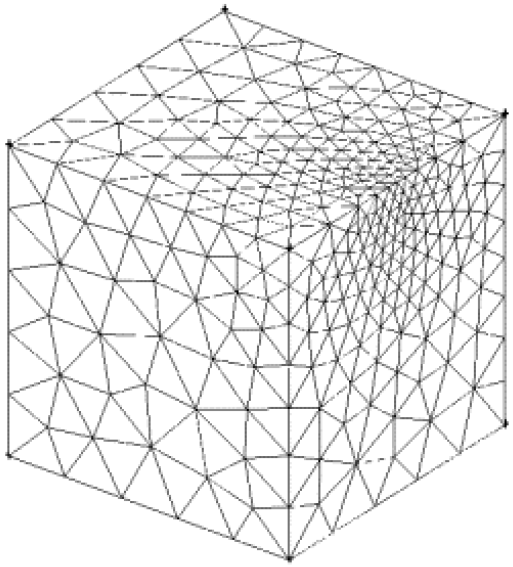
Méthodes itératives pour des systèmes linéaires

On sait aujourd'hui résoudre des systèmes linéaires creux
(i.e. dont la matrice possède une grande proportion de coefficients nuls)
qui sont de très grande taille (centaine de millions d'inconnues)



Exemple :

calcul de la température dans un cube, discrétisation de l'équation de la chaleur par différences finies

maillage 3D du cube :

un pas h=1/n donne $n^3 = N$ points

si n=100 $\Rightarrow N = 10^6$ inconnues

Principe d'une méthode itérative

Soit $A \in M_N(\mathbb{R})$ inversible et $b \in \mathbb{R}^N$.

On veut résoudre le système linéaire $Ax = b \quad (x \in \mathbb{R}^N)$

Une méthode itérative construit une suite récurrente $(x_k)_{k\geq 0}$ telle que :

$(\lim_{k\to\infty} x_k = x) \Rightarrow Ax = b$

La suite $(x_k)_{k\geq 0}$ sera solution d'une récurrence linéaire $x_{k+1} = Q\,x_k + y$

$(Q \in M_N(\mathbb{R}),\ y \in \mathbb{R}^N)$ car le système de départ $Ax = b$ est linéaire

$\Rightarrow$ Avantage pour grands systèmes creux : on ne manipule pas A

mais seulement les vecteurs x_k et une fonction $x_k \mapsto x_{k+1}$

nécessitant de stocker un nombre de coefficients $<< N^2$

Principe d'une méthode itérative

Soit un système linéaire $Ax = b$ ($A \in M_N(\mathbb{R})$ inversible, $b \in \mathbb{R}^N$)

et une méthode itérative linéaire associée à ce système :

$$x_{k+1} = Q\,x_k + y \qquad (Q \in M_N(\mathbb{R}),\ y \in \mathbb{R}^N)$$

Définition :

Une méthode itérative linéaire est convergente si $\lim_{k\to\infty} x_k = x$ pour toute condition initiale $x_0 \in \mathbb{R}^N$

Remarques :

- Si ce n'est pas le cas, on montre que la convergence a lieu pour des conditions initiales dans un sous-espace affine de dimension <N (donc de mesure nulle)
- En pratique, il faut définir un critère d'arrêt

Méthodes itératives avec splitting de A

Soit $A \in M_N(\mathbb{R})$ inversible et $b \in \mathbb{R}^N$.

On veut résoudre le système linéaire $Ax = b \quad (x \in \mathbb{R}^N)$

Décomposition ou "splitting" de A :

$A = M - N$

M est supposée inversible

On considère le schéma itératif :

$M\ x_{k+1} = N\ x_k + b \qquad (S)$

Si $\lim_{k\to\infty} x_k = x$ alors $M\ x = N\ x + b$ *i.e.* $Ax = b$

Méthodes avec splitting de A

Schéma itératif :

$$M\ x_{k+1} = N\ x_k + b \qquad (S)$$

Conditions à satisfaire :

- M inversible
- (S) facile à résoudre (pour chaque itération)
  Par exemple : M diagonale ou triangulaire, diagonale par blocs…
- (S) doit être convergent, avec convergence la plus rapide possible

Le choix du splitting a une grande influence sur la vitesse de CV

Celle-ci est liée aux valeurs propres (au rayon spectral) de $M^{-1}\ N$

Valeurs propres: définitions, propriétés

Soit $A \in M_N(\mathbb{C})$

$\lambda \in \mathbb{C}$ est valeur propre (vp) de A s'il existe un vecteur non nul $v \in \mathbb{C}^N$ tel que : $Av = \lambda v$

v est un "vecteur propre" de A associé à la valeur propre λ

L'ensemble de ces vecteurs propres $\cup\{0\}$ = "espace propre associé à λ"

C'est un sous-space vectoriel de $\mathbb{C}^N$, dimension=multiplicité géométrique de vp λ

Propriétés :

$\lambda \in \mathbb{C}$ valeur propre de $A \Leftrightarrow \exists v \in \mathbb{C}^N, v \neq 0, (\lambda I - A)v = 0 \Leftrightarrow \lambda I - A$ non inversible $\Leftrightarrow \det(\lambda I - A) = 0$

Donc les vp λ sont les racines du polynôme caractéristique $P_A(\lambda) = \det(\lambda I - A)$

$P_A(\lambda) = \det(\lambda I - A)$ est un polynôme complexe de degré N $(N \geq 1)$

$= (\lambda - \lambda_1)(\lambda - \lambda_2)\cdots(\lambda - \lambda_N)$

$\lambda_i \in \mathbb{C}$ sont les racines de P_A : $P_A(\lambda_i) = 0,\ 1 \leq i \leq N$

Spectre de $A \in M_N(\mathbb{C})$ (noté Sp(A)) = ensemble des valeurs propres de A

Les vp de A sont les racines de P_A, donc $\mathrm{Sp}(A) = \{\lambda_i,\ 1 \leq i \leq N\}$

Certaines racines λ_i peuvent être égales.

Multiplicité de la racine λ de P_A = "multiplicité algébrique" de la vp λ

multiplicité algébrique ≥ multiplicité géométrique

Rayon spectral de A = module maximal des valeurs propres de A

Notation : $\rho(A) = \max_{\lambda \in \mathrm{Sp}(A)} |\lambda| = \max_{1 \leq i \leq N} |\lambda_i|$

Exemples de méthodes avec splitting de A

Schéma itératif : $M\ x_{k+1} = N\ x_k + b,$ $\qquad M - N = A$ matrice du système

Méthode de Richardson stationnaire (RS) :

$$M = \frac{1}{\alpha} I, \qquad N = \frac{1}{\alpha} I - A, \quad \alpha \text{ paramètre non nul}$$

$$x_{k+1} = (I - \alpha A)\, x_k + \alpha\, b$$

Théorème : la méthode RS converge si et seulement si $\rho(I - \alpha A) < 1$.
Cela équivaut à ce que toutes les valeurs propres (complexes) de A appartiennent au disque ouvert de centre α^{-1} et rayon $|\alpha|^{-1}$

Le paramètre α optimal, i.e. qui maximise la vitesse de convergence, est celui qui minimise $\rho(I - \alpha A)$

Méthode de Jacobi :

$D\,x_{k+1} = (D - A)\,x_k + b$

on suppose $a_{ii} \neq 0 \forall i = 1,\ldots,N$

$D = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{NN})$

Ici $M = D, \quad N = D - A, \quad M - N = A$

$x_{k+1} = J\,x_k + D^{-1}\,b$

matrice de Jacobi : $J = I - D^{-1}A$

Théorème :

la méthode de Jacobi converge si et seulement si $\rho(J) < 1$.

Théorème : si A est à diagonale strictement dominante alors la méthode de Jacobi converge.

Méthode de Gauss-Seidel (GS) :

$$(L+D)\,x_{k+1} = -U\,x_k + b$$

on suppose $a_{ii} \neq 0 \forall i = 1,\ldots,N$

$$D = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{NN})$$

$$L = \begin{pmatrix} 0 & & \\ a_{ij} & \ddots & \\ (i>j) & & 0 \end{pmatrix} \quad U = \begin{pmatrix} a_{ij}\,(j>i) & & \\ & \ddots & \\ & & 0 \end{pmatrix}$$

Ici $A = L + D + U$,

$M = L + D, \quad N = -U, \quad M - N = A$

On définit la matrice de Gauss-Seidel : $\mathrm{G} = -(L+D)^{-1}U$

Théorème : la méthode GS converge si et seulement si $\rho(G) < 1$.

Théorème :

si A est à diagonale strictement dominante alors la méthode GS converge

Théorème :

si A est symétrique définie positive alors la méthode GS converge.

Méthode de relaxation :

$$(L+\frac{1}{\omega}D)\,x_{k+1}=(\frac{1-\omega}{\omega}D-U)\,x_k+b$$

Paramètre de relaxation : ω

(si ω=1 : Gauss-Seidel,

si $\omega<1$: sous-relaxation

si $\omega>1$: sur-relaxation

ou "SOR")

$$L=\begin{pmatrix} 0 & & \\ a_{ij} & \ddots & \\ (i>j) & & 0 \end{pmatrix} \qquad U=\begin{pmatrix} 0 & & a_{ij}\,(j>i) \\ & \ddots & \\ & & 0 \end{pmatrix}$$

$$D=\mathrm{diag}(a_{11},a_{22},\ldots,a_{NN}),\ a_{ii}\neq 0\,\forall i$$

Ici: $M=L+\frac{1}{\omega}D$, $\quad N=\frac{1-\omega}{\omega}D-U$, $\quad M-N=L+D+U=A$

Théorème :

La méthode de relaxation converge si et seulement si $\rho(Q_\omega)<1$,

où $Q_\omega=(L+\frac{1}{\omega}D)^{-1}(\frac{1-\omega}{\omega}D-U)$. Le paramètre de relaxation ω optimal,

i.e. qui maximise la vitesse de convergence, est celui qui minimise $\rho(Q_\omega)$

Méthode de relaxation : pour $A = L + D + U$,

$$(L + \frac{1}{\omega} D)\, x_{k+1} = (\frac{1-\omega}{\omega} D - U)\, x_k + b \quad \Leftrightarrow \quad (D + \omega L)\, x_{k+1} = ((1-\omega)D - \omega U)\, x_k + \omega b$$

Théorème :

Une condition nécessaire pour que la méthode de relaxation converge est $\omega \in]0,2[$

Théorème :

Si A est à diagonale strictement dominante et $\omega \in]0,1]$ alors la méthode de relaxation converge

Théorème :

Pour toute matrice A symétrique définie positive, la méthode de relaxation converge si et seulement si $\omega \in]0,2[$

Méthodes avec splitting de A: étude de convergence

Equation à résoudre : $M\ x = N\ x + b, \qquad M - N = A$ matrice du système

Schéma itératif : $M\ x_{k+1} = N\ x_k + b \qquad$ (S)

L'erreur $e_k = x_k - x$ vérifie : $M\ e_{k+1} = N\ e_k$

Donc $e_{k+1} = Q\ e_k$ avec $Q = M^{-1}N$

Solution : $e_k = Q^k e_0$

La solution du schéma itératif est donc

$x_k = x + Q^k e_0$

c'est à dire: $x_k = x + (M^{-1}N)^k (x_0 - x)$

Convergence $\Leftrightarrow\ \forall e_0, \lim_{k\to\infty} Q^k e_0 = 0 \Leftrightarrow\ \lim_{k\to\infty} Q^k = 0$

$\Rightarrow$ étudier la limite de puissances de matrice

Pour l'étude générale de la limite de puissances de matrice,

nous allons utiliser un lien entre

normes matricielles subordonnées $\|A\| = \underset{x\neq 0}{\text{Sup}} \dfrac{\|Ax\|}{\|x\|}$

et rayon spectral $\rho(\text{A}) = \underset{\lambda\in \text{Sp(A)}}{\text{Max}} |\lambda|$

Propriété 1 :

$\rho(\text{A}) \leq \|A\|$ pour toute norme matricielle subordonnée

Remarque :

vrai plus généralement pour toute norme matricielle sous-multiplicative

Propriété 2 : Soient $A \in M_N(\mathbb{C})$ et $\varepsilon > 0$.

Il existe une norme subordonnée $\| \ \|$ sur $M_N(\mathbb{C})$ telle que $\|A\| \leq \rho(\text{A}) + \varepsilon$

Théorème :

Soit $Q \in M_N(\mathbb{C})$. $\lim_{k\to\infty} Q^k = 0$ si et seulement si $\rho(Q)<1$

condition nécessaire : si $\rho(Q) \geq 1$ alors : ($\| \ \|$ désigne une norme subordonnée)

$\|Q^k\| \geq \rho(Q^k) = (\rho(Q))^k$ ne tend pas vers 0

condition suffisante : si $\rho(Q)<1$ alors : ($\| \ \|$ désigne une norme subordonnée)

$\|Q\| \leq \rho(Q)+\varepsilon < 1$ pour $\varepsilon > 0$ assez petit, d'où $\|Q^k\| \leq \|Q\|^k \leq (\rho(Q)+\varepsilon)^k \to 0$

à vitesse exponentielle (qq soit la norme)

Remarque si $\rho(Q)<1$:

Plus $\rho(Q)$ est petit, plus la CV $\|Q^k\| \to 0$ est rapide

Plus $\rho(Q)$ est proche de 1, plus la CV $\|Q^k\| \to 0$ est lente

Méthodes avec splitting de A: étude de convergence

Equation à résoudre : A $x = b$

Splitting : $A = M - N$, M inversible

*S*chéma itératif : $M\ x_{k+1} = N\ x_k + b$ (S)

$\Rightarrow\ x_k = x + (M^{-1}N)^k (x_0 - x)$. L'étude précédente implique :

Théorème :

(S) converge si et seulement si $\rho(M^{-1}N) < 1$

Convergence d'autant plus rapide que $\rho(M^{-1}N)$ est petit

Retour sur méthodes de Richardson stationnaire, Jacobi, Gauss-Seidel, relaxation

Schéma itératif : $M\ x_{k+1} = N\ x_k + b, \qquad M - N = A$ matrice du système

Richardson stationnaire: $M = \frac{1}{\alpha} I, \quad N = \frac{1}{\alpha} I - A, \quad \alpha$ paramètre non nul

$M^{-1}N = I - \alpha A$

convergence $\Leftrightarrow \rho(M^{-1}N) < 1$ c'est à dire $\rho(I - \alpha A) < 1$

Jacobi: $M = D, \quad N = D - A, \quad D$ partie diagonale de A

$M^{-1}N = I - D^{-1}A = J$ matrice de Jacobi

convergence $\Leftrightarrow \rho(M^{-1}N) < 1$ c'est à dire $\rho(J) < 1$

Schéma itératif : $M\ x_{k+1} = N\ x_k + b, \qquad M - N = A$ matrice du système

Gauss-Seidel: $M = L + D\,, \quad N = -U$

$A = L + D + U$ (parties triangulaire inférieure, diagonale, triang sup)

$M^{-1}N = -(L+D)^{-1}U = G$ matrice de Gauss-Seidel

convergence $\Leftrightarrow \rho(M^{-1}N) < 1$ c'est à dire $\rho(G) < 1$

Relaxation: $M = L + \frac{1}{\omega} D\,, \quad N = \frac{1-\omega}{\omega} D - U$

$A = L + D + U$

ω paramètre de relaxation ($\omega = 1$ pour Gauss-Seidel)

$$M^{-1}N = (L + \frac{1}{\omega} D)^{-1}(\frac{1-\omega}{\omega} D - U) = (\omega L + D)^{-1}((1-\omega)D - \omega U) = Q_\omega$$

convergence $\Leftrightarrow \rho(M^{-1}N) < 1$ c'est à dire $\rho(Q_\omega) < 1$

Estimations d'erreur et critères d'arrêt

Test d'arrêt classique basé sur le résidu $r_k = A\,x_k - b$:

$$\frac{\|r_k\|}{\|b\|} \le \varepsilon \qquad (\varepsilon \text{ désigne une tolérance relative})$$

Attention : même si ε est petit, l'erreur sur la solution peut être grande si A est mal conditionnée, i.e. si son conditionnement cond(A) est grand :

$\mathrm{cond}(A) = \|A\|\|A^{-1}\|$, où $\|\ \|$ est une norme matricielle subordonnée

$A\,x = b$ et $A\,x_k = b + r_k$ entraînent :

$$\frac{\|x_k - x\|}{\|x\|} \le \mathrm{cond}(A)\,\frac{\|r_k\|}{\|b\|} \le \mathrm{cond}(A)\,\varepsilon$$

Variante : $\|r_k\| \le \eta$ (η désigne une tolérance absolue)

$$\|x_k - x\| \le \|A^{-1}\|\|r_k\| \le \|A^{-1}\|\eta$$

Estimations d'erreur et critères d'arrêt

Autres estimations d'erreur :

utiliser les différences entre itérés successifs $\|x_k - x_{k-1}\|$

On suppose que $Q = M^{-1}N$ vérifie $\rho(Q) < 1$. Dans ce cas :

*$Q - I$ est inversible

*il existe une norme matricielle subordonnée telle que $\|Q\| < 1$

Remarque : ce qui suit reste vrai dans le cas d'une norme matricielle sous-multiplicative compatible avec la norme vectorielle: $\|Qy\| \leq \|Q\|\|y\|$

On relie l'erreur $x_k - x$ à la différence entre deux itérés successifs :

L'erreur $e_k = x_k - x$ vérifie $e_k = Q\,e_{k-1}$

donc $(Q - I)e_k = Q(e_k - e_{k-1}) = Q(x_k - x_{k-1})$

donc $\|x_k - x\| \leq \|(Q-I)^{-1}\|\ \|Q\|\ \|x_k - x_{k-1}\|$ $\Rightarrow$ estimer $\|(Q-I)^{-1}\|$

Estimations d'erreur et critères d'arrêt

Formule de Neumann : $(I-Q)^{-1} = \sum_{k=0}^{\infty} Q^k$ lorsque $\rho(Q) < 1$

et $\left\|(I-Q)^{-1}\right\| \leq (1-\|Q\|)^{-1}$ si $\|Q\| < 1$ (norme sous-multiplicative)

Donc l'erreur $e_k = x_k - x$ vérifie

$$\|x_k - x\| \leq \left\|(Q-I)^{-1}\right\| \|Q\| \|x_k - x_{k-1}\| \leq (1-\|Q\|)^{-1} \|Q\| \|x_k - x_{k-1}\|$$

Si $\dfrac{\|x_k - x_{k-1}\|}{\|x_k\|} \leq \varepsilon$ alors $\dfrac{\|x_k - x\|}{\|x_k\|} \leq C_Q \varepsilon$ avec $C_Q = (1-\|Q\|)^{-1}\|Q\|$

Remarque : $\rho(Q) \leq \|Q\| < 1$, donc $C_Q >> 1$ lorsque $\rho(Q) \approx 1$

Estimation du nombre d'itérations

Nous avons vu que l'erreur $e_k = x_k - x$ vérifie pour $Q = M^{-1}N$:

$e_{k+1} = Q\, e_k \quad \Rightarrow \quad e_k = Q^k e_0$

On suppose $\rho(Q) < 1$ (méthode itérative convergente)

et on considère une norme matricielle subordonnée

(ou sous-multiplicative et compatible avec la norme vectorielle)

telle que $\|Q\| < 1$

Majoration de l'erreur: $\|e_k\| \leq \|Q\|^k \ \|e_0\|$

On fixe une tolérance d'erreur ε (on suppose $\varepsilon < \|e_0\|$)

En calculant un nombre d'itérés k assez grand, avec $\|Q\|^k \ \|e_0\| \leq \varepsilon$

on aura alors: $\|e_k\| \leq \varepsilon$

Remarque:

dans la condition $\|Q\|^k \ \|e_0\| \le \varepsilon$, l'erreur initiale $\|e_0\| = \|x_0 - x\|$ est inconnue, mais elle interviendra souvent uniquement comme une constante multiplicative dans un équivalent de k.

Si besoin on peut aussi majorer $\|e_0\|$:

$\|e_0\| = \|x_0 - x\| = \|A^{-1}(Ax_0 - b)\|$ donne $\|e_0\| \le \|A^{-1}\| \|(Ax_0 - b)\|$

$(Q - I)e_0 = e_1 - e_0 = x_1 - x_0$ donne

$\|e_0\| \le \|(Q - I)^{-1}\| \|(x_1 - x_0)\| \le (1 - \|Q\|)^{-1} \|(x_1 - x_0)\|$

On a $\|Q\|^k \ \|e_0\| \le \varepsilon$ (avec $\varepsilon < \|e_0\|$) si et seulement si

$k \ge \dfrac{\ln(\|e_0\| / \varepsilon)}{|\ln(\|Q\|)|} = k_{\min}$ dont on souhaite obtenir une approximation

Nous avons montré que si $k \geq k_{\min} = \dfrac{\ln(\|e_0\| / \varepsilon)}{|\ln(\|Q\|)|}$ alors $\|e_k\| \leq \varepsilon$

On considère typiquement la tolérance d'erreur ε et $\|e_0\|$ fixés,
et on approche $k_{\min}$ (majoration, encadrement, équivalent)
en estimant $\|Q\|$.

Calcul classique :
pour une famille de matrices A donc la taille n tend vers l'infini :

– calculer un équivalent (ou ordre) de $k_{\min}$ à partir d'une estimation de $\|Q\|$,
– estimer le nb d'opérations arithmétiques élémentaires à chaque itération,
– déduire le nb d'opérations nécessaires pour résoudre le système $Ax = b$
(équivalent ou ordre quand $n \to \infty$)

Exemple: matrice tridiagonale $A \in M_n(\mathrm{IR})$
obtenue dans un schéma différences finies (cf chapitre 1):

$$A = \begin{pmatrix} 2+h^2c_1 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & 2+h^2c_i & \ddots & 0 \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2+h^2c_n \end{pmatrix} \text{ avec } h = \frac{1}{n+1}$$

et $c_i = c(ih)$ $(1 \le i \le n)$, c fonction continue > 0 sur $[0,1]$,

on note $c_{\min} = \min_{[0,1]}(c)$ $\quad (c_{\min} > 0)$

A est à diagonale strict. dominante donc la méth. de Jacobi converge
On évalue le coût (nombre d'opérations arithmétiques élémentaires)
de la méthode de Jacobi quand $n \to \infty$

$$A = \begin{pmatrix} 2+h^2c_1 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & 2+h^2c_i & \ddots & 0 \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2+h^2c_n \end{pmatrix}$$

diagonale strictement dominante donc (cf TD 3, exercice 1)

la matrice de Jacobi $J = I - D^{-1}A$ vérifie $\|\|J\|\|_\infty < 1$

Lorsque $h \to 0$ $(n \to \infty)$:

$$\|\|J\|\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |J_{ij}| = \frac{2}{2+h^2(c_{\min}+o(1))} = 1 - \frac{c_{\min}}{2}h^2 + o(h^2)$$

Lorsque $h \to 0$ $(n \to \infty)$: $\quad \|\|J\|\|_\infty = 1 - \dfrac{c_{\min}}{2} h^2 + o(h^2)$

Nombre d'itérations pour atteindre tol. d'erreur ε : $\quad (\ a := \ln(\|e_0\| / \varepsilon))$

$$k_{\min} = \frac{a}{\left|\ln(\|\|J\|\|_\infty)\right|} \sim \frac{2a}{c_{\min} h^2} \sim \frac{2a}{c_{\min}} n^2 \text{ lorsque } n \to \infty$$ (CV très lente)

Coût d'une itération $x_{k+1} = J\ x_k + D^{-1}\ b$:

$O(n)$ opérations car $J_{i,j} = 0$ si $j \neq i-1, i+1$

La résolution du système $Ax = b$ par la méthode de Jacobi nécessite donc $O(n^3)$ opérations quand $n \to \infty$

$\Rightarrow$ méthode pas appropriée car nous verrons que ce système (matrice tridiagonale, à diagonale strictement dominante) peut être résolu en $O(n)$ opérations par une méthode directe.

Cas des puissances de matrice diagonalisable

Définition :

$M \in M_n(\mathbb{C})$ est diagonalisable s'il existe une base de $\mathbb{C}^n$ formée de vecteurs propres de M

$M \in M_n(\mathrm{IR})$ est diagonalisable sur IR s'il existe une base de IR^n de vecteurs propres de M

Reformulation matricielle: $M = PDP^{-1}$

avec $P \in M_n(\mathrm{IK})$ (IK = $\mathbb{C}$ ou IR) inversible (colonnes = vecteurs propres de M),

$D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, $\quad \mathrm{Sp}(M) = \{\lambda_i,\ 1 \le i \le n\} \subset \mathrm{IK}$

Dans une méthode itérative avec splitting $A = M - N$ et

$Q = M^{-1}N$ diagonalisable, i.e. $Q = PDP^{-1}$,

l'erreur $e_k = x_k - x = Q^k e_0$ s'écrit plus simplement avec:

$$Q^k = PD^kP^{-1} \quad , \quad D^k = \mathrm{diag}(\lambda_1^k, \lambda_2^k, \ldots, \lambda_n^k)$$

Pour un choix approprié de norme vectorielle, nous allons en déduire :

$$\|e_k\| \le \rho(Q)^k \ \|e_0\|$$

Proposition:

Si $M \in M_n(\mathbb{C})$ est diagonalisable alors il existe une norme matricielle subordonnée telle que $\|M\| = \rho(M)$

Preuve: $M = PDP^{-1}$ avec $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$, $\mathrm{Sp}(M) = \{\lambda_i,\ 1 \le i \le n\}$

Soit la norme vectorielle : $\|x\| = \|P^{-1}x\|_2$, où $\|y\|_2 = \left(\sum_i |y_i|^2\right)^{1/2} = \overline{y}^T y$

et la norme matricielle subordonnée $\|M\| = \underset{x \neq 0}{\mathrm{Sup}} \dfrac{\|Mx\|}{\|x\|} \ge \rho(M)$

Alors pour tout vecteur x :

$\|Mx\| = \|P^{-1}Mx\|_2 = \|DP^{-1}x\|_2 = \left(\sum_i |\lambda_i y_i|^2\right)^{1/2}$ avec $y = P^{-1}x$

d'où $\|Mx\| \le \rho(M)\|y\|_2 = \rho(M)\|x\|$ pour tout vecteur x, donc $\|M\| \le \rho(M)$

Les deux inégalités en sens opposés $\Rightarrow$ $\|M\| = \rho(M)$

Remarques:

si $M \in M_n(\mathrm{IR})$ symétrique ($M^T = M$), alors $\mathrm{Sp}(M) = \{\lambda_i,\ 1 \le i \le n\} \subset \mathrm{IR}$

et M est diagonalisable dans une base orthonormée: $M = PDP^{-1}$

avec $P^T P = I$ (P matrice orthogonale), $D = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$.

On a alors $\|x\| = \|x\|_2$ dans le calcul précédent:

$$\|x\| = \|P^{-1}x\|_2 = \|P^T x\|_2$$

$$= \sqrt{(P^T x)^T P^T x} = \sqrt{x^T P P^T x} \text{ avec } PP^T = I$$

$$= \sqrt{x^T x} = \|x\|_2$$

La norme matricielle choisie est donc subordonnée à $\|\ \|_2$: $\|M\| = \|M\|_2 = \rho(M)$

L'identité $\|M\|_2 = \rho(M)$ est vraie plus généralement lorsque $M \in M_n(\mathbb{C})$ est une matrice normale, i.e. commute avec $\bar{M}^T$, car cette propriété équivaut à :

$M = PDP^{-1}$ avec D diagonale, $P \in M_n(\mathbb{C})$ unitaire: $\bar{P}^T P = I$

Etant donné une méthode itérative avec splitting $A = M - N$ où $Q = M^{-1}N$ est diagonalisable, il existe donc des normes vectorielle et matricielle subordonnée telles que:

$$\|Q\| = \rho(Q),$$

avec lesquelles l'estimation d'erreur $\|e_k\| \le \|Q\|^k \ \|e_0\|$ devient:

$$\|e_k\| \le \rho(Q)^k \ \|e_0\|$$

Exemple: méthode de Jacobi pour $A \in M_n(\mathrm{IR}) \Rightarrow Q = J = I - D^{-1}A$

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & \ddots & \ddots & & \vdots \\ 0 & \ddots & 2 & \ddots & 0 \\ \vdots & & \ddots & \ddots & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix} \Rightarrow \rho(J) = \rho(I - \frac{1}{2}A) = \cos\left(\frac{\pi}{n+1}\right) < 1$$

et J symétrique réelle

Nb d'itérations pour avoir $\|e_k\|_2 \le \rho(J)^k \|e_0\|_2 \le \varepsilon$:

$k \ge \dfrac{a}{|\ln(\rho(J))|} \sim C n^2$ lorsque $n \to \infty$ (CV très lente)

Cas des matrices tridiagonales

Hypothèse (H) : système linéaire $Ax = b$,
matrice A inversible, tridiagonale ($a_{ij} = 0$ si $|i-j| \geq 2$), $a_{ii} \neq 0$

$J = I - D^{-1}A$, $G = -(L+D)^{-1}U$ matrices de Jacobi et Gauss-Seidel
($A = L + D + U$: parties triangulaire inférieure, diagonale, triang sup)

Théorème: sous l'hypothèse (H), les méthodes de Jacobi et Gauss-Seidel convergent ou divergent simultanément,
et $\rho(G) = \rho(J)^2$

Remarque: on considère souvent que la méthode de GS nécessite environ deux fois moins d'itérations que celle de Jacobi, en considérant que l'erreur à la $k$ième itération est de l'ordre de $\rho(G)^k = \rho(J)^{2k}$

cas des matrices tridiagonales

Hypothèse (H) :

La matrice A est inversible, tridiagonale ($a_{ij} = 0$ si $|i-j| \geq 2$), $a_{ii} \neq 0$

Théorème (paramètre de relaxation optimal):

On fait l'hypothèse (H) sur la matrice A.

On suppose de plus que les valeurs propres de J sont réelles $\in (-1,1)$.

Alors la méthode de relaxation converge si et seulement si $\omega \in]0,2[$.

De plus, le paramètre de relaxation optimal (qui minimise $\rho(Q_\omega)$) est

$$\omega_J = \frac{2}{1+\sqrt{1-\rho(J)^2}} = \frac{2}{1+\sqrt{1-\rho(G)}} \in [1,2[$$

avec

$$\min_\omega \rho(Q_\omega) = \rho(Q_{\omega_J}) = \omega_J - 1 \in [0,1[$$

cas des matrices tridiagonales

Remarques:

- $\omega_J = \dfrac{2}{1+\sqrt{1-\rho(J)^2}} > 1$ si $\rho(J) \neq 0$

  (Gauss-Seidel ne donne pas la vitesse optimale de CV)

- Si $\rho(J) \approx 1$ alors $\omega_J \approx 2$

- Il existe des méthodes pour approcher numériquement $\rho(J)$ et donc ω_J

cas des matrices tridiagonales

Théorème: si la matrice A est tridiagonale symétrique définie positive alors le théorème du paramètre de relaxation optimal s'applique.

Preuve: comme A est symétrique définie positive, A est inversible et

$a_{ii} = e_i^T A e_i > 0$ (e_i : ième vecteur base canonique)

La matrice A vérifie donc l'hypothèse (H). Il reste à montrer que les valeurs propres de $J = I - D^{-1}A$ sont réelles $\in (-1,1)$.

Les coefficients de D étant > 0, on définit $D^{1/2} = \mathrm{diag}(\sqrt{a_{ii}})$,

d'inverse $D^{-1/2} = \mathrm{diag}(1/\sqrt{a_{ii}})$

$D^{-1}A = D^{-1/2}\left(D^{-1/2} A D^{-1/2}\right)D^{1/2}$ est semblable à la matrice $D^{-1/2} A D^{-1/2}$

cas des matrices tridiagonales

(suite de la preuve)

$D^{-1}A = D^{-1/2}\left(D^{-1/2}AD^{-1/2}\right)D^{1/2}$

$D^{-1/2}AD^{-1/2}$ est sym. définie positive donc ses vp sont réelles >0

($x^T D^{-1/2}AD^{-1/2}x = (D^{-1/2}x)^T A(D^{-1/2}x) > 0 \;\forall x \neq 0$ car $D^{-1/2}$ inversible)

donc $\mathrm{Sp}(D^{-1}A)=\mathrm{Sp}(D^{-1/2}AD^{-1/2})>0$

Donc les valeurs propres de $J = I - D^{-1}A$ sont réelles,

et $\mathrm{Sp}(\mathrm{J}) = 1 - \mathrm{Sp}(D^{-1}A) < 1$

Comme A est symétrique définie positive, la méthode de Gauss-Seidel converge.

Donc, A étant tridiagonale, $\rho(\mathrm{J})=\sqrt{\rho(\mathrm{G})}<1$. Donc $\mathrm{Sp}(\mathrm{J}) \subset]-1,1[$

Exemple :

$$A = \begin{pmatrix} 2 & -1 & 0 & \cdots & 0 \\ -1 & 2 & -1 & 0 & \vdots \\ 0 & -1 & \ddots & \ddots & 0 \\ \vdots & 0 & \ddots & 2 & -1 \\ 0 & \cdots & 0 & -1 & 2 \end{pmatrix}$$

$A \in M_n(\mathbb{R})$

est symétrique définie positive

Ses valeurs propres valent :

$$\lambda_m = 4\sin^2\left(\frac{m\pi}{2(n+1)}\right), \quad m = 1, \ldots, n.$$

$\rho(\mathrm{J})=\cos(\pi h)$ avec $h = 1/(n+1) \Rightarrow$ paramètre de relaxation optimal:

$$\omega_J = \frac{2}{1+\sqrt{1-\rho(J)^2}} = \frac{2}{1+\sin(\pi h)}$$

avec

$$\rho(Q_{\omega_J}) = \omega_J - 1 = \frac{1-\sin(\pi h)}{1+\sin(\pi h)} = 1 - 2\pi h + O(h^2) \text{ quand } h \to 0 \ (n \to \infty)$$

Le paramètre de relaxation optimal $\omega_J = \dfrac{2}{1+\sin(\pi h)}$ donne

$\rho(Q_{\omega_J}) = 1 - 2\pi h + O(h^2)$ quand $h \to 0$ $(n \to \infty)$

Si erreur $\|e_k\| \approx \rho(Q_{\omega_J})^k$, le nb d'itérations pour avoir $\|e_k\| \le \varepsilon$ est :

$$k \ge \frac{a}{\left|\ln(\rho(Q_{\omega_J}))\right|} \sim C\,n \text{ lorsque } n \to \infty$$

En comparaison, on doit réaliser $O(n^2)$ itérations avec les méthodes de Jacobi et Gauss-Seidel car $\rho(\mathrm{J})=1-O(h^2)$ et $\rho(\mathrm{G})=1-O(h^2)$

Coût d'une itération $M\ x_{k+1} = N\ x_k + b : O(n)$ opérations (M, N bidiagonales)
$\Rightarrow$ résoudre le système tridiagonal nécessite $O(n^2)$ opérations
avec le paramètre de relaxation optimal ω_J