

Méthodes itératives pour la solution de systèmes linéaires

8 novembre 2000

1 Méthodes de relaxation pour la solution de systèmes linéaires

1.1 Compléments d'analyse matricielle

1.1.1 Rayon spectral d'une matrice

Définition 1 Soit A une matrice de $\mathbb{C}^{N \times N}$. On sait que le polynôme caractéristique de A a N racines complexes. On appelle spectre de A et on note $sp(A)$ l'ensemble formé de ses N valeurs propres (comptées avec leur multiplicité). On appelle rayon spectral de A et on note $\rho(A)$ le réel positif ou nul défini par

$$\rho(A) = \max_{\lambda \in sp(A)} |\lambda|. \quad (1)$$

Proposition 1 ;

1. Soit $\| \cdot \|$ une norme matricielle sur $\mathbb{C}^{N \times N}$ ayant la propriété multiplicative

$$\|AB\| \leq \|A\| \|B\| \quad (*)$$

alors

$$\rho(A) \leq \|A\|$$

2. De plus pour toute matrice A et pour tout réel positif ϵ , il existe une norme matricielle $\| \cdot \|$ subordonnée à une norme vectorielle telle que

$$\|A\| \leq \rho(A) + \epsilon.$$

Démonstration

Démontrons le premier point. Soit $\| \cdot \|$ une norme matricielle sur $\mathbb{C}^{N \times N}$ ayant la propriété multiplicative (*). On peut alors définir une norme vectorielle sur \mathbb{C}^N qu'on note $||| \cdot |||$ par

$$|||v||| = \|M_v\|, \quad \text{où } M_v = (v, \dots, v)$$

exercice : vérifier qu'il s'agit bien d'une norme sur \mathbb{C}^n .

Soit λ une valeur propre de A telle que $|\lambda| = \rho(A)$ et v_λ le vecteur propre correspondant : on a

$$\rho(A) |||v||| = |||Av||| \leq \|(Av, \dots, Av)\| = \|A(v, \dots, v)\| \leq \|A\| \|(v, \dots, v)\| = \|A\| |||v|||,$$

l'inégalité venant de la propriété (*). On a donc

$$\rho(A) \leq \|A\|.$$

Pour le deuxième point, on sait qu'il existe une matrice unitaire U et une matrice triangulaire supérieure R telles que $A = URU^*$. La matrice triangulaire R s'écrit

$$R = \begin{pmatrix} \lambda_1 & r_{12} & \dots & r_{1d} \\ & \lambda_2 & r_{23} & \dots & r_{2d} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_{d-1} & r_{d-1d} \\ & & & & \lambda_d \end{pmatrix}$$

Soit alors la matrice $D_\eta = \text{diag}(1, \eta, \dots, \eta^{d-1})$. On considère la norme $||| \cdot |||$ définie par

$$|||M||| = \|D_\eta^{-1}U^*MUD_\eta\|_\infty$$

On a

$$D_\eta^{-1}U^*AUD_\eta = D_\eta^{-1}RD_\eta = \begin{pmatrix} \lambda_1 & \eta r_{12} & \dots & \eta^{d-1} r_{1d} \\ & \lambda_2 & \eta r_{23} & \dots & \eta^{d-2} r_{2d} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_{d-1} & \eta r_{d-1d} \\ & & & & \lambda_d \end{pmatrix}$$

Donc $|||A||| = \max_{1 \leq i \leq d} (|\lambda_i| + \sum_{j>i} \eta^{j-i} |r_{ij}|)$. Pour $\epsilon > 0$ fixé, on peut choisir η assez petit pour que $|||A||| \leq \rho(A) + \epsilon$.

Enfin on vérifiera facilement que $||| \cdot |||$ est la norme subordonnée à la norme vectorielle

$$|||v||| = \|D_\eta^{-1}U^*v\|_\infty$$

1.1.2 Suite des puissances d'une matrice

On dit qu'une suite de matrices $(A_n)_n$ dans $\mathbb{C}^{N \times M}$ tend vers 0, si pour une norme matricielle $\| \cdot \|$ définie sur $\mathbb{C}^{N \times M}$, la suite $(\|A_n\|)_n$ tend vers 0. Comme toutes les normes sont équivalentes, cette convergence a alors lieu pour toutes les normes. Si $N = M$, en prenant les normes subordonnées aux normes vectorielles, on voit que

$$A_n \rightarrow 0 \Leftrightarrow \forall v \in \mathbb{C}^N, Av \rightarrow 0.$$

On déduit de la proposition 1 le théorème:

Théorème 1 Soit $A \in \mathbb{C}^{N \times N}$. Une condition nécessaire et suffisante pour que la suite des puissances de A , $(A^n)_n$, converge vers 0 est que

$$\rho(A) < 1. \tag{2}$$

Démonstration

Soit une matrice A telle que $A_n \rightarrow 0$. Il est clair que $\rho(A^n) = \rho^n(A)$. Mais d'après la proposition 1, $\rho^n(A) \leq \|A^n\|$, pour toute norme ayant la propriété (*). Donc $\rho^n(A) \rightarrow 0$, ce qui implique que $\rho(A) < 1$.

Réciproquement, soit une matrice telle que $\rho(A) < 1$. D'après la proposition 1, il existe une

norme matricielle $\| \cdot \|$ subordonnée à une norme vectorielle, telle que $\|A\| < 1$. Comme $\| \cdot \|$ a la propriété (*), $\|A^n\| \leq \|A\|^n \rightarrow 0$.

On a de plus une caractérisation du rayon spectral d'une matrice :

Théorème 2 *Pour toute norme matricielle ayant la propriété (*),*

$$\rho(A) = \lim_{n \rightarrow \infty} \|A^n\|^{\frac{1}{n}}$$

1.2 Méthodes itératives de relaxation

1.2.1 Principe général

Pour des systèmes linéaires de grande taille, les méthodes directes (du type Gauss ou Choleski), peuvent s'avérer trop coûteuses en temps de calcul ou en place mémoire. L'idée est alors de ne plus chercher à résoudre exactement le système linéaire mais d'approcher sa solution par une suite de vecteurs, construite à l'aide d'une formule de récurrence simple.

Soit donc un système linéaire

$$Ax = b, \tag{3}$$

où $A \in \mathbb{C}^{d \times d}$ est inversible, $x, b \in \mathbb{C}^d$. Le principe des méthodes itératives présentées ici est de d'écrire A comme la différence de deux matrices :

$$A = M - N, \tag{4}$$

où

1. M est inversible
2. le système linéaire $My = c$ peut être résolu simplement, avec un coût de calcul faible : typiquement M sera diagonale ou triangulaire.

On va alors approcher la solution de (3) par la suite (x_n) définie par récurrence à partir de x_0 qu'on choisit, et de la formule

$$x_{n+1} = M^{-1}(b + Nx_n) \tag{5}$$

Remarque 1 *On n'a pas besoin de calculer M^{-1} , mais juste de savoir calculer la solution de $Mx_{n+1} = b + Nx_n$.*

Observation 1 *Si la suite converge vers y alors $y = x$. En effet, si la suite converge, on a à la limite, $My = b + Ny$ ou de manière équivalente $Ay = b$. Comme la solution de (3) est unique, $x = y$.*

Considérons l'erreur à l'étape n ,

$$e_n = x - x_n.$$

On a

$$\left. \begin{array}{l} Mx_{n+1} = b + Nx_n \\ Mx = b + Nx \end{array} \right\} \Rightarrow e_{n+1} = M^{-1}Ne_n \Rightarrow e_{n+1} = (M^{-1}N)^{n+1}e_0$$

On appelle $M^{-1}N$ la matrice d'itération de la méthode. On a démontré le résultat

Proposition 2 *La suite donnée par x_0 et (5) converge vers x pour tout choix de x_0 si et seulement si la matrice d'itération vérifie*

$$\rho(M^{-1}N) < 1. \tag{6}$$

Démonstration

La suite donnée par x_0 et (5) converge vers x pour tout choix de x_0 si et seulement si $(M^{-1}N)^n e_0 \rightarrow 0$ pour tout e_0 , ce qui équivaut à dire que $(M^{-1}N)^n \rightarrow 0$. D'après le théorème 1, ceci a lieu si et seulement si $\rho(M^{-1}N) < 1$.

1.2.2 Une condition suffisante dans le cas où A est hermitienne, définie positive

Théorème 3 *Soit A une matrice hermitienne, définie positive, et M, N deux matrices telles que $A = M - N$, M soit inversible et $M^* + N$ soit elle aussi définie positive. Alors $\rho(M^{-1}N) < 1$.*

Démonstration

Comme A est symétrique définie positive, on peut considérer la norme vectorielle définie par

$$\|v\|_A^2 = v^* A v,$$

et la norme matricielle subordonnée. On a pour tout vecteur v ,

$$\begin{aligned} \|M^{-1}Nv\|_A^2 &= v^*(M - A)^*(M^*)^{-1}AM^{-1}(M - A)v \\ &= v^*Av + v^*A^*(M^*)^{-1}AM^{-1}Av - v^*AM^{-1}Av - v^*A^*(M^*)^{-1}Av \end{aligned}$$

Montrons que $v^*A^*(M^*)^{-1}AM^{-1}Av - v^*AM^{-1}Av - v^*A^*(M^*)^{-1}Av < 0$ si $v \neq 0$. En effet

$$\begin{aligned} &v^*A^*(M^*)^{-1}AM^{-1}Av - v^*AM^{-1}Av - v^*A^*(M^*)^{-1}Av \\ &= v^*A^*(M^*)^{-1}AM^{-1}Av - v^*A^*(M^*)^{-1}MM^{-1}Av - v^*A^*(M^*)^{-1}M^*M^{-1}Av \\ &= v^*A^*(M^*)^{-1}(A - M - M^*)M^{-1}Av \\ &= -v^*A^*(M^*)^{-1}(M^* + N)M^{-1}Av \end{aligned}$$

qui est strictement négatif dès que $M^{-1}Av \neq 0 \rightarrow v \neq 0$, car $(M^* + N)$ est symétrique définie positive et A et M sont inversibles. Donc si $v \neq 0$, $\|M^{-1}Nv\|_A < \|v\|_A$. On en déduit que $\|M^{-1}N\|_A < 1$, ce qui achève la démonstration.

1.3 La méthode de Jacobi

On considère une matrice inversible A dont la diagonale D est inversible. La méthode de Jacobi consiste à choisir $M = D$ et $N = D - A$. La matrice d'itération $\rho(\mathcal{L}_J)$ de la méthode de Jacobi s'écrit $\rho(\mathcal{L}_J) = I - D^{-1}A$. On a les résultats suivants

Proposition 3 *Si A est à diagonale strictement dominante, i.e.*

$$\forall i, \quad |a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

alors la méthode de Jacobi converge pour tout choix de x_0 .

Démonstration

Si A est à diagonale strictement dominante, on a $\|M^{-1}N\|_\infty = \|D^{-1}(D - A)\|_\infty < 1$.

Proposition 4 *Si A et $2D - A$ sont hermitiennes définies positives, alors la méthode de Jacobi converge pour tout choix de x_0 .*

Démonstration

Si A est hermitienne définie positive, alors D l'est aussi, et on peut utiliser la méthode de Jacobi. De plus, $M + N = 2D - A$ est aussi hermitienne définie positive. On peut appliquer le théorème

3.

Algorithme La i ème coordonnée de x_{n+1} est donnée par

$$(x_{n+1})_i = \frac{b_i - \sum_{j \neq i} a_{ij}(x_n)_j}{a_{ii}}.$$

Voici la boucle de la méthode de Jacobi : le test d'arrêt est ici du type $\|x^{n+1} - x^n\| \leq \epsilon$, mais d'autres tests sont évidemment possibles.

```
while( err<eps)
{
  w=x;
  x=b;
  for(int i=0; i<x.size();i++)
  {
    for(int j=0; j<i;j++)
      x(i)-=a(i,j)*w(j);
    for(int j=i+1; j<v.size();j++)
      x(i)-=a(i,j)*w(j);
    x(i)=x(i)/a(i,i);
  }
  e=w-x;
  err=norm(e);
}
```

Quand la matrice admet une décomposition par bloc :

$$A = \begin{pmatrix} A_{11} & \dots & A_{1P} \\ \vdots & & \vdots \\ A_{i1} & \dots & A_{iP} \\ \vdots & & \vdots \\ A_{P1} & \dots & A_{PP} \end{pmatrix}$$

où les blocs diagonaux A_{ii} sont des matrices carrées (les blocs non diagonaux ne doivent pas nécessairement l'être), et si les blocs diagonaux sont tous inversibles, une méthode dite de Jacobi par blocs consiste à prendre $D = \text{diag}(A_{11}, \dots, A_{PP})$. Elle nécessite de savoir résoudre les systèmes avec les blocs A_{ii} .

1.4 La méthode de Gauss-Seidel

On considère une matrice inversible A dont la diagonale D est inversible. On note $A = D - L - U$, où $-L$ (respectivement $-U$) est la partie triangulaire inférieure strictement (respectivement supérieure) de A . La méthode de Gauss-Seidel consiste à choisir $M = D - L$ et $N = U$. La matrice d'itération $\rho(\mathcal{L}_{GS})$ de la méthode de Gauss-Seidel s'écrit $\rho(\mathcal{L}_{GS}) = I - (D - L)^{-1}A$. On a les résultats suivants ;

Proposition 5 *Si A est à diagonale strictement dominante, alors la méthode de Gauss-Seidel converge pour tout choix de x_0 .*

Démonstration On s'intéresse à la solution de $My = Nx$ pour x donné : on a

$$a_{ii}y_i = \sum_{j < i} a_{ij}x_j + \sum_{j > i} a_{ij}y_j.$$

On considère i_0 tel que $|y_{i_0}| = \|y\|_\infty$: on a

$$|a_{i_0 i_0}| |y_{i_0}| \leq \sum_{j < i_0} |a_{i_0 j}| \|y\|_\infty + \sum_{j > i_0} |a_{i_0 j}| \|x\|_\infty.$$

Si A est à diagonale strictement dominante,

$$|a_{i_0 i_0}| - \sum_{j < i_0} |a_{i_0 j}| > \sum_{j > i_0} |a_{i_0 j}|,$$

donc

$$\sup_{\substack{x \neq 0 \\ My = Nx}} \frac{\|y\|_\infty}{\|x\|_\infty} < 1,$$

ce qui veut dire que $\|M^{-1}N\|_\infty < 1$.

Proposition 6 Si A est hermitienne définie positive, alors la méthode de Gauss-Seidel converge pour tout choix de x_0 .

Démonstration

Si A est hermitienne définie positive, alors D l'est aussi, et $L = U^*$. Donc $M^* + N = D - L^* + U = D$ est hermitienne définie positive. On peut appliquer le théorème 3.

Algorithme La i ème coordonnée de x_{n+1} est donnée par

$$(x_{n+1})_i = \frac{b_i - \sum_{j < i} a_{ij}(x_{n+1})_j - \sum_{j > i} a_{ij}(x_n)_j}{a_{ii}}$$

Dans la méthode de Gauss-Seidel, dès que la i ème coordonnée de x_{n+1} est calculée, la i ème coordonnée de x_n devient inutile : on peut écraser la i ème coordonnée de x_n et la remplacer par la i ème coordonnée de x_{n+1} dès que celle-ci est calculée.

Voici la boucle de la méthode de Gauss-Seidel : le test d'arrêt est toujours du type $\|x^{n+1} - x^n\| \leq \epsilon$.

```
while( err < eps)
{
    e=x;
    for(int i=0; i<x.size();i++)
    {
        x(i)=b(i);
        for(int j=0; j<i;j++)
            x(i)-=a(i,j)*x(j);
        for(int j=i+1; j<x.size();j++)
            x(i)-=a(i,j)*x(j);
        x(i)=x(i)/a(i,i);
    }
    e=e-x;
    err=norm(e);
}
```

La méthode de Jacobi est complètement parallélisable, ce qui n'est pas le cas de la méthode de Gauss-Seidel.

Comme pour la méthode de Jacobi, on peut généraliser la méthode de Gauss-Seidel à une matrice par blocs, dont les blocs diagonaux sont tous carrés et inversibles.

1.5 Méthodes SOR (successive over relaxation)

La méthode de Gauss-Seidel est très facile à programmer mais sa convergence peut être très lente pour certains systèmes : on la modifie en introduisant un paramètre $\omega \neq 0$ dit paramètre de relaxation et en choisissant $M = \frac{1}{\omega}D - L$ et $N = U + (1 - \frac{1}{\omega})D$. La matrice M est inversible si la diagonale D est inversible. Pour $\omega = 1$, on retrouve la méthode de Gauss-Seidel. Pour $\omega < 1$ on parle de sous-relaxation. Pour $\omega > 1$ on parle de sur-relaxation. Un calcul facile montre que la matrice d'itération de cette méthode est

$$\mathcal{L}_\omega = (I - \omega D^{-1}L)^{-1}((1 - \omega)I + \omega D^{-1}U). \quad (7)$$

Proposition 7 *Si A est hermitienne définie positive, la méthode de relaxation avec paramètre ω converge pour tout x_0 si*

$$0 < \omega < 2. \quad (8)$$

Démonstration

Si A est hermitienne définie positive, alors D l'est aussi, et $L = U^*$. Donc $M^* + N = (\frac{2}{\omega} - 1)D - L^* + U = (\frac{2}{\omega} - 1)D$ est hermitienne définie positive dès que $0 < \omega < 2$. On peut appliquer le théorème 3.

Remarque 2 *On peut aussi relaxer la méthode de Jacobi en prenant $M = \frac{1}{\omega}D$ et $N = \frac{1}{\omega}D - A$. Si A est hermitienne définie positive, sous quelles conditions sur ω $M^* + N$ est elle définie positive?*

Théorème 4 *Si $0 \leq \omega$ ou si $\omega \geq 2$, La méthode SOR ne converge pas vers la solution x pour tout choix initial x_0 , Si les inégalités sont strictes, elle diverge.*

Démonstration

D'après (7), le déterminant de \mathcal{L}_ω est $(1 - \omega)^d$. Mais

$$|\rho(\mathcal{L}_\omega)| < 1 \Rightarrow |\det(\mathcal{L}_\omega)| < 1$$

car le déterminant est le produit des valeurs propres. Donc,

$$\omega \geq 2 \text{ ou } \omega \leq 0 \Rightarrow |(1 - \omega)^d| \geq 1 \Rightarrow |\rho(\mathcal{L}_\omega)| \geq 1$$

implique que la méthode SOR ne converge pas vers la solution pour tout choix initial.

Remarque 3 *La proposition 7 donne donc en fait une condition nécessaire et suffisante.*

Algorithme

Voici la boucle de la méthode SOR

```
while( err<eps)
{
  e=x;
  for(int i=0; i<x.size();i++)
  {
    x(i)=b(i)-(1-1/omega)*a(i,i)*x(i);
    for(int j=0; j<i;j++)
      x(i)-=a(i,j)*x(j);
    for(int j=i+1; j<x.size();j++)
```

```

    x(i)-=a(i,j)*x(j);
    x(i)=omega*x(i)/a(i,i);
}
e=e-x;
err=norm(e);
}

```

1.6 Comparaisons des méthodes pour des matrices tridiagonales

Théorème 5 Si A est tridiagonale, on a

$$\rho(\mathcal{L}_{GS}) = \rho(\mathcal{L}_J)^2 \quad (9)$$

Démonstration

Soit A une matrice de $\mathbb{C}^{d \times d}$ tridiagonale :

$$A = \begin{pmatrix} a_1 & b_1 & 0 & \dots & 0 \\ c_1 & a_2 & b_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & c_{d-2} & a_{d-1} & b_{d-1} \\ 0 & \dots & & 0 & c_{d-1} & a_d \end{pmatrix} \quad (10)$$

Le nombre complexe λ est valeur propre de \mathcal{L}_J si et seulement si $\det(D^{-1}(D - A) - \lambda I) = 0$ ou encore si $\det((1 - \lambda)D - A) = 0$, c'est à dire

$$\det \begin{pmatrix} \lambda a_1 & b_1 & 0 & \dots & 0 \\ c_1 & \lambda a_2 & b_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & c_{d-2} & \lambda a_{d-1} & b_{d-1} \\ 0 & \dots & & 0 & c_{d-1} & \lambda a_d \end{pmatrix} = 0 \quad (11)$$

D'autre part μ est valeur propre de \mathcal{L}_{GS} si et seulement si $\det((D - L)^{-1}(D - L - A) - \mu I) = 0$ ou encore si $\det((1 - \mu)(D - L) - A) = 0$, c'est à dire

$$\det \begin{pmatrix} \mu a_1 & b_1 & 0 & \dots & 0 \\ \mu c_1 & \mu a_2 & b_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & \mu c_{d-2} & \mu a_{d-1} & b_{d-1} \\ 0 & \dots & & 0 & \mu c_{d-1} & \mu a_d \end{pmatrix} = 0 \quad (12)$$

Supposons $\lambda \neq 0$, (11) est équivalente à

$$\det[\text{diag}(\lambda, \lambda^2, \dots, \lambda^d) \begin{pmatrix} \lambda a_1 & b_1 & 0 & \dots & 0 \\ c_1 & \lambda a_2 & b_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & c_{d-2} & \lambda a_{d-1} & b_{d-1} \\ 0 & \dots & & 0 & c_{d-1} & \lambda a_d \end{pmatrix} \text{diag}(\lambda^{-1}, \lambda^{-2}, \dots, \lambda^{-d})] = 0 \quad (13)$$

Ce produit de trois matrices vaut :

$$\lambda^{-1} \begin{pmatrix} \lambda^2 a_1 & b_1 & 0 & \dots & 0 \\ \lambda^2 c_1 & \lambda^2 a_2 & b_2 & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & & \lambda^2 c_{d-2} & \lambda^2 a_{d-1} & b_{d-1} \\ 0 & \dots & & 0 & \lambda^2 c_{d-1} & \lambda^2 a_d \end{pmatrix}$$

ce qui veut dire d'après (12) que λ^2 est valeur propre de \mathcal{L}_{GS} . Donc si $\lambda \neq 0$, λ est valeur propre de \mathcal{L}_J si et seulement si λ^2 est valeur propre de \mathcal{L}_{GS} , ce qui montre (9).

On a enfin un théorème plus précis dans le cas où A est de plus définie positive :

Théorème 6 *Si A est tridiagonale et hermitienne définie positive, les méthodes de Jacobi et de Gauss-Seidel convergent, et la méthode SOR converge si et seulement si $0 < \omega < 2$. Le paramètre optimal est*

$$\omega_{\text{opt}} = \frac{2}{1 + \sqrt{1 - \rho^2(\mathcal{L}_J)}} > 1, \quad (14)$$

et

$$\rho(\mathcal{L}_{\omega_{\text{opt}}}) = \frac{1}{\omega_{\text{opt}}}. \quad (15)$$

2 Méthodes de descente pour la résolution de systèmes linéaires

2.1 Principe des méthodes de descente

2.1.1 Minimisation de fonctions quadratiques

Le but est de construire des méthodes itératives pour résoudre un système linéaire

$$Ax = b, \quad (16)$$

en lui associant un problème de minimisation équivalent et en construisant une suite minimisante : on va travailler avec des matrices réelles symétriques et définies positives, mais on pourrait tout généraliser au cas de matrices hermitiennes définies positives. Considérons la forme quadratique F sur \mathbb{R}^d définie par

$$F(x) = \frac{1}{2}x^T Ax - x^T b. \quad (17)$$

La fonction F est continue et $\lim_{\|x\| \rightarrow \infty} F(x) = +\infty$. La fonction admet donc un minimum dans \mathbb{R}^d . La fonction F est de plus différentiable, et son gradient vaut

$$\forall y \in \mathbb{R}^n, \quad DF(y) = Ay - b. \quad (18)$$

Exercice Démontrer l'assertion précédente.

Comme la solution de (16) est unique car A est inversible, le gradient de F ne s'annule qu'en un seul point qui réalise le minimum de F .

Exercice Montrer que F est strictement convexe, c'est à dire que

$$F(\alpha x + (1 - \alpha)y) - \alpha F(x) - (1 - \alpha)F(y) \leq -\frac{1}{2}\alpha(1 - \alpha)\lambda_{\min}(A)$$

où $\lambda_{\min}(A)$ est la plus petite valeur propre de A .

On voit donc que le problème (16) est équivalent à la minimisation de F . L'idée va donc de construire des suites minimisantes de F pour approcher la solution de (16).

2.1.2 Méthodes de descente

Une méthode de descente consiste à construire une suite minimisante sous la forme

$$x_{n+1} = x_n + \alpha_n p_n, \quad (19)$$

où $p_n \in \mathbb{R}^d$, $p_n \neq 0$ et où le scalaire α_n est choisi pour que $F(x_{n+1}) < F(x_n)$. La convergence de la méthode dépend bien sûr des choix des p_n et α_n .

Définition 2 L'erreur à l'étape n est le vecteur

$$e_n = x - x_n. \quad (20)$$

On appelle résidu à l'étape n le vecteur

$$r_n = b - Ax_n = Ae_n. \quad (21)$$

Remarque 4 Le résidu à la n ème étape vérifie $r_n = -DF(x_n)$.

2.1.3 Choix optimal de α_n pour p_n fixé

Supposons choisie la direction p_n : on peut choisir α_n de manière à minimiser la fonction ϕ de \mathbb{R}^+ dans \mathbb{R} donnée par

$$\phi(t) = F(x_n + tp_n).$$

Cette fonction a un minimum unique pour

$$\alpha_{opt} = \frac{r_n^T p_n}{r_n^T A p_n}, \quad (22)$$

et on a

$$x_{n+1} = x_n + \frac{r_n^T p_n}{r_n^T A p_n} p_n$$

Proposition 8 Pour tout p_n , et si on choisit $\alpha_n = \alpha_{opt}$, on a

$$r_{n+1}^T p_n = 0. \quad (23)$$

2.2 Méthodes de gradient

2.2.1 Principe des méthodes de gradient

L'idée des méthodes de gradient va être de choisir comme direction de descente le gradient de F en x_n , ou (voir Remarque 4), de manière équivalente, le résidu r_n :

$$\begin{aligned} x_{n+1} &= x_n - \rho_n DF(x_n) \\ &= x_n + \rho_n r_n. \end{aligned} \quad (24)$$

Pour tout $x \in \mathbb{R}^n$, il existe un nombre réel $\rho_{\max}(x) > 0$, tel que

$$\rho \in]0, \rho_{\max}(x)[\Leftrightarrow F(x - \rho DF(x)) < F(x).$$

Exercice Démontrer cette assertion.

Le pas ρ_n doit donc être un réel positif choisi tel que $F(x_{n+1}) < F(x_n)$.

Remarque 5 On peut généraliser ces méthodes à des fonctions strictement convexes et différentiables.

2.2.2 Interprétation géométrique en dimension deux

Soit A une matrice d'ordre deux symétrique et définie positive. On sait que A est diagonalisable dans une base orthonormale. Quitte à changer les coordonnées on peut supposer que A est diagonale et que $b = 0$:

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}. \quad (25)$$

La fonction F s'écrit alors $F(x) = \lambda_1 x_1^2 + \lambda_2 x_2^2$, et les lignes de niveaux de F : $F(x) = r^2$ sont les ellipses concentriques

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 = r^2. \quad (26)$$

Le gradient de F au point x est orthogonal à l'ellipse d'équation (26) passant par x . La Figure 1 donne un exemple des premiers itérés d'une méthode de gradient.

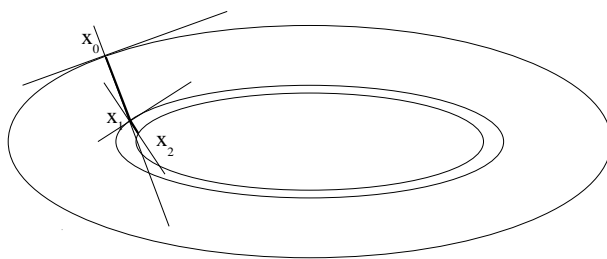


FIG. 1 – deux itérés d'une méthode de gradient

2.2.3 Méthodes du gradient à pas fixe

Si A est défini positif, on peut obtenir une méthode convergente en fixant ρ_n à une valeur bien choisie: la méthode du gradient à pas fixe consiste à construire la suite récurrente

$$x_{n+1} = x_n - \alpha(Ax_n - b).$$

L'erreur vérifie

$$e_{n+1} = (I - \alpha A)e_n = (I - \alpha A)^{n+1}e_0,$$

ce qui montre que la méthode du gradient à pas fixe converge si et seulement si

$$\rho(I - \alpha A) < 1$$

où $\tau(\alpha) = \rho(I - \alpha A)$ est le rayon spectral de $I - \alpha A$. On appelle $\tau(\alpha)$ le taux de convergence de la méthode. On a donc

Théorème 7 Si A est symétrique définie positive, la méthode du gradient à pas fixe converge vers la solution x de (16) si et seulement si

$$\alpha < \frac{2}{\lambda_{\max}(A)}. \quad (27)$$

De plus la valeur de α minimisant le taux de convergence $\tau(\alpha)$ est

$$\alpha_{opt} = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}.$$

et le taux de convergence vaut alors

$$\tau_{opt} = \frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}$$

Démonstration La matrice d'itérations $(I - \alpha A)$ est diagonalisable et ses valeurs propres sont $1 - \alpha\lambda$ où λ est valeur propre de A . La condition nécessaire et suffisante s'obtient facilement en écrivant la condition $\rho(I - \alpha A) < 1$.

On a pour toute valeur propre λ de A ,

$$1 - \alpha\lambda_{\max} \leq 1 - \alpha\lambda_{\max} \leq 1 - \alpha\lambda_{\min}$$

Donc $\rho(I - \alpha A) = \max(|1 - \alpha\lambda_{\max}(A)|, |1 - \alpha\lambda_{\min}(A)|)$. En traçant le graphe des deux fonctions

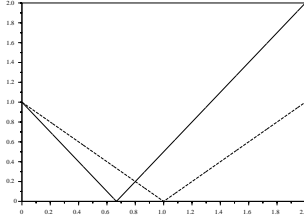


FIG. 2 – graphes des fonctions $\alpha \rightarrow |1 - \alpha\lambda_{\max}(A)|$ et $\alpha \rightarrow |1 - \alpha\lambda_{\min}(A)|$

$\alpha \rightarrow |1 - \alpha\lambda_{\max}(A)|$ et $\alpha \rightarrow |1 - \alpha\lambda_{\min}(A)|$, (voir Figure 2), on voit que

$$\rho(I - \alpha A) = \begin{cases} 1 - \alpha\lambda_{\min}(A) & \text{pour } \alpha \leq \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}, \\ \alpha\lambda_{\max}(A) - 1 & \text{pour } \alpha \geq \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}, \end{cases}$$

On voit aussi que le minimum de $\rho(I - \alpha A)$ est atteint pour $\alpha_{opt} = \frac{2}{\lambda_{\min}(A) + \lambda_{\max}(A)}$ et qu'il vaut

$$\tau_{opt} = \frac{\lambda_{\max}(A) - \lambda_{\min}(A)}{\lambda_{\max}(A) + \lambda_{\min}(A)} = \frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}$$

Remarque 6 On voit donc que la méthode du gradient à pas fixe converge d'autant plus lentement que le conditionnement de A est grand. En effet, avec le choix optimal pour ρ , il faut de l'ordre de

$$\frac{|\log(\epsilon)|}{\left| \log\left(\frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}\right) \right|}$$

itérations pour réduire l'erreur d'un facteur ϵ ; quand $\text{cond}_2(A) \gg 1$, le nombre d'itérations est donc de l'ordre de

$$|\log(\epsilon)| \frac{\text{cond}_2(A)}{2}.$$

2.2.4 Méthode du gradient à pas optimal

Comme dans § 2.1.3, on peut construire

$$x_{n+1} = x_n - \rho_n(Ax_n - b),$$

en choisissant ρ_n de manière à minimiser la fonction ϕ de \mathbb{R}^+ dans \mathbb{R} donnée par

$$\phi(t) = F(x_n - t(Ax_n - b)).$$

Cette fonction à un minimum unique pour

$$\rho_{opt} = \frac{\|Ax_n - b\|_2^2}{\|Ax_n - b\|_A^2} = \frac{\|r_n\|_2^2}{\|r_n\|_A^2}, \quad (28)$$

où

$$\|y\|_A^2 = y^T A y.$$

Il est important de noter que dans la méthode du gradient à pas optimal, d'après (23), les résidus successifs sont orthogonaux :

$$r_{n+1}^T r_n = 0. \quad (29)$$

Théorème 8 *Si A est symétrique et définie positive, la méthode du gradient à pas optimal converge vers la solution x de (16).*

Démonstration La suite des $J(x_n)$ est décroissante par construction, et bornée inférieurement par $J(x)$, donc converge. On en déduit que

$$J(x_{n+1}) - J(x_n) \rightarrow 0,$$

ce qui implique que

$$-\rho_n \|r_n\|_2^2 + \frac{1}{2} \rho_n^2 r_n^T A r_n \rightarrow 0,$$

et comme $\rho_n = \frac{\|r_n\|_2^2}{r_n^T A r_n}$, on trouve finalement que

$$\frac{\|r_n\|_2^4}{r_n^T A r_n} \rightarrow 0.$$

Comme A est définie positive, ceci implique que $r_n \rightarrow 0$. Toujours grâce au caractère défini positif de A , on en déduit que $e_n \rightarrow 0$.

Remarque 7 *Cette démonstration se généralise à la méthode du gradient à pas optimal appliqué à la minimisation d'une fonction F fortement convexe et de gradient Lipchitzien.*

Pour déterminer la vitesse de convergence de la méthode de gradient à pas optimal, on utilise l'inégalité de Kantorovitch

Lemme 1 *Soient d réels strictement positifs,*

$$0 < \ell_1 < \dots < \ell_i < \dots < \ell_d$$

et d réels positifs β_i tels que $\sum_1^d \beta_i = 1$. On note $\ell = \sum_1^d \beta_i \ell_i$. On a

$$\sum_1^d \frac{\beta_i}{\ell_i} \leq \frac{\ell_1 + \ell_d - \ell}{\ell_1 \ell_d}, \quad (30)$$

et

$$\frac{1}{\ell \sum_1^d \frac{\beta_i}{\ell_i}} \geq \frac{4\ell_1 \ell_d}{(\ell_1 + \ell_d)^2}. \quad (31)$$

Démonstration Pour prouver (30), on doit montrer que

$$\sum_1^d \beta_i \left(\frac{1}{\ell_i} + \frac{\ell_i}{\ell_1 \ell_d} \right) \leq \frac{\ell_1 + \ell_d}{\ell_1 \ell_d}$$

Pour cela, on voit que la fonction qui à $x \in [\ell_1, \ell_d]$ associe $\frac{1}{x} + \frac{x}{\ell_1 \ell_d}$ atteint son maximum en $x = \ell_1$ et en $x = \ell_d$ et le maximum vaut $\frac{1}{\ell_1} + \frac{1}{\ell_d}$. On conclut en utilisant le fait que $\sum_{i=1}^d \beta_i = 1$. Après, (31) s'obtient facilement en cherchant le minimum de $\ell \frac{\ell_1 + \ell_d - \ell}{\ell_1 \ell_d}$ sur l'intervalle $[\ell_1, \ell_d]$.

Proposition 9 (Kantorovitch) Soit $A \in \mathbb{R}^{d \times d}$ symétrique définie positive dont les valeurs propres vérifient $0 < \lambda_{\min} = \lambda_1 < \dots < \lambda_i < \dots < \lambda_d = \lambda_{\max}$. On a

$$\inf_{y \neq 0} \frac{\|y\|_2^4}{(y^T A y) (y^T A^{-1} y)} = \frac{4\lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}.$$

Démonstration On note $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_d$ les valeurs propres de A et $(v_i)_{1 \leq i \leq d}$ une base orthonormale de vecteurs propres : $A v_i = \lambda_i v_i$. Pour $y \in \mathbb{R}^d$, on note $\beta_i = \frac{(y^T v_i)^2}{\|y\|_2^2}$:

$$\sum_{i=1}^d \beta_i = 1.$$

On a aussi

$$\frac{(A y, y)}{\|y\|_2^2} = \sum_{i=1}^d \beta_i \lambda_i \quad \text{et} \quad \frac{(A^{-1} y, y)}{\|y\|_2^2} = \sum_{i=1}^d \frac{1}{\lambda_i} \beta_i.$$

et on applique le lemme précédent, et on obtient

$$\inf_{y \neq 0} \frac{\|y\|_2^4}{(y^T A y) (y^T A^{-1} y)} \geq \frac{4\lambda_{\min} \lambda_{\max}}{(\lambda_{\min} + \lambda_{\max})^2}.$$

Enfin, cet infimum est atteint par $y = v_1 + v_d$.

Théorème 9 Pour la méthode du gradient à pas optimal, on a l'estimation

$$\|e_n\|_A \leq \left(\frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1} \right)^n \|e_0\|_A$$

Démonstration On a $\rho_n = \frac{\|r_n\|_2^2}{r_n^T A r_n}$, $e_{n+1} = e_n - \frac{\|r_n\|_2^2}{r_n^T A r_n} r_n$, et $r_{n+1} = r_n - \frac{\|r_n\|_2^2}{r_n^T A r_n} A r_n$. Donc

$$\begin{aligned} \|e_{n+1}\|_A^2 &= e_{n+1}^T A e_{n+1} = e_{n+1}^T r_{n+1} \\ &= e_n^T r_n - \frac{\|r_n\|_2^4}{r_n^T A r_n} \\ &= \left(1 - \frac{\|r_n\|_2^4}{(r_n^T A r_n)(r_n^T A^{-1} r_n)}\right) e_n^T A e_n \\ &\leq \frac{(\lambda_{\max} - \lambda_{\min})^2}{(\lambda_{\max} + \lambda_{\min})^2} e_n^T A e_n \\ &= \left(\frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}\right)^2 \|e_n\|_A^2 \end{aligned}$$

Interprétation géométrique en dimension deux On reprend la matrice A donnée par (25). D'après (29), on peut construire graphiquement la suite des itérés, car le point x_{n+1} est à la fois sur la droite de direction r_n passant par x_n , et sur l'ellipse d'équation (26) tangente par cette droite. La Figure 3 donne un exemple des premiers itérés d'une méthode de gradient à pas optimal. Dire que la matrice A est mal conditionnée, c'est dire que les lignes de niveaux

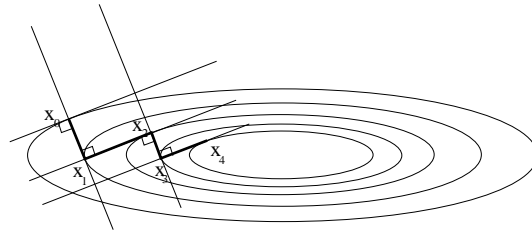


FIG. 3 – méthode de gradient à pas optimal

de F sont des ellipses très allongées, ou encore à fort rapport d'aspect. Dans ce cas, on voit que la suite des x_n se rapproche de sa limite x en zigzaguant beaucoup, et on comprend que la convergence est lente.