

Remarque 1.6.10

$$\|U^{ex} - U^{(n)}\|_2^2 = \sum_{i=1}^{n} |U_i^{ex} - U_i^{(n)}|^2 \leq n \|u^{ex} - u^{(n)}\|_\infty^1,$$

donc

$$\|U^{ex} - U^{(n)}\|_2 \leq \sqrt{n} C n^{-2} \|f^{(2)}\|_\infty \leq C n^{-3/2} \|f^{(2)}\|_\infty \xrightarrow[n \to +\infty]{} 0.$$

Chapitre 2

Méthodes directes

1 Remarques préliminaires

Soient $A \in \mathcal{GL}_n(\mathbf{K})$ et $b \in \mathbf{K}^n$. Pour résoudre $Ax = b$ par formule de Cramer, et calculs de déterminants de taille n, la complexité de calcul est en $O(nn!)$. Sur un supercalculateur à 200 petaflops ($= 200.10^{15}$ opérations flottants par seconde), pour $n = 50$, il faut 2.10^{61} années. Pour $n = 12$, il faut environ une année. En réalité, on peut trouver des algorithmes en $O(n^3)$ opérations. Pour $n = 10000$, votre machine peut résoudre en quelques secondes. Le cas particulier des matrices triangulaires mène à deux algorithmes :

- l'algorithme de descente, pour $A \in T_n^l(\mathbf{K})$, en $O(n^2)$ opérations,
- l'algorithme de remontée, pour $A \in T_n^u(\mathbf{K})$, de même complexité.

2 Factorisation LU

Rappel : Opérations du pivot de Gauss.

- Opérer sur les lignes équivaut à multiplier à gauche par A.
- Opérer sur les colonnes équivaut à multiplier à droite par A.
- La matrice $D(\lambda, i) = I_n + (\lambda - 1)E_{i,i}$, où $\lambda \notin \{0, 1\}$, est la matrice de dilatation de la $i^{\text{ème}}$ ligne par λ, $L_i \leftarrow \lambda L_i$.
- La matrice $P(i, j) = I_n - E_{i,i} - E_{j,j} + E_{i,j} + E_{j,i}$ est la matrice de permutation, qui permute les $i^{\text{ème}}$ et $j^{\text{ème}}$ lignes, $(L_i, L_j) \leftarrow (L_j, L_i)$.
- La matrice $T(\lambda, i, j) = I_n + \lambda E_{i,j}$, où $i \neq j$ et $\lambda \neq 0$, est la matrice de transvection, $L_i \leftarrow L_i + \lambda L_j$.

Remarque 2.2.1 Le calcul du rang, déterminant, noyau, image peuvent se faire à partir de ces opérations en examinant les invariants de ces opérations. Par exemple,

$$\operatorname{Ker} A = \operatorname{Ker}(T(\lambda, i, j)A), \quad \operatorname{Im} A = \operatorname{Im}(AT(\lambda, i, j)).$$

Pour résoudre $Ax = b$, 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 $i^{\text{ème}}$ vecteur de la base canonique.

Définition 2.2.2 (Mineurs principaux dominants) *Soit $A \in \mathcal{M}_n(\mathbf{K})$, les mineurs principaux dominants de A sont les $\det((A_{i,j})_{1 \leq i,j \leq k})$, pour $1 \leq k \leq n$.*

Remarque 2.2.3

- Toute matrice dont les mineurs principaux dominants sont non nuls est inversible.
- La réciproque est fausse. Par exemple,
$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}$$
est inversible, mais le mineurs principaux dominants de taille 1 est nul.
- Toute matrice symétrique définie positive réelle a ses mineurs principaux dominants tous non nuls. C'est une conséquence de la réduction des formes bilinéaires. En réalité, si A est symétrique, il y a équivalence entre A symétrique définie positive et ses mineurs principaux dominants sont strictement positifs.

Lemme 2.2.4 *Soit $A \in \mathcal{M}_n(\mathbf{K})$ dont tous les mineurs principaux dominants sont non nuls. Alors pour tous $\lambda \in \mathbf{K}$, pour tous $i < j$, $T(\lambda, i, j)A$ a aussi ses mineurs principaux dominants non nuls.*

PREUVE. Soit $k \in \{1, \dots, n\}$,

$$T(\lambda, i, j)A = \begin{pmatrix} T_k & 0 \\ \tilde{T} & T_{n-k} \end{pmatrix} \begin{pmatrix} A_k & B_k \\ C_k & D_k \end{pmatrix} = \begin{pmatrix} T_k A_k & * \\ * & * \end{pmatrix}.$$

Le mineurs principaux dominants de taille k est $\det(T_k A_k)$, avec

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

donc $\det T_k = 1$, d'où $\det(T_k A_k) = \det A_k \neq 0$. □

Théorème 2.2.5 *Soit $A \in \mathcal{M}_n(\mathbf{K})$ dont les mineurs principaux dominants sont tous non nuls. Alors il existe un unique couple $(L, U) \in T_n^l(\mathbf{K}) \times T_n^u(\mathbf{K})$ tel que*

$$\begin{cases} \forall i \in \{1, \dots, n\}, \quad L_{i,i} = 1 \\ A = LU \end{cases}.$$

PREUVE. Unicité :

$$A = L_1 U_1 = L_2 U_2 \iff \underbrace{L_2^{-1} L_1}_{\in T_n^l(\mathbf{K})} = \underbrace{U_2 U_1^{-1}}_{\in T_n^u(\mathbf{K})}.$$

Les coefficients diagonaux de $L_2^{-1} L_1$ étant tous égaux à 1, on a $L_2^{-1} L_1 = I_n$, d'où $L_1 = L_2$ et $U_1 = U_2$.

Existence : Par l'algorithme effectif de construction. Soit $k \in \{1, \dots, n-1\}$, supposons déterminées des matrices $T_1, \dots, T_{k-1}$ chacune produit de transvections inférieures telles que

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

Par le lemme, pour tous $j \in \{1, \dots, k\}$, $a_{j,j}^{(k)} \neq 0$. En particulier, le pivot $a_{k,k}^{(k)} \neq 0$. On peut faire l'opération $L_i \leftarrow L_i - x_{i,k}^{(k+1)} L_k$, avec $x_{i,k}^{(k)} := a_{i,k}^{(k)} / a_{k,k}^{(k)}$ pour tous $i \geq k+1$. On pose

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

alors

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

On conclut par récurrence sur k. Comment obtenir $A = LU$? On pose $U := T_{n-1} \dots T_1 A \in T_n^u(\mathbf{K})$, ainsi que $L := (T_{n-1} \dots T_1)^{-1} \in T_n^l(\mathbf{K})$ de diagonale 1. On en déduit l'existence. U est construite dans la récurrence. L s'obtient simplement :

$$L = \begin{pmatrix} 1 & & & 0 \\ & \ddots & & \\ & & \ddots & \\ (x_{i,j}^{(j)}) & & & 1 \end{pmatrix}.$$

Pourquoi ?

$$T_k = \prod_{i=k+1}^{n} (I_n - x_{i,n}^{(k)} E_{i,k}),$$

$$T_{k-1} = \prod_{i=k+1}^{n} (I_k + x_{i,n}^{(k)} E_{i,k}) = I_n + \sum_{i=k+1}^{n} x_{i,k}^{(k)} E_{i,k}.$$

Par exemple, $T_k^{-1} T_{k+1}^{-1}$ fait intervenir des produits $E_{i,k} E_{j,k+1}$ avec $i \geq k+1$ et $j \geq k+2$,

$$\begin{aligned} T_k^{-1} T_{k+1}^{-1} &= \left(I + \sum_{i \geq k+1} x_{i,k}^{(k)} E_{i,k} \right) \left(I + \sum_{j \geq k+2} x_{j,k+1}^{(k+1)} E_{j,k+1} \right) \\ &= I + \sum_{i \geq k+1} x_{i,k}^{(k)} E_{i,k} + \sum_{j \geq k+1} x_{j,k+1}^{(k+1)} E_{j,k+1} + 0. \end{aligned}$$

et ce produit est nul. On obtient ainsi

$$L = T_1^{-1} \dots T_{n-1}^{-1} = I_n + \sum_{j=1}^{n-1} \sum_{i=j+1}^{n} x_{i,j}^{(j)} E_{i,j}.$$

□

Conséquence : Pour résoudre un ou plusieurs systèmes $Ax = b$ (pour différents b, par exemple pour A^{-1}), on peut

- précalculer $A = LU$, en $O(2n^3/2)$ opérations,
- résoudre $Ly = b$ et $Ux = y$, chacun en $O(n^2)$ opérations.

On a « stocké »les opérations du pivot du second memebre b dans la matrice L.

Concernant le choix des pivots, si l'un des mineurs principaux dominants est nul, supposons le $k^{\text{ème}}$ le plus petit non nul, alors l'algorithme s'arrêt à la $k^{\text{ème}}$ étape car $a_{k,k}^{(k)} = 0$. Mais si A est inversible, il existe $a_{i,k}^{(k)} \neq 0$ pour $i \geq k+1$ (sans que $\operatorname{rg} A \leq n-1$). On pourrait permuter les lignes k et la i pour utiliser $a_{i,k}^{(k)}$ comme pivot. Cela donne une factorisation de $A = PLU$, où P est une matrice de permutation.

Si un pivot est très petit, une difficulté numérique se pose sur l'arithmétique approchée. L'erreur de calcul sur $a_{i,k}^{(k)}$ par $|a_{k,k}^{(k)}| >> 1$. On peut alors établir une stratégie de pivot en permutant pour limiter cet effet.

- Pivot partiel : Soit avec la ligne i de sorte que $\max_{i \geq k} |a_{i,k}^{(k)}|$ est réalisé sur la diagonale.
- Pivot total : On suppose $|a_{i,j}^{(k)}|$ maximal pour $i, j \geq k$. On permute lignes et colonnes pour utiliser cette valeur comme pivot, $A = PLUQ$.

On peut citer l'exemple des matrices bandes / creuses, dans lesquelles seule un groupe de sous-diagonales est non nul. On appelle *demi-largeur de bande* l'entier p tel que p sous-diagonales soient non nulles de chaque côté de la diagonale. La complexité de la factorisation LU est alors en $O(np^2)$ opérations, et les matrices L et U ont le même profil (seulement p sous-diagonales non nulles).

3 Décomposition de Cholesky

Théorème 2.3.1 (Décomposition de Cholesky) *Soit $A \in S_n^{++}(\mathbf{R})$, il existe une unique $B \in T_n^l(\mathbf{R})$ telle que*

$$\begin{cases} \forall i \in \{1, \dots, n\}, \quad B_{i,i} > 0, \\ A = BB^T. \end{cases}.$$

PREUVE. Existence : A admet une unique factorisation $LU = A$, avec $L_{i,i} = 1$ pour tous $1 \leq i \leq n$.

$$U = \begin{pmatrix} * & & * \\ & \ddots & \\ 0 & & * \end{pmatrix} = D\tilde{U},$$

avec $D = \operatorname{diag}((u_{i,i})_{1 \leq i \leq n})$, $\tilde{U} \in T_n^u(\mathbf{R})$, $\tilde{U}_{i,i} = 1$ pour tous $1 \leq i \leq n$. Comme $A = LU = A^T$, et que A^T admet une factorisation LU, qui est $U^T L^T$. On en déduit que $\tilde{U}^T D^T L^T = LD\tilde{U}$. Par unicité, $L = \tilde{U}^T$, $\tilde{U} = L^T$ et $D = D^T$, et donc $A = LDL^T$. Pour $x \neq 0$, comme A est symétrique définie positive, on a

$$0 < \langle Ax, x \rangle = \langle LDL^T x, x \rangle = \langle DL^T x, L^T x \rangle,$$

donc

$$\forall y \neq 0, \quad \langle Dy, y\rangle > 0,$$

ce qui montre que $D \in D_n^{++}(\mathbf{R})$. On peut donc écrire $A = LD\tilde{U} = L\sqrt{D}\sqrt{D}\tilde{U}$. On pose $B := L\,\mathrm{diag}((\sqrt{D_{i,i}})_{1\leq i\leq n}) \in T_n^l(\mathbf{R})$, avec $B_{i,i} = L_{i,i}\sqrt{D_{i,i}} > 0$. Ainsi, $A = BB^T$.

Unicité : laissé en exercice. □

Algorithme : $A = BB^T$, équations d'inconnues $(B_{i,j})$.

$$\forall j \in \{1, \ldots, n\}, \quad b_{j,j}^2 + \sum_{k=1}^{j-1} b_{j,k}^2 = a_{j,j}.$$

On pose

$$b_{j,j} := \sqrt{a_{j,j} + \sum_{k=1}^{n} b_{j,k}^2},$$

et donc

$$\forall i \geq j+1, \quad b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j+1} b_{j,k}b_{i,k}}{b_{j,j}}.$$

Cette méthode est en $O(n^3/3)$ opérations.

Chapitre 3

Méthodes itératives

1 Généralités

Soit $A \in \mathcal{GL}_n(\mathbf{K})$, on cherche à résoudre $Ax = b$. On choisit deux matrices $M, N \in \mathcal{M}_n(\mathbf{K})$ telles que

$$\begin{cases} A = M - N \\ M \text{ est facile à inverser (algorithmiquement peu coûteuse à inverser).} \end{cases}.$$

On définit alors pour un $x_0 \in \mathbf{K}^n$ donné la suite récurrente

$$\forall k \geq 1, \quad Mx_{k+1} = Nx_k + b.$$

Définition 3.1.1 (Méthode itérative convergente) *On dit qu'une méthode itérative est convergente si pour tout $x_0 \in \mathbf{K}^n$, la suite $(x_k)_k$ converge.*

Remarque 3.1.2 A étant inversible, la seule limite possible est la solution $x = A^{-1}b$.

Définition 3.1.3 (Vocabulaire)

- *Résidu à l'étape k : $r_k := b - Ax_k$.*
- *Erreur à l'étape k : $e_k := x - x_k$.*

On a

$$x_k \xrightarrow[k\to+\infty]{} x \iff e_k \xrightarrow[k\to+\infty]{} 0 \iff r_k \xrightarrow[k\to+\infty]{} 0.$$

Remarque 3.1.4 $e_k = x - x_k = A(b - Ax_k) = A^{-1}r_k$, donc $\|e_k\| \leq \|A^{-1}\|\|r_k\|$. Cela peut servir à définir un critère d'arrêt à partir de $\|r_k\|$ et de $\|A^{-1}\|$.

Théorème 3.1.5 *La méthode itérative pour $A = M - N$ est convergente si et seulement si $\rho(M^{-1}N) < 1$.*

Preuve. On a

$$\begin{cases} \forall k \geq 0, \quad x_{k+1} = M^{-1}Nx_k + M^{-1}b \\ x = M^{-1}Nx + M^{-1}b \end{cases},$$

donc pour tout $k \in \mathbf{N}$, $e_{k+1} = M^{-1}Ne_k$, soit $e_k = (M^{-1}N)^k e_0$. On souhaite que pour tout $e_0 \in \mathbf{K}^n$, $e_k \to_{(k\to+\infty)} 0$, donc que $(M^{-1}N)^k \to_{(k\to+\infty)} 0$. C'est équivalent au fait que $\rho(M^{-1}N) < 1$. □

Remarque 3.1.6 La vitesse de convergence est au plus géométrique, et est liée à la quantité $\rho(M^{-1}N)$:

$$\forall \varepsilon > 0, \exists \|\cdot\| \text{ norme sur } \mathbf{K}^n, \quad \|M^{-1}N\| \leq \rho(M^{-1}N) + \varepsilon.$$

Ainsi,

$$\forall k \in \mathbf{N}, \quad \|(M^{-1}N)^k\|\|e_0\| \leq (\rho(M^{-1}N) + \varepsilon)^k \|e_0\|,$$

avec pour ε petit, $\rho(M^{-1}N) + \varepsilon < 1$. Pour une norme donnée, par exemple $\|\cdot\|_2$ sur $\mathbf{K}^n$, par équivalence des normes en dimension finie, il existe $C_\varepsilon > 0$ tel que

$$\forall k \in \mathbf{N}, \quad \|e_k\|_2 \leq C(\rho(M^{-1}N) + \varepsilon)^k \|e_0\|_2.$$

On appelle *taux de convergence* la quantité $-\log(\rho(M^{-1}N))$. La convergence est d'autant plus rapide (pour le x_0 le pire) que le taux est grand.

2 Méthodes usuelles

2.1 Méthode de Richardson

Soit $\alpha \in \mathbf{R}^*$, on pose $M := \alpha^{-1} I_n$ et $N := \alpha^{-1} I_n - A$. La suite récurrence est

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b = x_k + \alpha(b - Ax_k),$$

et $M^{-1}N = I_n - \alpha A$. Par conséquent,

$$\sigma(M^{-1}N) = \{1 - \alpha\lambda, \lambda \in \sigma(A)\}.$$

Supposons que $A \in S_n^{++}(\mathbf{R})$, alors $\sigma(A) \subset \mathbf{R}_+^*$. On note $0 < \lambda_1 \leq \cdots \leq \lambda_n$ les valeurs propres de A. On a

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

avec $\alpha_* := 2/(\lambda_1 + \lambda_n)$. En particulier,

$$\rho(M^{-1}N) < 1 \iff \alpha \in \left]0, \frac{2}{\lambda_n}\right[= \left]0, \frac{2}{\rho(A)}\right[.$$

Mieux, le rayon spectral est minimal lorsque $\alpha = \alpha_*$. On a alors

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

Si $\mathrm{cond}_2(A) >> 1$, cette quantité approche 1 par valeurs inférieures. Si $\mathrm{cond}_2(A) \simeq 1$, elle approche 0.

2.2 Méthode de Jacobi

On choisit $M := D := \mathrm{diag}(A)$. Il est nécéssaire de supposer $D \in \mathcal{GL}_n(\mathbf{R})$. La matrice d'itération est $M^{-1}N = D^{-1}(D - A) = I_n - D^{-1}A$. En pratique, l'algorithme « en place » prend la forme :

Soit $x^{(0)} \in \mathbf{R}^n$,

$$\forall k \geq 0, \forall i \in \{1, \ldots, n\}, \quad x_i^{(k+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{j \neq i} a_{i,j} x_j^{(k)} \right).$$

Idée : x vérifie

$$\forall i \in \{1, \ldots, n\}, \quad a_{i,i} \underbrace{x_i}_{\text{inconnu}} + \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{i,j} \underbrace{x_j}_{\text{supposé connu}} = b_i.$$

2.3 Méthode de Gauss-Seidel

Soit $A \in \mathcal{M}_n(\mathbf{R})$, on note $D \in \mathcal{M}_n(\mathbf{R})$ sa diagonale, $-E \in T_n^l(\mathbf{R})$ sa sous-diagonale et $-F \in T_n^u(\mathbf{R})$ sa sur-diagonale, de sorte que $A = D - E - F$. On pose

$$\begin{cases} M & := & D - E \\ N & := & F. \end{cases}$$

M est inversible si et seulement si $D \in \mathcal{GL}_n(\mathbf{R})$.

Algorithme : On pose pour tous $k \geq 0$, $i \in \{1, \ldots, n\}$,

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

2.4 Méthode de relaxation (SOR)

Soit $\omega \in \mathbf{R}_+^*$, avec les notations précédentes, on pose

$$\begin{cases} M & := & \dfrac{1}{\omega} D - E \\ N & := & \dfrac{1 - \omega}{\omega} D + F. \end{cases}$$

M est inversible si et seulement si $D \in \mathcal{GL}_n(\mathbf{R})$.

3 Critères explicites de convergence

3.1 Matrices à diagonale dominante

Définition 3.3.1 (Matrice à diagonale strictement dominante) *Soit $A \in \mathcal{M}_n(\mathbf{C})$, on dit qu'elle est à diagonale strictement dominante si*

$$\forall i \in \{1, \dots, n\}, \quad |a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|.$$

Remarque 3.3.2 Toute telle matrice a une diagonale inversible. Les méthodes précédentes peuvent donc s'appliquer, i. e. on peut bien définir la suite récurrente $(x_k)_k$.

Proposition 3.3.3 *Si $A \in \mathcal{M}_n(\mathbf{C})$ est à diagonale strictement dominante, alors $A \in \mathcal{GL}_n(\mathbf{C})$.*

Preuve. Soit $x \in \operatorname{Ker} A$, alors

$$\forall i \in \{1, \dots, n\}, \quad \sum_{j=1}^{n} a_{i,j} x_j = 0,$$

donc

$$\forall i \in \{1, \dots, n\}, \quad a_{i,i} x_i = -\sum_{j \neq i} a_{i,j} x_j.$$

Soit $i \in \{1, \dots, n\}$, par inégalité triangulaire,

$$|a_{i,i}||x_i| \leq \sum_{j \neq i} |a_{i,j}||x_j| \leq \left(\sum_{j \neq i} |a_{i,j}| \right) \|x\|_\infty.$$

Or il existe $i \in \{1, \dots, n\}$ tel que $|x_i| = \|x\|_\infty$, et donc

$$\underbrace{\left(\sum_{j \neq i} |a_{i,j}| - |a_{i,i}| \right)}_{<0} \|x\|_\infty \geq 0,$$

d'où $x = 0$. □

A-t-on convergence des précédentes méthodes pour de telles matrices ?

Théorème 3.3.4 *Soit $A \in \mathcal{M}_n(\mathbf{C})$ à diagonale strictement dominante. Le méthodes de Jacobi, Gauss-Seidel et relaxation convergent pour $\omega \in]0, 1]$.*

Preuve. On se contentera de donner la preuve pour Jacobi. On a avec les notations précédentes, $M^{-1}N = I_n - D^{-1}A$. Soit $\lambda \in \sigma(M^{-1}N)$ tel que $|\lambda| = \rho(M^{-1}N)$ et $x \in \mathbf{C}^n$ un vecteur propre associé. Alors $M^{-1}Nx = \lambda x$, donc $Nx = \lambda Mx$, soit $(D - A)x = \lambda Dx$. On considère l'indice $i \in \{1, \dots, n\}$ tel que $|x_i| = \|x\|_\infty \neq 0$. On a

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

Ainsi,

$$|\lambda| \leq \frac{\displaystyle\sum_{j \neq i} |a_{i,j}||x_j|}{|a_{i,i}||x_i|} \leq \underbrace{\frac{\displaystyle\sum_{j \neq i} |a_{i,j}|}{a_{i,i}}}_{<1 \text{ par hypothèse}} \underbrace{\frac{\|x\|_\infty}{|x_i|}}_{=\|x\|_\infty}.$$

Donc $\rho(M^{-1}N) < 1$ et la méthode converge. □

3.2 Matrices $H_n^{++}(\mathbf{C})$

Théorème 3.3.5 *Soit $A \in H_n^{++}(\mathbf{C})$, on écrit $A = M - N$ avec $M \in \mathcal{GL}_n(\mathbf{C})$. Alors $M^* + N$ est hermitienne.*

Si de plus $M^* + N \in H_n^{++}(\mathbf{C})$, alors $\rho(M^{-1}N) < 1$ (et la méthode itérative converge).

Preuve. $(M^* + N)^* = M + N^* = (A + N) + N^* = (A^* + N^*) + N = M^* + N$, donc cette matrice est hermitienne.

Supposons que $M^* + N \in H_n^{++}(\mathbf{C})$. Il suffit de trouver une norme subordonnée sur $\mathcal{M}_n(\mathbf{C})$ tele que $\|M^{-1}N\| < 1$. On va utiliser la norme induite par A, induite par le produit scalaire

$$\forall x, y \in \mathbf{C}^n, \quad \langle x, y\rangle_A = \langle Ax, y\rangle.$$

Considérons $x \in \mathbf{C}^n$ tel que $\|x\|_A = 1$ et $\|M^{-1}Nx\|_1 = \|M^{-1}N\|_A$. On a

$$\begin{aligned}
\|M^{-1}Nx\|_1^2 &= \langle AM^{-1}Nx, M^{-1}Nx\rangle \\
&= \langle AM^{-1}(M-A)x, M^{-1}(M-A)x\rangle \\
&= \langle Ax - Ax, x - y\rangle \qquad (y := M^{-1}Ax) \\
&= \underbrace{\langle Ax, x\rangle}_{=\|x\|_A=1} - \underbrace{\langle Ay, x\rangle}_{\substack{=\langle y, Ax\rangle \\ =\langle y, My\rangle \\ =\langle M^*y, y\rangle}} \underbrace{-\langle \underbrace{Ax}_{=My}, y\rangle + \langle Ay, y\rangle}_{=-\langle Ny, y\rangle} \\
&= 1 - \langle (M^* + N)y, y\rangle \qquad (y \neq 0 \text{ car } M^{-1}A \in \mathcal{GL}_n(\mathbf{C}), x \neq 0) \\
&< 1.
\end{aligned}$$

□

Exemple 3.3.6 On considère la matrice A_n de l'opposé du Laplacien. On sait que $A_n \in S_n^{++}(\mathbf{R})$. On lui applique la méthode de Jacobi. Posons $M := 2I_n$, alors $M^* + N = -A_n + 4I_n$, mais $\sigma(A_n) \subset]0, 4[$, d'où $\sigma(M^* + N) \subset]0, 4[\subset \mathbf{R}_+^*$. Par conséquent, $M^* + N \in S_n^{++}(\mathbf{R})$ et la méthode de Jacobi converge.

Exemple 3.3.7 Pour la méthode de Gauss-Seidel, $M^* + N = 2I_n$ et la méthode converge aussi.

4 Résolution de systèmes linéaires au sens des moindres carrés

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

1. *carré* si $n = p$,
2. *surdéterminé* si $n > p$,
3. *sous-déterminé* si $n < p$,
4. de *rang maximal* si $\operatorname{rg} A = \min(, np)$,
5. *incompatible* si $b \notin \operatorname{Im} A$,
6. *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.

Mais que se passe-t-il si le système est incompatible ?

Définition 3.4.1 (Résolution au sens des moindres carrés) *Résoudre le système $Ax = b$ au sens des moindres carrés, c'est trouver un vecteur $x \in \mathbf{K}^p$ minimisant la quantité $\|Ax - b\|_2^2$.*

4.1 Existence et unicité

Théorème 3.4.2 *L'ensemble des solutions du système $Ax = b$ au sens des moindres carrés est*

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

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 $v \in \operatorname{Im} A$, d'après le théorème de Pythagore,

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

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

4.2 Équation normale

Lemme 3.4.3 *Les solutions du système $Ax = b$ au sens des moindres carrés sont exactement les solutions de l'équation $A^*Ax = A^*b$.*

PREUVE. D'après le lemme précédent, il faut montrer que $\Gamma_1 := \{x \in \mathbf{K}^p \mid Ax = p(b)\}$ est égal à $\Gamma_2 := \{x \in \mathbf{K}^p \mid A^*Ax = A^*b\}$.

Soit $x \in \Gamma_1$, alors $A^*Ax = A^*p(b)$. Montrons que $A^*p(b) = A^*b$, i. e. que $p(b) - b \in \operatorname{Ker} A^*$. On sait que $\operatorname{Ker} A^* = (\operatorname{Im} A)^\perp$, or $p(b) - b \in (\operatorname{Im} A)^\perp$, d'où $A^*p(b) = A^*b$, et $x \in \Gamma_2$.

Soit $x \in \Gamma_2$. Comme $p(b) - b \in \operatorname{Ker} A^*$, on a

$$A^*Ax = A^*(b - p(b) + p(b)) = A^*p(b),$$

d'où $Ax - p(b) \in \operatorname{Ker} A^*$. Mais $\operatorname{Ker} A^* = (\operatorname{Im} A)^\perp$, et $Ax - p(b) \in \operatorname{Im} A$. Par conséquent, $Ax - p(b) = 0$, donc $x \in \Gamma_1$.

Ainsi, $\Gamma_1 = \Gamma_2$. □

Remarque 3.4.4 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.

4.3 Exemple de la régression linéaire

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

$$\forall i \in \{1, \dots, n\}, \quad ax_i + b = y_i.$$

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

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

L'équation normale fait intervenir la matrice et le vecteur

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

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

$$\begin{cases} a &= \dfrac{n\sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n\sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i^2\right)^2} = \dfrac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ b &= \bar{y} - a\bar{x} \end{cases},$$

avec

$$\bar{x} := \frac{1}{n}\sum_{i=1}^n x_i^2, \quad \text{et} \quad \bar{y} = \frac{1}{n}\sum_{i=1}^n y_i^2.$$

4.4 Factorisation QR par les matrices de Householder

On veut utiliser de symétries orthogonales pour produire une factorisation QR.

Remarque 3.4.5 Si $Q \in H_n(\mathbf{C}) \cap U_n(\mathbf{C})$, alors $Q^{-1} = Q^* = Q$.

Définition 3.4.6 (Matrice de Householder) *On appelle matrice de Householder associée à un vecteur $v \in \mathbf{R}^n$ la matrice*

$$H(v) := \begin{cases} I_n & si \quad v = 0 \\ I_n - 2\dfrac{vv^T}{v^Tv} & sinon \end{cases}.$$

Proposition 3.4.7 *Soit $v \in \mathbf{R}^n$, 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* $\mathrm{Vect}(v)$,
3. *pour tout $e \in \mathbf{R}^n$ tel que $e^*e = 1$, on a*

$$H(v + \|v\|e)v = -\|v\|e, \quad et \quad H(v - \|v\|e)v = \|v\|e.$$

On obtient ainsi un nouvel algorithme d'élimination. Soit $A \in \mathcal{M}_{n,p}(\mathbf{R})$. On note $a^{(1)} \in \mathbf{R}^n$ la première colonne de A et $e_1 \in \mathbf{R}^n$ le premier vecteur de la base canonique de $\mathbf{R}^n$. On pose $H^{(1)} := H(a^{(1)} - \|a^{(1)}\|e_1)$. Alors

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

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

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

où T est triangulaire supérieure et $a^{(k)} \in \mathbf{R}^{n-k+1}$ 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

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

Remarque 3.4.8

- Si $n = p$ et A est inversible, on obtient $A = QR$ avec $R \in T_n^u(\mathbf{R})$ et $Q := H^{(1)} \dots H^{(n)} \in O_n(\mathbf{R})$. On a également $R \in \mathcal{GL}_n(\mathbf{R})$.
- Si $n > p$ et $\operatorname{rg} A = p$, on obtient $A = QR$, avec $Q \in O_n(\mathbf{R})$ et $R \in \mathcal{M}_{n,p}(\mathbf{R})$. La matrice R peut être mise sous la forme

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

  avec $R_1 \in T_p^u(\mathbf{R}) \cap \mathcal{GL}_p(\mathbf{R})$, 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 $x \in \mathbf{R}^p$, on a

  $$\|Ax - b\|^2 = \|QRx - b\|^2 = \|Rx - Q^*b\|^2 = \|R_1x - Q_1^*b\|^2 + \|Q_2^*b\|^2.$$

  On trouve donc la solution au sens des moindres carrés qui est $x = R_1^{-1}Q_1^*b$, et le résidu est $\|Ax - b\| = \|Q_2^*b\|$.
- Si $n > p$ et $r := \operatorname{rg} A < p$, 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 $R_1 \in T_r^u(\mathbf{R})$ et $R_2 \in \mathcal{M}_{n,p-r}(\mathbf{R})$. De plus, la matrice Q se met sous la forme $Q = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}$, avec $Q_1 \in \mathcal{M}_{n,r}(\mathbf{R})$. Alors pour tout $x = (x_1, x_2) \in \mathbf{R}^r \times \mathbf{R}^{p-r}$, 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 $x_1 = R_1^{-1}Q_1^*b - R_1^{-1}R_2x_2$, i. e. pour

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

  Un défaut d'unicité apparaît par la présence du vecteur $x_2 \in \mathbf{R}^{p-r}$ en paramètre, qui est traité en choissant la solution de norme minimale.

4.5 Décomposition en valeurs singulières

Définition 3.4.9 (Valeurs singulières) *Soit $A \in \mathcal{M}_{n,p}(\mathbf{C})$. Les valeurs singulières de A sont les réels positifs $\sigma_i := \sqrt{\lambda_i}$, où les λ_i sont les valeurs propres réelles non nulles de A^*A.*

Remarque 3.4.10 Comme les matrices A et A^*A ont même noyau, elles ont le même rang, donc la matrice A^*A admet exactement $r := \operatorname{rg} A$ valeurs singulières, comptées avec leurs multiplicités.

Théorème 3.4.11 *Soit $A \in \mathcal{M}_{n,p}(\mathbf{C})$ de rang $r \leq \min(n,p)$. Alors A ademt exactement r valeurs singulières comptées avec leurs multiplicités, et il existe $U \in U_n(\mathbf{C})$ et $V \in U_p(\mathbf{C})$ telles que $A = U\Sigma V^*$, avec*

$$\Sigma := \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0) \in \mathcal{M}_{n,p}(\mathbf{R}),$$

où les réels σ_i sont les valeurs singulières de A.

Définition 3.4.12 (Pseudo-inverse) *On appelle pseudo-inverse d'une matrice $A \in \mathcal{M}_{n,p}(\mathbf{C})$ la matrice*

$$A^+ := V\Sigma^+U^*, \quad \textit{avec} \quad \Sigma^+ := \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0) \in \mathcal{M}_{p,n}(\mathbf{R}),$$

en reprenant les notations du théorème précédent.

Remarque 3.4.13 La pseudo-inverse ne dépend pas du choix de la décomposition en valeurs singulières.

Appliquons ceci à la méthode des moindres carrés. On écrit $A = U\Sigma V^*$ la décomposition en valeurs singulières de A. Soit $x \in \mathbf{C}^p$. Alors

$$\|Ax - b\| = \|U^*Ax - U^*b\| = \|\Sigma V^*x - U^*b\|.$$

On pose $c := U^*b \in \mathbf{C}^n$, et $y := V^*x \in \mathbf{C}^p$. 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 $i \in \{1, \dots, r\}$, i. e. $y = \Sigma^+c + y_0$, avec $y_0 := (0, \dots, 0, y_{r+1}, \dots, y_p) \in \mathbf{C}^p$. D'où $x = A^+b + Vy_0$. Parmis ces solutions, le vecteur x est de norme minimale si et seulement si le vecteur y l'est, avec

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

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

5 Méthodes variationelles

5.1 Principe

5.2 Algorithme du gradient à pas fixe

5.3 Algorithme du gradient à pas optimal

On considère une suite $(x_j)_j$ de $\mathbf{R}^n$ définie par

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

Petit calcul : Soit $A \in S_n^{++}(\mathbf{R})$. On va poser $A^{1/2} \in S_n^{++}(\mathbf{R})$ vérifiant $(A^{1/2})^2 = A$. Ainsi,

$$\forall y \in \mathbf{R}^n, \quad \|y\|_A^2 = \langle Ay, y\rangle = \langle A^{1/2}y, A^{1/2}y\rangle = \|A^{1/2}y\|_2^2,$$

donc

$$\forall y \in \mathbf{R}^n, \quad f(y) = \frac{1}{2}(\langle Ay, y\rangle - 2\langle Ax, y\rangle) = \frac{1}{2}(\langle A(y-x), y-x\rangle - \langle Ax, x\rangle) = \frac{1}{2}(\|y-x\|_A^2 - \|x\|_A^2),$$

où $Ax = b$.

Proposition 3.5.1 *Pour tout espace affine $\mathcal{E}$ de $\mathbf{R}^n$,*

$$\operatorname*{argmin}_{\mathcal{E}} f = \operatorname*{argmin}_{y\in\mathcal{E}} \|y-x\|_A^2$$

est le projeté A-orthogonal de x sur $\mathcal{E}$.

PREUVE. On a

$$\forall j \in \mathbf{N}, \quad \|x_{j+1}-x\|_A \leq \frac{k-1}{k+1}\|x_j - x\|_A,$$

avec $K := \operatorname{cond}_2 A$. Notons pour $j \in \mathbf{N}$, $e_j := x_h - x$. Alors

$$\|e_{j+1}\|_A = \min_{\alpha\in\mathbf{R}} \| \underbrace{x_j - \alpha(Ax_j - b) - x}_{:=e_{j+1}^{(R)}} \|_A \leq \|x_j - \alpha_R(Ax_j - b) - x\|_A,$$

où α_R est la constante optimale dans la méthode itérative de Richardson, $\alpha_R := 2/(\lambda_{\min}+\lambda_{\max})$. Et

$$\|e_{j+1}^{(R)}\|_A = \|A^{1/2}e_{j+1}^{(R)}\| = A^{1/2}RA^{-1/2}A^{1/2}e_j\|_2 \leq \|\underbrace{A^{1/2}RA^{-1/2}}_{\in\S_n(\mathbf{R})}\|_2 \|A^{1/2}e_j\|_2 = \underbrace{\rho(A^{1/2}RA^{-1/2})}_{=\rho(R)=\frac{K-1}{K+1}}\|e_j\|_A,$$

où $R := (I - \alpha_R A)$. □

5.4 Espaces de Krylov

Définition 3.5.2 (Espace de Krylov) *Soient $r \in \mathbf{R}^n$, $j \in \mathbf{N}$ et $A \in \mathcal{M}_n(\mathbf{R})$. On appelle espace de Krylov de A associé à r et d'ordre j l'ensemble*

$$K_J := \operatorname{Vect}(r, Ar, \ldots, A^{j-1}r),$$

noté parfois $K_j(A,r)$. On pose par convention $K_0 := \{0\}$.

Proposition 3.5.3 *Soient $r \in \mathbf{R}^n$, $A \in \mathcal{M}_n(\mathbf{R})$ et la suite $(K_j)_j$ vérifiant*

$$K_0 \subset K_1 \subset \cdots \subset K_l = K_{l+1} = \ldots,$$

pour un certain $l \leq n$. On a donc

$$\dim K_j = \begin{cases} j & \text{pour} \quad j \leq l \\ l & \text{pour} \quad j > l. \end{cases}$$

Exercice 3.5.4 Considérer l'ensemble $\{j \in \mathbf{N} \mid (r, Ar, \ldots, A^{j-1}r\}$ est une famille libre$\}$, et l son maximum.

Remarque 3.5.5

- $\forall j \in \mathbf{N}, \quad K_j(A,r) = \{P(A)r \mid P \in \mathbf{R}_{j-1}[X]\}$.
- Les méthodes GPF et GPO vérifient la propriété : en posant pour $j \in \mathbf{N}$, $r_j := b - Ax_j$, on a

$$\begin{cases} r_j \in K_{j+1}(A, r_0) \\ x_j \in x_0 + K_j(A, r_0) \end{cases}$$

Ceci se démontre par récurrence, le cas $j = 0$ étant vérifié. Pour l'hérédité, on utilise le fait que pour $j \in \mathbf{N}$,

$$x_{j+1} = \underbrace{x_j}_{\in x_0+K_j} + \alpha_j \underbrace{r_j}_{\in K_{j+1}},$$

donc

$$r_{j+1} = \underbrace{r_j}_{\in K_{j+1}} - \alpha_j \underbrace{Ar_j}_{\in K_{j+2}} \in K_{j+2}(A, r_0).$$

Pourrait-on modifier les algorithmes pécédents de sorte que $x_j = x = A^{-1}b$ pour j assez grand (typiquement, n), avec toujours $x_j \in x_0 + K_j(A, r_0)$, i. e. construire $P \in \mathbf{R}[X]$ tel que $x - x_0 \in P(A)r_0$?

Proposition 3.5.6 *Soient $A \in \mathcal{GL}_n(\mathbf{R})$, $x_0 \in \mathbf{R}^n$ et $K_l(A, r_0)$ l'espace de Krylov maximal associé à $r_0 := b - Ax_0$. Alors $x = A^{-1}b \in x_0 + K_l(A, r_0)$, un sous-espace vectoriel affine de dimension $l \leq n$.*

PREUVE. La famille liée $(r_0, Ar_0, \ldots, A^l r_0)$ par définition de l'entier l. Alors il existe $\alpha_0, \ldots, \alpha_{l-1} \in \mathbf{R}$ tels que

$$A^l r_0 = \alpha_0 r_0 + \cdots + \alpha_{l-1}A^{l-1}r_0.$$

Pour montrer que $\alpha_0 \neq 0$, on procède par l'absurde. Si $\alpha_0 = 0$, alors

$$A(A^{l-1}r_0) = A(\alpha_1 r_0 + \cdots + \alpha_{l-1}a^{l-2}r_0),$$

ce qui est impossible car $(r_0, \ldots, A^{l-1}r_0)$ est libre. Ainsi,

$$x = \frac{1}{\alpha_0}(A^{l-1}r_0 - \alpha_1 r_0 - \cdots - A^{l-2}r_0) + x_0 \in x_0 + K_l(A, r_0).$$

□

5.5 Algorithme du gradient conjugué

Soit $A \in S_n^{++}(\mathbf{R})$. On veut construire récursivement les projetés A-orthogonaux de $x = A^{-1}b$ sur les sous-espaces affines croissants $x_0 + K_j(A, r_0)$, pour $j \in \mathbf{N}$ à partir d'un choix $x_0 \in \mathbf{R}^n$, et ce sans connaître x. Par la proposition précédente, le $n^{\text{ème}}$ itéré (au plus tard) coïncide avec x. l'algorithme converge exactement en l étapes. Mais comment construire un telle suite $(x_j)_j$, de sorte que

$$x_j = \underset{x_0+K_j(A,r_0)}{\operatorname{argmin}} f.$$

Pour obtenir x_{j+1} à partir de x_j, on va utiliser les directions A-orthogonales des espaces de Krylov.

Soient $x_0 \in \mathbf{R}^n$, $r_0 := b - Ax_0$, $p_0 := r_0$, la direction de descente. La famille (p_0) est une base A-orthogonale de $K_1(A, r_0)$. L'objectif est de construire $(p_0, \dots, p_{l-1})$ une base A-orthogonale de $K_l(A, r_0)$ en drapeau. Connaissant x_j, on cherche $x_{j+1} = x_j + \alpha_j p_j$, pour un α_j bien choisi. On sait que $x_j = p_{\mathcal{E}_j}(x)$, avec $\mathcal{E}_j := x + K_j(A, r_0)$. Alors

$$\forall i \leq j-1, \quad \langle x_{j+1} - x, p_i\rangle_A = \underbrace{\langle x_{j+1} - x_j, p_i\rangle_A}_{=0 \text{ car } \langle pj,\pi\rangle_A=0} + \underbrace{\langle x_j - x, p_i\rangle_A}_{=0 \text{ car } x_j=p_{\mathcal{E}_j}(x)} = 0.$$

Pour $i = j$,

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

Ainsi, on pose

$$\alpha_j := \frac{\langle r_j, p_j\rangle}{\|p_j\|_A^2},$$

avec $r_j := b - Ax_j$.

Comment trouver $(p_j)_j$? On utilise l'agorithme de Gram-Schmidt pour déduire p_j de $(p_0, \dots, p_{j-1})$ base de K_j et de r_j qui complète cette base en une base de K_{j+1}. Cette orthogonalisation est « courte » :

$$p_j = r_j - \sum_{i=0}^{j-1} \beta_i p_i, \quad \text{avec} \quad \beta_i = \frac{\langle r_j, p_i\rangle}{\|p_i\|_A^2}.$$

Pour $i \leq j-2$, $\langle r_j, p_i\rangle_A = \langle r_j, Ap_i\rangle_2$, avec $Ap_i \in K_{j-1}$. or $r_j = b - Ax_j \in K_j^{\perp_A}$, donc $\langle r_j, pi\rangle_A = 0$. Donc $p_j = r_j - \beta_{j-1}p_{j-1}$, avec

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

Chapitre 4

Approximation spectrale

1 Motivation

Soit $A \in \mathcal{M}_n(\mathbf{K})$, $\mathbf{K} = \mathbf{R}$ ou $\mathbf{C}$. Comment trouver les éléments propres de A ? Ce problème pose plusieurs difficultés.

- Pour n grand, on n'a pas d'espoir de résoudre algébriquement.
- L'équation est non linéaire.
- Si on ne connaît A qu'approximativement, le résultat est-il suffisant ?

2 Méthodes numériques

2.1 Méthode de la puissance

Soit $A \in \mathcal{M}_n(\mathbf{R})$, pour trouver la valeur propre $\lambda \in \mathbf{C}$ tel que $|\lambda| = \rho(A)$, on s'appuie sur l'observation vague $A^k x \simeq \lambda^k x$ pour k grand (c'est faux). Plus précisément, A^k est obtenue à partir des sous-espaces propres ou caractéristiques de A et des λ_i^k, $\lambda_i \in \sigma(A)$.

On considère l'algorithme suivant. Soit $x_0 \in \mathbf{C}^n$, on pose pour tout $k \in \mathbf{N}$,

$$x_{k+1} := \frac{Ax_k}{\|Ax_k\|_2}, \quad v_k := \langle x_k, Ax_k \rangle.$$

Théorème 4.2.1 *Soit $A \in \mathcal{M}_n(\mathbf{C})$, de valeurs propres $\lambda_1, \ldots, \lambda_d$, avec*

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_d|.$$

On dit que λ_1 est dominante. Soit $x_0 \notin \oplus_{i=2}^d \operatorname{Ker}((A - \lambda_i I_n)^{m_i})$. On peut interpréter ceci en disant que x_0 possède une composante normale dans $\operatorname{Ker}((A - \lambda_1 I_n)^{m_1})$ dans $\mathbf{C}^n = \oplus_{i=1}^d \operatorname{Ker}((A - \lambda_i I_n)^{m_i})$.

Alors les suites $(x_k)_k$ et $(v_k)_k$ sont bien définies et on a

$$\lim_{k \to +\infty} v_k = \lambda_1,$$

et

$$\lim_{k \to +\infty} q_k = e \in \operatorname{Ker}(A - \lambda_1 I_n),$$

où

$$q_k := \left(\frac{\bar{\lambda_1}}{|\lambda_1|} \right)^k x_k.$$

De plus, $(x_k)_k$ « converge »vers $\operatorname{Ker}(A - \lambda_1 I_n)$.