4.4 Méthodes itératives de résolution d'un système linéaire

Référence : L. Dumas, *Modélisation à l'oral de l'agrégation*, Ellipses, 1999.

Leçons concernées : 157, 162, 226, 233.

Soit $A \in \mathrm{GL}_n(\mathbb{R})$ et soit $b \in \mathbb{R}^n$. On s'intéresse au système $Ax = b$.

Définition 1. On suppose que $A = M - N$ où $M \in \mathrm{GL}_n(\mathbb{R})$ et $N \in \mathcal{M}_n(\mathbb{R})$. On dit que la méthode itérative associée à (M, N) converge si toute suite récurrente de la forme $u_{k+1} = M^{-1}(Nu_k + b)$ pour $k \geqslant 0$ et $u_0 \in \mathbb{R}^n$ converge vers u tel que $Au = b$.

Théorème 2. *La méthode associée à (M, N) converge si et seulement si $\rho(M^{-1}N) < 1$.*

La preuve repose entièrement sur le lemme suivant.

Lemme 3. *Soit $A \in \mathcal{M}_n(\mathbb{C})$ et soit $\varepsilon > 0$, alors il existe une norme subordonnée $||.||$ telle que*

$$||A|| \leqslant \rho(A) + \varepsilon.$$

Démonstration. On trigonalise A dans $\mathbb{C}$: soit $P \in \mathrm{GL}_n(\mathbb{C})$ telle que $P^{-1}AP = T = (t_{i,j})_{1 \leqslant i,j \leqslant n}$ soit triangulaire supérieure. On note également $D_\delta = \mathrm{diag}(1, \delta, \dots, \delta^{n-1})$. On a[1]

$$\begin{aligned} T_\delta = D_\delta^{-1} T D_\delta &= \begin{pmatrix} 1 & & & (0) \\ & \delta^{-1} & & \\ & & \ddots & \\ (0) & & & \delta^{1-n} \end{pmatrix} \begin{pmatrix} t_{1,1} & t_{1,2} & \cdots & t_{1,n} \\ & t_{2,2} & \cdots & t_{2,n} \\ & & \ddots & \vdots \\ (0) & & & t_{n,n} \end{pmatrix} \begin{pmatrix} 1 & & & (0) \\ & \delta & & \\ & & \ddots & \\ (0) & & & \delta^{n-1} \end{pmatrix} \\ &= \begin{pmatrix} t_{1,1} & \delta t_{1,2} & \cdots & \delta^{n-1} t_{1,n} \\ & t_{2,2} & \cdots & \delta^{n-2} t_{2,n} \\ & & \ddots & \vdots \\ (0) & & & t_{n,n} \end{pmatrix} = (\delta^{j-i} t_{i,j})_{1 \leqslant i,j \leqslant n}. \end{aligned}$$

On définit alors la norme $||.||$ sur $\mathbb{R}^n$ par $||x|| = ||(PD_\delta)^{-1}x||_\infty$ et on note $|||.|||$ la norme subordonnée associée. On voit alors facilement que pour $B \in \mathcal{M}_n(\mathbb{C})$, $|||B||| = |||(PD_\delta)^{-1}BPD_\delta|||_\infty$, et on sait que si $B = (b_{i,j})_{i,j} \in \mathcal{M}_n(\mathbb{C})$, $|||B|||_\infty = \sup_{1 \leqslant i \leqslant n} \sum_{j=1}^n |b_{i,j}|$. Ainsi, si $\delta > 0$ est tel que pour tout $i \in [|1, n|]$, $\sum_{j=i+1}^n \delta^{j-i} |t_{i,j}| < \varepsilon$, alors

$$|||A||| = |||T_\delta|||_\infty = \sup_{1 \leqslant i \leqslant n} \sum_{j=1}^n \delta^{j-i} |t_{i,j}| \leqslant \rho(A) + \varepsilon.$$

□

1. On peut également voir cette égalité en remarquant que D_δ est une matrice de passage, exprimer l'endomorphisme T dans cette nouvelle base et utiliser les formules de changement de bases.

Démonstration (Théorème). Soit $u \in \mathbb{R}^n$ tel que $Au = b$, de telle sorte que $Mu = Nu + b$, et soit $e_k = u_k - u$. On a alors,

$$e_{k+1} = M^{-1}(Nu_k + b) - M^{-1}Nu - M^{-1}b = M^{-1}Ne_k$$

et donc par récurrence immédiate, $e_k = (M^{-1}N)^k e_0$. Deux cas se présentent alors.

- Ou bien $\rho(M^{-1}N) < 1$, et dans ce cas, d'après le lemme, il existe une norme subornée $||.||$ telle que $||M^{-1}N|| < 1$. On a alors, si on note aussi $||.||$ la norme sur $\mathbb{R}^n$ associée à $||.||$, $||e_k|| \leqslant ||(M^{-1}N)^k|| \, ||e_0|| \leqslant ||M^{-1}N||^k ||e_0|| \underset{k \to +\infty}{\longrightarrow} 0$ et donc $(u_k)_k$ converge vers u.
- Ou bien $\rho(M^{-1}N) \geqslant 1$. Dans ce cas on note λ une valeur propre complexe de module supérieur ou égal à 1 et $v = v_1 + iv_2 \in \mathbb{C}^n$ le vecteur propre associé. On a alors $(M^{-1}N)^k \, v = \lambda^k \, v$ et donc la méthode ne converge pas pour $u_0 = u + v_1$ ou $u_0 = u + v_2$, puisqu'alors $e_k = \Re(\lambda^k \, v)$ ou $e_k = \Im(\lambda^k \, v)$ et $\lambda^k v$ ne converge pas vers 0, donc au moins sa partie réelle ou sa partie imaginaire ne converge pas vers 0.

□

Définition 4. Soit $A \in \mathrm{GL}_n(\mathbb{R})$ telle que $a_{i,i} \neq 0$ pour tout i. On note $D = \mathrm{diag}(a_{1,1}, \ldots, a_{n,n})$, $-E = (a_{i,j}\mathbb{1}_{i>j})$ et$-F = (a_{i,j}\mathbb{1}_{i<j})$ les parties diagonale, triangulaire inférieure stricte et triangulaire supérieure stricte de A. On définit alors les méthodes

- de Jacobi, où $M = D$ et $N = D - A$, et on pose $J = D^{-1}(D - A)$
- et de Gauss-Seidel, où $M = D - E$ et $N = F$ et on pose $\mathcal{L}_1 = (D - E)^{-1}F$.

Proposition 5. *Si A est une matrice tridiagonale, alors $\rho(\mathcal{L}_1) = (\rho(J))^2$.*

Démonstration. On commence par introduire la matrice $A(\mu)$ pour $\mu \neq 0$ définie par

$$A(\mu) = \begin{pmatrix} a_1 & \mu^{-1} c_2 & & (0) \\ \mu \, b_2 & a_2 & \ddots & \\ & \ddots & \ddots & \mu^{-1} c_n \\ (0) & & \mu \, b_n & a_n \end{pmatrix}$$

où $A(1) = A$. On remarque que si $Q(\mu) = \mathrm{diag}(1, \mu, \ldots, \mu^{n-1})$, $A(\mu) = Q(\mu)AQ(\mu)^{-1}$, et donc $\det(A(\mu)) = \det(A)$.

Les valeurs propres de J sont les racines du polynôme caractéristique $p_J(\lambda) = \det(D^{-1}(E + F) - \lambda I)$ qui sont aussi les racines du polynôme $q_J(\lambda) = \det(\lambda D - E - F)$. De même, les valeurs propres de $\mathcal{L}_1$ sont les racines du polynôme $p_{\mathcal{L}_1}(\lambda) = \det((D - E)^{-1}F - \lambda I)$ qui sont aussi celles du polynôme $q_{\mathcal{L}_1}(\lambda) = \det(\lambda D - \lambda E - F)$. Maintenant si $\lambda \in \mathbb{C}^*$, $q_{\mathcal{L}_1}(\lambda^2) = \det(\lambda^2 D - \lambda^2 E - F) = \lambda^n \det(\lambda D - \lambda E - \lambda^{-1} F) = \lambda^n \det(\lambda D - E - F) = \lambda^n q_J(\lambda)$ d'après le résultat préliminaire. Ainsi, les valeurs propres non nulles de $\mathcal{L}_1$ sont exactement les carrés des valeurs propres non nulles de J, d'où le résultat. □

Remarque. On justifie ici la formule utilisée pour la norme infinie subordonnée. Soit $A \in \mathcal{M}_n(\mathbb{R})$, alors, si $x \in \mathbb{R}^n$,

$$||Ax||_\infty = \max_{1 \leqslant i \leqslant n} \Big| \sum_{j=1}^n a_{i,j} x_j \Big| \leqslant \Big(\max_{1 \leqslant i \leqslant n} \sum_{j=1}^n |a_{i,j}| \Big) ||x||_\infty.$$

Ainsi, $||A||_\infty \leqslant \max_{1 \leqslant i \leqslant n} \sum_{j=1}^n |a_{i,j}|$. D'autre part, si i_0 est tel que $\sum_{j=1}^n |a_{i_0,j}| = \max_{1 \leqslant i \leqslant n} \sum_{j=1}^n |a_{i,j}|$, on définit $x = (x_j)_j$ par $x_j = a_{i_0,j}/|a_{i_0,j}|$ si $a_{i_0,j} \neq 0$ et $x_j = 1$ sinon, et on a alors $||x||_\infty = 1$, et

$$||Ax||_\infty = \max_{1 \leqslant i \leqslant n} \Big| \sum_{j=1}^n a_{i,j} x_j \Big| \geqslant \Big| \sum_{j=1}^n a_{i_0,j} x_j \Big| = \sum_{j=1}^n |a_{i_0,j}| = \Big(\max_{1 \leqslant i \leqslant n} \sum_{j=1}^n |a_{i,j}| \Big) ||x||_\infty$$

d'où le résultat.