Dumas, modélisation à l'oral de l'agrégation

Méthodes itératives de résolution d'un système linéaire

Soient $A \in GL_n(\mathbb{R})$ et $b \in \mathbb{R}^n$. On cherche à résoudre le système linéaire $Ax = b$.

Soient $M \in GL_n(\mathbb{R})$ et $N \in \mathcal{M}_n(\mathbb{R})$ telles que $A = M - N$. On dit que la méthode itérative associée au couple (M,N) converge si la suite récurrente définie par $\begin{cases} u_{k+1} = M^{-1}(Nu_k + b) \\ u_0 \in \mathbb{R}^n \end{cases}$ converge pour tout $u_0 \in \mathbb{R}^n$.

<u>Théorème :</u> La méthode itérative associée au couple (M,N) converge si et seulement si $\rho(M^{-1}N) < 1$.

<u>Démonstration :</u>

• <u>Lemme :</u> Soient $A \in \mathcal{M}_n(\mathbb{C})$ et $\varepsilon > 0$. Il existe une norme subordonnée $|||\cdot|||$ telle que $|||A||| \leq \rho(A) + \varepsilon$.

→ <u>Preuve du lemme :</u> On commence par trigonaliser A : on fixe $P \in GL_n(\mathbb{C})$ et $T = \begin{pmatrix} t_{11} & \cdots & t_{1n} \\ 0 & \ddots & \vdots \\ & & t_{nn} \end{pmatrix} \in \mathcal{M}_n(\mathbb{C})$ telles que $T = P^{-1}AP$. On note $(e_1, \dots, e_n)$ la base canonique de $\mathbb{C}^n$.

Soit $\delta > 0$ à ajuster. Pour tout $i \in [\![1, n]\!]$, on pose $e'_i = \delta^{n-i} e_i$. On pose $D_\delta = \mathrm{diag}(1, \delta, \dots, \delta^{n-1})$.

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

On a donc $T_\delta := D_\delta^{-1} T D_\delta = \begin{pmatrix} t_{11} & \delta t_{12} & \cdots & \delta^{n-1} t_{1n} \\ 0 & \ddots & \ddots & \vdots \\ & & & \delta t_{n-1,n} \\ & & & t_{nn} \end{pmatrix}$. On définit la norme $\|\cdot\|$ sur $\mathbb{R}^n$ en posant, pour tout $x \in \mathbb{R}^n$, $\|x\| = \|(PD_\delta)^{-1} x\|_\infty$, et on note $|||\cdot|||$ la norme subordonnée associée. Pour tout $B \in \mathcal{M}_n(\mathbb{R})$, on a $|||B||| = |||(PD_\delta)^{-1} B PD_\delta|||_\infty$.

On prend $\delta > 0$ tel que, pour tout $i \in [\![1, n-1]\!]$, on ait $\sum_{j=i+1}^{n} \delta^{j-i} |t_{ij}| \leq \varepsilon$.

On a alors $|||A||| = |||T_\delta|||_\infty \leq \rho(A)+\varepsilon$, ce qui achève la preuve du lemme.

On retourne à la preuve du théorème. Soit $u \in \mathbb{R}^n$ tel que $Au = b$.

Pour tout $k \in \mathbb{N}$, on pose $e_k = u_k - u$, où $u_0 \in \mathbb{R}^n$ et (u_k) est comme plus haut.

Pour tout $k \in \mathbb{N}$, on a, $e_{k+1} = u_{k+1} - u = (M^{-1}N u_k + M^{-1}b) - (M^{-1}Nu + M^{-1}b)$

$$= M^{-1}N e_k$$

donc $e_k = (M^{-1}N)^k e_0$.

- <u>Si $\rho(M^{-1}N) < 1$</u> : on fixe, par le lemme, une norme subordonnée $|||\cdot|||$ telle que $|||M^{-1}N||| < 1$. On a $\|e_k\| \leq |||M^{-1}N|||^k \|e_0\| \xrightarrow[k\to+\infty]{} 0$.

- <u>Si $\rho(M^{-1}N) \geq 1$</u> : Soient λ une valeur propre complexe de $M^{-1}N$ avec $|\lambda| \geq 1$ et $\tilde{u} = \tilde{u}_1 + i\tilde{u}_2$ un vecteur propre associé (avec $\tilde{u}_1, \tilde{u}_2 \in \mathbb{R}^n$). On a, pour tout $k \in \mathbb{N}$, $(M^{-1}N)^k \tilde{u} = \lambda^k \tilde{u}$.
La méthode itérative ne converge pas pour $u_0 = u + \tilde{u}_1$.

<u>Définition</u> : On définit :

- la méthode de Jacobi comme la méthode itérative associée à la décomposition
$\begin{cases} M = \mathrm{diag}(a_{1,1}, \dots, a_{n,n}) = D \\ N = D - A \end{cases}$. On pose alors $J = M^{-1}N = I - D^{-1}A$;

- la méthode de Gauss-Seidel comme la méthode itérative associée à la décomposition
$\begin{cases} M = \mathrm{diag}(a_{1,1}, \dots, a_{n,n}) + A_{inf} = D - E \\ N = -A_{sup} = F \end{cases}$, où A_{inf} (resp. A_{sup}) est la partie triangulaire inférieure (resp. supérieure) stricte de A. On pose alors $\mathcal{L}_1 = M^{-1}N = (D-E)^{-1}F$.

<u>Théorème</u> : Si A est tridiagonale, on a $\rho(\mathcal{L}_1) = \rho(J)^2$. En particulier, pour de telles A, les méthodes de Jacobi et de Gauss-Seidel convergent dans les mêmes cas, et la méthode de Gauss-Seidel converge plus rapidement.