Chapitre 2

Les méthodes de descente

2.1 Principes des méthodes de descente

2.1.1 Choix de la fonctionnelle à minimiser

Soit A une matrice symétrique définie positive $n \times n$. Trouver la solution $\bar{x}$ de $Ax = b$ est équivalent à trouver le vecteur qui minimise la fonctionnelle J :

$$J(x) = (Ax, x) - 2\,(b, x),$$

où (.,.) représente le produit scalaire dans $\mathbb{R}^n$.

Théorème 2.1.1 *La solution $\bar{x}$ de $Ax = b$ est le vecteur pour lequel $J(x)$ atteint son minimum et on a :*

$$J(\bar{x}) = -(b, A^{-1}b).$$

Démonstration.
Soit

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

$(A\bar{x}, \bar{x})$ est une constante. Par conséquent, puisque $E(x) > 0$ si $x \neq \bar{x}$, et $E(\bar{x}) = 0$, alors $\bar{x}$ minimise $J(x)$.

$$J(\bar{x}) = -(A\bar{x}, \bar{x}) = -(b, A^{-1}b).$$

D'autre part le vecteur qui minimise J annule le gradient g de J (car J est une fonctionnelle quadratique et définie positive).

$$\begin{aligned} J(x) &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}\, x_i\, x_j - 2\sum_{i=1}^{n} b_i\, x_i. \\ \frac{\partial}{\partial x_k} J(x) &= 2\sum_{i=1}^{n} a_{ki}\, x_i - 2\, b_k. \end{aligned}$$

donc $g(\bar{x}) = 2(A\bar{x} - b)$. $\square$

Posons $g(x) = 2(Ax - b) = -2\,r(x)$ où $r(x) = b - Ax = A\,\bar{x} - Ax$ est le vecteur résidu du système $Ax = b$.

Il est équivalent de minimiser J ou E définie dans le théorème 2.1.1. Si on pose $x - \bar{x} = e(x)$, on a :

$$E(x) = (A\,e(x), e(x)).$$

Puisque A est symétrique et définie positive, alors (Ax, y) est un produit scalaire et $E(x) = \| e(x) \|_A^2$, avec $\| e \|_A = (Ae, e)^{1/2}$ norme associée à ce produit scalaire. Le minimum de E est nul et est atteint en $\bar{x}$.

$E(x)$ peut aussi s'exprimer en fonction du résidu $r(x) = A\,\bar{x} - Ax$:

$$E(x) = (r(x), A^{-1}r(x)).$$

Pour minimiser la fonctionnelle E, les méthodes de "descente" donnent x_{k+1} à partir de x_k en choisissant à la $(k+1)^{ème}$ itération une direction de descente $p_k \neq 0$ (c'est un vecteur de $\mathbb{R}^n$) et un scalaire α_k avec

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

de manière que $E(x_{k+1}) < E(x_k)$.

2.1.2 Choix optimal de α_k dans une direction fixée p_k

On suppose la direction p_k fixée.

Le choix local optimal de α_k est obtenu lorsqu'à chaque itération, on minimise $E(x_{k+1})$ dans la direction p_k :

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

Or

$$\begin{aligned} E(x_k + \alpha\, p_k) &= \big(A(x_k + \alpha\, p_k - \bar{x}), x_k + \alpha\, p_k - \bar{x}\big) \\ &= E(x_k) - 2\alpha(r_k, p_k) + \alpha^2\,(Ap_k, p_k). \end{aligned} \tag{2.1}$$

On a un trinôme du second degré en α, dont le terme de plus haut degré (Ap_k, p_k) est strictement positif $\forall p_k \neq 0$, puisque A est définie positive. Son minimum est atteint pour

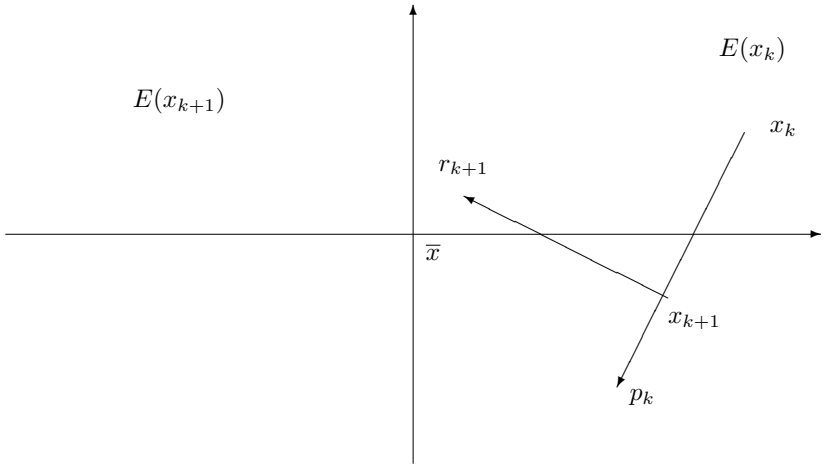
$$\alpha_k = \frac{(r_k, pk)}{(Ap_k, p_k)}. \tag{2.2}$$

Propriété 2.1.2 *$\forall p_k \neq 0$, pour α_k optimal, on a les deux relations suivantes :*
i) $\forall k \geq 0$, $r_{k+1} = r_k - \alpha_k\, A\, p_k$,
ii) $(p_k, r_{k+1}) = 0$.

Démonstration.
i) $r_{k+1} = b - Ax_{k+1} = b - A\,(x_k + \alpha_k\, p_k) = r_k - \alpha_k\, A\, p_k$.
ii) $(p_k, r_{k+1}) = (p_k, r_k) - \alpha_k\,(p_k, Ap_k) = 0$ quand on remplace α_k par (2.2). □

Interprétation géométrique dans R^n des méthodes de descente.

$E(x) = cste > 0$ est l'équation d'un hyperellipsoïde. On obtient une famille d'hyperellipsoïdes concentriques autour du minimum $\bar{x}$ de la fonctionnelle ; elles représentent les courbes de niveau.



Le vecteur p_k est tangent à l'hyperellipsoïde $E(x_{k+1})$.

A partir de (2.1) et (2.2)

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

Or $E(x_k) = (r_k, A^{-1}r_k)$. Donc $E(x_{k+1}) = E(x_k)(1-\gamma_k)$ avec

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

$\gamma_k > 0$ sauf si $p_k = 0$ (cas que l'on élimine), ou si $r_k = 0$ (alors x_k est la solution $\bar{x}$).

Lemme 2.1.3 *$\forall p_k \neq 0$, pour α_k optimal local, on a la relation suivante valable pour $k \geq 0$:*

$$\gamma_k = \frac{(r_k, p_k)^2}{(Ap_k, p_k)\,(A^{-1}r_k, r_k)} \geq \frac{1}{K(A)}\left(\frac{r_k}{\parallel r_k \parallel_2}, \frac{p_k}{\parallel p_k \parallel_2}\right)^2,$$

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

Démonstration.
Rappelons que dans le cas d'une matrice A symétrique définie positive nous avons

$$\mathrm{cond}_2(A) = \parallel A \parallel_2 \parallel A^{-1} \parallel_2 = K(A) = \frac{\max\limits_i \lambda_i}{\min\limits_i \lambda_i},$$

où les λ_i sont les valeurs propres de A.

$$(Ap_k, p_k) \leq \lambda_1 \parallel p_k \parallel_2^2,$$

où λ_1 est le $\max_i \lambda_i$. En effet, si on associe aux λ_i une base orthonormée de vecteurs propres u_i, alors

$$p_k = \sum_{i=1}^{n} a_i\, u_i.$$

$$(Ap_k, p_k) = (\sum_{i=1}^{n} a_i\, A\, u_i, \sum_{i=1}^{n} a_i\, u_i) = (\sum_{i=1}^{n} a_i\, \lambda_i\, u_i, \sum_{i=1}^{n} a_i\, u_i) = \sum_{i=1}^{n} a_i^2\, \lambda_i$$

en tenant compte du fait que $(u_i, u_j) = \delta_{ij}$.
Comme $\sum_{i=1}^{n} a_i^2 = \| p_k \|_2^2$, on obtient la majoration proposée.
De même $(A^{-1} r_k, r_k) \leq \frac{1}{\lambda_n} \| r_k \|_2^2$ où λ_n est la plus petite valeur propre de A.
Donc γ_k vérifie bien la relation proposée. □

Ce lemme va permettre un choix des directions de descente.

Théorème 2.1.4 *Pour α_k optimal local, toute direction p_k qui vérifie $\forall k \geq 0$*

$$\left(\frac{r_k}{\| r_k \|_2}, \frac{p_k}{\| p_k \|_2} \right)^2 \geq \mu > 0, \tag{2.3}$$

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

Démonstration.
Dans ce cas : $E(x_{k+1}) \leq E(x_k)(1 - \frac{\mu}{K(A)})$. D'où : $E(x_k) \leq (1 - \frac{\mu}{K(A)})^k E(x_0)$.
Or $0 < \mu \leq 1$ (on applique l'inégalité de Cauchy-Schwarz à (2.3)) et $K(A) \geq 1$. Par conséquent $0 \leq 1 - \frac{\mu}{K(A)} < 1$. Donc $\lim_{k \to \infty} E(x_k) = 0$.
Or par un raisonnement similaire à celui du lemme 2.1.3 on a

$$E(x_k) \geq \lambda_n \| x_k - \bar{x} \|_2^2 \text{ avec } \lambda_n > 0,$$

ce qui implique que

$$\lim_{k \to \infty} \| x_k - \bar{x} \|_2 = 0.$$

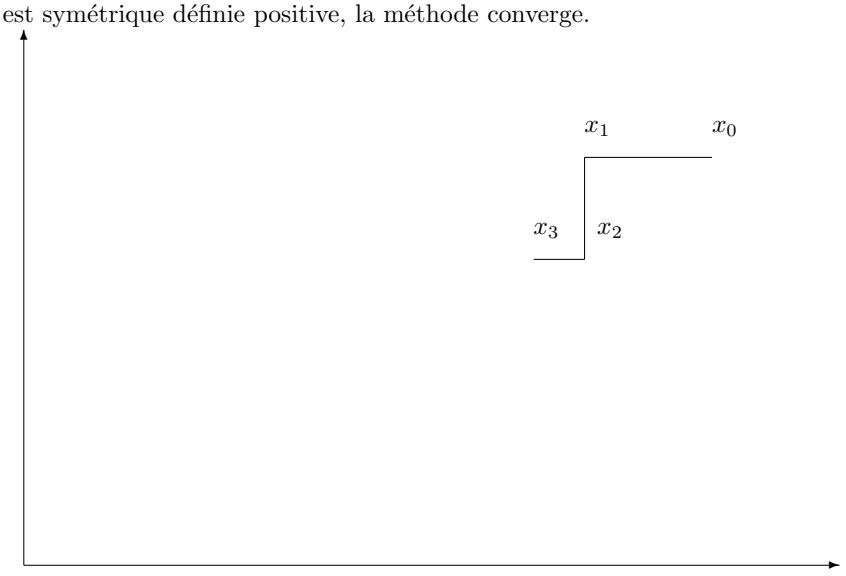
□

Ce théorème montre donc que p_k doit être non orthogonal à r_k. Il en résulte un premier choix évident : $p_k = r_k$, ce qui entraîne que $\mu = 1$.

N.B. : La méthode de Gauss-Seidel est une méthode de descente. La direction de descente est successivement $e_1, e_2, ..., e_n, e_1, e_2 \ldots$ etc. Dans ce cas

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k e_i, \\ \alpha_k &= \frac{(r_k, e_i)}{(Ae_i, e_i)} = \frac{(b - Ax_k, e_i)}{a_{ii}} \end{aligned}$$

et si A est symétrique définie positive, la méthode converge.



2.2 Les méthodes de gradient

2.2.1 La méthode du gradient à paramètre optimal ($p_k = r_k$).

$$\alpha_k = \frac{\| r_k \|^2}{(A\, r_k, r_k)} \text{ car } p_k = r_k.$$

D'où

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

On utilise l'inégalité de Kantorovitch.

Lemme 2.2.1 *Inégalité de Kantorovitch. Si A est hermitienne définie positive, alors*

$$\forall x \neq 0,\ 1 \leq \frac{(Ax, x)\,(A^{-1}x, x)}{\| x \|_2^4} \leq \frac{\left(K(A)^{1/2} + K(A)^{-1/2}\right)^2}{4}$$

avec K(A) = nombre conditionnement de la matrice A :

$$K(A) = \frac{\lambda_1}{\lambda_n} = \frac{\max\limits_i \lambda_i}{\min\limits_i \lambda_i}.$$

Démonstration.
On rapporte $\mathbb{C}^n$ à la base orthonormée de vecteurs propres $\{u_i\}_{i=1}^n$ de A.
Tout x s'exprime sous la forme :

$$x = \sum_{i=1}^{n} \alpha_i\, u_i.$$

Nous obtenons alors

$$\begin{aligned}
(Ax, x) &= \sum_{i=1}^{n} |\, \alpha_i \,|^2\ \lambda_i, \\
(A^{-1}x, x) &= \sum_{i=1}^{n} |\, \alpha_i \,|^2\ \frac{1}{\lambda_i}, \\
(x, x) &= \sum_{i=1}^{n} |\, \alpha_i \,|^2\ .
\end{aligned}$$

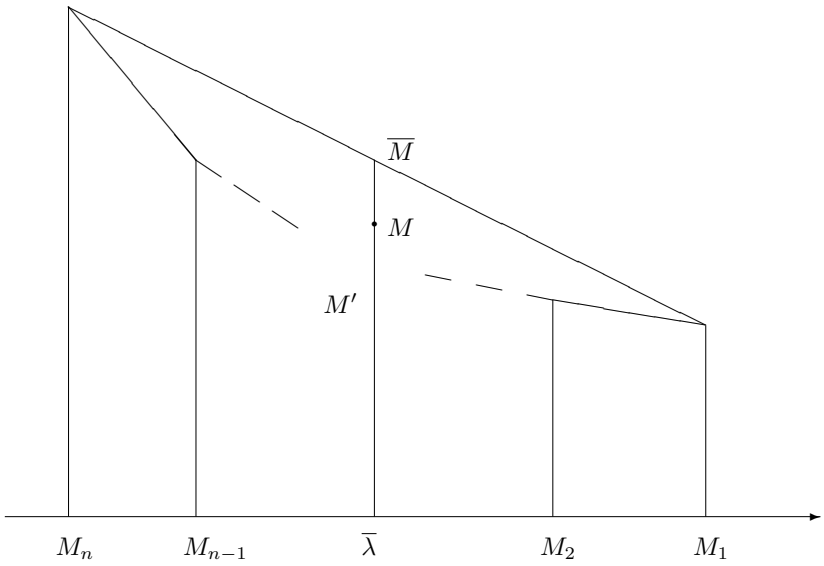
Donc

$$\frac{(Ax, x)\,(A^{-1}x, x)}{(x, x)^2} = \frac{\sum_{i=1}^n |\, \alpha_i \,|^2\ \lambda_i}{\sum_{i=1}^n |\, \alpha_i \,|^2}\ \frac{\sum_{j=1}^n |\, \alpha_j \,|^2\ \dfrac{1}{\lambda_j}}{\sum_{j=1}^n |\, \alpha_j \,|^2}.$$

Posons $\beta_i = \dfrac{|\, \alpha_i \,|^2}{\sum_{j=1}^n |\, \alpha_j \,|^2}$, alors $\sum_{i=1}^n \beta_i = 1$. Par conséquent l'expression précédente s'écrit

$$\Big(\sum_{i=1}^{n} \beta_i\, \lambda_i\Big)\ \Big(\sum_{j=1}^{n} \beta_j\, \frac{1}{\lambda_j}\Big).$$

Si M_i est le point du plan de coordonnées $(\lambda_i, \frac{1}{\lambda_i})$, alors $M = \sum_{i=1}^n \beta_i\, M_i$ est une combinaison linéaire convexe des points M_i. $(\lambda, \frac{1}{\lambda})$ est la branche d'hyperbole équilatère $y = \frac{1}{x}$. M appartient à l'enveloppe convexe des M_i, c'est-à-dire se trouve dans le polygone d'arêtes M_1M_2, M_2M_3, …, $M_{n-1}M_n$ et M_1M_n.



Pour l'hyperbole la corde M_1M_n est au-dessus de toutes les autres arêtes. M est en dessous de $\overline{M}\ \big(\bar{\lambda}, y(\bar{\lambda})\big)$ situé sur la corde M_1M_n et est au-dessus de $M'\ (\bar{\lambda}, \frac{1}{\bar{\lambda}})$ situé sur l'hyperbole. Donc

$$\bar{\lambda}\,\frac{1}{\bar{\lambda}} \leq \big(\sum_{i=1}^{n} \beta_i\,\lambda_i\big)\,\big(\sum_{j=1}^{n} \beta_j\,\frac{1}{\lambda_j}\big) \leq \bar{\lambda}\,y(\bar{\lambda}) = \bar{\lambda}\left(\frac{\lambda_1+\lambda_n-\bar{\lambda}}{\lambda_1\,\lambda_n}\right).$$

D'où

$$1 \leq \big(\sum_{i=1}^{n} \beta_i\,\lambda_i\big)\,\big(\sum_{j=1}^{n} \beta_j\,\frac{1}{\lambda_j}\big) \leq \max_{\lambda_n\leq\bar{\lambda}\leq\lambda_1}\left(\bar{\lambda}\,\big(\frac{\lambda_1+\lambda_n-\bar{\lambda}}{\lambda_1\,\lambda_n}\big)\right).$$

Ce maximum est atteint pour $\bar{\lambda} = \frac{\lambda_1+\lambda_n}{2}$ et vaut $\frac{(\lambda_1+\lambda_n)^2}{4\,\lambda_1\,\lambda_n}$.
On a finalement

$$1 \leq \frac{(Ax,x)\,(A^{-1}x,x)}{(x,x)^2} \leq \frac{(\lambda_1+\lambda_n)^2}{4\,\lambda_1\,\lambda_n} = \frac{\big(K(A)+1\big)^2}{4\,K(A)}$$

en tenant compte du fait que $\lambda_1 = \lambda_n\,K(A)$. $\square$

A l'aide de l'inégalité précédente on a

$$\frac{\parallel r_k \parallel^4}{(A\,r_k,r_k)\,(A^{-1}\,r_k,r_k)} \geq \frac{4\,\lambda_1\,\lambda_n}{(\lambda_1+\lambda_n)^2} = \frac{4\,K(A)}{\big(K(A)+1\big)^2}.$$

Alors

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

D'où

$$E(x_{k+1}) \leq E(x_0)\left(\frac{K(A)-1}{K(A)+1}\right)^{2k+2}.$$

Or $E(x_k) \geq \lambda_n \parallel x_k - \bar{x} \parallel_2^2$. Par conséquent

$$\parallel x_k - \bar{x} \parallel_2 \leq \beta \left(\frac{K(A)-1}{K(A)+1} \right)^k \text{ avec } \beta = \left(\frac{E(x_0)}{\lambda_n} \right)^{1/2}.$$

D'où le théorème :

Théorème 2.2.2 *La méthode du gradient à paramètre local optimal est convergente. La rapidité de convergence dépend de* $\frac{K(A)-1}{K(A)+1}$.

N.B. : Plus $K(A)$ est proche de 1, et plus la méthode convergera vite.

Quand $K(A) = 1$, alors toutes les valeurs propres sont égales. $A = \lambda I$ et $E(x) = \lambda \parallel x - \bar{x} \parallel^2$. Lorsque $E(x) =$ cste, on a l'équation d'une sphère. Quel que soit le point de la sphère, le gradient pointe vers le centre. On a convergence en une itération.

Si $K(A)$ est grand, alors λ_1 et λ_n sont très différents. L'hyperellipsoïde est très aplati et la convergence lente.

Pour avoir $\frac{E(x_k)}{E(x_0)} \leq \varepsilon$, il suffit d'avoir $\left(\frac{K(A)-1}{K(A)+1} \right)^{2k} \leq \varepsilon$, ce qui donne $k \simeq \frac{K(A)}{4} \text{Log} \frac{1}{\varepsilon}$. On obtient cet ordre de grandeur en écrivant un développement limité de l'expression précédente en puissances de $\frac{1}{K(A)}$. Le nombre d'itérations est proportionnel à $K(A)$.

2.2.2 La méthode du gradient à paramètre constant.

(Méthode de Richardson).

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

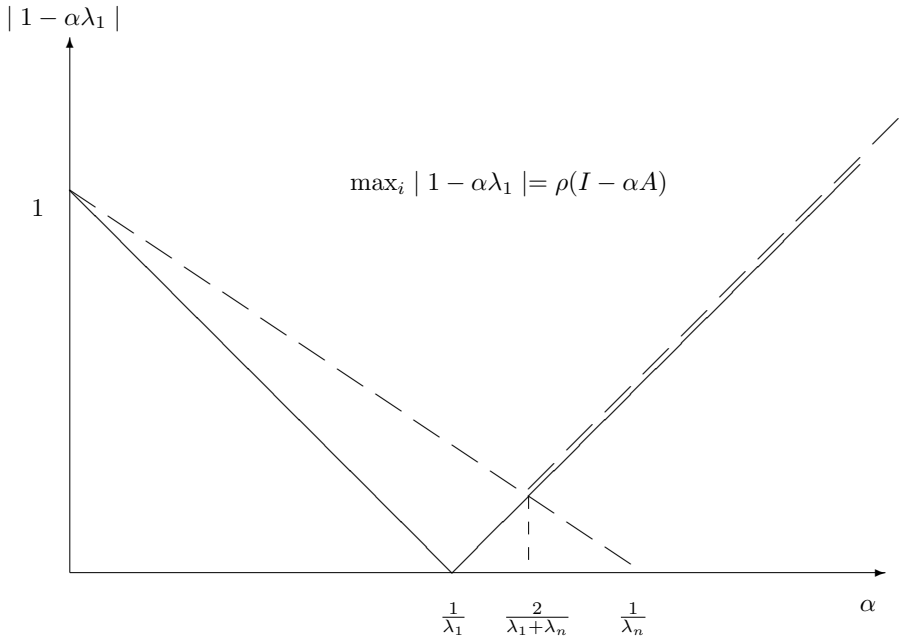
$$\begin{aligned} x_{k+1} &= x_k + \alpha\, r_k, \\ r_k &= b - A\, x_k = A(\bar{x} - x_k). \end{aligned}$$

L'erreur à la $(k+1)^{ème}$ itération est égale à e_{k+1}.

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

D'où $e_{k+1} = (I - \alpha\, A)^{k+1} e_0$.

Donc une C.N.S. de convergence est que $\rho(I - \alpha\, A) < 1$. Alors $\mid 1 - \alpha\, \lambda_i \mid < 1$ pour $i = 1, 2, \ldots, n$. et par conséquent $0 < \alpha < \frac{2}{\lambda_i}$. Si λ_1 est la plus grande valeur propre, alors $0 < \alpha < \frac{2}{\lambda_1}$. Le meilleur choix de α est celui qui minimise $\rho(I - \alpha\, A)$. Or $\rho(I - \alpha\, A) = \max_i \mid 1 - \alpha\, \lambda_i \mid = \max(\mid 1 - \alpha\, \lambda_1 \mid, \mid 1 - \alpha\, \lambda_n \mid)$.



α est solution de $1-\alpha\,\lambda_1 = \alpha\,\lambda_n - 1$. Par conséquent $\alpha_{opt} = \dfrac{2}{\lambda_1+\lambda_n}$. Alors $\rho(I-\alpha_{opt}\,A) = \dfrac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n} = \dfrac{K(A)-1}{K(A)+1}$.

N.B. : Il faut connaître λ_1 et λ_n, ce qui n'est pas le cas en pratique. Le facteur de réduction de l'erreur est de l'ordre de $\frac{K(A)-1}{K(A)+1}$.

2.3 Les méthodes de gradient conjugué

Hestenes et Stiefel (1952).

2.3.1 Introduction

On choisit α_k = minimum local, alors $(p_{k-1}, r_k) = 0$. On cherche p_k dans le plan (r_k, p_{k-1}).

On pose $p_k = r_k + \beta_k\, p_{k-1}$. β_k sera déterminé de telle façon que le facteur de réduction de l'erreur soit le plus grand possible. Or $E(x_{k+1}) = E(x_k)(1-\gamma_k)$. On choisit β_k pour que γ_k soit maximum.

Comme $(r_k, p_k) = (r_k, r_k) + \beta_k(r_k, p_{k-1}) = \| r_k \|_2^2$ (on prend $p_0 = r_0$ ($\beta_0 = 0$) pour que la relation précédente soit vraie $\forall k \geq 0$), γ_k sera maximum, si $(A\,p_k, p_k)$ est minimum.

$$\begin{aligned}(A\,p_k, p_k) &= \Big(A(r_k + \beta_k\,p_{k-1}), r_k + \beta_k\,p_{k-1}\Big)\\ &= \beta_k^2\,(A\,p_{k-1}, p_{k-1}) + 2\,\beta_k\,(A\,p_{k-1}, r_k) + (A\,r_k, r_k).\end{aligned}$$

Le trinôme est minimum si $\beta_k = -\dfrac{(A\,p_{k-1}, r_k)}{(A\,p_{k-1}, p_{k-1})}$.

Cette valeur de β_k correspond aussi au point d'annulation de la dérivée. On obtient donc :

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

Définition 2.3.1 *Deux vecteurs u et v qui vérifient $(A\,u,v)=0$ sont dits A-conjugués.*

Comme A est symétrique définie positive, $(A\,u,v)$ est un produit scalaire $(u,v)_A$. Par conséquent deux vecteurs A-conjugués sont orthogonaux pour ce produit scalaire.

Propriété 2.3.2 *Si $r_i \neq 0$ pour $i=0,\ldots,k$, alors*
i) $(r_{k+1},r_k)=0$ pour $k\geq 0$,
ii) $\beta_0=0$, $\beta_k=\frac{\|r_k\|^2}{\|r_{k-1}\|^2}$ pour $k\geq 1$.

Démonstration.
i) $(r_{k+1},r_k)=(r_k-\alpha_k\,A\,p_k,r_k)=\|\,r_k\,\|^2-\alpha_k\,(A\,p_k,r_k)$.

$$(A\,p_k,r_k)=(A\,p_k,p_k)-\beta_k\,(A\,p_k,p_{k-1})=(A\,p_k,p_k)\text{ car }(A\,p_k,p_{k-1})=0.$$

De plus $\alpha_k=\dfrac{(r_k,p_k)}{(A\,p_k,p_k)}=\dfrac{\|\,r_k\,\|^2}{(A\,p_k,p_k)}$. D'où le résultat.
ii) $A\,p_{k-1}=\frac{1}{\alpha_{k-1}}(r_{k-1}-r_k)$ d'après la propriété 2.1.2. D'où $(A\,p_{k-1},r_k)=\frac{-1}{\alpha_{k-1}}\,\|\,r_k\,\|^2$. Enfin

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

Le rapport des deux expressions précédentes donne la valeur proposée de β_k. □

Remarque 2.3.3 *Comme $2\,r_k=-g_k$, où $-g_k$ est le gradient de la fonctionnelle, alors* $\beta_k=\dfrac{\|\,g_k\,\|^2}{\|\,g_{k-1}\,\|^2}$.

2.3.2 L'algorithme

$$\text{On initialise : }\begin{cases} x_0 \\ p_0=r_0=b-A\,x_0. \end{cases}$$

$$\text{Pour } k=0,1,\ldots\begin{cases} \alpha_k=\dfrac{\|\,r_k\,\|^2}{(A\,p_k,p_k)}, \\ x_{k+1}=x_k+\alpha_k\,p_k, \\ r_{k+1}=r_k-\alpha_k\,A\,p_k, \\ \beta_{k+1}=\dfrac{\|\,r_{k+1}\,\|^2}{\|\,r_k\,\|^2}, \\ p_{k+1}=r_{k+1}+\beta_{k+1}\,p_k. \end{cases}$$

Complexité : Si c est le nombre moyen de coefficients non nuls par ligne de A, le nombre d'opérations est le suivant :

	* et /	+ −
$q=Ap$	$N\,c$	$N(c-1)$
(q,p)	N	$N-1$
α	1	
x	N	N
r	N	N
$\|\,r\,\|^2$	N	$N-1$
β	1	
p	N	N
	$(c+5)\,N+2$	$(c+4)N-2$

Si k qui est le nombre d'itérations, est égal à N, on a environ $2\,c\,N^2$ opérations. Si $c = N$, on a $2\,N^3$ opérations, ce qui est important. (Dans la méthode de Cholesky ce nombre est égal à $\frac{N^3}{3}$).

Grâce au préconditionnement de A, le nombre d'itérations sera très inférieur à N. Cette méthode est alors une des mieux adaptées à la résolution de systèmes linéaires dont la matrice est symétrique définie positive et creuse.

N.B. : $x_{k+1} = x_k + \alpha_k\, p_k = x_k + \alpha_k\, r_k + \frac{\alpha_k\,\beta_k}{\alpha_{k-1}}\,(x_k - x_{k-1})$ puisque $p_k = r_k + \beta_k\, p_{k-1} = r_k + \beta_k\,\frac{(x_k - x_{k-1})}{\alpha_{k-1}}$.
Donc $x_{k+1} = x_{k-1} + \big(1 + \frac{\alpha_k\,\beta_k}{\alpha_{k-1}}\big)(x_k - x_{k-1}) + \alpha_k\, r_k$.

Si on pose $\gamma_{k+1} = 1 + \frac{\alpha_k\,\beta_k}{\alpha_{k-1}}$, alors

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

et x_{k+1} est déterminé à partir de x_k et x_{k-1}.

2.3.3 Propriétés de l'algorithme

Théorème 2.3.4 *Dans la méthode du gradient conjugué, si on choisit*

$$p_0 = r_0 = b - Ax_0,$$

alors $\forall k \geq 1$ *et si* $r_i \neq 0$, $0 \leq i \leq k$,

$$\begin{aligned}
(r_k, p_i) &= 0 \text{ pour } i \leq k-1, && (2.4)\\
V(r_0, \ldots, r_k) &= V(r_0, A\,r_0, \ldots, A^k\,r_0), && (2.5)\\
V(p_0, \ldots, p_k) &= V(r_0, A\,r_0, \ldots, A^k\,r_0), && (2.6)\\
(p_k, A\,p_i) &= (A\,p_k, p_i) = 0 \text{ pour } i \leq k-1, && (2.7)\\
(r_k, r_i) &= 0 \text{ pour } i \leq k-1. && (2.8)
\end{aligned}$$

V désigne le sous espace vectoriel de R^n engendré par les vecteurs placés en argument.

Démonstration.
Si $r_i = 0$, alors $x_i = \bar{x}$. Donc la restriction à $r_i \neq 0$ n'est pas contraignante.

On effectue une démonstration par récurrence.
i) $(p_{k-1}, r_k) = 0$, $(A\,p_{k-1}, p_k) = 0$ et $(r_k, r_{k-1}) = 0$ impliquent que (2.4), (2.7), (2.8) sont vraies pour $k = 1$. D'autre part : $p_0 = r_0$, $p_1 = r_1 + \beta_1\, p_0$, ce qui entraîne $V(r_0, r_1) = V(p_0, p_1)$.
Enfin $r_1 = r_0 - \alpha_0\, A\,p_0$ et $\alpha_0 = \frac{\|r_0\|^2}{(A\,r_0, r_0)} \neq 0$, alors $A\,p_0 = \frac{r_0 - r_1}{\alpha_0}$.
D'où $V(p_0, A\,p_0) = V(r_0, A\,r_0) = V(r_0, r_1)$.
Donc (2.5) et (2.6) sont vraies pour $k = 1$.

ii) On suppose les relations vraies pour k et on les démontre pour $k+1$. Alors $(r_{k+1}, p_i) = (r_k, p_i) - \alpha_k\,(A\,p_k, p_i) = 0$ pour $i \leq k-1$ et $(r_{k+1}, p_k) = 0$ impliquent que (2.4) soit vraie. D'autre part : $(r_{k+1}, r_k) = 0$ et $(r_{k+1}, r_i) = (r_k, r_i) - \alpha_k\,(A\,p_k, r_i) = 0$ pour $i \leq k-1$ impliquent que (2.7) soit vraie.

Ensuite $r_k \in V(r_0, A\,r_0, \ldots, A^k\,r_0)$ d'après (2.5).
$A\,p_k \in A\,V(r_0, A\,r_0, \ldots, A^k\,r_0)$ d'après (2.6).
$A\,V(r_0, A\,r_0, \ldots, A^k\,r_0) = V(A\,r_0, \ldots, A^{k+1}\,r_0) \subset V(r_0, A\,r_0, \ldots, A^{k+1}\,r_0)$.
Par conséquent $r_{k+1} = r_k - \alpha_k\, A\,p_k \in V(r_0, A\,r_0, \ldots, A^{k+1}\,r_0)$, ce qui entraîne que $V(r_0, \ldots, r_{k+1}) \subset V(r_0, A\,r_0, \ldots, A^{k+1}\,r_0)$.

D'autre part dim $V(r_0,\ldots,r_{k+1}) = k+2$. Donc
dim$V(r_0, A\,r_0,\ldots,A^{k+1}\,r_0) = k+2$ et $r_{k+1} \notin V(r_0, A\,r_0,\ldots,A^k\,r_0)$.
$V(r_0,\ldots,r_k) \oplus V(r_{k+1}) = V(r_0, A\,r_0,\ldots,A^k\,r_0) \oplus V(A^{k+1}r_0)$.
$V(A^{k+1}r_0) \subset V(r_0,\ldots,r_{k+1})$.
On montre de façon identique que :

$$V(p_0,\ldots,p_{k+1}) = V(r_0, A\,r_0,\ldots,A^{k+1}\,r_0).$$

$(A\,p_{k+1}, p_k) = 0$ puisque c'est la condition vérifiée par deux directions successives qui sont A-conjugées.

$$\begin{aligned}(p_{k+1}, A\,p_i) &= (r_{k+1}, A\,p_i) + \beta_{k+1}\,(p_k, A\,p_i) \text{ pour } i \le k-1\\ &= (r_{k+1}, A\,p_i).\end{aligned}$$

Or $A\,p_i \in V(r_0,\ldots,A^{i+1}\,r_0) = V(p_0,\ldots,p_{i+1})$.
Donc $(r_{k+1}, p_{i+1}) = 0 => (r_{k+1}, A\,p_i) = 0$. □

Définition 2.3.5 *L'espace vectoriel $\kappa_k = V(r_0,\ldots,A^{k-1}\,r_0)$ est appelé espace de Krylov.*

Les $r_i = 0,\ldots,k-1$ forment une base orthogonale de cet espace.

Théorème 2.3.6

$$E(x_k) \le E(x)\ \forall x \in x_0 + \kappa_k.$$

Démonstration.

$$x_k = x_0 + \sum_{i=0}^{k-1} \alpha_i\,p_i \in x_0 + \kappa_k.$$

$$\begin{aligned}E(x_k) = \min_{x\in x_0+\kappa_k} E(x) &\iff E(x_k) \le E(x_k + y)\ \forall y \in \kappa_k\\ &\iff (E'(x_k), y) = 0\ \forall y \in \kappa_k\\ &\iff (2r_k, y) = 0.\end{aligned}$$

Or $(r_k, r_i) = 0\ \forall i \le k-1$. D'où le résultat. □

Corollaire 2.3.7 *(Théorème de Stiefel) L'algorithme du gradient conjugué converge en au plus n itérations.*

Démonstration.
Ou $r_k = 0$ pour $k \le n-1$. On a alors convergence en k itérations.
Ou r_n est orthogonal à $p_0,\ldots,p_{n-1}$ qui sont n vecteurs linéairement indépendants, car ils sont orthogonaux pour le produit scalaire $(A\,x, y)$, et par conséquent $r_n = 0$. □

Pratiquement, à cause des erreurs d'arrondis, les relations de A-conjugaison ne sont pas exactement vérifiées. On a alors une méthode itérative.

On va d'abord montrer que le facteur de convergence dépend de $K(A)$. Puis on introduira le préconditionnement pour améliorer la convergence.

Théorème 2.3.8 *x_k obtenu à la $k^{ème}$ itération vérifie*

$$E(x_k) \le 4\left(\frac{\sqrt{K(A)}-1}{\sqrt{K(A)}+1}\right)^{2k} E(x_0).$$

2.4 Préconditionnement d'une matrice

2.4.1 Principe

On remplace la résolution de $A\,x = b$ par celle de $C^{-1}\,A\,x = C^{-1}b$. C^{-1} doit être choisi avec l'objectif que $K(C^{-1}\,A) << K(A)$.
En théorie, le meilleur choix est donc $C^{-1} = A^{-1}$. Dans ce cas $K(C^{-1}\,A) = 1$. En pratique, on devra trouver C^{-1} le plus proche de A^{-1}, sans que les calculs de C^{-1} soient trop coûteux.

2.4.2 L'algorithme du gradient conjugué préconditionné

On ne peut appliquer directement l'algorithme du gradient conjugué à $C^{-1}A$, car il faut que $C^{-1}A$ soit symétrique, ce qui est faux en général, même si C^{-1} est symétrique.

Si C^{-1} est symétrique définie positive, on peut définir $C^{-\frac{1}{2}}$ symétrique et définie positive telle que $(C^{-\frac{1}{2}})^2 = C^{-1}$.

Or $C^{\frac{1}{2}}\,(C^{-1}\,A)\,C^{-\frac{1}{2}} = C^{-\frac{1}{2}}\,A\,C^{-\frac{1}{2}}$ est symétrique définie positive. De plus $C^{-1}\,A$ est semblable à $C^{-\frac{1}{2}}\,A\,C^{-\frac{1}{2}}$. Donc, au lieu d'utiliser le système $C^{-1}\,Ax = C^{-1}\,b$, on prend $C^{\frac{1}{2}}\,(C^{-1}\,A)\,C^{-\frac{1}{2}}\,C^{\frac{1}{2}}x = C^{-\frac{1}{2}}\,b$.

On pose $y = C^{\frac{1}{2}}\,x$. On doit alors trouver y tel que $C^{\frac{1}{2}}\,(C^{-1}\,A)C^{-\frac{1}{2}}\,y = C^{-\frac{1}{2}}\,b$.

La méthode du gradient conjugé est appliquée à ce nouveau système de matrice $\widetilde{A} = C^{-\frac{1}{2}}\,A\,C^{-\frac{1}{2}}$, c'est-à-dire :

i. minimiser $\widetilde{E}(y) = \big(\widetilde{A}\,(y-\overline{y}), y-\overline{y}\big)$ où $\overline{y} = C^{\frac{1}{2}}\overline{x}$ est la solution de $\widetilde{A}\,y = C^{-\frac{1}{2}}\,b$.

ii. rendre les directions de descente $\widetilde{A}$-conjuguées.

Or on cherche $\overline{x}$ et non $\overline{y}$. On simplifiera alors l'algorithme.

$$\begin{aligned}
\widetilde{A} &= C^{-\frac{1}{2}}\,A\,C^{-\frac{1}{2}},\\
y_k &= C^{\frac{1}{2}}\,x_k,\\
\tilde{r}_k &= C^{-\frac{1}{2}}\,b - \widetilde{A}\,y_k = C^{-\frac{1}{2}}\,r_k \text{ avec } r_k = b - A\,x_k.
\end{aligned}$$

On pose $\tilde{p}_k = C^{\frac{1}{2}}\,p_k$.

Algorithme appliqué à $\widetilde{A}$	Idem en tenant compte des relations précédentes
$\tilde{\alpha}_k = \dfrac{\parallel \tilde{r}_k \parallel^2}{(\widetilde{A}\,\tilde{p}_k, \tilde{p}_k)}$	$\tilde{\alpha}_k = \dfrac{(C^{-1}r_k, r_k)}{(A\,p_k, p_k)}$
$y_{k+1} = y_k + \tilde{\alpha}_k\,\tilde{p}_k$	$x_{k+1} = x_k + \tilde{\alpha}_k\,p_k$
$\tilde{r}_{k+1} = \tilde{r}_k - \tilde{\alpha}_k\,\widetilde{A}\tilde{p}_k$	$r_{k+1} = r_k - \tilde{\alpha}_k\,A\,p_k$
$\tilde{\beta}_{k+1} = \dfrac{\parallel \tilde{r}_{k+1} \parallel^2}{\parallel \tilde{r}_k \parallel^2}$	$\tilde{\beta}_{k+1} = \dfrac{(C^{-1}r_{k+1}, r_{k+1})}{(C^{-1}r_k, r_k)}$
$\tilde{p}_{k+1} = \tilde{r}_{k+1} + \tilde{\beta}_{k+1}\,\tilde{p}_k$	$p_{k+1} = C^{-1}r_{k+1} + \tilde{\beta}_{k+1}\,p_k$

D'où l'algorithme du gradient conjugué préconditionné :

$$\text{Initialisations :} \begin{cases} x_0 \text{ donné,} \\ r_0 = b - A\,x_0, \\ C\,p_0 = r_0, \\ z_0 = p_0. \end{cases}$$

$$\text{Pour } k = 0, 1, \ldots \begin{cases} \alpha_k = \dfrac{(r_k, z_k)}{(A\,p_k, p_k)}, \\ x_{k+1} = x_k + \alpha_k\,p_k, \\ r_{k+1} = r_k - \alpha_k\,A\,p_k, \\ C\,z_{k+1} = r_{k+1}, \\ \beta_{k+1} = \dfrac{(r_{k+1}, z_{k+1})}{(r_k, z_k)}, \\ p_{k+1} = z_{k+1} + \beta_{k+1}\,p_k. \end{cases}$$

A chaque itération il faut résoudre $C\,z = r$. Il est donc nécessaire que cette résolution soit facile.

On utilisera des préconditionnements tels que $C = T\,T^T$ avec T matrice triangulaire inférieure.

2.4.3 Le préconditionnement SSOR d'Evans

A est décomposée en $A = D - E - E^T$. On prend la matrice de préconditionnement d'Evans :

$$C = \frac{1}{\omega(2-\omega)}(D - \omega E)\,D^{-1}(D - \omega E)^T.$$

ω est un paramètre réel compris entre 0 et 2 ($0 < \omega < 2$).

D est bien définie positive, donc on peut définir $D^{\frac{1}{2}}$. On a $C = T\,T^T$ où $T = \dfrac{(D - \omega E)D^{-\frac{1}{2}}}{\sqrt{\omega(2-\omega)}}$.

Dans le préconditionnement SSOR d'Evans pour le problème du Laplacien sur un carré :

$$\begin{cases} -\Delta u &= f \text{ dans } \Omega =]0,1[\times]0,1[, \\ u &= g \text{ sur } \Gamma \text{ frontière de } \Omega, \end{cases}$$

on a $K(A) = O(\frac{1}{h^2})$ et $K(C^{-1}A) = O(\frac{1}{h})$.

2.4.4 Le préconditionnement basé sur la factorisation incomplète de Cholesky

$A = L\,L^T$.

Les méthodes $IC(n)$.

On commence par $IC(0)$.

Pour calculer T tel que que $C = T\,T^T$ soit voisin de A, on impose a priori la structure de T qui dans la méthode $IC(0)$ est la même que celle de la partie triangulaire inférieure de A, c'est-à-dire

$$t_{ij} = 0 \text{ si } a_{ij} = 0.$$

Pour trouver la valeur de $t_{ij} \neq 0$, on impose la condition

$$(A - T\,T^T)_{ij} = 0 \text{ si } a_{ij} \neq 0.$$

Par exemple, dans le problème du Laplacien dans un carré, A est symétrique pentadiagonale.

$$A = \begin{pmatrix}
\ddots & \ddots & & \ddots & & & & & \\
\ddots & \ddots & \ddots & & \ddots & & & 0 & \\
& \ddots & \ddots & \ddots & & \ddots & & & \\
\ddots & & \ddots & \ddots & \ddots & 0 & \ddots & & \\
& \ddots & & \ddots & \ddots & \ddots & & \ddots & \\
& & \ddots & 0 & \ddots & \ddots & \ddots & & \ddots \\
& & & \ddots & & \ddots & \ddots & \ddots & \\
& 0 & & & \ddots & & \ddots & \ddots & \ddots \\
& & & & & \ddots & & \ddots & \ddots
\end{pmatrix}.$$

$$T = \begin{pmatrix}
\ddots & & & & & & & & \\
\ddots & \ddots & & & & & & & \\
& \ddots & \ddots & & & & & 0 & \\
\ddots & & \ddots & \ddots & & & & & \\
& \ddots & & \ddots & \ddots & & & & \\
& & \ddots & 0 & \ddots & \ddots & & & \\
& & & t_{i,i-m} & & t_{i,i-1} & t_{ii} & & \\
& 0 & & & \ddots & & \ddots & \ddots & \\
& & & & & \ddots & & \ddots & \ddots
\end{pmatrix}.$$

Si on calcule $C = T\,T^T$, alors C a deux diagonales supplémentaires par rapport à A.

$$T\,T^T = \begin{pmatrix}
\ddots & \ddots & & \ddots & \ddots & & & & \\
\ddots & \ddots & \ddots & & \ddots & \ddots & & 0 & \\
& \ddots & \ddots & \ddots & & \ddots & \ddots & & \\
\ddots & & \ddots & \ddots & \ddots & 0 & \ddots & \ddots & \\
\ddots & \ddots & & \ddots & \ddots & \ddots & & \ddots & \ddots \\
& \ddots & \ddots & 0 & \ddots & \ddots & \ddots & & \ddots \\
& & \ddots & \ddots & & \ddots & \ddots & \ddots & \\
& 0 & & \ddots & \ddots & & \ddots & \ddots & \ddots \\
& & & & \ddots & \ddots & & \ddots & \ddots
\end{pmatrix}$$

Donc $R = T\,T^T - A$ est une matrice symétrique qui possède deux diagonales.

$$r_{ij} \neq 0 \text{ si } j = i - m + 1 \text{ et } i + m - 1.$$

L'algorithme du calcul des t_{ij} est très simple dans ce cas. Si on suppose connues les colonnes de T jusqu'à $i-1$, alors la colonne i de T s'obtient par :

$$t_{ii}^2 = a_{ii} - t_{i,i-m}^2 - t_{i,i-1}^2,$$

puis

$$t_{ij} = \frac{a_{ji}}{t_{ii}} \text{ pour } j = i + 1 \text{ et } i + m.$$

Il faut donc exécuter 2 produits, 2 divisions , 2 additions et 1 extraction de racine carrée par colonne.

A cause de l'extraction de la racine carrée, la décomposition incomplète n'existe pas toujours. Si A est a diagonale strictement dominante, elle existe. Sinon on peut modifier la méthode en effectuant la décomposition incomplète de la matrice $A(\alpha) = A + \alpha D$, où D est la matrice diagonale de A. α est un scalaire à choisir.

Les extensions de cette méthode

Si on choisit T dans l'exemple ci-dessus

$$T = \begin{pmatrix} \ddots & & & & & & & & \\ \ddots & \ddots & & & & & & & \\ \ddots & \ddots & \ddots & & & & 0 & & \\ & \ddots & \ddots & \ddots & & & & & \\ \ddots & & \ddots & \ddots & \ddots & & & & \\ & \ddots & & \ddots & \ddots & \ddots & & & \\ & & \ddots & 0 & \ddots & \ddots & \ddots & & \\ & & & \ddots & & \ddots & \ddots & \ddots & \\ & 0 & & & \ddots & & \ddots & \ddots & \ddots \\ & & & & & \ddots & & \ddots & \ddots & \ddots \end{pmatrix}$$

On a rajouté une diagonale supplémentaire par rapport à la structure de la partie triangulaire de A (cette nouvelle matrice T est à comparer à la matrice T obtenue dans la méthode $IC(0)$).

On peut définir de la même façon la méthode $IC(2), \ldots, IC(n)$. Plus n est grand et plus C est proche de A. Mais on aura un coût important des calculs. Dans la pratique, on utilise $n = 0, 1$ ou 2.

Il n'y a aucune estimation théorique de $K(C^{-1}A)$ pour ces méthodes qui, en pratique, se revèlent efficaces.

Les méthodes MIC (n).

Les méthodes MIC sont des méthodes IC légérement modifiées pour fournir de meilleurs conditionnements.

Méthode MIC (0) (Dupont, Kendall, Rachford).

On donne à $R = C - A$ la structure $R = \hat{R} + D$ où D est une matrice diagonale strictement positive dont le choix dépend des conditions aux limites :

(Ex : conditions de type Dirichlet : $D = h^2$ diag (A)).

$\hat{R}$ est une matrice semi-définie négative avec $\sum_j \hat{r}_{ij} = 0$ pour $1 \leq i \leq N$.

$$\hat{R} = \begin{pmatrix} \ddots & & \ddots & & & & \\ & \ddots & & \ddots & & 0 & \\ & & \ddots & & \ddots & & \\ \ddots & & \ddots & 0 & \ddots & & \\ & \ddots & & \ddots & & \ddots & \\ & & \ddots & 0 & \ddots & & \ddots \\ & & \ddots & & \ddots & & \\ & 0 & & \ddots & & \ddots & \\ & & & & \ddots & & \ddots \end{pmatrix}$$

$\hat{R}$ est alors une matrice tridiagonale avec $\hat{r}_{ii} = -(\hat{r}_{i,i-m+1} + \hat{r}_{i,i+m-1})$.

Si on identifie $C = T\,T^T = A + R$, on a :

$C_{ii} = t_{ii}^2 + t_{i,i-1}^2 + t_{i,i-m}^2 = a_{ii}(1+h^2) - \hat{r}_{i,i-m+1} - \hat{r}_{i+m-1,i}$,

$C_{i,i-1} = t_{i,i-1}\,t_{i-1,i-1} = a_{i,i-1}$,

$C_{i,i-m+1} = t_{i,i-m}\,t_{i-m+1,i-m} = \hat{r}_{i,i-m+1}$,

$C_{i,i-m} = t_{i,i-m}\,t_{i-m,i-m} = a_{i,i-m}$.

On en déduit colonne après colonne les éléments de T et de $\hat{R}$:

$t_{ii}^2 = a_{ii}(1+h^2) - \hat{r}_{i,i-m+1} - \hat{r}_{i+m-1,i} - t_{i,i-1}^2 - t_{i,i-m}^2$,

$t_{i+1,i} = \dfrac{a_{i+1,i}}{t_{ii}}$,

$\hat{r}_{i+m-1,i} = t_{i+m-1,i-1}\,t_{i,i-1}$,

$t_{i+m,i} = \dfrac{a_{i+m,i}}{t_{ii}}$.

La factorisation n'est valable que si t_{ii}^2 est positif.

Méthode $MIC(1)$

T a une diagonale supplémentaire par rapport à la partie triangulaire inférieure de A. $MIC(2)$ a deux diagonales supplémentaires.

$$T = \begin{pmatrix} \ddots & & & & & & & & \\ \ddots & \ddots & & & & & & & \\ \ddots & \ddots & \ddots & & & & 0 & & \\ & \ddots & \ddots & \ddots & & & & & \\ \ddots & & \ddots & \ddots & \ddots & & & & \\ & \ddots & & \ddots & \ddots & \ddots & & & \\ & & \ddots & 0 & \ddots & \ddots & \ddots & & \\ & & & \ddots & & \ddots & \ddots & \ddots & \\ & 0 & & & \ddots & & \ddots & \ddots & \ddots \\ & & & & & \ddots & & \ddots & \ddots & \ddots \end{pmatrix}.$$

$$T\,T^T = \begin{pmatrix}
\ddots & \ddots & \ddots & & \ddots & \ddots & & & & & \\
\ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & & 0 & & \\
\ddots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & & & \\
 & \ddots & \ddots & \ddots & \ddots & \ddots & 0 & \ddots & \ddots & & \\
\ddots & & \ddots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots & \\
\ddots & \ddots & & \ddots & \ddots & \ddots & \ddots & \ddots & & \ddots & \ddots \\
 & \ddots & \ddots & 0 & \ddots & \ddots & \ddots & \ddots & \ddots & & \ddots \\
 & & \ddots & \ddots & & \ddots & \ddots & \ddots & \ddots & \ddots & \\
 & 0 & & \ddots & \ddots & & \ddots & \ddots & \ddots & \ddots & \ddots \\
 & & & & \ddots & \ddots & & \ddots & \ddots & \ddots & \ddots \\
 & & & & & \ddots & \ddots & & \ddots & \ddots & \ddots
\end{pmatrix}.$$