1 Résolution des systèmes linéaires, méthodes itératives

28/10 : début

Références

[S] M. Schatzman. *Analyse numérique : Cours et exercices pour la licence.* InterEditions 1991.

[D] D. Serre. *Les matrices.* Dunod, Masson Sciences 2001.

[CM] P.G. Ciarlet. *Introduction à l'analyse numérique matricielle et à l'optimisation.* Dunod 1998, Masson 1990.

1.1 Introduction

On connaît avec Gauss ou LU une méthode pour résoudre $Ax = b$ de taille n en $O(n^3)$ opérations arithmétiques. Mais

- Si la matrice est creuse, avec par exemple au plus $m << n^2$ coefficients non nuls, le coût du calcul du produit Ax (par l'algorithme le plus évident) passe de $O(n^2)$ à $O(m)$. Une méthode itérative avec une évaluation de $f(x) = Ax-b$ à chaque itération peut alors être moins couteuse.
- Quand à la précision, on a vu qu'avec les méthodes directes, la succession des $O(n^3)$ opérations, même avec les diverses stratégies de pivotage, peut poser un problème de stabilité numérique (qui se conjugue à l'instabilité naturelle représentée par un mauvais conditionnement). On va voir qu'une méthode itérative qui converge est forcément stable, qu'elle corrige d'elle-même peu à peu par la suite les erreurs (par exemple numériques) qu'elle introduit à chaque itération.
- Dans nombre de cas, on doit résoudre successivement une suite de systèmes linéaires $A_k x_k = b_k$ où A_{k+1} et b_{k+1} (qui peuvent dépendre de x_k) diffèrent peu de A_k et b_k. Il est alors naturel de penser que x_k est déjà une bonne approximation de x_{k+1}, et qu'il suffit de le corriger un peu. Un exemple typique est la méthode Newton pour résoudre un système de n équations non linéaires à n inconnues.

On doit résoudre $f(x) = Ax-b$, système de n équations linéaires à n inconnues. On le fait par une méthode de point fixe, $x_{k+1} = \Phi(x_k)$, avec Φ linéaire (plutôt affine) de la forme $\Phi(x) = Tx + c = x - M^{-1}f(x)$, avec M matrice inversible telle que le système $My = f$ soit facile à résoudre. On a $T = I - M^{-1}A = M^{-1}N$ avec $N = M - A$ et donc $A = M - N$. Et $c = M^{-1}b$. Le théorème de point fixe pour Φ contractant dans X métrique complet se présente dans ce cas particulier (T linéiare) ainsi :

Théorème. *$\mathbb{K}$ est le corps, $\mathbb{R}$ ou $\mathbb{C}$. On se donne dans $\mathbb{K}^{n\times n}$ les matrices $A = M - N$, M inversible et $T = M^{-1}N$. Les assertions suivantes sont équivallentes*

1. *Pour tout b et x_0 dans $\mathbb{K}^n$, la suite $x_{k+1} = Tx_k + M^{-1}b$ converge vers $A^{-1}b$.*
2. *La suite T^k converge vers 0.*
3. *$c(T) < 1$, avec $c(T) = \inf N(T)$, l'infimum pris sur toutes les normes N sur $\mathbb{K}^{n\times n}$ induites par les normes N sur $\mathbb{K}^n$. Autrement dit, il existe N norme sur $\mathbb{K}^n$ telle que T (et Φ) soit strictement contractante sur $\mathbb{K}^n$ pour cette norme.*

4. *$r(T) < 1$, avec $r(T) = \lim_k N(T^k)^{1/k}$, pour toute norme sur $\mathbb{K}^{n\times n}$, induite ou non, matricielle ou non. Dans le cas des normes matricielles (*ie $N(AB) \leq N(A)N(B)$*), cette limite est en fait un* inf.
5. *$\rho(T) < 1$ où $\rho(T)$ est le rayon spectral, le plus grand des modules des valeurs propres dans $\mathbb{C}$ de T. Autrement dit pour tout μ tel que $|\mu|\rho(T) < 1$, $I - \mu T$ est inversible (dans $\mathbb{C}^{n\times n}$), et il existe μ avec $|\mu|\rho(T) = 1$ tel que $I - \mu T$ n'est pas inversible.*

Remarque. Pour tout scalaire $\mu \in \mathbb{K}$, on vérifie facilement $c(\mu T) = |\mu|c(T)$, et de même pour r et ρ. Donc le théorème assure que $c(T) = r(T) = \rho(T)$ pour toute matrice T.

Dém. Commençons par montrer que si N est une norme matricielle, pour toute matrice T la suite $N(T^k)^{1/k}$ converge vers son inf. On note $v_k = \ln N(T^k)$. Parce que N, est matricielle, $v_{p+q} \leq v_p + v_q$. Fixons p. Pour tout k la division euclidienne donne $k = pq + r$ avec $r < p$, donc $v_k \leq v_{pq} + v_r \leq qv_p + \max_{r<p} v_r = (k-r)v_p/p + \max_{r<p} v_r$. On divise par k puis on prend la lim sup quand $k \to \infty$: $\limsup_k v_k/k \leq v_p/p$. On prend l'inf sur p et l'exponentielle, et c'est fini. Si maintenant N' est une autre norme sur $\mathbb{K}^{n\times n}$, comme les normes sont équivallentes, et que $C^{1/k} \to 1$ pour tout $C > 0$, on a bien $\lim_k N(T^k)^{1/k} = \lim_k N'(T^k)^{1/k}$.

1⇒2. On prend $b = 0$ et x_0 successivement égale à $e_1, \ldots, e_n$, donc $X_0 = I$, $X_{k+1} = TX_k$, donc $X_k = T^k$.

2⇒1. $(x_{k+1} - A^{-1}b) = T(x_k - A^{-1}b)$.

3⇒1. C'est le théorème de point fixe usuel

3⇒2 et 4. Les normes induites sont matricielles. Pour l'une norme des normes induites, $N(T) < 1$ et $N(T^k) \leq N(T)^k \to 0$. .

4⇒2 et 3. Pour toute norme sur $\mathbb{K}^{n\times n}$, à partir d'un certain rang $N(T^k) \leq c^k$ avec $c < 1$. Si N est induite sur $\mathbb{K}^{n\times n}$ par une norme (encore notée) N sur $\mathbb{K}^n$ alors $N'(x) = \sum_k N(T^k x)$ converge, pour tout x, et fournit une norme N' sur $\mathbb{K}^n$, équivallente et supérieure à N. On vérifie que $N'(T) < 1$.

2⇒4. Soit N une norme induite. À partir d'un certain rang, $N(T^k) \leq c < 1$. Donc il existe k tel que $r(T) \leq N(T^k)^{1/k} < 1$.

2⇒5. Soit μ une valeur propre de module $\rho(T)$ et x un vecteur (complexe) propre associé (donc non nul). $N(T^k x) = \rho(T)^k N(x) \to 0$. Donc $\rho(T) < 1$.

5⇒2. On sait que dans $\mathbb{C}$, $T = D + N$ avec D diagonalisable, N nilpotente ($N^{j+1} = 0$) et $ND = DN$. Le spectre de T est celui de D. Si $k \geq j$, on calcule avec le binôme $T^k = D^{k-j} \sum_{i=0}^{j} C_k^i D^{j-i} N^i$. En écrivant C_k^i comme un polynôme en k de degré $\leq i$, et en rassemblant les termes de la sommes selon les puissances en k, on obtient $T^k = D^{k-j} \sum_{i=0}^{j} k^i U_i$ pour certaines matrices U_i. Pour toute norme matricielle, $N(T^k) \leq N(D^{k-j}) N(\sum_{i=0}^{j} k^i U_i) = O(\rho(T)^k) O(k^j)$.

5⇔4 La fonction $\mu \mapsto (I - \mu T)^{-1}$ est développable en série entière au voisinage de chacun des points de l'ouvert des $\mu \in \mathbb{C}$ tels que $1/\mu$ n'est pas valeur propre de T, complémentaire du fermé F contenant les inverses des valeurs propres non nulles. Donc le rayon de convergence d'une telle série entière au voisinage de μ_0 est la distance de μ_0 à F. Or, pour $\mu_0 = 0$, on trouve la série $\sum_k \mu^k T^k$, de rayon de convergence $1/r(T)$ et la distance à F est $1/\rho(T)$. □

28/10 : manque de temps. Vu seulement 4⇒2 et 3.

Remarque. Je suis preneur d'une preuve directe de 2⇒3

1.2 Méthodes courantes

On décompose $A = D - E - F$ ou $D(i,i) = A(i,i)$, $E(i,j) = -A(i,j)$ si $i > j$ (sous la diagonale) et $F(i,j) = -A(i,j)$ si $i < j$ (au dessus de la diagonale). On suppose toujours D inversible (les $A(i,i)$ non nuls).

Jacobi $A = D - (E+F)$, $T = J := D^{-1}(E+F)$. Durant la $k+1$-eme itération, on calcul dans y le vecteur x_{k+1} à partir de x_k dans x, puis on met à jour :

pour $i = 1 : n$

$$y(i) = \left(b(i) - \sum_{j=1}^{i-1} a(i,j)x(j) - \sum_{j=i+1}^{n} a(i,j)x(j) \right) / a(i,i)$$

$x = y$

Gauss-Seidel $A = (D-E) + F$, $T = G := (D-E)^{-1}F$. Durant la $k+1$-eme itération, on calcul en place, dans x, le vecteur x_{k+1} à partir de x_k par

pour $i = 1 : n$

$$x(i) = \left(b(i) - \sum_{j=1}^{i-1} a(i,j)x(j) - \sum_{j=i+1}^{n} a(i,j)x(j) \right) / a(i,i)$$

Relaxation C'est une combinaison linéaire des deux précedentes. Soit $\omega \in \mathbb{C}^*$ et $\theta = 1/\omega$. $A = (\theta D - E) - ((\theta-1)D + F)$, $T = L_\omega := (\theta D - E)^{-1}((\theta-1)D + F) = (D - \omega E)^{-1}((1-\omega)D + \omega F)$. Durant la $k+1$-eme itération, on calcul en place, dans x, le vecteur x_{k+1} à partir de x_k par

pour $i = 1 : n$

$$x(i) = (1-\omega)x(i) + \omega \left(b(i) - \sum_{j=1}^{i-1} a(i,j)x(j) - \sum_{j=i+1}^{n} a(i,j)x(j) \right) / a(i,i)$$

1.3 Résultats de convergence

Théorème. *Si la methode de relaxation converge, alors* $|\omega - 1| < 1$.

Remarque. On choisira donc toujours le paramètre dans le disque complexe de centre 1 et rayon 1.

Dém. Si la méthode converge, alors $1 > \rho(L_\omega)^n > |\det(L_\omega)| = |(1-\omega)^n|$. (Le determinant d'une matrice triangulaire est le produit de ses coefficients diagonaux). □

Théorème. *Si A est à diagonale strictement dominante (*ie $|A(i,i)| > \sum_{j \neq i} |A(i,j)|$ *pour $i = 1 : n$), alors les méthode de Jacobi, Gauss-Seidel, et relaxation pour $\omega \in]0,1]$ convergent.*

Dém. $J = D^{-1}(E+F)$ vérifie ($J(i,i) = 0$ et) $\|J\|_\infty = \max_i \sum_j |J(i,j)| < 1$. Donc Jacobi converge.

28/10 : fin

$(D-\omega E)L_\omega = (1-\omega)D+\omega F$. Donc si μ est valeur propre de L_ω, alors $\mu+\omega-1$ est valeur propre de $B = \omega D^{-1}(\mu E+F)$. Donc $|\mu+\omega-1| \leq \rho(B) \leq \|B\|_\infty < |\omega||\mu|$ pourvu que $|\mu| \geq 1$. Mais alors $|\mu|-|\omega-1| < |\omega||\mu|$ et donc $1-|\omega| \leq |\mu|(1-|\omega|) < |\omega-1|$ ce qui est l'inégalité triangulaire stricte. Si on a supposé que $\omega \in [0,1]$, on en déduit que toute valeur propre μ de L_ω est de module < 1, *ie* $\rho(L_\omega) < 1$. □

Théorème. *Si A est hermitienne ($A^* = A$) définie ($x^*Ax = 0 \Rightarrow x = 0$) positive ($x^*Ax \geq 0$), ainsi que $M^* + N$ (avec $A = M - N$), alors la méthode converge.*

Si A est hermitienne définie positive, alors la méthode de relaxation converge si (et seulement si) $|\omega - 1| < 1$.

Dém. Noter que A hermitienne entraine $M^* + N = M^* + M - A$ hermitienne. $N^2(x) = x^*Ax$ définit une norme sur $\mathbb{C}^n$. On a $Tx = x - M^{-1}Ax = x - y$ donc $Ax = My$, $x^*A = y^*M^*$. Pour $x \neq 0$, on a $N^2(Tx) = (Tx)^*A(Tx) = x^*Ax - y^*Ax - x^*Ay + y^*Ay = N^2(x) - y^*(M + M^* - A)y < N^2(x)$ car par hypothèse $M^* + N$ est définie positive et $y \neq 0$. Donc $N(T) < 1$.

A hermitienne entraine $E^* = F$ et D hermitienne. $M^* + N = (\frac{1}{\omega} + \frac{1}{\overline{\omega}} - 1)D = \frac{1-|\omega-1|^2}{|\omega|^2}D$. A définie positive entraine D définie positive, donc $M^* + N$ aussi dès que $|\omega - 1| < 1$. □

Théorème. *On suppose A tridiagonale (éventuellement par bloc) et D inversible. Alors $\rho(G) = \rho(J)^2$*

Si de plus les valeurs propres de J sont réelles et $\rho(J) < 1$, ce qui arrive si A est hermitienne définie positive, alors, pour les $\omega \in \mathbb{R}$, la convergence équivaut à $0 < \omega < 2$, l'infimum de $\rho(L_\omega)$ est atteint pour $\omega_0 = \frac{2}{1+\sqrt{1-\rho(J)^2}} \geq 1$ et vaut $\omega_0 - 1$.

Exercice. Chercher les rayons spectraux des matrices J, G et L_ω pour A tridiagonale de taille n, avec 2 sur la diagonale et -1 sur la sous- et la sur- diagonale. En déduire le gain (en nombre d'itérations) de la relaxation de paramètre optimal, par rapport à Gauss-Seidel.