Notes de cours de l'ISIMA, première année
http://www.isima.fr/~leborgne

Introduction à la méthode du gradient conjugué

Gilles Leborgne

22 février 2018

On souhaite résoudre le problème $A.\vec{x} = \vec{b}$ à l'aide d'une méthode itérative lorsque A est une matrice symétrique.

La méthode itérative du gradient conjugué est une méthode de Gauss–Seidel relativement à une base A-orthogonale, la base A-orthogonale choisie étant construite à l'aide de la méthode de Gram–Schmidt à partir de la base particulière donnée par les résidus successifs.

Plan : méthode de Gauss–Seidel, puis méthode de Gram–Schmidt, puis méthode du gradient pour motiver le choix de la base de départ sur laquelle on applique Gram–Schmidt, puis la méthode du gradient conjugué.

Table des matières

1 Méthode de descente de Gauss–Seidel

1.1 Méthode de Gauss–Seidel

On note $(\vec{e_i})_{i=1,\dots,n}$ la base canonique de $\mathbb{R}^n$. On veut trouver $\vec{x} = \sum_{i=1}^n x_i \vec{e_i} \overset{\text{noté}}{=} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ tel que :

$$A.\vec{x} = \vec{b}, \tag{1.1}$$

où $A = [a_{ij}]_{\substack{1\le i\le n \\ 1\le j\le n}}$ est une matrice réelle donnée et $\vec{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$ un vecteur donné dans $\mathbb{R}^n$.

Pour un $\vec{x^0} \in \mathbb{R}^n$ qcq, on note :

$$\text{le résidu} = \vec{r^0} \overset{\text{def}}{=} A.\vec{x^0} - \vec{b}, \quad \text{soit} \quad \begin{pmatrix} r_1^0 \\ \vdots \\ r_n^0 \end{pmatrix} = \begin{pmatrix} \sum_{j=1}^n a_{1j}x_j^0 - b_1 \\ \vdots \\ \sum_{j=1}^n a_{nj}x_j^0 - b_n \end{pmatrix}. \tag{1.2}$$

Si $\vec{r^0} = 0$ (résidu nul), alors $\vec{x^0}$ est la solution cherchée.

La méthode de Gauss–Seidel consiste, à partir "d'un point de départ" $\vec{x^0}$ donné a priori, à calculer le résidu $\vec{r^0}$, puis à annuler ce résidu composante par composante en ne modifiant qu'une composante de $\vec{x^0}$ à la fois.

• 1ère étape : annulation de $r_1^0 = \sum_{j=1}^n a_{1j}x_j^0 - b_1$ en ne modifiant que x_1^0. Notons x_1^1 la valeur correspondante, où donc, dès que $a_{11} \neq 0$:

$$a_{11}x_1^1 + \sum_{j=2}^n a_{1j}x_j^0 - b_1 = 0, \quad \text{donc} \quad x_1^1 = \frac{1}{a_{11}}(b_1 - \sum_{j=2}^n a_{1j}x_j^0), \quad \text{et} \quad \vec{x^{0+\frac{1}{n}}} \overset{\text{def}}{=} \begin{pmatrix} x_1^1 \\ x_2^0 \\ \vdots \\ x_n^0 \end{pmatrix}. \tag{1.3}$$

Donc $\vec{x^{0+\frac{1}{n}}} \in \mathbb{R}^n$ est le point qui vérifie donc $(A.\vec{x^{0+\frac{1}{n}}} - \vec{b})_1 = 0$.

• 2ème étape : on prend comme nouveau point de départ le point $\vec{x^{0+\frac{1}{n}}}$. Le résidu est maintenant $\vec{r^{0+\frac{1}{n}}} = A.\vec{x^{0+\frac{1}{n}}} - \vec{b}$. Et on veut annuler $(\vec{r^{0+\frac{1}{n}}})_2 = a_{21}x_1^1 + \sum_{j=2}^n a_{2j}x_j^0 - b_2$ en ne modifiant que $(\vec{x^{0+\frac{1}{n}}})_2$. Notons x_1^2 la valeur correspondante, où donc, dès que $a_{22} \neq 0$:

$$a_{21}x_1^1 + a_{22}x_2^1 + \sum_{j=3}^n a_{2j}x_j^0 - b_2 = 0, \quad \text{donc} \quad x_2^2 = \frac{1}{a_2}(b_2 - a_{21}x_1^1 - \sum_{j=3}^n a_{2j}x_j^0). \tag{1.4}$$

Et notons $\vec{x^{0+\frac{2}{n}}} = (x_1^1, x_2^1, x_3^0, \dots, x_n^0)^T \in \mathbb{R}^n$, point qui vérifie $(\vec{r^{0+\frac{2}{n}}})_2 \overset{\text{def}}{=} (A.\vec{x^{0+\frac{2}{n}}} - \vec{b})_2 = 0$. (Mais on n'a pas $(\vec{r^{0+\frac{2}{n}}})_1 \neq 0$ en général...)

• Et on itère le procédé : à l'étape n on a obtenu le point $\vec{x^1} \overset{\text{def}}{=} (x_1^1, x_2^1, x_3^1, \dots, x_n^1)^T \in \mathbb{R}^n$ (fin de l'étape n).

• Puis on recommence à partir de $\vec{x^1}$ pour obtenir $\vec{x^2} \overset{\text{def}}{=} (x_1^2, x_2^2, x_3^2, \dots, x_n^2)^T \in \mathbb{R}^n$ (fin de l'étape $2n$), ...

• Et on itère le procédé : à la fin de l'étape kn on a obtenu le point $\vec{x^k}$ $(k \geq 1)$.

• Et on s'arrête dès que par exemple $||\vec{x^k} - \vec{x^{k-1}}|| < \varepsilon$ pour une précision ε souhaitée.

Proposition 1.1 *Si A est une matrice symétrique définie positive, alors la suite $(\vec{x^k})_{k\in\mathbb{N}}$ converge vers la solution $\vec{x}$ de (1.1), quel que soit le point $\vec{x^0}$ pris comme point de départ de la méhtode itérative de Gauss–Seidel.*

Preuve. A l'aide de l'expression matricielle de l'algorithme, voir § suivant, proposition 1.3. ∎

1.2 Expression matricielle de la méthode de Gauss–Seidel

Notons $D = \text{diag}(A) = \text{diag}(a_{11}, \ldots, a_{nn})$ la matrice diagonale relative à $A = [a_{ij}]$, notons $-E$ la matrice sous-diagonale de A (triangulaire inférieure stricte), et $-F$ est la matrice sur-diagonale de A (triangulaire supérieure stricte). Donc :

$$A = D - E - F, \quad \text{et} \quad A.\vec{x} = \vec{b} \quad \text{ssi} \quad (D-E).\vec{x} = F.\vec{x} + \vec{b}. \tag{1.5}$$

Proposition 1.2 *Quand $a_{ii} \neq 0$ pour tout i, l'algorithme de Gauss–Seidel s'écrit à l'étape $(k+1)n$, $k \geq 0$:*

$$(D-E).\vec{x}^{k+1} = F.\vec{x}^k + \vec{b}. \tag{1.6}$$

(Et la suite $(\vec{x}^k)_{k\in\mathbb{N}}$ convergera vers $\vec{x}$ solution de $A.\vec{x} = \vec{b}$, i.e. de $(D-E)\vec{x} = F\vec{x} + \vec{b}$, voir proposition suivante.)

Preuve. À la fin de l'étape k, posons $\vec{y}^k = F.\vec{x}^k + \vec{b}$, et il s'agit de calculer $\vec{x}^{k+1}$ solution de, cf. (1.6) :

$$(D-E).\vec{x}^{k+1} = \vec{y}^k. \tag{1.7}$$

C'est un système de n équations à n inconnues les composantes $(x^{k+\frac{i}{n}})_{i=1,\ldots,n}$ de $\vec{x}^{k+1}$, avec $D-E$ triangulaire inférieure : la première équation donne directement $x^{k+\frac{1}{n}}$, d'où la deuxième équation donne $x^{k+\frac{2}{n}}$, d'où ..., d'où $\vec{x}^{k+1}$ (méthode de descente.) ∎

Proposition 1.3 *Si A est une matrice symétrique définie positive, alors la suite $(\vec{x}^k)_{k\in\mathbb{N}^*}$ converge vers la solution $\vec{x}$ de $A.\vec{x} = \vec{b}$, quel que soit le point $\vec{x}^0$ pris comme point de départ de la méthode itérative de Gauss–Seidel.*

Preuve. A est définie positive, donc inversible. Soit $\vec{x} = A^{-1}.\vec{b}$ la solution. On a $(D-E).\vec{x} = F.\vec{x}+\vec{b}$, cf. (1.5). Et par construction, $(D-E).\vec{x}^{k+1} = F.\vec{x}^k + \vec{b}$. Donc $(D-E).(\vec{x}^{k+1} - \vec{x}) = F.(\vec{x}^k - \vec{x})$, soit $(\vec{x}^{k+1} - \vec{x}) = B.(\vec{x}^k - \vec{x})$ où on a posé $B = (D-E)^{-1}.F$.

• Vérifions que $(D-E)$ est inversible. A définie positive implique $a_{ii} = \vec{e}_i^T.A.\vec{e}_i > 0$. Donc $\det(D) = \prod a_{ii} > 0$ pour tout i. Et $\det(D-E) = \det(D) > 0$, d'où $(D-E)$ est inversible.

• Vérifions que le rayon spectral de B est < 1, voir annexe B. A est symétrique, donc $F = E^T$, et $B = (D-E)^{-1}.E^T$. On "symétrise en partie" la matrice B : notons $D^{\frac{1}{2}} = \text{diag}(\sqrt{a_{11}}, ..., \sqrt{a_{nn}})$. On a $E = D^{\frac{1}{2}}.L.D^{\frac{1}{2}}$ où $L \stackrel{\text{def}}{=} D^{-\frac{1}{2}}.E.D^{-\frac{1}{2}}$. Donc $(D-E) = D^{\frac{1}{2}}.(I-L).D^{\frac{1}{2}}$. Et on a $E^T = D^{\frac{1}{2}}.L^T.D^{\frac{1}{2}}$. Donc $B = D^{-\frac{1}{2}}.(I-L)^{-1}.D^{-\frac{1}{2}}.E^T = D^{-\frac{1}{2}}.(I-L)^{-1}.L^T.D^{\frac{1}{2}}$.

Posons $B_1 = D^{\frac{1}{2}}.B.D^{-\frac{1}{2}} = (I-L)^{-1}.L^T$. On va utiliser l'inégalité $\frac{X}{1-X} < 1$ pour $X < \frac{1}{2}$ (immédiat car $X < 1-X$ pour $X < \frac{1}{2}$).

B_1 et B ont mêmes valeurs propres car $B_1.\vec{v} = \lambda\vec{v}$ ssi $D^{\frac{1}{2}}.B.D^{-\frac{1}{2}}.\vec{v} = \lambda\vec{v}$, ssi $B.D^{-\frac{1}{2}}.\vec{v} = \lambda D^{-\frac{1}{2}}.\vec{v}$, i.e. $\vec{v}$ est vecteur propre de B_1 ssi $D^{-\frac{1}{2}}.\vec{v}$ est vecteur propre de B pour la même valeur propre.

Soit λ une valeur propre non nulle de B_1 (si $\lambda = 0$ alors $\lambda < 1$ le rayon spectral souhaité), associée à un vecteur propre $\vec{v}$ tel que $||\vec{v}|| = 1$. Donc $B_1.\vec{v} = \lambda\vec{v}$, donc $L^T.\vec{v} = \lambda(I-L).\vec{v}$ et donc $\vec{v}^T.L^T.\vec{v} = \lambda(1 - \vec{v}^T.L.\vec{v})$, donc $\lambda = \frac{\vec{v}^T.L^T.\vec{v}}{1-\vec{v}^T.L.\vec{v}}$. Montrons $\vec{v}^T.L^T.\vec{v} < \frac{1}{2}$: on aura bien $|\lambda| < 1$.

On a $\vec{v}^T.L.\vec{v} = \vec{v}^T.(D^{-\frac{1}{2}}.E.D^{-\frac{1}{2}}).\vec{v}$, avec A définie positive donc $0 < (D^{-\frac{1}{2}}.\vec{v})^T.A.(D^{-\frac{1}{2}}.\vec{v}) = \vec{v}^T.D^{-\frac{1}{2}}.A.D^{-\frac{1}{2}}.\vec{v} = \vec{v}^T.(D^{-\frac{1}{2}}.(D-E-E^T).D^{-\frac{1}{2}}).\vec{v} = ||\vec{v}||^2 - 2\vec{v}^T.(D^{-\frac{1}{2}}.E.D^{-\frac{1}{2}}).\vec{v}$, d'où $\vec{v}^T.L.\vec{v} < \frac{1}{2}||\vec{v}||^2$. Et ici $||\vec{v}|| = 1$.

Le rayon spectral de B étant < 1, la méthode de Gauss–Seidel converge. ∎

Remarque 1.4 La méthode de Gauss-Seidel relaxée s'écrit :

$$M_\omega.\vec{x}^{k+1} = N_\omega.\vec{x}^k + \vec{b} \qquad \text{où} \qquad M_\omega = \frac{1}{\omega}(D - \omega E) \quad \text{et} \quad N_\omega = (\frac{1-\omega}{\omega}D + F).$$

Pour $\omega = 1$ on a la méthode de Gauss-Seidel. Pour $0 < \omega < 2$ la méthode relaxée converge. On parle de sous-relaxation pour $\omega < 1$ et de sur-relaxation pour $\omega > 1$. La méthode de sur-relaxation est conseillée (par exemple avec $\omega = 1,8$).

Si on note $B_\omega = M_\omega^{-1}.N_\omega$, le coefficient ω est calculé de telle sorte que le rayon spectral de la matrice B_ω soit le plus faible possible, de manière à ce que la convergence soit la plus rapide possible. Voir Golub et Van Loan [2]. ∎

1.3 Différentielle et gradient de f

Soit une fonction $f \in C^1(\mathbb{R}^n;\mathbb{R})$.

Comme $f \in C^1$, f admet un développement limité au premier ordre au voisinage de tout point $\vec{x}_0$: il existe une application linéaire $L_{\vec{x}_0} \overset{\text{noté}}{=} df(\vec{x}_0) : \mathbb{R}^n \to \mathbb{R}$, appelée la différentielle de f en $\vec{x}_0$, telle que :

$$\forall \vec{p} \in \mathbb{R}^n, \quad f(\vec{x}_0 + h\vec{p}) - f(\vec{x}_0) = h\, df(\vec{x}_0).\vec{p} + o(h). \tag{1.8}$$

Et pour $\vec{p}$ vecteur fixé, la valeur :

$$df(\vec{x}_0).\vec{p} = \lim_{h\to 0} \frac{f(\vec{x}_0 + h\vec{p}) - f(\vec{x}_0)}{h}. \tag{1.9}$$

est appelée la dérivée de f en $\vec{x}_0$ dans la direction $\vec{p}$.

Avec le produit scalaire cartésien $(\cdot,\cdot)_{\mathbb{R}^n}$, le théorème de représentation de Riesz permet de représenter la forme linéaire $df((\vec{x}_0)$ à l'aide d'un vecteur noté $\vec{\nabla} f(\vec{x}_0)$ et appelé le gradient de F en $\vec{x}_0$:

$$\forall \vec{v} \in \mathbb{R}^n, \quad df(\vec{x}).\vec{v} = (\vec{\nabla} f(\vec{x}), \vec{v})_{\mathbb{R}^n}. \tag{1.10}$$

Donc :

$$f(\vec{x}_0 + h\vec{p}) - f(\vec{x}_0) = h\,(\vec{\nabla} f(\vec{x}_0), \vec{p})_{\mathbb{R}^n} + o(h), \quad \text{et} \quad (\vec{\nabla} f(\vec{x}_0), \vec{p})_{\mathbb{R}^n} = \lim_{h\to 0} \frac{f(\vec{x}_0 + h\vec{p}) - f(\vec{x}_0)}{h}.$$

1.4 Cas f quadratique

C'es le cas, pour A matrice $n \times n$ de $f : \mathbb{R}^n \to \mathbb{R}$ définie par :

$$f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x} + \vec{c}, \tag{1.11}$$

et on cherche le minimum de f. Par dérivation le vecteur $\vec{c}$ disparaît : pour alléger l'écriture on prend $\vec{c} = \vec{0}$ dans la suite.

Proposition 1.5 *On a $\vec{\nabla} f(\vec{x}) = \frac{1}{2}(A + A^T).\vec{x} - \vec{b}$. En particulier, si A est une matrice symétrique, on a :*

$$\vec{\nabla} f(\vec{x}) = A.\vec{x} - \vec{b},$$

Preuve. Par définition de la dérivation dans une direction $\vec{v}$ donnée, cf. (1.9), on a :

$$\begin{aligned} df(\vec{x}).\vec{v} &= \lim_{h\to 0} \frac{\frac{1}{2}h\vec{v}^T.A.\vec{x} + \frac{1}{2}\vec{x}^T.A.\vec{v} + \frac{1}{2}h^2\vec{v}^T.A.\vec{v} - h\vec{b}^T.\vec{v}}{h} \\ &= \frac{1}{2}\vec{x}^T.A^T.\vec{v} + \frac{1}{2}\vec{x}^T.A.\vec{v} - \vec{b}^T.\vec{v} = \frac{1}{2}((A + A^T).\vec{x} - \vec{b})^T.\vec{v} = (\vec{\nabla} f(\vec{x}), \vec{v})_{\mathbb{R}^n}, \end{aligned}$$

où on a utilisé (1.10). Et si A est symétrique, $A = A^T$. ∎

Dans la suite, on supposera souvent A symétrique définie positive. Et dans ce cas, le minimum de f existe, est unique, et est donné par le point $\vec{x}$ tel que $\vec{\nabla} f(\vec{x}) = 0$, i.e. tel que :

$$A.\vec{x} = \vec{b}. \tag{1.12}$$

C'est bien le problème initial (1.1) qu'on avait à résoudre.

1.5 Gauss–Seidel : point de vue descente le long des vecteurs de la base canonique

Cas f donnée par (1.11) avec A symétrique.

On se donne un point $\vec{x}^0 \in \mathbb{R}^n$, et on cherche un point noté $\vec{x}^{0+\frac{1}{n}}$ tel que :

$$f(\vec{x}^{0+\frac{1}{n}}) \leq \min_{\rho} f(\vec{x}^0 - \rho\vec{e}_1).$$

On aura $\vec{x}^{0+\frac{1}{n}} = \vec{x}^0 - \rho\vec{e}_1$ qui réalise le minimum de f sur la droite $\vec{x}^0 + \text{Vect}\{\vec{e}_1\}$ (droite parallèle au premier axe des coordonnées passant par le point $\vec{x}^0$). Pour ce on définit $\varphi^0 : \mathbb{R} \to \mathbb{R}$ par :

$$\varphi^0(\rho) = f(\vec{x}^0 - \rho\vec{e}_1), \tag{1.13}$$

et on minimise φ^0. On a :

$$(\varphi^0)'(\rho) = df(\vec{x}^0 - \rho\vec{e}_1).(-\vec{e}_1) = -\vec{\nabla}f(\vec{x}^0 - \rho\vec{e}_1)^T.\vec{e}_1 = -(A.(\vec{x}^0 - \rho\vec{e}_1) - \vec{b})^T.\vec{e}_1, \tag{1.14}$$

et le minimum est donné pour ρ tel que $(\varphi^0)'(\rho) = 0$, i.e. pour ρ tel que :

$$a_{11}(x_1^0 - \rho) + a_{12}x_2^0 + \ldots + a_{1n}x_n^0 - b_1 = 0, \tag{1.15}$$

d'où ρ. Et on a retrouvé (1.3) avec $x_0^1 = x_1^0 - \rho$ donnant le point $\vec{x}^{0+\frac{1}{n}}$.

De même pour les étapes suivantes en posant $\varphi^{k+\frac{i-1}{n}}(\rho) = f(\vec{x}^{k+\frac{i-1}{n}} + \rho\vec{e}_i)$.

La méthode Gauss–Seidel est donc bien une méthode de descente.

1.6 Gauss–Seidel et convergence en n étapes pour la sphère

On rappelle que si A une matrice symétrique définie positive, alors A est diagonalisable dans une b.o.n., ce qui s'écrit $A = P.D.P^{-1}$ où $D = \text{diag}(\lambda_1, ..., \lambda_n)$ est la matrice diagonale des valeurs propres, et $P^{-1} = P^T$.

Soit $I = \text{diag}(1, 1, ..., 1)$ la matrice identité.

Proposition 1.6 *Si $\lambda \neq 0$ et si $A = \lambda I$ (matrice sphérique), alors, partant d'un point $\vec{x}_0$ donné quelconque, l'algorithme de Gauss–Seidel converge en n étapes, i.e. le point $\vec{x}^{0+\frac{n}{n}} = \vec{x}^1$ est solution de $\lambda I.\vec{x} = \vec{b}$, i.e. $\vec{x}^1 = \frac{1}{\lambda}\vec{b}$.*

Et quelque soit base orthonormée $(\vec{p}_i)_{i=1,...,n}$ et le changement de base $\vec{e}_i \to \vec{p}_i$ pour tout $i = 1, .., n$, la matrice sphérique A est inchangée (reste sphérique), et l'algorithme de Gauss–Seidel converge en n étapes.

(Signification : la matrice A est la matrice d'un endomorphismes $L : \mathbb{R}^n \to \mathbb{R}^n$ dans la base $(\vec{e}_i)$, on note $[L]_{|(\vec{e})} = A$; et la matrice de L dans une base $(\vec{p}_i)$ est notée $[L]_{|(\vec{p})}$. Et la proposition affirme que si $[L]_{|(\vec{e})}$ est sphérique alors $[L]_{|(\vec{p})}$ est sphérique dès que $(\vec{p}_i)$ est une b.o.n.)

Preuve. L'algorithme de Gauss–Seidel s'écrit $(D - E).\vec{x}^1 = F.\vec{x}^0 + \vec{b}$, i.e. ici $\lambda I.\vec{x}^1 = \vec{b}$, d'où $\vec{x}^1 = \frac{1}{\lambda}\vec{b}$.

Soit $P = ([\vec{p}_1] \ \ldots \ [\vec{p}_n])$ la matrice dont les colonnes sont données par les coordonnées des vecteurs $\vec{p}_i$ dans la base canonique. Comme $[\vec{p}_i]^T.[\vec{p}_j] = \delta_{ij}$ pour tout i, j, on a $P^T.P = I$. D'où $P^{-1} = P^T$, et par changement de base, la matrice $A = \lambda I$ est transformée en $B = P^{-1}.A.P = \lambda P^{-1}.I.P = \lambda I = A$: la matrice est inchangée. ▟

Noter qu'une matrice $A = \lambda I$ est appelée matrice sphérique, car si $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} = \frac{\lambda}{2}||\vec{x}||^2_{\mathbb{R}^n}$, alors une surface de niveau $\{\vec{x} : f(\vec{x}) = \text{constante}\}$ est une sphère.

Proposition 1.7 *Si $A = \text{diag}(\lambda_1, \ldots, \lambda_n)$ est une matrice diagonale où tous les $\lambda_i \neq 0$, alors, partant d'un point $\vec{x}_0$ donné quelconque, l'algorithme de Gauss–Seidel converge en n étapes, i.e. le point $\vec{x}^{0+\frac{n}{n}} = \vec{x}^1$ est solution de $\text{diag}(\lambda_1, \ldots, \lambda_n).\vec{x} = \vec{b}$, i.e. $(\vec{x}^1)_i = \frac{1}{\lambda_i}(\vec{b})_i$.*

Preuve. L'algorithme s'écrit $(D - E).\vec{x}^1 = F.\vec{x}^0 + \vec{b}$, i.e. ici $\text{diag}(\lambda_1, \ldots, \lambda_n).\vec{x}^1 = \vec{b}$. ▟

Noter qu'une matrice $A = \text{diag}(\lambda_1, \ldots, \lambda_n)$ est appelée matrice elliptique quand tous les $\lambda_i > 0$, car si $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} = \frac{1}{2}\sum_i \lambda_i x_i^2$, alors une surface de niveau $\{\vec{x} : f(\vec{x}) = \text{constante}\}$ est un ellipsoïde (dont les axes principaux sont les axes de coordonnées).

Remarque 1.8 La méthode du gradient conjugué consistera à rendre l'ellipsoïde sphérique. ▟

Remarque 1.9 Mais attention : si on fait un changement de base orthonormée $(\vec{p}_i)_{i=1,...,n}$, la matrice diagonale A devient $B = P^{-1}.A.P$ qui n'est plus diagonale (bien que symétrique), et la méthode de Gauss–Seidel ne converge plus en n étapes, contrairement au cas de la matrice sphérique.

Par exemple, soit $A = D = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ et soit $P = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$ la matrice de rotation d'angle $\theta \neq 0$ modulo 2π. On prend P comme matrice de passage : donc $\vec{p}_1 = \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$ et $\vec{p}_2 = \begin{pmatrix} -\sin\theta \\ \cos\theta \end{pmatrix}$. On a $B = P^{-1}.\begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.P$. Prenons $\theta = -\frac{\pi}{4}$ par exemple. Alors $B = \begin{pmatrix} \frac{3}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}$: n'est pas diagonale. ▟

1.7 Matrice elliptique et base A-orthogonale

1.7.1 Base A-orthogonale

Si A est une matrice symétrique définie positive, si $A = P.D.P^{-1}$ avec D diagonale et $P^{-1} = P^T$ (décomposition diagonale), alors on note $\sqrt{A} = A^{\frac{1}{2}} \stackrel{\text{def}}{=} P.\sqrt{D}.P^{-1}$ la matrice symétrique définie positive dite racine carrée de A, où $D^{\frac{1}{2}} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_n})$ (on vérifie trivialement que $\sqrt{A}.\sqrt{A} = A^{\frac{1}{2}}.A^{\frac{1}{2}} = A$).

Et alors la forme bilinéaire $(\cdot,\cdot)_A : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ définie par :

$$(\vec{v},\vec{w})_A \stackrel{\text{def}}{=} (A.\vec{v},\vec{w})_{\mathbb{R}^n} = \vec{w}^T.A.\vec{v} = \vec{v}^T.A.\vec{w} = (\sqrt{A}.\vec{v},\sqrt{A}.\vec{w})_{\mathbb{R}^n} \tag{1.16}$$

est un produit scalaire, et $||.||_A : \mathbb{R}^n \to \mathbb{R}_+$ définie par :

$$||\vec{v}||_A = \sqrt{(\vec{v},\vec{w})_A} = ||\sqrt{A}.\vec{v}||_{\mathbb{R}^n} \tag{1.17}$$

est la norme associée.

Définition 1.10 Soit A une matrice symétrique définie positive. Deux vecteurs $\vec{v},\vec{w} \in \mathbb{R}^n$ son dits A-conjugués (ou simplement conjugués si la matrice A est implicite) ssi ils sont orthogonaux pour le produit scalaire $(\cdot,\cdot)_A$, i.e. ssi ils sont $(\cdot,\cdot)_A$-orthogonaux :

$$\vec{v} \text{ et } \vec{w} \text{ } A\text{-conjugués} \quad \Leftrightarrow \quad (\vec{v},\vec{w})_A = 0 \quad \Leftrightarrow \quad (A.\vec{v},\vec{w})_{\mathbb{R}^n} = 0 \quad \Leftrightarrow \quad (\sqrt{A}.\vec{v},\sqrt{A}.\vec{w})_{\mathbb{R}^n} = 0. \tag{1.18}$$

En d'autres termes les directions $\vec{v}$ et $\vec{w}$ sont A-conjuguées ssi, une fois déformées par la transformation $\vec{v} \to \sqrt{A}\vec{v}$ elle sont orthogonales : $(\sqrt{A}.\vec{v},\sqrt{A}.\vec{w})_{\mathbb{R}^n} = 0$.

Définition 1.11 Soit $(\vec{p}_i)_{i=1,...,n}$ une base de $\mathbb{R}^n$. La base $(\vec{p}_i)_{i=1,...,n}$ est dite A-orthogonale ssi, pour tout $i,j = 1,...,n$:

$$i \neq j \quad \Rightarrow \quad (\vec{p}_i,\vec{p}_j)_A = 0 \quad (= \vec{p}_i^T.A.\vec{p}_j), \tag{1.19}$$

i.e. ssi les $\vec{p}_i$ sont deux à deux A-conjugués (i.e. $(\cdot,\cdot)_A$-orthogonaux). Et la base $(\vec{p}_i)_{i=1,...,n}$ est dite A-orthonormale ssi, pour tout $i,j = 1,...,n$:

$$(\vec{p}_i,\vec{p}_j)_A = \delta_{ij} \qquad (= \vec{p}_i^T.A.\vec{p}_j), \tag{1.20}$$

i.e. ssi elle est $(\cdot,\cdot)_A$-orthonormale, i.e. orthonormale relativement au produit scalaire $(\cdot,\cdot)_A$.

1.7.2 Ellipsoïde rendu sphérique

La proposition 1.7 indique que les directions des vecteurs propres sont des directions de descentes qui permettent une convergence en n étapes. Malheureusement le coût de calcul des vecteurs propres est de même ordre que le coût du calcul de l'inverse de la matrice : c'est d'ailleurs une technique usuelle de calcul de l'inverse, voir (1.27). Donc le calcul des vecteurs propres coûte trop cher.

L'idée est alors de rendre l'ellipsoïde sphérique, plus exactement de le faire apparaître sphérique, à l'aide d'un produit scalaire adapté, celui donné par la matrice A. Voir polycopié "Compléments" : Directions conjuguées : orthogonalité sur l'ellipse versus orthogonalité sur le cercle.

Et toute base A-orthogonale va alors permettre la convergence de Gauss–Seidel en n-étapes.

Soit :

$$f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} = \frac{1}{2}(\vec{x},\vec{x})_A = \frac{1}{2}||\vec{x}||_A^2 = \frac{1}{2}(A^{\frac{1}{2}}.\vec{x}, A^{\frac{1}{2}}.\vec{x})_{\mathbb{R}^n}. \tag{1.21}$$

Toute courbe de niveau de f est un ellipsoïde. On pose :

$$g(\vec{X}) \stackrel{\text{def}}{=} f(A^{-\frac{1}{2}}.\vec{X})) = f(\vec{x}) \qquad \text{quand} \qquad \vec{X} = A^{\frac{1}{2}}.\vec{x}. \tag{1.22}$$

Donc :

$$g(\vec{X}) = \frac{1}{2}\vec{X}^T.\vec{X} = \frac{1}{2}||\vec{X}||_{\mathbb{R}^n}^2, \tag{1.23}$$

et toute courbe de niveau de g est sphérique : on a transformé les ellipsoïdes en sphères.

Ici on a utilisé le changement de variables $\vec{x} \in \mathbb{R}^n \to \vec{X} = A^{\frac{1}{2}}.\vec{x} \in \mathbb{R}^n$.

Donc, si $(\vec{P}_i)_{i=1,\dots,n}$ est une base orthonormée de $\mathbb{R}^n$, i.e. t.q. pour tout i,j :

$$(\vec{P}_i, \vec{P}_j)_{\mathbb{R}^n} = \delta_{ij}, \tag{1.24}$$

alors l'algorithme de Gauss–Seidel appliqué à g converge en n-étapes, voir proposition 1.6.

Retour à f : soit $\vec{X} = \sum_i X_i\vec{P}_i$. À l'aide du changement de variables inverse $\vec{x} = A^{-\frac{1}{2}}.\vec{X}$, on définit les vecteurs transformés, pour tout i :

$$\vec{p}_i = A^{-\frac{1}{2}}.\vec{P}_i. \tag{1.25}$$

Comme $\vec{P}_i = A^{\frac{1}{2}}.\vec{p}_i$, les $\vec{p}_i$ forment une base A-orthonormée :

$$(\sqrt{A}.\vec{p}_i, \sqrt{A}.\vec{p}_j)_{\mathbb{R}^n} = \delta_{ij} = (\vec{p}_i,\vec{p}_j)_A. \tag{1.26}$$

C'est une telle base $(\vec{p}_i)$ qu'on va choisir pour f : si on descend dans les directions d'une telle base, l'algorithme de Gauss–Seidel converge en n étapes, car descendre le long de $\vec{p}_i$ pour f c'est descendre le long de $\vec{P}_i$ pour g.

On va voir qu'on n'aura pas besoin de calculer $\sqrt{A}$ (trop coûteux) pour trouver une base $(\vec{p}_i)$ A-orthonormée, voir corollaire 1.15.

1.8 Inverse de A

Proposition 1.12 *Soit A une matrice symétrique définie positive. Si $(\vec{p}_i)$ est une base A-orthonormale, alors A^{-1} est somme des n matrices élémentaires $\vec{p}_k.\vec{p}_k^T$:*

$$A^{-1} = \sum_{k=1}^n \vec{p}_k.\vec{p}_k^T. \tag{1.27}$$

(C'est trivial si $A = I$.)

Preuve. A est inversible car définie positive. Soit $B = \sum_{k=1}^n \vec{p}_k.\vec{p}_k^T$.

On a $B.A = \sum_{k=1}^n \vec{p}_k.\vec{p}_k^T.A$, d'où $B.A.\vec{p}_j = \sum_{k=1}^n \vec{p}_k.\vec{p}_k^T.A.\vec{p}_j = \sum_{k=1}^n \vec{p}_k\delta_{kj} = \vec{p}_j$ pour tout j : on a obtenu $B.A.\vec{p}_j = \vec{p}_j$ pour tout j.

Et donc $B.A.\vec{x} = \vec{x}$ pour tout $\vec{x}$, donc $B.A = I$, donc $B = A^{-1}$ (multiplier par A^{-1} à droite). ▪

Exercice 1.13 Montrer que $A.B = I$ où B est la matrice de la preuve précédente (sans utiliser la démonstration précédente qui montre $B.A = I$).

Réponse. Comme $(\vec{p}_j)$ est une base de $\mathbb{R}^n$, $(A\vec{p}_j)$ est également une base car A est inversible. Et on a $A.B = \sum_{k=1}^n A.\vec{p}_k.\vec{p}_k^T$, et donc $A.B.(A.\vec{p}_j) = \sum_{k=1}^n A.\vec{p}_k.\vec{p}_k^T.A.\vec{p}_j = \sum_{k=1}^n A.\vec{p}_k.\delta_{kj}$, donc, pour tout j on a $A.B.(A.\vec{p}_j) = A.\vec{p}_j$. Donc $A.B.\vec{x} = \vec{x}$ pour tout $\vec{x}$, i.e. $A.B = I$. ▪

Corollaire 1.14 *Si A est une matrice symétrique définie positive, si $(\vec{p}_i)_{i=1,\dots,n}$ est une base A-orthogonale, alors la solution $\vec{x}$ vérifiant $A.\vec{x} = \vec{b}$ est donnée par :*

$$\vec{x} = \sum_{k=1}^n \alpha_k.\vec{p}_k \qquad \text{où} \qquad \alpha_k = \frac{\vec{p}_k^T.\vec{b}}{||\vec{p}_k||_A^2} \qquad (= \frac{(\vec{p}_k,\vec{b})_{\mathbb{R}^n}}{(\vec{p}_k,\vec{p}_k)_A}). \tag{1.28}$$

(Les α_k sont les composantes de $\vec{x}$ sur la base $(\vec{p}_i)_{i=1,\dots,n}$.) (Trivial si $A = I$.)

Preuve. On a $\vec{x} = A^{-1}.\vec{b} = \sum_{k=1}^n \frac{\vec{p}_k.\vec{p}_k^T}{||\vec{p}_k||_A^2}.\vec{b} = \sum_{k=1}^n \frac{\vec{p}_k^T.\vec{b}}{||\vec{p}_k||_A^2}.\vec{p}_k$. ▪

1.9 Gauss-Seidel généralisé et convergence en n étapes pour l'ellipsoïde

Corollaire 1.15 *Soit A est une matrice symétrique définie positive. Soit $(\vec{p}_j)$ une base A-orthogonale donnée. Soit $\vec{x}_0 \in \mathbb{R}^n$ donné (point de départ de la méthode itérative). On pose $\vec{r}_0 = A.\vec{x}_0 - \vec{b}$ (résidu initial). Alors la solution $\vec{x}$ du problème $A.\vec{x} = \vec{b}$ est donnée par :*

$$\vec{x} = \vec{x}_0 - \sum_{k=1}^{n} \alpha_k \vec{p}_k, \qquad où \qquad \forall k \geq 1, \quad \alpha_k = \frac{\vec{p}_k^T.\vec{r}_0}{||\vec{p}_k||_A^2} \quad (= \frac{\vec{p}_k^T.A.\vec{x}_0 - \vec{p}_k^T.\vec{b}}{\vec{p}_k^T.A.\vec{p}_k}). \tag{1.29}$$

D'où l'algorithme itératif de résolution de $A.\vec{x} = \vec{b}$: on part de $\vec{x}_0$ et on obtient la solution $\vec{x} = \vec{x}_n$ par l'intérmédiaire des n-étapes :

$$\vec{x}_1 = \vec{x}_0 - \alpha_1 \vec{p}_1, \ldots, \quad \vec{x}_k = \vec{x}_{k-1} - \alpha_k \vec{p}_k, \quad \ldots, \quad \vec{x}_n = \vec{x}_{n-1} - \alpha_n \vec{p}_n. \tag{1.30}$$

C'est la méthode de Gauss-Seidel généralisée après choix d'une base $(\vec{p}_i)$ A-orthogonale.

Preuve. Par définition de $\vec{r}_0$ on a $\vec{b} = A\vec{x}_0 - \vec{r}_0$, et $A.\vec{x} = \vec{b}$ équivaut à $A.\vec{x} = A.\vec{x}_0 - \vec{r}_0$, i.e. à $A.(\vec{x} - \vec{x}_0) = -\vec{r}_0$. Et donc $\vec{x} - \vec{x}_0 = -A^{-1}.\vec{r}_0 = -\sum_{k=1}^n \frac{\vec{p}_k.\vec{p}_k^T}{||\vec{p}_k||_A^2}.\vec{r}_0 = -\sum_{k=1}^n \frac{\vec{p}_k^T.\vec{r}_0}{||\vec{p}_k||_A^2}\,\vec{p}_k$, i.e. (1.29). ∎

Corollaire 1.16 *Et si on pose $\vec{r}_k = A.\vec{x}_k - \vec{b}$ pour tout $k = 1, \ldots, n$, alors :*

$$\vec{r}_k = \vec{r}_{k-1} - \alpha_k A.\vec{p}_k \qquad et \qquad \vec{r}_k = \vec{r}_\ell - \sum_{i=\ell+1}^{k} \alpha_i A.\vec{p}_i, \quad \forall 0 \leq \ell \leq k-1, \tag{1.31}$$

et en particulier $\vec{r}_k = \vec{r}_0 - \sum_{i=1}^k \alpha_i A.\vec{p}_i$. D'où :

$$(\vec{r}_k, \vec{p}_i)_{\mathbb{R}^n} = 0, \quad \forall 1 \leq i \leq k, \tag{1.32}$$

i.e. le nouveau résidu est orthogonal aux k premiers vecteurs $\vec{p}_1, ..., \vec{p}_k$ de la base A-orthogonale :

$$\vec{r}_k \in (\mathrm{Vect}\{\vec{p}_1, ..., \vec{p}_k\})^\perp, \quad \forall k \geq 1. \tag{1.33}$$

En particulier $\vec{r}_n = \vec{0}$, i.e. $A.\vec{x}_n - \vec{b} = 0$, et $\vec{x}_n$ est la solution. (Il est fondamental de prendre une base $(\vec{p}_i)_{i=1,...,n}$ qui est A-orthogonale.) Et on a également :

$$\alpha_k = \frac{\vec{p}_k^T.\vec{r}_{k-1}}{||\vec{p}_k||_A^2}. \tag{1.34}$$

Preuve. $\vec{x}_k = \vec{x}_{k-1} - \alpha_k \vec{p}_k$ donne $A.\vec{x}_k = A.\vec{x}_{k-1} - \alpha_k A.\vec{p}_k$, d'où $\vec{r}_k = \vec{r}_{k-1} - \alpha_k A.\vec{p}_k$. D'où $\vec{r}_k = (\vec{r}_{k-2} - \alpha_{k-1} A.\vec{p}_{k-1}) - \alpha_k A.\vec{p}_k = \vec{r}_{k-2} - \sum_{i=k-1}^k \alpha_i A.\vec{p}_i$, ..., d'où $\vec{r}_k = \vec{r}_\ell - \sum_{i=\ell+1}^k \alpha_i A.\vec{p}_i$, $\forall 0 \leq \ell \leq k-1$, et finalement $\vec{r}_k = \vec{r}_0 - \sum_{i=1}^k \alpha_i A.\vec{p}_i$.

D'où $(\vec{r}_k, \vec{p}_k)_{\mathbb{R}^n} = (\vec{r}_0, \vec{p}_k)_{\mathbb{R}^n} - 0 - \alpha_k \vec{p}_k^T.A.\vec{p}_k = 0$ car α_k est donné par (1.29).

D'où $(\vec{r}_k, \vec{p}_{k-1})_{\mathbb{R}^n} = (\vec{r}_{k-1}, \vec{p}_{k-1})_{\mathbb{R}^n} - \alpha_k (A.\vec{p}_k, \vec{p}_{k-1})_{\mathbb{R}^n} = 0 + 0 = 0$.

D'où $(\vec{r}_k, \vec{p}_{k-2})_{\mathbb{R}^n} = (\vec{r}_{k-2} - \alpha_k A.\vec{p}_k - \alpha_{k-1} A.\vec{p}_{k-1}, \vec{p}_{k-2})_{\mathbb{R}^n} = (\vec{r}_{k-2}, \vec{p}_{k-2})_{\mathbb{R}^n} + 0 + 0 = 0$, puis, ..., puis $(\vec{r}_k, \vec{p}_1)_{\mathbb{R}^n} = (\vec{r}_1 - \sum_{i=2}^k \alpha_i A.\vec{p}_i, \vec{p}_1)_{\mathbb{R}^n} = 0 + 0 = 0$.

Enfin, on obtient $(\vec{r}_{k-1}, \vec{p}_k)_{\mathbb{R}^n} = (\vec{r}_0, \vec{p}_k)_{\mathbb{R}^n} - \sum_{i=1}^{k-1} (A.\vec{p}_i, \vec{p}_k) = (\vec{r}_0, \vec{p}_k)$, d'où $\alpha_k = \frac{\vec{p}_k^T.\vec{r}_0}{||\vec{p}_k||_A^2} = \frac{\vec{p}_k^T.\vec{r}_{k-1}}{||\vec{p}_k||_A^2}$. ∎

1.10 Gauss-Seidel généralisé = point de vue descente le long de vecteurs A-conjugués

On transforme le problème $A.\vec{x} = \vec{b}$ en le problème : trouver le minimum de la fonction $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ (cas A symétrique définie positive).

Proposition 1.17 *Soit A une matrice symétrique définie positive et $(\vec{p}_i)$ une base A-orthogonale donnée. Appliqué au problème $A.\vec{x} = \vec{b}$, l'algorithme de Gauss-Seidel généralisé, d'annulation des composantes sur la base $(\vec{p}_i)$, est un algorithme de descente le long des vecteurs $\vec{p}_i$.*

Et (1.32) indique que le dernier résidu $\vec{r}_k$ est orthogonal aux k premières directions de descentes $(\vec{p}_1, ..., \vec{p}_k)$.

Preuve. On a $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ et on se donne un point $\vec{x}^0$. On pose :

$$\varphi^0(\alpha) = f(\vec{x}^0 - \alpha\vec{p}_1) \tag{1.35}$$

qu'on minimise pour obtenir $\vec{x}^{0+\frac{1}{n}} = \vec{x}^0 - \alpha_1\vec{p}_1$. Puis séquentiellement on pose :

$$\varphi^{0+\frac{i-1}{n}}(\alpha) = f(\vec{x}^{0+\frac{i-1}{n}} - \alpha\vec{p}_i) \tag{1.36}$$

qu'on minimise pour obtenir $\vec{x}^{0+\frac{i}{n}} = \vec{x}^{0+\frac{i-1}{n}} - \alpha_i\vec{p}_i$.

Le calcul des α_i est donné par :

$$0 = (\varphi^{0+\frac{i-1}{n}})'(\alpha_i) = -\vec{\nabla}f(\vec{x}^{0+\frac{i-1}{n}} - \alpha_i\vec{p}_i)^T.\vec{p}_i = -((\vec{x}^{0+\frac{i-1}{n}} - \alpha_i\vec{p}_i)^T.A - \vec{b}^T).\vec{p}_i,$$

et donc :

$$\alpha_i||\vec{p}_i||_A^2 = (\vec{x}^{0+\frac{i-1}{n}})^T.A.\vec{p}_i - \vec{b}^T.\vec{p}_i = (\vec{r}^{0+\frac{i-1}{n}})^T.\vec{p}_i. \tag{1.37}$$

On a bien la méthode de Gauss–Seidel généralisée du § précédent. ∎

2 Algorithme de Gram–Schmidt

À partir d'une base quelconque $(\vec{v}_1, ..., \vec{v}_n)$ de $\mathbb{R}^n$ on veut construire une base $(\vec{p}_1, ..., \vec{p}_n)$ A-orthogonale, i.e. une base de directions conjuguées relativement à la matrice A.

(Classiquement, à partir d'une base quelconque $(\vec{v}_i)$ on veut construire une base orthonormée, i.e. que Gram–Schmidt classique traite le cas $A = I$.)

On pose (initialisation) :

$$\vec{p}_1 = \vec{v}_1, \tag{2.1}$$

Puis on construit $\vec{p}_2$ à l'aide de $\vec{p}_1$ et $\vec{v}_2$ en posant :

$$\vec{p}_2 = \vec{v}_2 - \beta_{12}\vec{p}_1,$$

où β_{12} est calculé pour que $(\vec{p}_2, \vec{p}_1)_A = 0$. On veut donc $0 = (\vec{v}_2, \vec{p}_1)_A - \beta_{12}(\vec{p}_1, \vec{p}_1)_A$, i.e. :

$$\beta_{12} = \frac{(\vec{v}_2, \vec{p}_1)_A}{(\vec{p}_1, \vec{p}_1)_A} = \frac{\vec{v}_2^T.A.\vec{p}_1}{\vec{p}_1^T.A.\vec{p}_1}.$$

Puis successivement, on définit $\vec{p}_j$ à l'aide de $\vec{p}_1, ..., \vec{p}_{j-1}$ et $\vec{v}_j$ par :

$$\vec{p}_j = \vec{v}_j - \sum_{i=1}^{j-1}\beta_{ij}\vec{p}_i. \tag{2.2}$$

où les β_{ij} pour $1 \leq i < j$ sont calculés de telle sorte que $(\vec{p}_j, \vec{p}_\ell)_A = 0$ pour tout $\ell < j$. On obtient : $0 = (\vec{v}_j, \vec{p}_\ell)_A - \beta_{\ell j}(\vec{p}_\ell, \vec{p}_\ell)_A$ pour tout $\ell < j$, et donc :

$$\beta_{ij} = \frac{(\vec{v}_j, \vec{p}_i)_A}{(\vec{p}_i, \vec{p}_i)_A} = \frac{\vec{p}_i^T.A.\vec{v}_j}{\vec{p}_i^T.A.\vec{p}_i}, \qquad \forall i, j : 1 \leq i < j \leq n. \tag{2.3}$$

On a ainsi obtenue la base $(\vec{p}_1, ..., \vec{p}_n)$ A-orthogonale. En particulier, la base $(\frac{\vec{p}_j}{||\vec{p}_j||_A})$ est A-orthonornale.

L'algorithme est donc :

1. Initialisation : $\vec{p}_1 = \vec{v}_1$,
2. Etape $j \geq 2$:
   $\beta_{ij} = \frac{\vec{v}_j^T.A.\vec{p}_i}{\vec{p}_i^T.A.\vec{p}_i}$ pour $i < j$, i.e. (2.3).
   $\vec{p}_j = \vec{v}_j - \sum_{i=1}^{j-1}\beta_{ij}\vec{p}_i$, i.e. (2.2).

Remarque 2.1 Dans le cas général ci-dessus, l'algorithme créé des vecteurs $\vec{p}_i$ qui deviennent très rapidement non orthogonaux (accumulation d'erreurs). On préfère alors utiliser un algorithme modifié, voir Golub et Van Loan. Pour l'application à la méthode du gradient conjugué, on verra que les β_{ij} sont tous nuls sauf les $\beta_{i,i+1}$, et la perte d'orthogonalité est moindre.

On peut noter que l'algorithme donne $[\underline{\underline{v}}] = [\underline{\underline{p}}].[\underline{\underline{\beta}}]$, cf (2.2) qui donne cette égalité matricielle colonne par colonne, où $[\underline{\underline{v}}]$ est la matrice dont les colonnes sont les vecteurs $\vec{v}_j$, $[\underline{\underline{p}}]$ est la matrice dont les colonnes sont les vecteurs $\vec{p}_j$, et $[\underline{\underline{\beta}}]$ est la matrice triangulaire supérieure avec des 1 sur la diagonale et les β_{ij} pour $j > i$ donnés par (2.3). ∎

3 Méthode du gradient

On note $(\cdot,\cdot)_{\mathbb{R}^n}$ le produit scalaire canonique de $\mathbb{R}^n$.

3.1 Pourquoi descendre le long du gradient

On veut trouver un minimum d'une fonction $f \in C^1(\mathbb{R}^n;\mathbb{R})$: trouver $\vec{x}_\infty \in \mathbb{R}^n$ tel que :

$$f(\vec{x}_\infty) = \inf_{\vec{x}\in\mathbb{R}^n} f(\vec{x}). \tag{3.1}$$

Comme $f \in C^1$, on dispose de sa différentielle, cf. (1.8). Et ici on utilise le produit scalaire cartésien : on dispose donc du gradient, cf. (1.10).

D'où, pour minimiser f : on part d'un point $\vec{x}_0$ donné.
On cherche un point $\vec{x}_1$ tel que $f(\vec{x}_1) < f(\vec{x}_0)$.
Soit $\vec{p} \in \mathbb{R}^n$ fixé (on posera $\vec{x}_1 = \vec{x}_0 + h\vec{p}$).
Le taux de variation vérifie $\frac{f(\vec{x}_0+h\vec{p})-f(\vec{x}_0)}{h} = (\vec{\nabla}f(\vec{x}_0),\vec{p})_{\mathbb{R}^n} + o(1)$, cf. (1.3).
Donc à la limite $h \to 0$, on choisit $\vec{p}$ de direction t.q. $(\vec{\nabla}f(\vec{x}_0),\vec{p})_{\mathbb{R}^n}$ soit minimal (maximal en valeur absolue). Le théorème de Cauchy–Schwarz qui montre qu'on doit choisir $\vec{p} \parallel \vec{\nabla}f(\vec{x}_0)$, et plus précisément :

$$\vec{p} = -\rho\vec{\nabla}f(\vec{x}_0) \quad \text{où} \quad \rho > 0. \tag{3.2}$$

Quitte à modifier la longueur de $\vec{p}$ (la direction de descente), prenons $\vec{p} = -\vec{\nabla}f(\vec{x}_0)$. Ainsi, le taux de variation négatif le plus élevé en valeur absolue est donné par :

$$\frac{f(\vec{x}_0 - h\vec{\nabla}f(\vec{x}_0)) - f(\vec{x}_0)}{h} = -||\vec{\nabla}f(\vec{x}_0)||^2_{\mathbb{R}^n} + o(1).$$

On cherche alors $\vec{x}_1$ sur la droite $\vec{x}_0 + \mathrm{Vect}\{\vec{\nabla}f(\vec{x}_0)\}$, i.e. $\vec{x}_1$ de la forme :

$$\vec{x}_1 = \vec{x}_0 - \rho_0\vec{\nabla}f(\vec{x}_0), \tag{3.3}$$

où $\rho_0 > 0$ est à déterminer, tel que $f(\vec{x}_1) < f(\vec{x}_0)$: plus précisément, on choisit un ρ_0 qui approche "bien" le minimum de $g(\rho) = f(\vec{x}_1(\rho))$: par exemple avec la méthode de gradient à pas optimal, ou la méthode de gradient à pas fixe, voir la suite.

Ensuite on part de $\vec{x}_1 = \vec{x}_0 - \rho_0\vec{\nabla}f(\vec{x}_0)$ et on descend le long de $\vec{\nabla}f(\vec{x}_1)$ pour obtenir $\vec{x}_2 = \vec{x}_1 - \rho_1\vec{\nabla}f(\vec{x}_1)$. Et on itère le procédé pour obtenir la suite $(\vec{x}_k)$ définie par récurrence par :

$$\vec{x}_{k+1} = \vec{x}_k - \rho_k\vec{\nabla}f(\vec{x}_k). \tag{3.4}$$

3.2 Cas $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$

Pour A symétrique on a dans ce cas :

$$\vec{\nabla}f(\vec{x}) = A.\vec{x} - \vec{b}, \tag{3.5}$$

et la direction de descente à partir d'un point $\vec{x}$ est donc $\vec{p} = \vec{b} - A.\vec{x}$.

Et l'algorithme du gradient s'écrit :
1- $\vec{x}_0$ donné, donc $\vec{\nabla}f(\vec{x}_0) = A.\vec{x}_0 - \vec{b}$.
2- Etape k pour $k \geq 0$:
on calcul ρ_k (voir la suite),
on pose $\vec{x}_{k+1} = \vec{x}_k - \rho_k(A.\vec{x}_k - \vec{b})$.
3- On met un critère d'arrêt de type $\varepsilon > 0$ et $|f(\vec{x}_k) - f(\vec{x}_{k+1})| < \varepsilon$.

3.3 Méthode du gradient à pas optimal

On commence par la proposition :

Proposition 3.1 *Soit* $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ *où* A *est une matrice symétrique définie positive. Soit* $\vec{x} \in \mathbb{R}^n$ *donné. Soit* $\vec{v} \in \mathbb{R}^n$ *donné. Soit :*

$$\varphi(\rho) = f(\vec{x} - \rho\vec{v}). \tag{3.6}$$

(Donc φ *donne les valeurs de* f *le long de la droite passant par* $\vec{x}$ *de vecteur directeur* $\vec{v}$*.) Alors le minimum de* φ *est donné lorsque* ρ *vérifie :*

$$\rho = \frac{\vec{r}^T.\vec{v}}{\vec{v}^T.A.\vec{v}} \qquad où \qquad \vec{r} = A.\vec{x} - \vec{b} \qquad (= \vec{\nabla} f(\vec{x})). \tag{3.7}$$

ρ *est appelé le pas optimal donné pour la direction* $\vec{v}$ *choisie, et* $\vec{x}+\rho\vec{v}$ *est le point où* f *atteint son minimum sur la droite* $\vec{x} + \mathrm{Vect}\{\vec{v}\}$.

Preuve. On a :

$$\begin{aligned} f(\vec{x}-\rho\vec{v}) &= \frac{1}{2}(\vec{x}-\rho\vec{v})^T.A.(\vec{x}-\rho\vec{v}) - \vec{b}^T.(\vec{x}-\rho\vec{v}) \\ &= f(\vec{x}) + \rho(-\vec{v}^T.A.\vec{x} + \vec{b}^T.\vec{v}) + \frac{1}{2}\rho^2\vec{v}^T.A.\vec{v}, \end{aligned}$$

et le membre de droite est un polynôme de degré 2 en ρ dont le minimum est donné par (milieu des racines) $\rho = \frac{\vec{v}^T.A.\vec{x} - \vec{v}^T.\vec{b}}{\vec{v}^T.A.\vec{v}}$, i.e. (3.7). ∎

Corollaire 3.2 *Pour la méthode du gradient à pas optimal à l'étape* k *:*
avec $\vec{\nabla} f(\vec{x}_k) = A.\vec{x}_k - \vec{b} = \vec{r}_k$ *(= le résidu), on obtient :*

$$\rho_k = \frac{{\vec{r}_k}^T.\vec{r}_k}{{\vec{r}_k}^T.A.\vec{r}_k}. \tag{3.8}$$

Algorithme du gradient à pas optimal

1- $\vec{x}_0$ donné,
2- $\vec{r}_k = A.\vec{x}_k - \vec{b}$,
3- $\rho_k = \dfrac{{\vec{r}_k}^T.\vec{r}_k}{\vec{r}_k^T.A.\vec{r}_k}$,
4- $\vec{x}_{k+1} = \vec{x}_k + \rho_k\vec{r}_k$,
5- $f(\vec{x}_{k+1})$ calculé, et si $|f(\vec{x}_k) - f(\vec{x}_{k+1})| < \varepsilon$ on s'arrête sinon on recommence à l'étape 2.

3.4 Méthode du gradient à pas fixe

Proposition 3.3 *Pour* $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ *avec* A *matrice symétrique définie positive. On note* α *et* M *les plus petite et plus grande valeurs propres de* A*. Alors, si l'on choisit* ρ *tel que :*

$$0 < \rho < \frac{2\alpha}{M^2}, \tag{3.9}$$

l'algorithme de gradient à pas fixe converge. Et un choix intéressant de ρ *est :*

$$\rho = \frac{\alpha}{M^2}.$$

Preuve. On a $\alpha||\vec{v}||^2_{\mathbb{R}^n} \leq \vec{v}^T.A.\vec{v}$ et $||A.\vec{v}||_{\mathbb{R}^n} \leq M\,||\vec{v}||_{\mathbb{R}^n}$.

f étant quadratique définie positive, on a l'existence d'un minimum $\vec{x}_\infty$ et f étant dérivable, on a $\vec{\nabla} f(\vec{x}_\infty) = 0 = A.\vec{x}_\infty - \vec{b}$. D'où :

$$\begin{aligned} \vec{x}_{n+1}-\vec{x}_\infty &= \vec{x}_n - \rho\vec{\nabla} f(\vec{x}_n) - \vec{x}_\infty = (\vec{x}_n-\vec{x}_\infty) - \rho(\vec{\nabla} f(\vec{x}_n)-\vec{\nabla} f(\vec{x}_\infty)) \\ &= (\vec{x}_n-\vec{x}_\infty) - \rho A.(\vec{x}_n-\vec{x}_\infty), \end{aligned}$$

d'où :

$$\begin{aligned} ||\vec{x}_{n+1}-\vec{x}_\infty||^2_{\mathbb{R}^n} &= ||\vec{x}_n-\vec{x}_\infty||^2_{\mathbb{R}^n} - 2\rho(\vec{x}_n-\vec{x}_\infty, A.(\vec{x}_n-\vec{x}_\infty))_{\mathbb{R}^n} + \rho^2||A.(\vec{x}_n-\vec{x}_\infty)||^2_{\mathbb{R}^n} \\ &\leq (1 - 2\rho\alpha + M^2\rho^2)||\vec{x}_n-\vec{x}_\infty||^2_{\mathbb{R}^n}. \end{aligned}$$

D'où la convergence si $0 < 1 - 2\rho\alpha + M^2\rho^2 < 1$, i.e. dès que $0 < \rho < \frac{2\alpha}{M^2}$.

Et la convergence est a priori la plus rapide si $\tau = 1 - 2\rho\alpha + M^2\rho^2 = \tau(\rho)$ est le plus petit possible, i.e. quand $\tau'(\rho) = 0$, i.e. quand $\rho = \frac{\alpha}{M^2}$. ∎

Corollaire 3.4 *D'où l'algorithme du gradient à pas fixe :*
1- $\vec{x}_0$ donné, et calcul (en fait estimation) de α et M, puis de $\rho = \frac{\alpha}{M^2}$.
2- Calcul de $\vec{p}_k = \vec{\nabla} f(\vec{x}_k)$,
3- Calcul de $\vec{x}_{k+1} = \vec{x}_k - \rho\vec{p}_k$, puis de $f(\vec{x}_{k+1})$.
4- Si $|f(\vec{x}_{k+1}) - f(\vec{x}_k)| > \varepsilon$, où ε est une tolérance donnée, retour en 2-, sinon fin.

4 Méthode du gradient conjugué

4.1 Gradient conjugué

Soit $f \in C^1(\mathbb{R}^n;\mathbb{R})$ et on dispose du développement limité (1.8).

Soit A une matrice symétrique définie positive. Soit $(\cdot,\cdot)_A$ le produit scalaire associé. Et ici on utilise le produit scalaire $(\cdot,\cdot)_A$. Le théorème de représentation de Riesz, avec le produit scalaire $(\cdot,\cdot)_A$, indique qu'il existe un vecteur $\vec{\nabla}_A f(\vec{x}_0)$ t.q. :

$$\forall \vec{p} \in \mathbb{R}^n, \quad df(\vec{x}_0).\vec{p} = (\vec{\nabla}_A f(\vec{x}_0), \vec{p})_A. \tag{4.1}$$

Le vecteur $\vec{\nabla}_A f(\vec{x}_0)$ est appelé le gradient A-conjugué de f en $\vec{x}_0$, ou plus simplement le gradient conjugué de f en $\vec{x}_0$ si A est implicite.

Proposition 4.1 *Si f est C^1, alors on a :*

$$A.\vec{\nabla}_A f(\vec{x}_0) = \vec{\nabla} f(\vec{x}_0). \tag{4.2}$$

Preuve. On a $df(\vec{x}_0).\vec{p} = (\vec{\nabla}_A f(\vec{x}_0), \vec{p})_A = (\vec{\nabla} f(\vec{x}_0), \vec{p})_{\mathbb{R}^n}$, cf. (1.10) et (4.1), donc $\vec{p}^T.A.\vec{\nabla}_A f(\vec{x}_0) = \vec{p}^T.\vec{\nabla} f(\vec{x}_0)$, ce pour tout $\vec{p}$. ∎

4.2 Raisons géométriques

Reprendre ce qui a déjà été dit pour la méthode de Gauss–Seidel généralisée, qui permet le passage de la fonction $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x}$ (pour laquelle une courbe de niveau est un ellipsoïde) en la fonction déformée $g(\vec{X}) = \frac{1}{2}\vec{X}^T.I.\vec{X}$ où $\vec{X} = \sqrt{A}.\vec{x}$ (pour laquelle une courbe de niveau est sphérique).

On ne souhaitera pas passer par la fonction g : cela nécessiterait le calcul de $\sqrt{A}$. On remarque alors que $f(\vec{x}) = \frac{1}{2}(\vec{x},\vec{x})_A = \frac{1}{2}||\vec{x}||_A^2$, et donc la courbe de niveau R de f est donnée par les $\vec{x}$ t.q. $\frac{1}{2}||\vec{x}||_A^2 = R$. Autrement dit, le produit scalaire adapté au problème est le produit scalaire $(\cdot,\cdot)_A$ (pour lequel une courbe de niveau de f est sphérique).

4.3 Raisons analytiques

On veut trouver un minimum d'une fonction $f \in C^2(\mathbb{R}^n;\mathbb{R})$. Soit $\vec{x}_0$ donné. Le développement limité de f au second ordre (pour améliorer la méthode du gradient qui utilise le développement limité au premier ordre) :

$$\begin{aligned} f(\vec{x}_0+h\vec{p}) &= f(\vec{x}_0) + h\, df(\vec{x}_0).\vec{p} + \frac{h}{2} d^2 f(\vec{x}_0)(\vec{p},\vec{p}) + \ o(h^2) \\ &= f(\vec{x}_0) + h\underbrace{\Big(\vec{\nabla} f(\vec{x}_0)^T.\vec{p} + \frac{h}{2}\vec{p}^T.H(\vec{x}_0).\vec{p}\Big)}_{\overset{\text{noté}}{=}\ \psi_{\vec{x}_0}(\vec{p})} + \ o(h^2) \end{aligned} \tag{4.3}$$

où $d^2 f(\vec{x})$ est la différentielle seconde de f en $\vec{x}_0$, où $H(\vec{x}_0) = [\frac{\partial^2 f}{\partial x_i x_j}(\vec{x}_0)]$ est la matrice hessienne de f en $\vec{x}_0$ (représentant $d^2 f(\vec{x}_0)$ dans la base canonique), matrice symétrique car f est supposée C^2

(théorème de Schwarz), et où on a noté $\vec{\nabla} f(\vec{x}_0)^T$ la matrice transposée de la matrice colonne $[\vec{\nabla} f(\vec{x}_0)]$ représentant $\vec{\nabla} f(\vec{x}_0)$ dans la base canonique. On a donc, à h fixé :

$$\frac{f(\vec{x}_0+h\vec{p}) - f(\vec{x}_0)}{h} = \psi_{\vec{x}_0}(\vec{p}) + o(h), \qquad \text{où} \qquad \psi_{\vec{x}_0}(\vec{p}) = \vec{\nabla} f(\vec{x}_0)^T.\vec{p} + \frac{h}{2}\vec{p}^T.H(\vec{x}_0).\vec{p}. \tag{4.4}$$

Et $\psi_{\vec{x}_0}(\vec{p})$ donne une valeur de la pente moyenne entre $\vec{x}_0$ et $\vec{x}_0+h\vec{p}$ qui est "meilleure" que le simple gradient. Et, la matrice $H(\vec{x}_0)$ étant symétrique, on a :

$$\vec{\nabla}\psi_{\vec{x}_0}(\vec{p}) = \vec{\nabla} f(\vec{x}_0) + h\, H(\vec{x}_0).\vec{p}. \tag{4.5}$$

On veut un $\vec{p}$ t.q. la pente $\psi_{\vec{x}_0}(\vec{p})$ soit maximum en valeur absolue, donc t.q. $\vec{\nabla}\psi_{\vec{x}_0}(\vec{p}) = 0$, donc t.q. :

$$h\, H(\vec{x}_0).\vec{p} = -\vec{\nabla} f(\vec{x}_0). \tag{4.6}$$

Proposition 4.2 *Soit A une matrice symétrique définie positive. Soit $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$. Soit $\vec{x}_0$ fixé.*

Le vecteur $h\vec{p}$ qui réalise le minimum de $f(\vec{x}_0+h\vec{p}) - f(\vec{x}_0)$, vecteur de direction de descente optimale au second ordre, est (à une constante multiplicative positive près) l'opposé du gradient conjugué :

$$h\vec{p} = -\vec{\nabla}_A f(\vec{x}_0) \qquad (= A^{-1}.\vec{\nabla} f(\vec{x}_0)). \tag{4.7}$$

Preuve. Ici $H(\vec{x}_0) = A$, et (4.6) s'écrit $A.(h\vec{p}) = -\vec{\nabla} f(\vec{x}_0)$, et (4.2) donne $h\vec{p} = -\vec{\nabla}_A f(\vec{x}_0)$. Et ici le développement limité (4.3) est exact. ▟

On ne calculera pas nécessairement le gradient conjugué (nécessite la résolution de $A.\vec{\nabla}_A f(\vec{x}_0) = \vec{\nabla} f(\vec{x}_0)$ pour chaque $\vec{x}_0$), car grâce à la méthode de Gauss–Seidel généralisée il suffit de disposer d'une base A-orthogonale (calculée une fois pour toute).

4.4 Gradient conjugué = Gauss–Seidel généralisée...

La méthode de Gauss–Seidel généralisée donne :

Corollaire 4.3 *Si $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ avec A matrice symétrique définie positive, et si $(\vec{p}_j)$ est une base A-orthonormale, alors la méthode de descente à pas maximal dans les directions $\vec{p}_j$ converge en n étapes (au plus), i.e. converge après une descente successive le long des n vecteurs $\vec{p}_i$ à l'aide des formules (1.30).*

Preuve. C'est la méthode de Gauss–Seidel généralisé. ▟

4.5 ... avec construction de Gram–Schmidt à partir des résidus

Il reste à choisir une base initiale $(\vec{v}_j)$ de $\mathbb{R}^n$ à partir de laquelle on va construire une base $(\vec{p}_j)$ A-orthogonale à l'aide de Gram–Schmidt, voir § 2, la base canonique n'étant en général pas un "bon choix numérique" (rapide perte de l'A-orthogonalité par accumulation d'erreurs).

On préfère prendre $\vec{v}_j = \vec{\nabla} f(\vec{x}_{j-1}) = A\vec{x}_{j-1} - \vec{b}$ le résidu à l'étape j : c'est une direction de descente "naturelle". Et de plus on va voir que $(\vec{\nabla} f(\vec{x}_{j-1}))_{j=0,\dots,k-1}$ est une famille libre, quand $\vec{\nabla} f(\vec{x}_{k-1}) \neq \vec{0}$, pour laquelle les β_{ij} donnés dans (2.2) et (2.3) sont tous nuls sauf quand $j = i+1$. Noter que si pour un k on a $\vec{\nabla} f(\vec{x}_k) = \vec{0}$, alors on a atteint le minimum en ce point, et ce point donne la solution.

Proposition 4.4 Méthode du gradient conjugué. *Soit $\vec{x}_0 \in \mathbb{R}^n$ donné, $\vec{r}_0 = A.\vec{x}_0 - \vec{b}$ qu'on suppose non nul (sinon $\vec{x}_0$ est la solution cherchée et on s'arrête). Soit la construction par Gram–*

Schmidt de la suite A-orthogonale $(\vec{p}_j)$ à partir des résidus successifs, obtenue à l'aide de :

$$\vec{p}_1 = \vec{r}_0, \tag{4.8}$$

puis pour $k \geq 1$ tant que $\vec{r}_{k-1} \neq 0$ (sinon on s'arrête et $\vec{x}_{k-1}$ est la solution cherchée) :

$$\begin{aligned}
&\alpha_k = \frac{\vec{p}_k^T.\vec{r}_{k-1}}{\vec{p}_k^T.A.\vec{p}_k} && \text{(coeff de descente max : Gauss–Seidel généralisé)},\\
&\vec{x}_k = \vec{x}_{k-1} - \alpha_k\vec{p}_k && \text{(point le plus bas dans la direction } \vec{p}_k),\\
&\vec{r}_k = A.\vec{x}_k - \vec{b} && \text{(résidu correspondant)},\\
&\beta_{k-1,k} = \frac{\vec{r}_{k-1}^T.A.\vec{p}_k}{\vec{p}_k^T.A.\vec{p}_k} && \text{(coefficient pour la base de Gram–Schmidt)},\\
&\vec{p}_{k+1} = \vec{r}_k - \beta_{k-1,k}\vec{p}_k && \text{(prochaine direction de descente : Gram–Schmidt)},
\end{aligned} \tag{4.9}$$

et on s'arrête au premier indice k tel que $\vec{r}_k = 0$.

On a : $(\vec{r}_j)_{j=0,\ldots,k-1}$ est une suite orthogonale de $\mathbb{R}^n$:

$$(\vec{r}_k, \vec{r}_i)_{\mathbb{R}^n} = 0, \qquad \forall i \neq k. \tag{4.10}$$

On note $K_k = \text{Vect}\{\vec{r}_0, \ldots, \vec{r}_{k-1}\}$. On a $K_k = \text{Vect}\{\vec{p}_1, \ldots, \vec{p}_k\}$, et $K_k \supset \text{Vect}\{A.\vec{p}_1, \ldots, A.\vec{p}_{k-1}\}$ (et K_k est appelé espace de Krylov d'ordre k).

Et on a, pour tout $j = 0, \ldots, k-1$:

$$\begin{cases} (\vec{r}_j, \vec{p}_i)_{\mathbb{R}^n} = 0, & \forall i = 1, \ldots, j, \\ (\vec{r}_j, \vec{p}_i)_A = 0, & \forall i = 1, \ldots, j-1, \end{cases} \tag{4.11}$$

i.e. $\vec{r}_j$ est orthogonal à $\vec{p}_1, \ldots, \vec{p}_j$ et est A-orthogonal à $\vec{p}_1, \ldots, \vec{p}_{j-1}$.

Preuve. On a $\vec{x}_k = \vec{x}_{k-1} - \alpha_k\vec{p}_k$, cf (1.30) et (1.29), d'où, $A.\vec{x}_k = A.\vec{x}_{k-1} - \alpha_k A.\vec{p}_k$, d'où, pour $k \geq 1$:

$$\vec{r}_k = \vec{r}_{k-1} - \alpha_k A.\vec{p}_k \qquad \text{où} \qquad \alpha_k = \frac{\vec{p}_k^T.\vec{r}_0}{||\vec{p}_k||_A^2}. \tag{4.12}$$

Et donc $(\vec{r}_k, \vec{p}_i)_{\mathbb{R}^n} = (\vec{r}_0, \vec{p}_i)_{\mathbb{R}^n} - \sum_{j=1}^k \alpha_j(\vec{p}_i, \vec{p}_j)_A = (\vec{r}_0, \vec{p}_i) - \alpha_i||\vec{p}_i||_A^2 = 0$, ce pour $i \leq k$.

Donc $\vec{r}_k \perp \text{Vect}\{\vec{p}_1, \ldots, \vec{p}_k\}$. Et par construction des vecteurs $\vec{p}_j$ (Gram–Schmidt) on a $\text{Vect}\{\vec{p}_1, \ldots, \vec{p}_k\} = \text{Vect}\{\vec{r}_0, \ldots, \vec{r}_{k-1}\} \overset{\text{noté}}{=} K_k$. Donc $\vec{r}_k$ est othogonal à tous les $\vec{r}_j$ pour $j < k$ (pour le produit scalaire $(\cdot,\cdot)_{\mathbb{R}^n}$) : on a (4.10) et $(4.11)_1$.

Puis (4.12) indique par récurrence que $A.\vec{p}_k \in \text{Vect}\{\vec{r}_0, \ldots, \vec{r}_k\}$ et on a bien $\text{Vect}\{\vec{p}_1, \ldots, \vec{p}_k\} \supset \text{Vect}\{A.\vec{p}_1, \ldots, A.\vec{p}_{k-1}\}$. D'où $(4.11)_2$.

D'où avec (2.3) on obtient $\beta_{k-1,j} = 0$ pour tout $j > k$, d'où l'expression de $\vec{p}_{k+1}$ dans (4.9) à l'aide de (2.2) et (2.3).

De plus (4.12) donne $\vec{r}_{k-1} = \vec{r}_0 - \sum_{i=1}^{k-1} \alpha_i A.\vec{p}_i$, d'où $(\vec{r}_{k-1}, \vec{p}_k)_{\mathbb{R}^n} = (\vec{r}_0, \vec{p}_k)_{\mathbb{R}^n} - 0$, et donc $\alpha_k = \frac{\vec{p}_k^T.\vec{r}_{k-1}}{||\vec{p}_k||_A^2}$. ∎

Corollaire 4.5 *Et $\vec{x}_k$, $1 \leq k \leq m$, donné par (1.30) réalise le minimum de f sur l'espace affine $\vec{x}_0 + K_k$:*

$$f(\vec{x}_k) = \min_{\vec{x} \in K_k} f(\vec{x}).$$

Preuve. Puis $(\vec{\nabla} f(\vec{x}_k), \vec{p}_i)_{\mathbb{R}^n} = (A.\vec{x}_k - \vec{b}, \vec{p}_i) = (\vec{r}_k, \vec{p}_i) = 0$ pour tout $i \leq k$, donc f restreint à l'espace affine $\vec{x}_0 + K_k$ atteint son minimum en $\vec{x}_k$. ∎

A Annexe : méthode de gradient dans le cas non linéaire ou non symétrique

A.1 Fonction α-convexe

Définition A.1 Soit $\Omega \subset \mathbb{R}^n$, et $f \in C^1(\Omega;\mathbb{R})$. On dira que f est α-convexe ssi :

$$\exists \alpha > 0, \quad \forall \vec{x},\vec{y} \in K \quad : \quad (\vec{\nabla} f(\vec{y}) - \vec{\nabla} f(\vec{x}), \vec{y} - \vec{x})_{\mathbb{R}^n} \geq \alpha ||\vec{y} - \vec{x}||^2_{\mathbb{R}^n}. \tag{A.1}$$

On dit également que f est coercive (ou coercitive), elliptique, ou fortement convexe.

Exemple A.2 Si $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ avec A matrice $n \times n$ symétrique définie positive, alors f est $\lambda_{\min}$-convexe, où $\lambda_{\min}$ est la plus petite valeur propre de A. En effet, $\vec{\nabla} f(\vec{x}) = A.\vec{x} - \vec{b}$ donne $(\vec{\nabla} f(\vec{y}) - \vec{\nabla} f(\vec{x}), \vec{y} - \vec{x})_{\mathbb{R}^n} = (A(\vec{y} - \vec{x}), (\vec{y} - \vec{x}))_{\mathbb{R}^n}$. ▟

Exercice A.3 Montrer que l'α-convexité est équivalente à :

$$f(\vec{y}) - f(\vec{x}) \geq (\vec{\nabla} f(\vec{x}), \vec{y} - \vec{x})_{\mathbb{R}^n} + \frac{\alpha}{2} ||\vec{y} - \vec{x}||^2_{\mathbb{R}^n}, \tag{A.2}$$

i.e. que le graphe de f est localement (au voisinage de $\vec{x}$) au dessus du paraboloïde $g : \vec{y} \to g(\vec{y}) = f(\vec{x}) + (\vec{\nabla} f(\vec{x}), \vec{y} - \vec{x})_{\mathbb{R}^n} + \frac{\alpha}{2}(\vec{y} - \vec{x})^T.(\vec{y} - \vec{x})$.

Réponse. En permuttant $\vec{x}$ et $\vec{y}$ dans la formule (A.2) et en additionnant les deux formules, on obtient (A.1).

Réciproquement, si on suppose (A.1), on pose $\varphi(h) = f(\vec{x} + h(\vec{y} - \vec{x}))$, et on a :

$$\varphi'(h) = (\vec{\nabla} f(\vec{x} + h(\vec{y} - \vec{x})), \vec{y} - \vec{x})_{\mathbb{R}^n},$$

et donc avec (A.1) et $h > 0$:

$$\varphi'(h) - \varphi'(0) = (\vec{\nabla} f(\vec{x} + h(\vec{y} - \vec{x})) - \vec{\nabla} f(\vec{x}), \frac{h(\vec{y} - \vec{x})}{h} \geq \frac{1}{h}\alpha(h||\vec{y} - \vec{x}||)^2 = \alpha h ||\vec{y} - \vec{x}||^2.$$

D'où par intégration en h sur $[0,1]$:

$$\varphi(1) - \varphi(0) - \varphi'(0) \geq \alpha \frac{1}{2} ||\vec{y} - \vec{x}||^2,$$

i.e. (A.2).

Soit $g(\vec{y}) = f(\vec{x}) + (\vec{\nabla} f(\vec{x}), \vec{y} - \vec{x})_{\mathbb{R}^n} + \frac{\alpha}{2}(\vec{y} - \vec{x})^T.(\vec{y} - \vec{x})$ quadratique. Localement le développement limité de f indique que $f(\vec{y}) \geq g(\vec{y})$. ▟

Exercice A.4 Montrer que pour $f \in C^2(\mathbb{R}^n, \mathbb{R})$ et $d^2 f(\vec{x})$ représentée par la matrice hessienne $H_f(\vec{x}) = [\frac{\partial^2 f}{\partial x_i \partial x_j}(\vec{x})]$, alors l'$\alpha$-coercivité est équivalente à : pour tout $\vec{x}, \vec{v} \in \mathbb{R}^n$:

$$(H_f(\vec{x}).\vec{v}, \vec{v})_{\mathbb{R}^n} \geq \alpha ||\vec{v}||^2_{\mathbb{R}^n}, \tag{A.3}$$

i.e. la matrice $H_f(\vec{x})$ définit un produit scalaire (symétrique définie positive).

Réponse. On a $\vec{\nabla} f(\vec{x} + \vec{v}) - \vec{\nabla} f(\vec{x}) = H_f(\vec{x}).\vec{v} + o(\vec{v})$, et (A.1) implique (A.3).

Réciproquement, on a la formule de Taylor $f(\vec{x} + \vec{v}) = f(\vec{x}) + (\vec{\nabla} f(\vec{x}), \vec{v})_{\mathbb{R}^n} + \frac{1}{2}(H_f(\vec{x} + \theta\vec{v}).\vec{v}, \vec{v})_{\mathbb{R}^n}$ pour un $\theta \in [0,1]$. D'où (A.3) implique (A.1). ▟

A.2 Fonction M-lipschitzienne

Définition A.5 Pour $f \in C^1(\Omega;\mathbb{R})$, on dit que f est M-lipschitzienne sur Ω, ssi :

$$\exists M \in \mathbb{R}, \quad \forall \vec{x},\vec{y} \in K \quad : \quad ||\vec{\nabla} f(\vec{y}) - \vec{\nabla} f(\vec{x})||_{\mathbb{R}^n} \leq M ||\vec{y} - \vec{x}||_{\mathbb{R}^n}. \tag{A.4}$$

Exemple A.6 Si $f(\vec{x}) = \frac{1}{2}\vec{x}^T.A.\vec{x} - \vec{b}^T.\vec{x}$ avec A matrice $n \times n$ symétrique, alors f est M-Lipschitzienne avec $M = ||A||$ la norme de A. En effet, $\vec{\nabla} f(\vec{x}) = A.\vec{x} - \vec{b}$ donne $||\vec{\nabla} f(\vec{y}) - \vec{\nabla} f(\vec{x})||_{\mathbb{R}^n} = ||A(\vec{y} - \vec{x})||_{\mathbb{R}^n} \leq ||A||\,||\vec{y} - \vec{x}||_{\mathbb{R}^n}$ par définition de la norme matricielle. ▟

A.3 Méthode du gradient à pas fixe

La méthode du gradient à pas fixe (peu coûteuse) est : choisir un $\rho > 0$ donné, poser $\rho_n = \rho$ pour tout n, et calculer $\vec{x}_{n+1}$ à l'aide de (3.4).

Proposition A.7 *Soit $f \in C^1(\mathbb{R}^n;\mathbb{R})$ telle que f est α-convexe et M-Lipschitzienne.*
Alors, si l'on choisit ρ tel que :

$$0 < \rho < \frac{2\alpha}{M^2}, \tag{A.5}$$

l'algorithme de gradient à pas fixe converge. Un choix intéressant de ρ est :

$$\rho = \frac{\alpha}{M^2}.$$

Preuve. L'hypothèse de coercivité de f sur $\mathbb{R}^n$ assure l'existence d'un minimum $\vec{x}_\infty$ et f étant dérivable, on a $\vec{\nabla} f(\vec{x}_\infty) = 0$. D'où :

$$\vec{x}_{n+1} - \vec{x}_\infty = \vec{x}_n - \vec{x}_\infty - \rho \vec{\nabla} f(\vec{x}_n) = \vec{x}_n - \vec{x}_\infty - \rho(\vec{\nabla} f(\vec{x}_n) - \vec{\nabla} f(\vec{x}_\infty)),$$

d'où :

$$\begin{aligned} ||\vec{x}_{n+1} - \vec{x}_\infty||^2_{\mathbb{R}^n} &= ||\vec{x}_n - \vec{x}_\infty||^2_{\mathbb{R}^n} - 2\rho(\vec{x}_n - \vec{x}_\infty, \vec{\nabla} f(\vec{x}_n) - \vec{\nabla} f(\vec{x}_\infty))_{\mathbb{R}^n} + \rho^2 ||\vec{\nabla} f(\vec{x}_n) - \vec{\nabla} f(\vec{x}_\infty)||^2_{\mathbb{R}^n} \\ &\leq (1 - 2\rho\alpha + M^2\rho^2) ||\vec{x}_n - \vec{x}_\infty||^2_{\mathbb{R}^n}. \end{aligned}$$

D'où la convergence si $0 < 1 - 2\rho\alpha + M^2\rho^2 < 1$, i.e. dès que $0 < \rho < \frac{2\alpha}{M^2}$.

Et la convergence est a priori la plus rapide si $\tau = 1 - 2\rho\alpha + M^2\rho^2 = \tau(\rho)$ est le plus petit possible, i.e. quand $\tau'(\rho) = 0$, i.e. quand $\rho = \frac{\alpha}{M^2}$. ∎

A.4 Méthode du gradient à pas optimal

La méthode du gradient à pas optimal (plus efficace mais plus coûteuse que la précédente) est : pour $\vec{x}_n$ donné, trouver ρ_n tel que :

$$f(\vec{x}_n - \rho_n \vec{\nabla} f(\vec{x}_n)) = \inf_{\rho \in \mathbb{R}} f(\vec{x}_n - \rho \vec{\nabla} f(\vec{x}_n)), \tag{A.6}$$

puis poser $\vec{x}_{n+1} = \vec{x}_n - \rho_n \vec{\nabla} f(\vec{x}_n)$, comme en (3.4). On descend ainsi au maximum dans chaque direction de descente.

On pose :

$$\varphi(\rho) = f(\vec{x}_n - \rho \vec{\nabla} f(\vec{x}_n)), \tag{A.7}$$

et calculer ρ_n revient à minimiser φ.

Proposition A.8 *Soit $f \in C^1(\mathbb{R}^n;\mathbb{R})$ telle que f est α-convexe et M-Lipschitzienne. La méthode du gradient à pas optimal converge vers la solution de (3.1), et de plus deux directions successives de descente sont orthogonales :*

$$\vec{\nabla} f(\vec{x}_n) \perp \vec{\nabla} f(\vec{x}_{n+1}). \tag{A.8}$$

Preuve. La suite $(f(\vec{x}_n))$ est décroissante (par construction), et minorée par $f(\vec{x}_\infty)$.

On pose $\varphi(\rho) = f(\vec{x}_n - \rho \vec{\nabla} f(\vec{x}_n))$ qui est C^1. D'où $\varphi'(\rho) = -(\vec{\nabla} f(\vec{x}_n - \rho \vec{\nabla} f(\vec{x}_n)), \vec{\nabla} f(\vec{x}_n))_{\mathbb{R}^n}$, et on cherche ρ tel que $\varphi'(\rho) = 0$. On posera alors $\vec{x}_{n+1} = \vec{x}_n - \rho \vec{\nabla} f(\vec{x}_n)$, et on a obtenu $(\vec{\nabla} f(\vec{x}_{n+1}), \vec{\nabla} f(\vec{x}_n))_{\mathbb{R}^n} = 0$ i.e. (A.8).

On en déduit que, avec Cauchy–Schwarz et f lipschitzienne :

$$||\vec{\nabla} f(\vec{x}_n)||^2_{\mathbb{R}^n} = (\vec{\nabla} f(\vec{x}_n), \vec{\nabla} f(\vec{x}_n) - \vec{\nabla} f(\vec{x}_{n+1}))_{\mathbb{R}^n} \leq ||\vec{\nabla} f(\vec{x}_n)||_{\mathbb{R}^n} M \, ||\vec{x}_n - \vec{x}_{n+1}||_{\mathbb{R}^n}, \tag{A.9}$$

et donc :

$$||\vec{\nabla} f(\vec{x}_n)||_{\mathbb{R}^n} \leq M \, ||\vec{x}_n - \vec{x}_{n+1}||_{\mathbb{R}^n}.$$

De même, on déduit de (A.8) que $(\vec{\nabla} f(\vec{x}_{n+1}), \vec{x}_n - \vec{x}_{n+1})_{\mathbb{R}^n} = 0$ (car $\vec{x}_n - \vec{x}_{n+1} \,//\, \vec{\nabla} f(\vec{x}_n)$ par

construction de $\vec{x}_{n+1}$), d'où avec (A.2) :

$$f(\vec{x}_n) - f(\vec{x}_{n+1}) \geq (\vec{\nabla} f(\vec{x}_{n+1}), x_n - x_{n+1})_{\mathbb{R}^n} + \frac{\alpha}{2}||x_n - x_{n+1}||^2_{\mathbb{R}^n} = \frac{\alpha}{2}||x_n - x_{n+1}||^2_{\mathbb{R}^n}.$$

On en déduit que $\lim ||x_n - x_{n+1}||^2_{\mathbb{R}^n} = 0$. Et donc avec (A.9) que $||\vec{\nabla} f(\vec{x}_n)||^2_{\mathbb{R}^n} \to 0$. Enfin, l'$\alpha$-coercivité donne :

$$\begin{aligned}\alpha||\vec{x}_n - \vec{x}_\infty||^2_{\mathbb{R}^n} &\leq (\vec{\nabla} f(\vec{x}_n) - \vec{\nabla} f(\vec{x}_\infty), \vec{x}_n - \vec{x}_\infty)_{\mathbb{R}^n} = (\vec{\nabla} f(\vec{x}_n), \vec{x}_n - \vec{x}_\infty)_{\mathbb{R}^n} \\ &\leq ||\vec{\nabla} f(\vec{x}_n)||_{\mathbb{R}^n} ||x_n - x_{n+1}||_{\mathbb{R}^n},\end{aligned}$$

d'où $(\vec{x}_n - \vec{x}_\infty) \to 0$, et la suite converge. ∎

B Annexe : Rayon spectral, convergence

Définition B.1 On appelle rayon spectrale d'une matrice A de taille $n \times n$ la plus grande des valeurs propres en valeur absolue, i.e. le réel noté :

$$\rho(A) = \max\{|\lambda_i| : i = 1, \ldots, n,\ \lambda_i \text{ valeur propre de } A\}. \tag{B.1}$$

Proposition B.2 *Soit A une matrice $n \times n$ telle que $\rho(A) < 1$. Alors, pour tout $\vec{x} \in \mathbb{R}^n$, la suite $(A^k.\vec{x})_{k\in\mathbb{N}}$ est convergente vers 0.*

Preuve. On suppose $A \neq 0$ (sinon c'est trivial). Soit $A = P.T.P^{-1}$ une trigonalisation de A avec $P^{-1} = P^T$. Comme A et T ont mêmes valeurs propres, on a $\rho(A) = \rho(T)$. Et si $T^k.\vec{x} \longrightarrow_{k\to\infty} \vec{0}$, alors $A^k.\vec{x} \longrightarrow_{k\to\infty} \vec{0}$ En effet, $A^k = P.T^k.P^{-1}$ donne $(A^k.\vec{x}, \vec{y})_{\mathbb{R}^n} = (T^k.(P^{-1}.\vec{x}), (P^{-1}.\vec{y}))_{\mathbb{R}^n}$ puisque $P^T = P^{-1}$. Et comme P et une bijection, $T^k.(P^{-1}.\vec{x}) \longrightarrow_{k\to\infty} \vec{0}$. Donc pour tout $\vec{x}, \vec{y}$, on a $(A^k.\vec{x}, \vec{y})_{\mathbb{R}^n} \longrightarrow_{k\to\infty} 0$. D'où $A^k.\vec{x} \longrightarrow_{k\to\infty} \vec{0}$.

Montrons donc que $T^k.\vec{x} \longrightarrow_{k\to\infty} \vec{0}$ quand $\rho(T) < 1$. Posons $T = D + U$ avec D la matrice diagonale de T, i.e. la matrice diagonale des valeurs propres, et avec U matrice triangulaire supérieure stricte. En particulier, U est nilpotente d'ordre n, i.e. $U^n = 0$ est la matrice nulle. Et donc, pour $k \geq n$:

$$T^k = (D+U)^k = D^k + \binom{k}{1} D^{k-1}U + \ldots + \binom{k}{n-1} D^{k-(n-1)}U^{n-1}.$$

Soit $c = \sup_{i=1,\ldots,n-1} ||U^i||$. Comme $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$, on a $D^i = \mathrm{diag}(\lambda_1^i, \ldots, \lambda_n^i)$, et $||D^i|| = \rho(D)^i \leq \rho(D)^k$ pour tout $i \leq k$, car $\rho(D) = \rho(T) = \rho(A) < 1$. D'où, toujours avec $k > n$:

$$||T^k|| \leq c(1 + \binom{k}{1} + \ldots + \binom{k}{n-1})\rho(D)^k.$$

Or $(1 + \binom{k}{1} + \ldots + \binom{k}{n-1}) = P_n(k)$ où P_n est un polynôme de degré n. D'où :

$$||T^k|| = cP_n(k)\rho(D)^k = cP_n(k)e^{-\alpha k},$$

où $\alpha = -\log(\rho(D)) > 0$ car $0 < \rho(D) < 1$ (si $\rho(D) = 0$ c'est trivial car alors T est nilpotente), qui tend vers 0 quand $k \to \infty$. ∎

Remarque B.3 La démonstration précédente est simplifiable dans le cas A est diagonalisable : dans ce cas $A = PDP^{-1}$ avec $D = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ matrice diagonale des valeurs propres, et $A^k = PD^kP^{-1}$ avec $D^k = \mathrm{diag}(\lambda_1^k, ..., \lambda_n^k)$. Et comme $|\lambda_i| < 1$ pour tout $i = 1, ..., n$, on a $D^k.\vec{x} \longrightarrow_{k\to 0} \infty$, on en déduit que $A^k.\vec{x} \to 0$. ∎

Exercice B.4 Montrer par récurrence sur la taille n de la matrice que $T^k.\vec{x} \to 0$ quand $\rho(T) < 1$ et quand $\rho(T) = |\beta|$ où β est valeur propre de multiplicité 1.

Réponse. C'est immédiat si $n = 1$. Supposons que ce soit vrai pour une matrice T_n triangulaire supérieure $n \times n$. Soit T_{n+1} une matrice triangulaire supérieure $(n+1) \times (n+1)$. On a T_{n+1} de la forme :

$$T_{n+1} = \begin{pmatrix} T_n & \vec{b} \\ 0 & \beta \end{pmatrix},$$

où $\vec{b} \in \mathbb{R}^n$ et le 0 représente le vecteur ligne nul de $\mathbb{R}^n$, et où on a appelé β la plus grande valeur propre (en valeur absolue) de T_{n+1}.

D'où $T_{n+1}^2 = \begin{pmatrix} T_n^2 & (T_n + \beta I).\vec{b} \\ 0 & \beta^2 \end{pmatrix}$, D'où $T_{n+1}^3 = \begin{pmatrix} T_n^3 & (T_n^2 + T_n\beta + \beta^2 I).\vec{b} \\ 0 & \beta^3 \end{pmatrix}$, et par une récurrence immédiate :

$$T_{n+1}^{k+1} = \begin{pmatrix} T_n^{k+1} & (T_n^k + T_n^{k-1}\beta + \ldots + T_n\beta^{k-1} + \beta^k I).\vec{b} \\ & \beta^{k+1} \end{pmatrix} = \begin{pmatrix} T_n^{k+1} & (\sum_{i=0}^k T_n^{k-i}\beta^i).\vec{b} \\ 0 & \beta^{k+1} \end{pmatrix}.$$

Soit $\vec{x} \in \mathbb{R}^{n+1}$, $\vec{x} = \begin{pmatrix} \vec{y} \\ z \end{pmatrix}$ où $\vec{y} \in \mathbb{R}^n$. On obtient :

$$T_{n+1}^{k+1}.\vec{x} = \begin{pmatrix} T_n^{k+1}.\vec{y} + z(\sum_{i=0}^k T_n^i\beta^{k-i}).\vec{b} \\ \beta^{k+1}z \end{pmatrix},$$

et en particulier, avec "$(a+b)^2 \leq 2a^2 + 2b^2$" :

$$||T_{n+1}^{k+1}.\vec{x}||^2 \leq 2||T_n^{k+1}.\vec{y}||^2 + 2z^2||(\sum_{i=0}^k T_n^i\beta^{k-i}).\vec{b}||^2 + z^2\beta^{2(k+1)}.$$

Par hypothèse, on a $|\beta| < 1$ et $T_n^k \longrightarrow_{k\to\infty} 0$. Les premier et troisième termes du membre de droite tendent donc vers 0. Et pour le second, on a $\frac{\rho(T_n)}{\beta} < 1$ d'où :

$$\sum_{i=0}^k T_n^i\beta^{k-i} = \beta^k \sum_{i=0}^k \left(\frac{T_n}{\beta}\right)^i = \beta^k (I - \frac{T_n}{\beta})^{-1}(I - T_n^{k+1}).$$

En effet, la matrice $(I - \frac{T_n}{\beta})$ est inversible car triangulaire supérieure de diagonale non nulle, toutes les valeurs propres de $\frac{T_n}{\beta}$ étant < 1, et on a $(\sum_{i=0}^k \left(\frac{T_n}{\beta}\right)^i)(I - \frac{T_n}{\beta}) = (I - T_n^{k+1})$ (calcul immédiat). Et par hypothèse de récurrence, sachant $\beta^k \to 0$, on en déduit que le second terme tend vers 0 avec k. ∎

Références

[1] Ciarlet P.G. : *Introduction à l'analyse numérique matricielle et à l'optimisation.* Masson, 1992.

[2] Golub G., Van Loan C. : *Matrix Computations.* John Hopkins University Press, 1996.

[3] Lascaux P., Théodor R. : *Analyse numérique matricielle appliquée à l'art de l'ingénieur.* Masson, 1998.

[4] Shewchuk J. : An introduction to the Conjugate Gradient Method without Agonizing Pain. http ://www-2.cs.cmu.edu/ jrs/jrspapers.html.

[5] Strang G. : *Linear Algebra and its Applications.* Harcourt Brace (1988).