Méthodes itératives de résolution d'un système linéaire

Leçons : 157, 162, 226, 233

Soit $A \in \mathrm{GL}_n(\mathbb{R})$, $b \in \mathbb{R}^n$. On étudie le système $Ax = b$.

Définition 1

Si $(M,N) \in \mathrm{GL}_n(\mathbb{R}) \times \mathscr{M}_n(\mathbb{R})$ est tel que $A = M - N$, on dit que la méthode itérative associée à (M,N) converge si pour tout $u_0 \in \mathbb{R}^n$, la suite de premier terme u_0 et définie par $\forall k \in \mathbb{N}, u_{k+1} = M^{-1}(Nu_k + b)$ converge.

Théorème 2

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

Commençons par montrer un lemme :

Lemme 3

Soit $A \in \mathscr{M}_n(\mathbb{C})$, $\varepsilon > 0$. Alors il existe une norme subordonnée $|||\cdot|||$ telle que $|||A||| \leqslant \rho(A) + \varepsilon$.

Démonstration. Comme A est à coefficients dans $\mathbb{C}$, elle est trigonalisable : on se donne donc P inversible et $T = (t_{ij})_{1\leqslant i,j\leqslant n}$ triangulaire supérieure tels que $A = PTP^{-1}$.

Notons $(e_1, \dots, e_n)$ la base canonique de $\mathbb{C}^n$. Pour $\delta > 0$, on pose $e'_1 = \delta^{i-1}e_i$ et $D_\delta = \mathrm{Diag}(1, \delta, \cdots, \delta^{n-1})$.

On a donc

$$\forall j \in [\![1,n]\!], Te'_j = \delta^{j-1}Te_j = \delta^{j-1}\sum_{i=1}^{j} t_{ij}e_i = \sum_{i=1} \delta^{j-i}t_{ij}e'_i,$$

de sorte que

$$T_\delta := D_\delta^{-1}TD_\delta = \begin{pmatrix} t_{11} & \delta t_{12} & \dots & \delta^{n-1}t_{1n} \\ & \ddots & \ddots & \dots \\ (0) & & \ddots & \delta t_{n-1n} \\ & & & t_{nn} \end{pmatrix}.$$

On définit pour $x \in \mathbb{R}^n$, $\| x \| = \| (PD_\delta)^{-1}x \|_\infty$, et on note $|||\cdot|||$ la norme subordonnée associée. On vérifie aisément que $\forall B \in \mathscr{M}_n(\mathbb{R}), |||B||| = |||(PD_\delta)^{-1}BPD_\delta|||_\infty$.

Or (admis ici), pour tout $B = (b_{ij})_{i,j} \in \mathscr{M}_n(\mathbb{R})$, on a $|||B|||_\infty = \sup_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |b_{ij}|$. En choisissant $\delta > 0$ tel que pour tout $1 \leqslant i \leqslant n-1$, $\sum_{j=i+1}^{n} \delta^{j-i}|t_{ij}| \leqslant \varepsilon$, on obtient donc, puisque $\rho(A) = \sup_{1\leqslant i\leqslant n} |t_{ii}|$, $|||A||| = |||T_\delta|||_\infty \leqslant \rho(A) + \varepsilon$. □

Démonstration (du théorème)**.** Soit $u \in \mathbb{R}^n$ tel que $Au = b$, c'est à dire $Mu = Nu + b$. Posons $e_k = u_k - u$ en reprenant les notations du théorème. Alors

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

Ainsi, par une récurrence immédiate, $\forall k \in \mathbb{N}, e_k = (M^{-1}N)^k e_0$. Dès lors, deux cas se présentent :

- Si $\rho(M^{-1}N) < 1$, on fixe $\varepsilon = \dfrac{1-\rho(M^{-1}N)}{2}$ et le lemme nous fournit une norme subordonnée $|||\cdot|||$ telle que $|||M^{-1}N||| \leqslant \rho(M^{-1}N)+\varepsilon < 1$. Donc pour la norme $\|\cdot\|$ associée, on a pour tout k, $\| e_k \| \leqslant |||M^{-1}N|||^k \| e_0 \|$ donc $\lim_{k\to+\infty} e_k = 0$ si bien que $(u_k)_k$ converge vers u.
- Si $\rho(M^{-1}N) \geqslant 1$, soit λ valeur propre complexe de module supérieur ou égal à 1, et $\tilde{u} = \tilde{u}_1 + i\tilde{u}_2$ un vecteur propre associé. Comme pour tout k, $(M^{-1}N)^k\tilde{u} = \lambda^k\tilde{u}$, la méthode itérative ne converge pas pour $u_0 = u + \tilde{u}_1$.

$\square$

Décrivons maintenant quelques cas particuliers de méthodes itératives :

- Méthode de Jacobi : $M = \mathrm{Diag}(a_{11},\dots,a_{nn}) = D$ et $N = D-A$. On note $J = D^{-1}(D-A)$
- Méthode de Gauss-Seidel : $M = D-E$ où $D = \mathrm{Diag}(a_{11},\dots,a_{nn})$ et $E = -A_{\mathrm{inf}}$, partie triangulaire inférieure stricte de A. $N = -A_{\mathrm{sup}} = F$. On note $\mathscr{L}_1 = (D-E)^{-1}F$.
- Méthode de relaxation : $M = \dfrac{D}{\omega} - E$ et $N = \dfrac{1-\omega}{\omega}D + F$,

$$\mathscr{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right).$$

Proposition 4

Si A est une matrice tridiagonale, $\rho(\mathscr{L}_1) = (\rho(J))^2$. La méthode de Gauss-Seidel a donc une vitesse de convergence double de celle de la méthode de Jacobi.

Démonstration. Remarque préliminaire : introduisons pour $\mu \neq 0$:

$$A(\mu) = \begin{pmatrix} b_1 & \mu^{-1}c_2 & & (0) \\ \mu a_2 & b_2 & \ddots & \\ & \ddots & \ddots & \mu^{-1}c_n \\ (0) & & \mu a_n & b_n \end{pmatrix}$$

où $A = A(1)$. Alors $A(\mu) = Q(\mu)A(1)Q(\mu)^{-1}$ où $Q(\mu) = \mathrm{Diag}(\mu,\mu^2,\dots,\mu^n)$, donc $\det A(\mu) = \det A(1)$.

Les valeurs propres de J sont les racines du polynôme caractéristique

$$p_J(\lambda) = \det(D^{-1}(E+F) - \lambda I),$$

ce sont aussi celles de $q_J(\lambda) = \det(\lambda D - E - F)$. De même, les valeurs propres de $\mathscr{L}_1$ sont les racines de $p_{\mathscr{L}_1}(\lambda) = \det((D-E)^{-1}F - \lambda I)$, et celles de $q_{\mathscr{L}_1}(\lambda) = \det(\lambda D - \lambda E - F)$.

Mais selon la remarque préliminaire,

$$\forall \lambda \in \mathbb{C}^*, q_{\mathscr{L}_1}(\lambda^2) = \det(\lambda^2 D - \lambda^2 E - F) = \lambda^n \det(\lambda D - \lambda E - \lambda^{-1}F) = \lambda^n \det(\lambda D - E - F) = \lambda^n q_J(\lambda).$$

Donc les valeurs propres non nulles de $\mathscr{L}_1$ sont les carrés de valeurs propres non nulles de J, ce qui permet de conclure. $\square$

Proposition 5

Le rayon spectral de $\mathscr{L}_\omega$ est strictement supérieur à $|\omega - 1|$. La méthode de relaxation ne peut donc converger que si $\omega \in]0,2[$.

Démonstration. La matrice $\mathscr{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right)$ est trigonalisable comme produit de matrices trigonalisables et en notant $\lambda_1, \dots, \lambda_n$ ses valeurs propres avec multiplicité, on a

$$\prod_{i=1}^{n} \lambda_i = \det(\mathscr{L}_\omega) = \frac{\det\left(\frac{1-\omega}{\omega}D + F\right)}{\det\left(\frac{D}{\omega} - E\right)} = \frac{\prod_{i=1}^{n} \frac{1-\omega}{\omega} a_{ii}}{\prod_{i=1}^{n} \frac{a_i i}{\omega}} = (1-\omega)^n.$$

Donc $\rho(\mathscr{L}_\omega)^n \geqslant |\det(\mathscr{L}_\omega)| = |1-\omega|^n$ de sorte que $\rho(\mathscr{L}_\omega) \geqslant |\omega - 1|$. □

Remarque.
- Par des techniques similaires, on montre que si A est tridiagonale et J a un spectre réel, la méthode de Jacobi et la méthode de relaxation pour $0 < \omega < 2$ convergent ou divergent simultanément. De plus, $\omega_0 = \frac{1}{1+\sqrt{1-\rho(J)^2}}$ est un paramètre de relaxation tel que $\rho(\mathscr{L}_{\omega_0})$ est minimal.
- En 15 minutes, on peut difficilement faire tout le développement, la dernière proposition est là à titre culturel.

Référence : Philippe CIARLET (1988). *Introduction à l'analyse numérique et à l'optimisation*. Masson, p. 102