CHAPITRE 8

Méthodes de gradient conjugué

8.0. Introduction

A chaque itération d'une méthode de descente, on détermine un vecteur p_k et un scalaire α_k, ce qui permet de calculer x_{k+1} par :

$$x_{k+1} = x_k + \alpha_k p_k \quad , \tag{1}$$

avec l'objectif de minimiser une fonctionnelle.

Avant d'introduire l'importante méthode du gradient conjugué pour résoudre un système linéaire dont la matrice est symétrique et définie positive, il est nécessaire d'aborder sommairement les méthodes du gradient.

Comme ces méthodes convergent d'autant plus vite que le conditionnement de la matrice est faible, nous étudierons certaines techniques de préconditionnement pour accélérer la rapidité de la convergence. On obtient alors des méthodes dont la convergence est souvent beaucoup plus rapide que celles du type relaxation.

Pour terminer ce chapitre, nous étudierons sommairement quelques extensions des méthodes de gradient conjugué à des systèmes dont la matrice n'est plus nécessairement symétrique et définie positive.

8.1. Principe des méthodes de descente

8.1.1 Choix de la fonctionnelle à minimiser

Soit A une matrice symétrique et définie positive. On note l'équivalence entre la solution $\bar{x}$ de $Ax = b$ et le vecteur qui réalise le minimum de la fonctionnelle J définie par :

(2) $$\boxed{J(x) = (Ax|x) - 2\,(b|x)} \quad .$$

En effet, la fonctionnelle J est quadratique et définie positive.
Son minimum unique est obtenu en annulant le gradient g de J.

Or $g(x) = 2\,(Ax - b) = -\,2\,r(x)$ (cf. chapitre 1 ; 4, 5)

où $r(x) = b - Ax = A\bar{x} - Ax$ est le résidu du système $Ax = b$.

Il est encore équivalent de minimiser J ou de minimiser E défini par :

(3) $$\boxed{E(x) = (A(x - \bar{x}) \mid x - \bar{x}) = (A\,e(x) \mid e(x))} \quad \text{où } e(x) = x - \bar{x} \,,$$

car
$$\begin{aligned} E(x) &= (Ax \mid x) - 2\,(x \mid A\bar{x}) + (A\bar{x} \mid \bar{x}) \\ &= J(x) + (A\bar{x} \mid \bar{x}) \quad . \end{aligned}$$

Comme $(A\bar{x} \mid \bar{x})$ est une constante, E et J atteignent leur minimum au même point. A étant symétrique et définie positive $(Ax \mid y]$ est un produit scalaire et $E(x) = \| e(x) \|_A^2$ où $\|e\|_A = (Ae \mid e)^{1/2}$ est la norme associée à ce produit scalaire.

On remarque également que $E(x)$ peut s'exprimer en fonction du résidu $r(x) = A\bar{x} - Ax$ par

(4) $$\boxed{E(x) = \left(r(x) \mid A^{-1}\, r(x)\right)} \quad .$$

Pour minimiser la fonctionnelle E, les méthodes de "descente" sont construites en choisissant à la $k^{ième}$ itération une direction de descente $p_k \neq 0$ et un scalaire α_k de manière que $E(x_{k+1}) < E(x_k)$.

8.1.2 Choix optimal de α_k dans une direction fixée p_k

Supposons que la direction p_k soit fixée. *Le choix local optimal* de α_k consiste, à chaque itération, à choisir α_k de façon à minimiser $E(x_{k+1})$ dans la direction p_k .

Le α_k optimal est donc tel que :

$$E(x_k + \alpha_k\, p_k) = \min_{\alpha \in \mathbf{R}} E(x_k + \alpha\, p_k) \quad .$$

Or,
$$\begin{aligned} E(x_k + \alpha\, p_k) &= (A(x_k + \alpha\, p_k - \bar{x}) \mid x_k + \alpha\, p_k - \bar{x}) \\ &= E(x_k) - 2\,\alpha\,(r_k \mid p_k) + \alpha^2\,(A\,p_k \mid p_k) \quad , \end{aligned}$$

est un trinôme du second degré en α dont le terme de plus haut degré $(A\,p_k \mid p_k)$ est strictement positif quel que soit le choix de $p_k \neq 0$, car A est une matrice définie positive.

Son minimum est atteint pour :

(5) $$\boxed{\alpha_k = \frac{(r_k \mid p_k)}{(A\,p_k \mid p_k)}} \quad .$$

Alors :

(6) $$\boxed{x_{k+1} = x_k + \frac{(r_k \mid p_k)}{(A\,p_k \mid p_k)}\,p_k} \quad .$$

Proposition 1

Quel que soit le choix de $p_k \neq 0$ pour α_k optimal, on a les deux relations suivantes :

(7) $\forall\, k \geq 0$ $$\boxed{r_{k+1} = r_k - \alpha_k\, A\,p_k} \quad ,$$

(8) $$\boxed{(p_k \mid r_{k+1}) = 0} \quad .$$

Preuve :

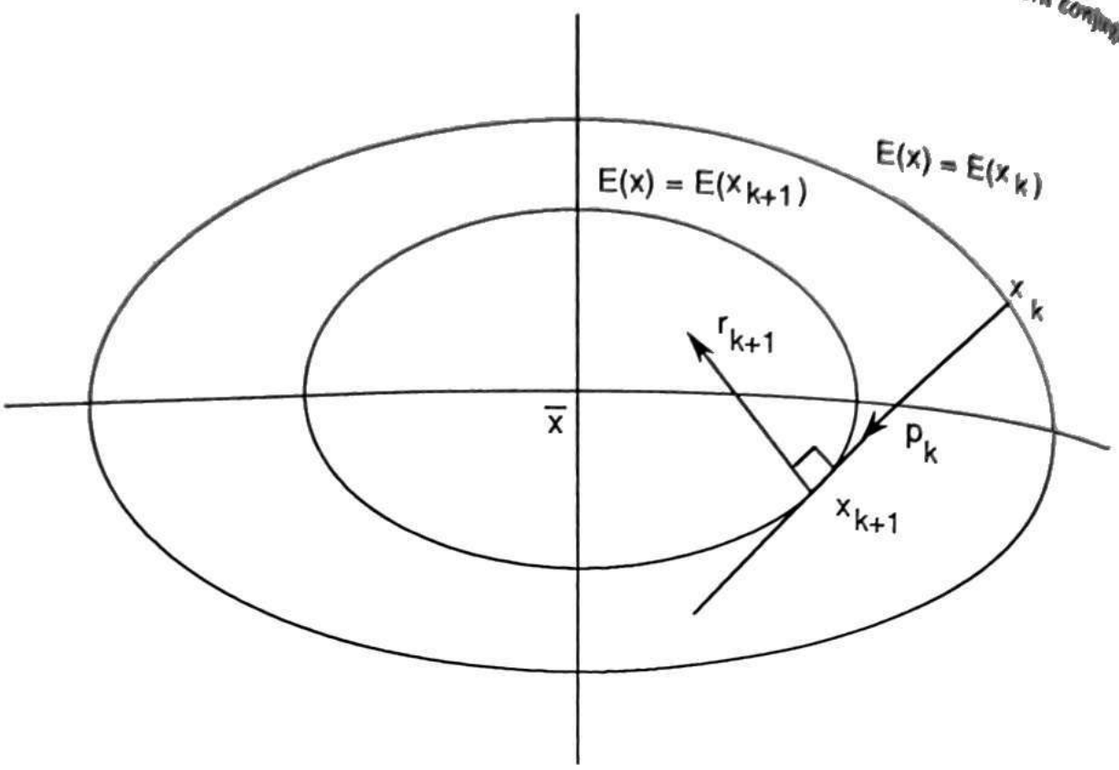
$$r_{k+1} = b - A\,x_{k+1} = b - A\,(x_k + \alpha_k\,p_k) = r_k - \alpha_k\,A\,p_k \quad .$$

$$(p_k \mid r_{k+1}) = (p_k \mid r_k - \alpha_k\,A\,p_k) = (p_k \mid r_k) - \frac{(r_k \mid p_k)}{(A\,p_k \mid p_k)}\,(p_k \mid A\,p_k) = 0 \quad . \qquad \blacksquare$$

Donnons une interprétation géométrique dans $\mathbf{R}^2$ des méthodes de descente.

$E(x)$ = constante positive est l'équation d'une ellipse.

Pour les différentes valeurs de (x_k), on obtient une famille d'ellipses $E(x) = E(x_k)$ concentriques autour du minimum $\bar{x}$ de la fonctionnelle, et qui représentent les courbes de niveau :



Le vecteur p_k est tangent à l'ellipse $E(x) = E(x_{k+1})$. Comme r_{k+1} est orthogonal à p_k, r_{k+1} est orthogonal à la tangente de la courbe de niveau.

Reportons la valeur de α_k dans $E(x_{k+1})$, on obtient :

$$E(x_{k+1}) = E(x_k) - \frac{(r_k \mid p_k)^2}{(A\, p_k \mid p_k)} ,$$

$$E(x_{k+1}) = E(x_k)\left[\, 1 - \frac{1}{E(x_k)} \, \frac{(r_k \mid p_k)^2}{(A\, p_k \mid p_k)}\right] ,$$

ou encore, puisque $E(x_k) = (r_k \mid A^{-1} r_k)$:

(9) $$\boxed{E(x_{k+1}) = E(x_k)(1 - \gamma_k) \quad \text{où} \quad \gamma_k = \frac{(r_k \mid p_k)^2}{(A\, p_k \mid p_k)(A^{-1} r_k \mid r_k)}} .$$

Sauf pour $p_k = 0$ (cas que l'on élimine) ou $r_k = 0$ (auquel cas, x_k est la solution cherchée), le nombre

$$\gamma_k = \frac{(r_k \mid p_k)^2}{(A\, p_k \mid p_k)(A^{-1} r_k \mid r_k)}$$

est toujours défini et positif car A étant une matrice symétrique et définie positive, on a $(A\, p_k \mid p_k) > 0$ et $(A^{-1} r_k \mid r_k) > 0$.

Lemme 2

Quel que soit le choix de $p_k \neq 0$ pour α_k optimal local, on a la relation suivante valable pour $k \geq 0$:

$$\gamma_k = \frac{(r_k \mid p_k)^2}{(A\,p_k \mid p_k)(A^{-1} r_k \mid r_k)} \geq \frac{1}{K(A)} \left(\frac{r_k}{\|r_k\|} \,\middle|\, \frac{p_k}{\|p_k\|} \right)^2, \tag{10}$$

où $K(A)$ est le nombre conditionnement de la matrice A.

Preuve :

On a $(A\,p_k \mid p_k) \leq \lambda_1 \|p_k\|^2$ où λ_1 est la plus grande valeur propre de A.

$$(A^{-1} r_k \mid r_k) \leq \frac{1}{\lambda_N} \|r_k\|^2 \quad \text{où } \lambda_N \text{ est la plus petite valeur propre de } A.$$

Donc :
$$\frac{(A\,p_k \mid p_k)(A^{-1} r_k \mid r_k)}{\|p_k\|^2 \;\; \|r_k\|^2} \leq \frac{\lambda_1}{\lambda_N} = K(A) \; .$$

D'où :
$$\gamma_k \geq \frac{1}{K(A)} \frac{(r_k \mid p_k)^2}{\|r_k\|^2 \|p_k\|^2} \; . \qquad \blacksquare$$

Ce lemme va nous permettre de choisir des directions de descente.

Théorème 3

Pour α_k optimal local, toute direction p_k qui vérifie, $\forall\ k \geq 0$:

$$\left(\frac{r_k}{\|r_k\|} \,\middle|\, \frac{p_k}{\|p_k\|} \right)^2 \geq \mu > 0 \qquad \text{où } \mu \text{ est indépendant} \tag{11}$$

de k, implique que la suite (x_k) converge vers la solution $\bar{x}$ qui minimise $E(x)$.

Preuve :

D'après (9) (10) et (11), on a :

$$E(x_{k+1}) \leq E(x_k)\left(1 - \frac{\mu}{K(A)}\right) .$$

Donc $$E(x_k) \leq \left(1 - \frac{\mu}{K(A)}\right)^k E(x_0) .$$

D'après (11) et l'inégalité de Schwarz, on a $0 < \mu \leq 1$.

Comme $K(A) \geq 1$, on a $0 \leq 1 - \frac{\mu}{K(A)} < 1$.

Donc : $$\lim_{k \to +\infty} E(x_k) = 0 .$$

Or : $$E(x_k) \geq \lambda_N \|x_k - \bar{x}\|^2 \quad \text{avec } \lambda_N > 0 ,$$

donc : $$\lim_{k \to +\infty} \|x_k - \bar{x}\|^2 = 0 .$$

Dans le cadre de α_k optimal local, ce théorème est une condition suffisante de convergence qui signifie que pour tout k , p_k doit être non orthogonal à r_k. Il en résulte un premier choix évident $p_k = r_k$, car alors :

$$\mu = \left(\frac{r_k}{\|r_k\|} \,\middle|\, \frac{p_k}{\|p_k\|}\right)^2 = 1 .$$

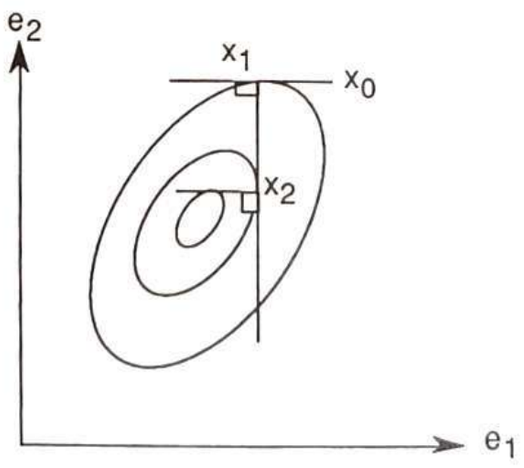
C'est ce que l'on va étudier au paragraphe suivant.

Remarque 4

Si l'on choisit comme direction de descente les vecteurs unitaires e_i des axes de coordonnées de l'espace à N dimensions dans l'ordre naturel e_1 , e_2 , ... , e_N puis on recommence cycliquement, on obtient la méthode de Gauss-Seidel.

En effet $$x_{k+1} = x_k + \alpha_k e_i \quad \text{où} \quad \alpha_k = \frac{(r_k \mid e_i)}{(Ae_i \mid e_i)} = \frac{(b - A x_k \mid e_i)}{(a_{ii})} .$$

On sait alors que lorsque la matrice A est symétrique et définie positive la méthode est convergente.



8.2. Les méthodes du gradient

Dans ce paragraphe, comme le suggère le résultat précédent, nous allons choisir le gradient (ou ce qui revient au même, le résidu) comme direction de descente, d'abord dans le cadre de α_k optimal local, mais aussi pour α_k constant.

8.2.1 La méthode du gradient à paramètre optimal

C'est la méthode où $x_{k+1} = x_k + \alpha_k r_k$,

avec

$$\boxed{\alpha_k = \frac{(r_k \mid p_k)}{(A\, p_k \mid p_k)} = \frac{\|r_k\|^2}{(A\, r_k \mid r_k)}} \quad . \tag{12}$$

Alors : $$E(x_{k+1}) = E(x_k)\left(1 - \frac{\|r_k\|^4}{(A\, r_k \mid r_k)(A^{-1} r_k \mid r_k)}\right) .$$

Grâce à l'inégalité de Kantorovitz (cf. chapitre 1 ; 4,3), on a :

$$\frac{\|r_k\|^4}{(A\, r_k \mid r_k)(A^{-1} r_k \mid r_k)} \geq \frac{4\,\lambda_1 \lambda_N}{(\lambda_1 + \lambda_N)^2} = \frac{4\,\frac{\lambda_1}{\lambda_N}}{\left(\frac{\lambda_1}{\lambda_N} + 1\right)^2} = \frac{4\,K(A)}{(K(A)+1)^2} .$$

Alors :

$$E(x_{k+1}) \leq E(x_k)\left(1 - \frac{4\,K(A)}{(K(A)+1)^2}\right) = E(x_k)\left(\frac{K(A)-1}{K(A)+1}\right)^2 .$$

Donc :

$$\boxed{E(x_k) \leq E(x_0)\left(\frac{K(A)-1}{K(A)+1}\right)^{2k}} \quad . \tag{13}$$

Comme $E(x_k) \geq \lambda_N \|x_k - \overline{x}\|^2$, on a :

$$\|x_k - \overline{x}\| \leq \beta \left(\frac{K(A)-1}{K(A)+1}\right)^k \quad \text{où} \quad \beta = \left(\frac{E(x_0)}{\lambda_N}\right)^{1/2} . \tag{14}$$

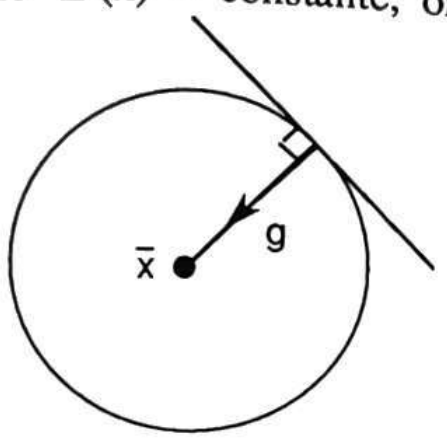
On a donc le théorème suivant :

Théorème 5

La méthode du gradient à paramètre local optimal est convergente. Sa rapidité de convergence dépend de : $\dfrac{K(A)-1}{K(A)+1}$.

Remarque 6

Plus $K(A)$ est proche de 1, plus la méthode convergera vite. Lorsque $K(A)=1$, toutes les valeurs propres de A sont égales. On a $A=\lambda I$ et $E(x)=\lambda\|x-\bar{x}\|^2$. Lorsque $E(x)$ = constante, on obtient l'équation d'une sphère.



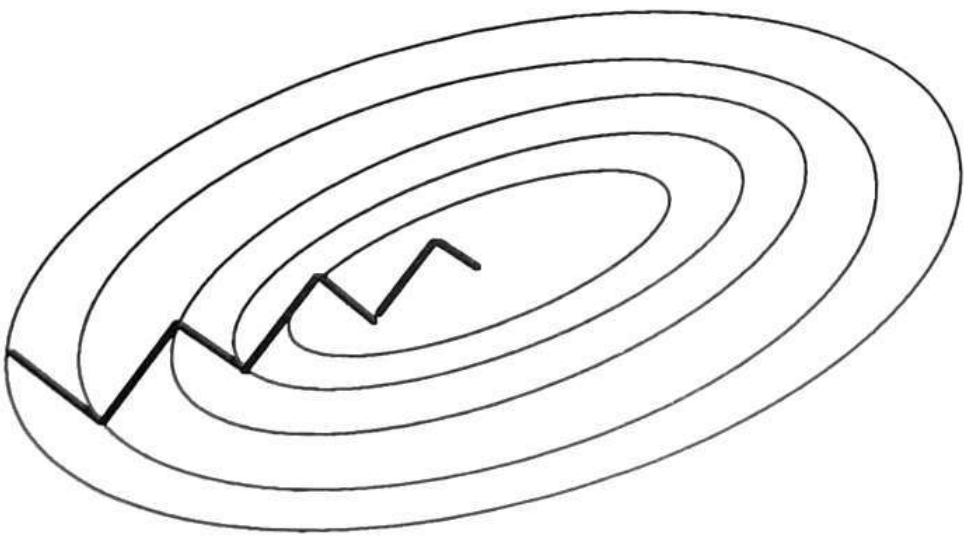
De n'importe quel point de la sphère, le gradient pointe vers le centre. On converge en une itération.

Par contre, lorsque $K(A)$ est grand, les valeurs propres extrêmes sont très différentes : l'ellipsoïde est alors très aplati.

La convergence est lente, pour que $\dfrac{E(x_k)}{E(x_0)}=\varepsilon$, il suffit que :

$$\left(\frac{K(A)-1}{K(A)+1}\right)^{2k}\leq\varepsilon\,,\quad\text{soit encore que :}\quad k\sim\frac{K(A)}{4}\ln\frac{1}{\varepsilon}\,.$$

Le nombre d'itérations est proportionnel à $K(A)$.



8.2.2 Un exemple

Considérons le problème simple suivant : résoudre le système :

$$\begin{cases} \frac{1}{2}x = 0 \\ \frac{c}{2}y = 0 \end{cases} \quad \text{qui s'écrit} \quad \begin{bmatrix} \frac{1}{2} & 0 \\ 0 & \frac{c}{2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

où c est une constante donnée supérieure à 1 .

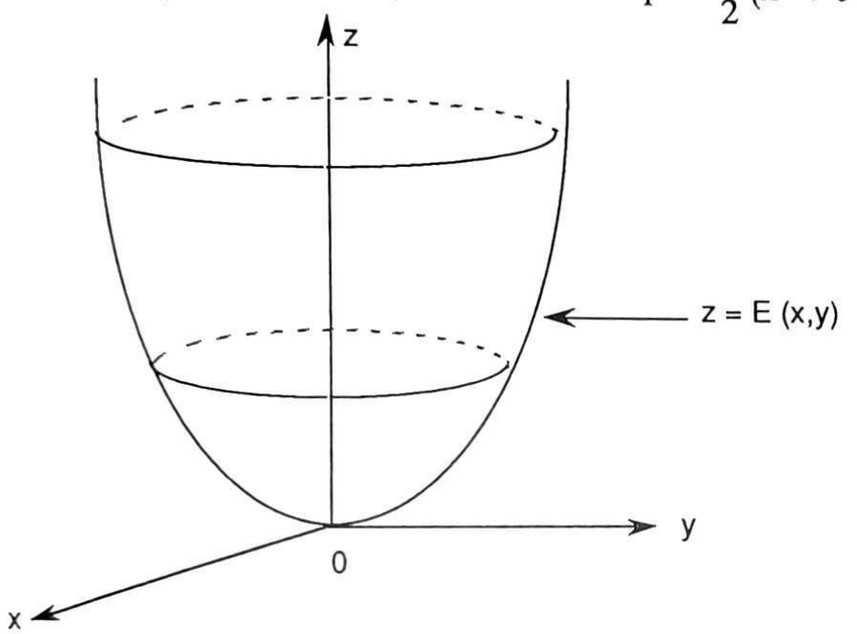
La solution est évidemment $x = y = 0$. Le nombre de conditionnement de A vaut $K(A) = c$ (dans ce paragraphe x et y désignent deux variables réelles).

Ce problème a pour solution le minimum de la fonctionnelle :

$$E(x,y) = \left(A\begin{pmatrix} x \\ y \end{pmatrix} \middle| \begin{matrix} x \\ y \end{matrix}\right) = \frac{1}{2}\left(x^2 + c\,y^2\right) .$$

Représentons géométriquement le problème de minimisation de $E(x,y)$.

$E(x,y) =$ constante positive est l'équation d'une ellipse : $\frac{1}{2}\left(x^2 + c\,y^2\right) = \text{Cte}$.



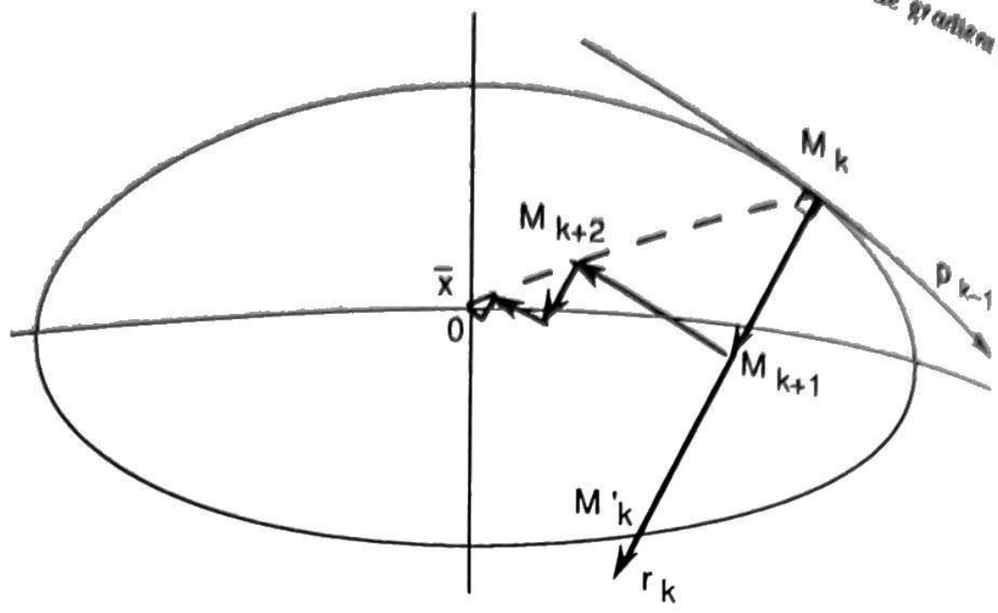
Pour différents choix de cette constante positive, on obtient une famille d'ellipses qui représentent les courbes de niveau de la fonctionnelle E .

Etudions la méthode du gradient à paramètre local optimal sur cet exemple. Le résidu $r = b - Ax$ est le vecteur $-\frac{1}{2}\begin{pmatrix} x \\ cy \end{pmatrix}$. Le paramètre est le nombre :

$$\alpha = \frac{\|r\|^2}{(A\,r|r)} = \frac{x^2 + c^2\,y^2}{\frac{x^2}{2} + c^3\,\frac{y^2}{2}} .$$

Connaissant le point $M_k\,(x_k\,,\,y_k)$, on construit le point $M_{k+1}\,(x_{k+1}\,,\,y_{k+1})$

$$M_{k+1}\begin{cases} x_{k+1} = x_k - \frac{\alpha_k}{2}\,x_k = \left(1 - \frac{\alpha_k}{2}\right)x_k = \frac{c^2\,(c-1)\,y_k^2}{x_k^2 + c^3\,y_k^2}\,x_k \ , \\ y_{k+1} = y_k - c\,\frac{\alpha_k}{2}\,y_k = \left(1 - c\,\frac{\alpha_k}{2}\right)y_k = \frac{(1-c)\,x_k^2}{x_k^2 + c^3\,y_k^2}\,y_k \ . \end{cases}$$



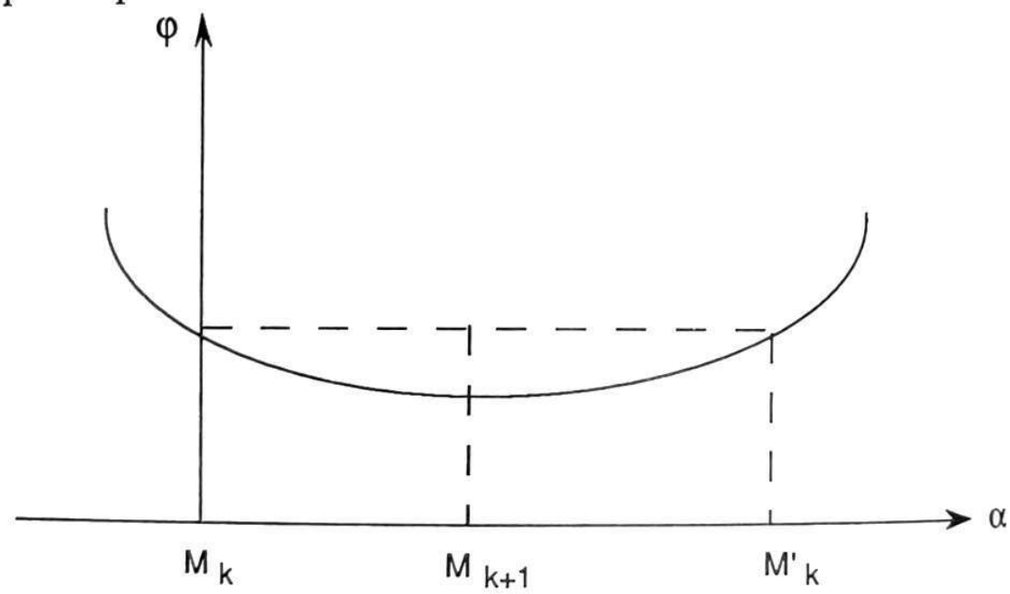
On peut construire géométriquement le point M_{k+1} à partir de M_k .

On sait que la direction r_k est orthogonale à celle de p_{k-1} .

Traçons l'ellipse passant par M_k . Soit M'_k son intersection avec la droite de direction r_k passant par M_k .

Alors M_{k+1} est le milieu de la corde $M_k M'_k$. En effet, soit $\varphi(\alpha) = E(M_k + \alpha r_k)$ qui est un trinôme du second degré en α .

α_k est le minimum de $\varphi(\alpha)$. Or, comme $E(M_k) = E(M'_k)$, le minimum est obtenu pour le point milieu



Démontrons que M_{k+2} et M_k sont alignés avec 0 .

En effet, posons $t_k = \dfrac{y_k}{x_k}$ qui est la pente de $0\,M_k$.

$$t_{k+1} = \frac{y_{k+1}}{x_{k+1}} = -\frac{1}{c^2\, t_k^2} \qquad t_k = -\frac{1}{c^2\, t_k} \ .$$

Donc :

$$t_{k+2} = -\frac{1}{c^2\, t_{k+1}} = -\frac{1}{-c^2 \dfrac{1}{c^2\, t_k}} = t_k \ .$$

Les itérés successifs sont situés sur deux droites passant par l'origine. Soit t la pente d'une de ces deux droites.

Il est commode pour évaluer le facteur moyen τ de réduction de l'erreur de calculer :

$$\tau^2 = \frac{y_{k+2}}{y_k} = \frac{x_{k+2}}{x_k} = \left(1 - \frac{\alpha_{k+1}}{2}\right)\left(1 - \frac{\alpha_k}{2}\right) = \frac{c^2(c-1)\left(\frac{1}{c^2 t}\right)^2}{1+c^3\left(\frac{1}{c^2 t}\right)^2} \cdot \frac{c^2(c-1)t^2}{1+c^3 t^2}$$

$$= \frac{(c-1)^2}{\left(1+\frac{c^3}{c^4 t^2}\right)(1+c^3 t^2)} = \frac{(c-1)^2}{\frac{1}{c t^2}(c t^2+1)(1+c^3 t^2)} .$$

Or :

$$\frac{1}{c t^2}(c t^2+1)(1+c^3 t^2) = c^3 t^2 + 1 + c^2 + \frac{1}{c t^2} = (c+1)^2 - 2c + c^3 t^2 + \frac{1}{c t^2}$$

$$= (c+1)^2 + c\left(c^2 t^2 - 2 + \frac{1}{c^2 t^2}\right) = (c+1)^2\left(1 + \frac{c}{(c+1)^2}\left(ct - \frac{1}{ct}\right)^2\right) .$$

D'où : $$\tau^2 = \frac{(c-1)^2}{(c+1)^2} \frac{1}{1+\frac{c}{(c+1)^2}\left(ct-\frac{1}{ct}\right)^2} .$$

Pour $t = \frac{1}{c}$, $\frac{1}{1+\frac{c}{(c+1)^2}\left(ct-\frac{1}{ct}\right)^2}$ est maximum,

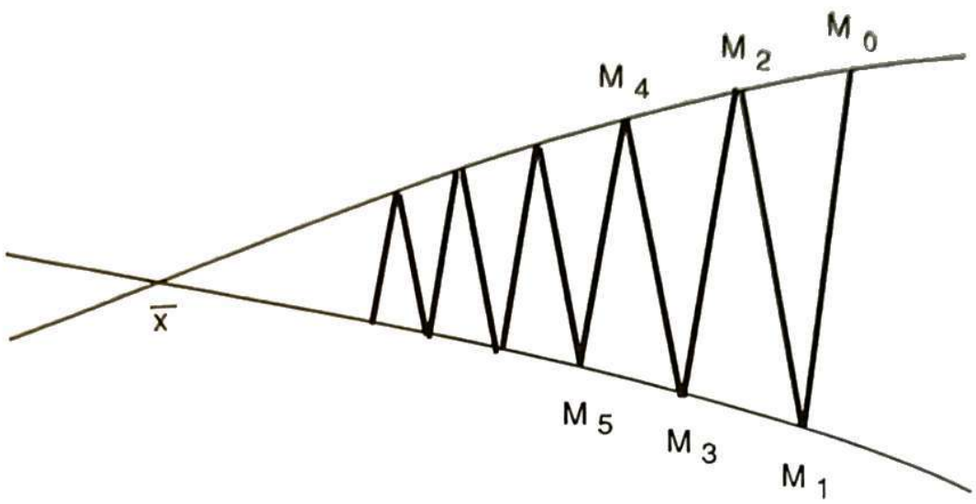
donc $\tau(t)$ est maximum et vaut :

$$\tau = \frac{c-1}{c+1} = \frac{K(A)-1}{K(A)+1} .$$

Par contre, si le point de départ est situé sur un des axes de l'ellipse ($t = 0$ ou $t = \infty$), alors la méthode converge en une seule itération.

On constate que le facteur de réduction de l'erreur peut être égal à $\frac{K(A)-1}{K(A)+1}$.

Sauf dans le cas où le point de départ est situé sur un des axes de l'ellipse, la convergence aura lieu en zigzag, avec un facteur de réduction de l'erreur d'autant plus voisin de 1 que $K(A)$ est grand.



Ces observations permettent d'envisager d'autres méthodes qui sont plus "performantes".
Si l'on veut diminuer le nombre d'opérations, on peut considérer la méthode où α_k serait constant au cours des itérations, ce qui se traduit dans l'exemple par :

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

$$y_{k+1} = y_k - \alpha\, c\, y_k = (1-\alpha\, c)\, y_k \ .$$

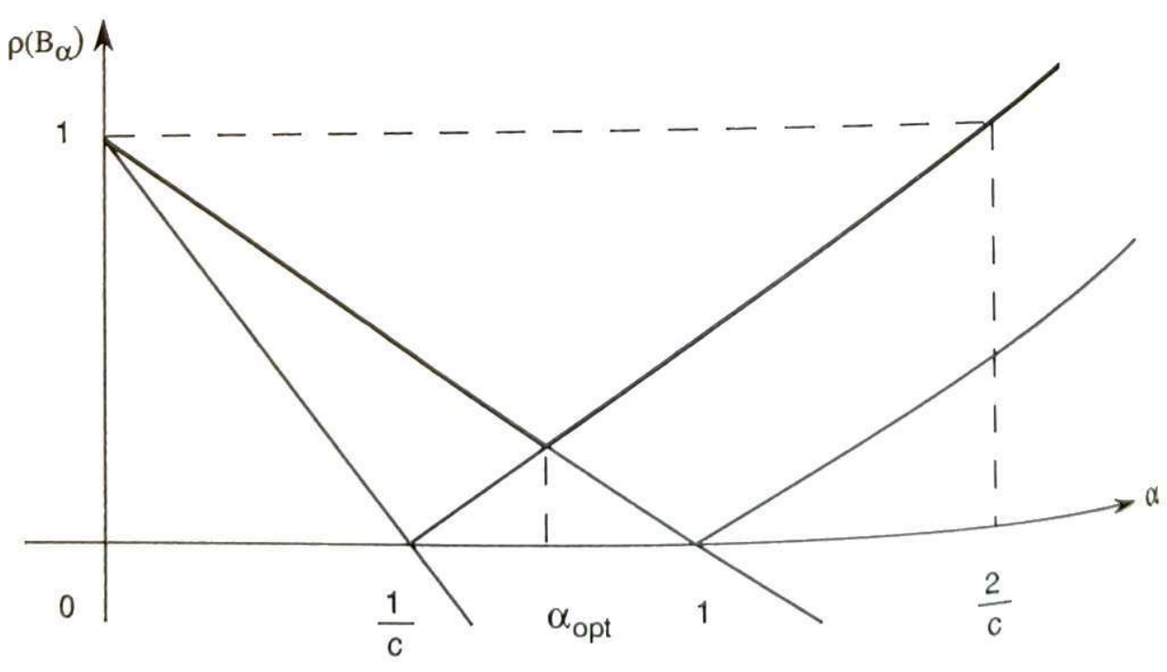
Pour déterminer α , on considère la matrice de l'itération

$$B_\alpha = \begin{bmatrix} 1-\alpha & 0 \\ 0 & 1-\alpha\, c \end{bmatrix} .$$

La condition nécessaire et suffisante de convergence est que :

$$\rho\,(B_\alpha) < 1 \ . \quad \text{Or} \quad \rho\,(B_\alpha) = \max\,(|\,1-\alpha\,|\,,\ |\,1-\alpha\, c\,|\,) \ .$$

Le choix optimal de α est celui qui rend $\rho(B_\alpha)$ le plus petit possible.



Sur le graphe, on constate que la convergence a lieu pour $0 < \alpha < \frac{2}{c}$.

Le choix optimal est :

$\alpha_{opt} = \frac{2}{1+c}$ pour lequel $\rho(B_{\alpha opt}) = 1 - \frac{2}{1+c} = \frac{c-1}{c+1} = \frac{K(A)-1}{K(A)+1}$.

On retrouve pour $ct = 1$ le même facteur de réduction de l'erreur que dans la méthode du gradient à paramètre optimal, ce qui prouve que α_k optimal peut être aussi mauvais que α constant optimal. Nous allons étudier le cas général de cette méthode au paragraphe suivant.

Comme les droites passant par M_k, de directions r_k, ne semblent pas donner entière satisfaction, on en cherchera de nouvelles. On souhaiterait qu'elles passent par le point qui rend minimum la forme quadratique, c'est-à-dire par le centre de l'ellipse.

Reprenons les notations habituelles.

Comme on a pour α_k optimal local $(p_k | r_{k+1}) = 0$ et que :

$$r_{k+1} = b - A x_{k+1} = A \bar{x} - A x_{k+1} ,$$

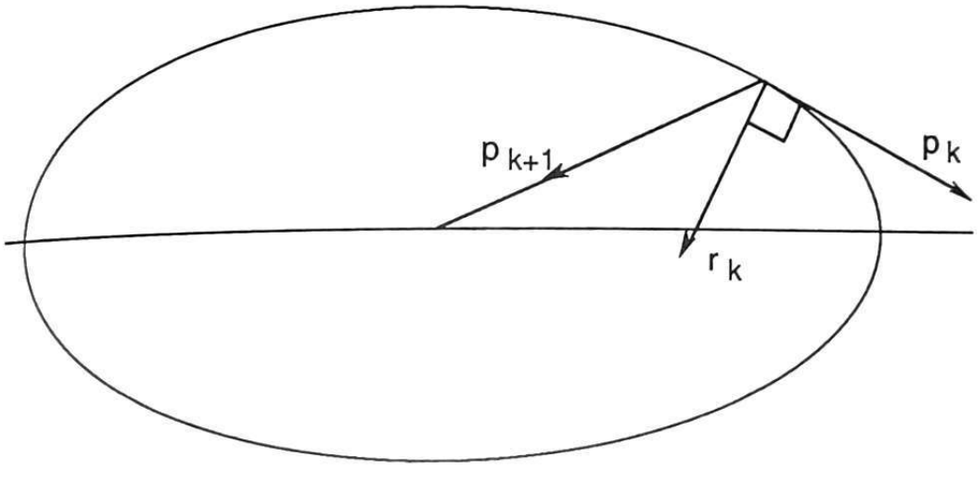
$$(p_k | A (\bar{x} - A x_{k+1})) = 0 .$$

on obtient :

Si on veut que $x_{k+2} = \bar{x}$, il faut que la direction $p_{k+1} = \frac{1}{\alpha_{k+1}}(x_{k+2} - x_{k+1})$

soit telle que : $(p_k | A p_{k+1}) = (A p_k | p_{k+1}) = 0$.

Nous étudierons cette condition au paragraphe sur le gradient conjugué et nous vérifierons alors que dans le cas d'une ellipse, on converge en deux itérations au plus.



8.2.3 *La méthode du gradient à paramètre constant (méthode de Richardson)*

Comme on l'a remarqué dans l'exemple précédent, il peut être inutile d'optimiser α_k à chaque itération, compte tenu de l'effet zigzag et du coût de calcul de α_k.

On prend toujours comme direction de descente celle du gradient c'est-à-dire celle du résidu au point considéré et on choisit α indépendant de k de façon que la suite des points (x_k) converge vers la solution $\bar{x}$.

$$x_{k+1} = x_k + \alpha\, r_k \ ,$$
$$r_k = b - A\, x_k = A\, \bar{x} - A\, x_k \ .$$

avec :

L'erreur à la $k+1^{ième}$ itération e_{k+1} peut s'exprimer en fonction de l'erreur à la $k^{ième}$ itération, en effet :

$$e_{k+1} = x_{k+1} - \bar{x} = x_k + \alpha\, r_k - \bar{x} = x_k + \alpha\,(A\,\bar{x} - A\,x_k) - \bar{x}$$
$$= (x_k - \bar{x}) - \alpha\,(A\,x_k - A\,\bar{x}) = (I - \alpha\, A)\, e_k \ .$$

Donc, on obtient :

$$\boxed{e_k = (I - \alpha\, A)^k\, e_0} \quad . \tag{15}$$

La condition nécessaire et suffisante de convergence (cf. chapitre 7, théorème 8) est que : $\rho = (I - \alpha\, A) < 1$.

Il faut et il suffit que les N valeurs propres positives λ_i de A vérifient :

$$|1 - \alpha\, \lambda_i| < 1 \ ,$$

c'est-à-dire $\forall\, i = 1, 2, \ldots, N \ , \quad 0 < \alpha < \dfrac{2}{\lambda_i}$.

Donc, en classant les valeurs propres de A par ordre décroissant :

$$0 < \lambda_N \leq \lambda_{N-1} \leq \ldots \leq \lambda_1 \ ,$$

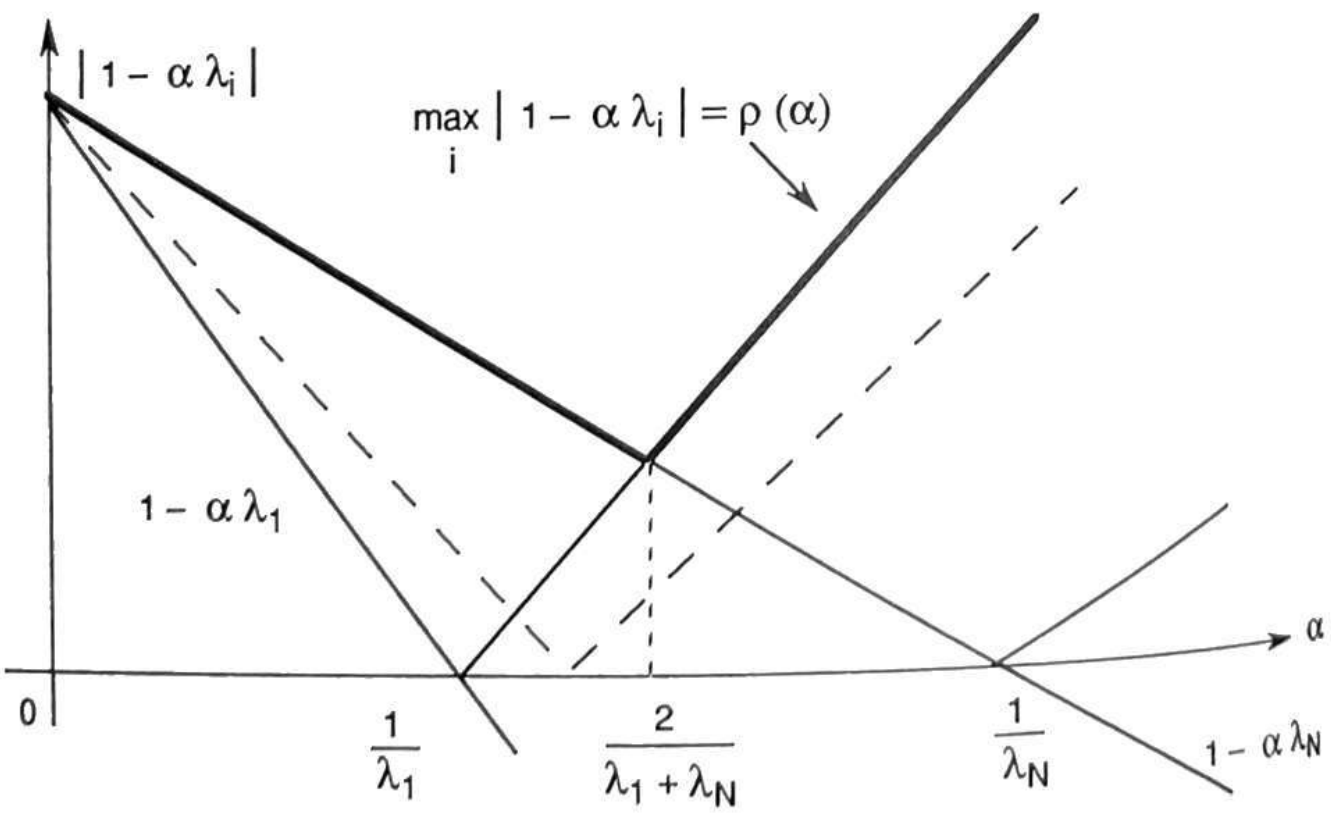
il faut et il suffit que :

$$\boxed{0 < \alpha < \frac{2}{\lambda_1}} \quad . \tag{16}$$

Le meilleur choix de α est celui qui minimise $\rho\,(I - \alpha\, A)$.

Or, $\rho\,(I - \alpha\, A) = \max\limits_i |1 - \alpha\, \lambda_i| = \max\left(|1 - \alpha\, \lambda_1|\,, |1 - \alpha\, \lambda_N|\right)$.

Comme on le voit sur la figure, α est la solution de $1 - \alpha\, \lambda_1 = \alpha\, \lambda_N - 1$.



(17)
$$\boxed{\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_N}} .$$

Pour cette valeur a_{opt}, on obtient $\rho(I - \alpha_{opt} A) = \dfrac{\lambda_1 - \lambda_N}{\lambda_N + \lambda_1} = \dfrac{\dfrac{\lambda_1}{\lambda_N} - 1}{\dfrac{\lambda_1}{\lambda_N} + 1}$,

(18)
$$\boxed{\rho(I - \alpha_{opt} A) = \frac{K(A) - 1}{K(A) + 1}} .$$

Remarque 7

En prenant α optimal dans la méthode du gradient à paramètre constant, le facteur de réduction de l'erreur est de l'ordre de $\dfrac{K(A) - 1}{K(A) + 1}$, comme dans le pire des cas de la méthode du gradient à paramètre optimal. (Cf. exemple 8, 2, 2). Même si on augmente peu le nombre d'itérations, il est nécessaire de connaître λ_1 et λ_N, ce qui n'est pas le cas en pratique.

Remarque 8

L'équation $Ax = b$ est équivalente, si la matrice diagonale $D = \text{diag}(A)$ est inversible à $D^{-1} Ax = D^{-1} b$. Posons $A' = D^{-1} A$ et $b' = D^{-1} b$.

Pour ce nouveau système, dont la matrice est telle que les éléments diagonaux sont égaux à 1, la méthode de Jacobi est définie par la relation :

$$x_{k+1} = (I - A') x_k + b' .$$

La méthode de Jacobi, dans ce cas, coïncide avec la méthode du gradient avec paramètre constant égal à 1.

8.3. Les méthodes du gradient conjugué

8.3.1 Introduction

On désire déterminer de nouvelles directions de descente p_k. On choisit α_k minimum local. Comme on l'a démontré en (8) $(p_{k-1} \mid r_k) = 0$, on va chercher p_k dans le plan formé par les deux directions orthogonales r_k et p_{k-1}.

Posons :

(19)
$$p_k = r_k + \beta_k p_{k-1} .$$

β_k va être déterminé afin que le facteur de réduction de l'erreur soit le plus grand possible dans E (x) .

Or, d'après (9) :

$$E(x_{k+1}) = E(x_k)\left(1 - \frac{(r_k \mid p_k)^2}{(A\,p_k \mid p_k)(A^{-1}\,r_k \mid r_k)}\right) .$$

Donc β_k sera choisi de façon que $\dfrac{(r_k \mid p_k)^2}{(A\,p_k \mid p_k)(A^{-1}\,r_k \mid r_k)}$ soit maximum.

Or : $(r_k \mid p_k) = (r_k \mid r_k + \beta_k\, p_{k-1}) = \|r_k\|^2 + \beta_k\,(r_k \mid p_{k-1}) = \|r_k\|^2$.

Donc :

(20) $$\boxed{(r_k \mid p_k) = \|r_k\|^2} .$$

On choisit $p_0 = r_0$ (donc $\beta_0 = 0$) pour que cette relation soit vérifiée quel que soit $k \geq 0$.

La détermination du maximum se réduit à minimiser $(A\,p_k \mid p_k)$.

$$(A\,p_k \mid p_k) = (A(r_k + \beta\, p_{k-1}) \mid r_k + \beta\, p_{k-1})$$

$$= \beta^2 (A\,p_{k-1} \mid p_{k-1}) + 2\,\beta\,(A\,p_{k-1} \mid r_k) + (A\,r_k \mid r_k)$$

Pour que ce trinôme soit minimum, il faut choisir β_k vérifiant :

$$\beta_k (A\,p_{k-1} \mid p_{k-1}) + (A\,p_{k-1} \mid r_k) = 0 .$$

Donc :

(21) $$\boxed{\beta_k = -\frac{(A\,p_{k-1} \mid r_k)}{(A\,p_{k-1} \mid p_{k-1})}} .$$

On en déduit que $(A\,p_{k-1} \mid r_k + \beta_k\, p_{k-1}) = 0$, c'est-à-dire :

(22) $$\boxed{(A\,p_{k-1} \mid p_k) = 0} .$$

Lorsque deux vecteurs u , v vérifient la relation $(A\,u \mid v) = 0$, on dit qu'ils sont *A–conjugués*. Comme A est symétrique et définie positive, $(A\,u \mid v)$ définit un produit scalaire. La relation pour deux vecteurs d'être A–conjugués signifie qu'ils sont orthogonaux pour ce produit scalaire.

Proposition 9

On a les relations suivantes, valables si $r_i \neq 0$, $i = 0$ à k :

(23) $$\boxed{(r_{k+1} \mid r_k) = 0} \qquad k \geq 0 ,$$

(24) $$\boxed{\beta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}} \qquad \text{pour } k \geq 1 .$$

Preuve :

$$(r_{k+1} \mid r_k) = (r_k - \alpha_k A p_k \mid r_k) = \|r_k\|^2 - \alpha_k (A p_k \mid r_k)$$

$$= \|r_k\|^2 - \alpha_k (A p_k \mid p_k - \beta_k p_{k-1})$$

$$= \|r_k\|^2 - \alpha_k (A p_k \mid p_k) + \alpha_k \beta_k (A p_k \mid p_{k-1}) = 0 ,$$

en tenant compte de : $$(A p_k \mid p_{k-1}) = 0 ,$$

ainsi que de : $$\alpha_k = \frac{(r_k \mid p_k)}{(A p_k \mid p_k)} = \frac{\|r_k\|^2}{(A p_k \mid p_k)} .$$

$$A p_{k-1} = \frac{1}{\alpha_{k-1}} (r_{k-1} - r_k) \qquad \text{d'après (7)}, \quad k \geq 1 ,$$

$$(A p_{k-1} \mid r_k) = \frac{1}{\alpha_{k-1}} (r_{k-1} - r_k \mid r_k) = -\frac{1}{\alpha_{k-1}} \|r_k\|^2 ,$$

$$(A p_{k-1} \mid p_{k-1}) = \frac{1}{\alpha_{k-1}} (r_{k-1} - r_k \mid p_{k-1}) = \frac{1}{\alpha_{k-1}} (r_{k-1} \mid p_{k-1}) = \frac{1}{\alpha_{k-1}} \|r_{k-1}\|^2 ,$$

donc : $$\beta_k = -\frac{(r_k \mid A p_{k-1})}{(p_{k-1} \mid A p_{k-1})} = \frac{\|r_k\|^2}{\|r_{k-1}\|^2} .$$

Comme $2 r_k = -g_k$ où g_k est le gradient de la fonctionnelle, on a également : $$\beta_k = \frac{\|g_k\|^2}{\|g_{k-1}\|^2} .$$ ■

8.3.2 L'algorithme

On initialise en choisissant x_0 et $p_0 = r_0$:

$$\begin{cases} x_0 , \\ p_0 = r_0 = b - A\,x_0 , \end{cases}$$

Pour $k = 0, 1 \ldots$

$$\text{(25)} \qquad \begin{cases} \alpha_k = \dfrac{\|r_k\|^2}{(A\,p_k \mid p_k)} , \\ x_{k+1} = x_k + \alpha_k\,p_k , \\ r_{k+1} = r_k - \alpha_k\,A\,p_k , \\ \beta_{k+1} = \dfrac{\|r_{k+1}\|^2}{\|r_k\|^2} , \\ p_{k+1} = r_{k+1} + \beta_{k+1}\,p_k . \end{cases}$$

Le test d'arrêt des itérations porte sur $\|r_k\|$ comme d'habitude. A chaque itération, le nombre d'opérations est le suivant : (c est le nombre moyen de coefficients non nuls par ligne de A)

	multiplications divisions	additions soustractions
calcul de $q = Ap$	N c	N(c – 1)
produit scalaire $(q\|p)$	N	N – 1
α	1	
x	N	N
r	N	N
calcul de $\|r\|^2$	N	N – 1
β	1	
p	N	N

soit au total (c + 5) N + 2 multiplications,
(c + 4) N – 2 additions.

Donc un nombre d'opérations voisin de $2\,cN$ par itération.

Le coût essentiel est celui du produit Ap .

Pour un nombre d'itérations k de l'ordre de N , on aboutit à un nombre d'opérations voisin de $2\,c\,N^2$ ce qui est relativement important notamment lorsque c est grand (si $c = N$, on a $2\,N^3$ opérations, alors que la méthode de Cholesky n'en nécessite que $\frac{N^3}{3}$) .

En fait, on va démontrer que grâce au préconditionnement de A, le nombre d'itérations sera nettement inférieur à N. Cette méthode est alors une des mieux adaptées à la résolution de système linéaire dont la matrice est symétrique, définie positive et creuse.

Au préalable, on va démontrer les résultats essentiels qui justifient le choix des directions de descente.

Remarque 10

En remplaçant $p_k = r_k + \beta_k p_{k-1} = r_k + \beta_k \left(\frac{x_k - x_{k-1}}{\alpha_{k-1}} \right)$ dans la relation

$$x_{k+1} = x_k + \alpha_k p_k \ ,$$

on obtient :

$$x_{k+1} = x_k + \alpha_k r_k + \frac{\alpha_k \beta_k}{\alpha_{k-1}} (x_k - x_{k-1}) \ ,$$

$$x_{k+1} = x_{k-1} + \left(1 + \frac{\alpha_k \beta_k}{\alpha_{k-1}} \right) (x_k - x_{k-1}) + \alpha_k r_k \ ,$$

soit en posant : $$\gamma_{k+1} = 1 + \frac{\alpha_k \beta_k}{\alpha_{k-1}} \ ,$$

$$x_{k+1} = x_{k-1} + \gamma_{k+1} (x_k - x_{k-1}) + \alpha_k (b - A x_k) \ .$$

On constate que x_{k+1} est calculé à partir de x_k et de x_{k-1}.

8.3.3 *Propriétés de l'algorithme*

Théorème 11

Dans la méthode du gradient conjugué, en choisissant $p_0 = r_0 = b - A x_0$,

on a les relations suivantes valables pour tout $k \geq 1$ à condition que

$r_i \neq 0$ pour $0 \leq i \leq k$.

(26) $(r_k \mid p_i) = 0$ pour $i \leq k-1$,

(27) $\mathcal{E}(r_0 , r_1 , \ldots, r_k) = \mathcal{E}(r_0 , A r_0 , \ldots, A^k r_0)$,

(28) $\mathcal{E}(p_0 , p_1 , \ldots, p_k) = \mathcal{E}(r_0 , A r_0 , \ldots, A^k r_0)$,

(29) $(p_k \mid A p_i) = (A p_k \mid p_i) = 0$ pour $i \leq k-1$,

(30) $(r_k \mid r_i) = 0$ pour $i \leq k-1$,