

# Introduction à l'Analyse Numérique Matricielle

Thierry Nkaoua  
Université de Marne La Vallée

# Chapitre 1

## Introduction

L'objet de ce cours est une introduction aux méthodes numériques pour la résolution des systèmes linéaires  $Ax = b$ . La plupart du temps,  $A$  sera réelle, mais pas toujours carrée.

On distingue deux grandes familles de méthodes :

- Les méthodes directes : on obtient la solution en un nombre fini d'opérations. Au bout d'un nombre fini d'opérations, on a la solution. Si elle est numériquement “mauvaise”, à cause des erreurs d'arrondis par exemple, on ne peut en général plus rien faire pour l'améliorer. Par contre, on a l'assurance de “terminer”.
- Les méthodes itératives : on construit une suite qui tend vers la solution. Si on n'est pas satisfait du résultat fourni, on peut “continuer”. Cette fois, la qualité de la convergence est primordiale.

# Chapitre 2

## Méthode de Gauss

### 2.1 Exemple

On veut résoudre le système linéaire suivant :

$$\begin{cases} 2x + 3y + z = 1 \\ x + y + z = 2 \\ x - y + 2z = 1 \end{cases}$$

On multiplie la première ligne par  $(-\frac{1}{2})$ , et on l'additionne aux lignes 2 et 3 :

$$\begin{cases} 2x + 3y + z = 1 \\ -\frac{y}{2} + \frac{z}{2} = \frac{3}{2} \\ -\frac{5}{2}y + \frac{3}{2}z = \frac{1}{2} \end{cases}$$

On multiplie la deuxième ligne par  $(-5)$  et on l'additionne à la troisième :

$$\begin{cases} 2x + 3y + z = 1 \\ -y + z = 3 \\ z = 7 \end{cases}$$

A ce stade, on a un système triangulaire, que l'on sait résoudre facilement :

$$\begin{cases} x = -9 \\ y = 4 \\ z = 7 \end{cases}$$

### 2.2 Description

Soit  $A$  la matrice de coefficients  $a_{i,j}$  de dimension  $m \times n$ . Décrivons la méthode de Gauss dans le cas général :

**1<sup>ère</sup> étape**

On garde la première ligne :  $a_{1,j}^{(2)} = a_{1,j}$  pour  $j \in \{1, \dots, n\}$ . On note  $A^{(2)}$  la nouvelle matrice. A la ligne  $i$  :

$$\begin{aligned} a_{i,j}^{(2)} &= a_{i,j} - \frac{a_{i,1}}{a_{1,1}} a_{1,j} \text{ pour } 2 \leq i \leq m \text{ et } 2 \leq j \leq n \\ b_i^{(2)} &= b_i - \frac{a_{i,1}}{a_{1,1}} b_1 \end{aligned}$$

Ici, on suppose qu'à chaque étape,  $a_{k,k}^{(k)} \neq 0$  (c'est le pivot de la méthode de Gauss.)

**$k^{\text{ième}}$  étape**

$$a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} a_{k,j}^{(k)}$$

$$k+1 \leq i \leq m$$

$$k+1 \leq j \leq n$$

Tant que le pivot ne s'annule pas, on continue. A la fin, on obtient un système triangulaire supérieur.

### 2.3 Méthode du pivot partiel

Dans la pratique, il se peut que l'un des pivots de la méthode de Gauss soit nul. Dans ce cas, on cherche dans la même colonne s'il y a une ligne où il n'est pas nul, et on fait une permutation entre les deux lignes. On ne choisit pas n'importe quelle ligne. En effet, dans un ordinateur, plus on divise par des nombres petits, plus on risque de faire des erreurs d'arrondis importantes. On choisit donc le nouveau pivot de manière à minimiser ces erreurs : on prend celui dont la valeur absolue est la plus grande. Echanger les lignes revient à renuméroter les équations. Comme on le verra plus loin, on utilise systématiquement la méthode du pivot partiel.

### 2.4 Méthode du pivot total

On peut aussi envisager de chercher le nouveau pivot dans tout le carré situé sous le pivot. On intervertit donc les lignes et les colonnes si nécessaire. En intervertissant les colonnes, on renumérote les inconnues. Cette méthode n'est pas très utilisée.

### 2.5 Décomposition LU

**Etape 1** : on a : ( $L_1$  carrée  $m \times m$ )

$$A^{(2)} = L_1 A$$

avec :

$$L_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -\frac{a_{2,1}}{a_{1,1}} & 1 & & \vdots \\ \vdots & & \ddots & 0 \\ -\frac{a_{m,1}}{a_{1,1}} & 0 & \dots & 1 \end{pmatrix}$$

$A^{(2)}$  a la même première ligne que  $A = A^{(1)}$ .

**Etape k** :

$$A^{(k+1)} = L_k A^{(k)}$$

avec :

$$L_k = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 0 & \ddots & & & \vdots \\ \vdots & & 1 & & \vdots \\ \vdots & & -\frac{a_{k+1,k}^{(k)}}{a_{k,k}^{(k)}} & \ddots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ \vdots & & \vdots & & \vdots \\ 0 & -\frac{a_{m,k}^{(k)}}{a_{k,k}^{(k)}} & 0 & \cdots & 1 \end{pmatrix}$$

$A^{(k+1)}$  a les même  $k$  premières lignes que  $A^{(k)}$ .

D'où à la fin :

$$A^{(m)} = L_{m-1}L_{m-2}\dots L_1A$$

avec  $A^{(m)}$  triangulaire supérieure. On pose :

$$U = A^{(m)}$$

et on a par un calcul élémentaire :

$$L_{m-1}\dots L_1 = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -\frac{a_{2,1}^{(1)}}{a_{1,1}^{(1)}} & \ddots & & & \vdots \\ \vdots & \ddots & 1 & & \vdots \\ \vdots & & -\frac{a_{k+1,k}^{(k)}}{a_{k,k}^{(k)}} & \ddots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ -\frac{a_{m,1}^{(1)}}{a_{1,1}^{(1)}} & \cdots & -\frac{a_{m,k}^{(k)}}{a_{k,k}^{(k)}} & 0 & \cdots & 1 \end{pmatrix}$$

Cette matrice est carrée inversible. On peut donc poser :

$$A = LU$$

avec  $L = L_1^{-1}\dots L_{m-1}^{-1}$  qu'un calcul élémentaire prouve être :

$$L = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ \frac{a_{2,1}^{(1)}}{a_{1,1}^{(1)}} & \ddots & & & \vdots \\ \vdots & \ddots & 1 & & \vdots \\ \vdots & & \frac{a_{k+1,k}^{(k)}}{a_{k,k}^{(k)}} & \ddots & 0 \\ \vdots & & \vdots & \ddots & \vdots \\ \frac{a_{m,1}^{(1)}}{a_{1,1}^{(1)}} & \cdots & \frac{a_{m,k}^{(k)}}{a_{k,k}^{(k)}} & 0 & \cdots & 1 \end{pmatrix}$$

**Theorem 1** *Si les pivots de la méthode de Gauss ne s'annulent pas, il existe  $L$  triangulaire inférieure avec une diagonale de 1 et  $U$  triangulaire supérieure telles que  $A = LU$ .*

A présent, regardons quand les pivots s'annulent. On note  $\tilde{A}_k$  une sous-matrice principale d'ordre  $k$  de  $A$  :

$$\begin{pmatrix} a_{1,1} & \cdots & a_{1,k} \\ \vdots & & \vdots \\ a_{k,1} & \cdots & a_{k,k} \end{pmatrix}$$

Si on utilise la méthode de Gauss sur  $\tilde{A}_k$ , on fait les mêmes opérations que pour  $A$ . Posons

$$\begin{aligned} \tilde{A}_k &= \tilde{L}_k \tilde{U}_k \\ \tilde{U}_k &= \begin{pmatrix} a_{1,1} & & & \alpha \\ & a_{2,2}^{(2)} & & \\ & & \ddots & \\ 0 & & & a_{k,k}^{(k)} \end{pmatrix} \end{aligned}$$

On regarde si  $\tilde{A}_k$  est inversible :

$$\begin{aligned} \det \tilde{A}_k &= \det \tilde{L}_k \det \tilde{U}_k \\ &= \prod_{i=1}^k a_{i,i}^{(i)} \end{aligned}$$

donc le pivot à l'étape  $k+1$  est non nul ssi tous les pivots des étapes précédentes sont non nuls, c'est à dire ssi  $\det \tilde{A}_k \neq 0$ , c'est à dire ssi  $\tilde{A}_k$  inversible. On a alors le théorème suivant :

**Theorem 2** *Les pivots de la méthode de Gauss ne s'annulent pas ssi toutes les matrices  $\tilde{A}_k$  sont inversibles.*

Si l'on dispose de  $A$  sous la forme  $LU$ , alors pour résoudre le système  $Ax = b$ , il suffit de résoudre :

$$\begin{cases} Ly = b \\ Ux = y \end{cases}$$

La résolution de  $Ax = b$  se ramène donc à la résolution de deux systèmes triangulaires. Si l'on doit résoudre plusieurs systèmes linéaires de même matrice, il suffit de faire une fois pour toutes la décomposition  $LU$  de  $A$ .

**Theorem 3** *La décomposition  $LU$  d'une matrice  $A$  inversible est unique.*

**Remark 1** *Il est implicitement dit que  $L$  est triangulaire inférieure avec une diagonale de 1 et  $U$  triangulaire supérieure.*

**Démonstration**

Soit

$$A = L_1 U_1 = L_2 U_2$$

avec  $L_1$  et  $L_2$  triangulaires inférieures de diagonale unité, et  $U_1$  et  $U_2$  triangulaires supérieures. Les  $L$  et  $U$  sont inversibles. On peut donc écrire :

$$\begin{aligned} L_2^{-1} L_1 U_1 &= U_2 \\ L_2^{-1} L_1 &= U_2 U_1^{-1} = I \end{aligned}$$

car  $L_2^{-1} L_1$  est triangulaire inférieure et  $U_2 U_1^{-1}$  est triangulaire supérieure, ce sont donc des matrices diagonales or les  $L$  ont pour diagonale  $I$ .



## 2.6 Stabilité numérique

On veut savoir si de petites erreurs d'arrondis peuvent entraîner des erreurs importantes sur la solution.

**Exemple 1** soit  $0 < \varepsilon \ll 1$

$$(S) \begin{cases} \varepsilon x_1 + x_2 & = 1 \\ x_1 + x_2 & = 2 \end{cases}$$

$$A = \begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix}$$

Décomposition LU de A

$$L = \begin{pmatrix} 1 & 0 \\ \frac{1}{\varepsilon} & 1 \end{pmatrix}$$

$$U = \begin{pmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{1}{\varepsilon} \end{pmatrix}$$

$$(S) \Leftrightarrow \begin{cases} \varepsilon x_1 + x_2 & = 1 \\ (1 - \frac{1}{\varepsilon})x_2 & = 2 - \frac{1}{\varepsilon} \end{cases} \Leftrightarrow \begin{cases} x_1 & = \frac{1-x_2}{\varepsilon} \\ x_2 & = \frac{1-2\varepsilon}{1-\varepsilon} \end{cases}$$

Si la précision de la machine est  $10^{-15}$  et  $\varepsilon = 10^{-20}$ , l'ordinateur trouve comme solution :  $x_2 = 1$ . On reporte dans la première équation :  $x_1 = 0$ . Ce qui voudrait dire que  $1 + 0 = 2$  ! (deuxième équation de (S)).

Recommençons donc avec la méthode du pivot partiel

On renumérote les équations :

$$(S') \begin{cases} x_1 + x_2 & = 2 \\ \varepsilon x_1 + x_2 & = 1 \end{cases}$$

$$A' = \begin{pmatrix} 1 & 1 \\ \varepsilon & 1 \end{pmatrix}$$

$$(S') \Leftrightarrow \begin{cases} x_1 + x_2 & = 2 \\ (1 - \varepsilon)x_2 & = 1 - 2\varepsilon \end{cases} \Leftrightarrow \begin{cases} x_1 & = 2 - x_2 \\ x_2 & = \frac{1-2\varepsilon}{1-\varepsilon} \end{cases}$$

Avec les mêmes hypothèses, on trouve :  $x_2 = 1$  et  $x_1 = 1$ , ce qui est une solution tout à fait acceptable !

Conclusion : dans la pratique, on utilise la méthode du pivot partiel.

**Exercice 1** Démontrez que le nombre d'opérations de la méthode de Gauss est en  $n^3$  et que le nombre d'opérations de la résolution d'un système triangulaire est en  $n^2$ .

## Chapitre 3

# Matrices symétriques définies positives

### 3.1 Rappels

**Theorem 4**  $\text{Ker}(A^t) = (\text{Im } A)^\perp$

**Démonstration**

$$\begin{aligned}x &\in \text{Ker}(A^t) \\ &\Leftrightarrow Ax = 0\end{aligned}$$

Mais on a le théorème. : Un vecteur est nul  $\Leftrightarrow$  il est orthogonal à tous les vecteurs, d'où :

$$\begin{aligned}x &\in \text{Ker}(A^t) \Leftrightarrow \forall y \in R^m, ({}^tAx, y) = 0 \\ &\Leftrightarrow \forall y \in R^m, (x, Ay) = 0\end{aligned}$$

$$\begin{aligned}\text{car } (Ax, y) &= (x, {}^tAy) \\ &\Leftrightarrow x \in (\text{Im } A)^\perp\end{aligned}$$

■

**Definition 1** Une matrice carrée est symétrique définie positive ssi  $A^t = A$  et  $\forall x \neq 0, (Ax, x) > 0$

**Theorem 5** Toute matrice symétrique n'a que des valeurs propres réelles et est diagonalisable en base orthonormée, c'est à dire qu'il existe une matrice  $P$  et une matrice  $D$  diagonale telles que  $D = P^{-1}AP$  ( $P$  orthogonale  $\Rightarrow P^t = P^{-1}$  d'où  $D = {}^tPAP$ .)

**Theorem 6** Soit  $A$  une matrice symétrique, alors elle est définie positive ssi ses valeurs propres sont  $> 0$ .

**Definition 2**

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2}$$

où à droite la norme est la norme 2 sur  $R^n$ . C'est la norme induite par la norme 2 de  $R^n$ .

**Definition 3** On définit de même les normes induite  $\infty$  ou 1 sur les matrices carrées comme les normes induites par les normes équivalentes sur  $R^n$ .

**Definition 4** On dit qu'une norme sur les matrices est matricielle ssi :

$$\|AB\| \leq \|A\| \|B\|$$

**Theorem 7** Les normes induites sont des normes matricielles.

**Definition 5** Pour  $A$  inversible on définit son conditionnement par :

$$\text{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2$$

**Theorem 8**

$$\|A\|_2 = \sqrt{\rho(A^t A)}$$

où  $\rho(M)$  désigne le rayon spectral de  $M$  (le plus grand module des valeurs propres de  $M$ )

**Démonstration**

$A^t A$  est symétrique. Donc ses valeurs propres sont réelles et  ${}^t A A$  est diagonalisable en base orthonormée :

$$\exists D = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} \text{ et } P \text{ orthogonale}$$

telles que :

$$PDP^t = A^t A$$

On a alors :

$$\begin{aligned} \frac{(Ax, Ax)}{(x, x)} &= \frac{({}^t A A x, x)}{(x, x)} = \frac{(PDP^t x, x)}{(x, x)} = \frac{(DP^t x, P^t x)}{(x, x)} \\ &= \frac{\sum_i \lambda_i x_i'^2}{\sum_i x_i'^2} \end{aligned}$$

où les  $x_i'$  sont les coordonnées de  $x$  dans la base de diagonalisation ( $x' = P^t x$ ). Tous les  $\lambda_i$  sont positifs puisque si  ${}^t A A u = \lambda u$ , alors  $({}^t A A u, u) = (\lambda u, u) = \lambda \|u\|^2 = (A u, A u) \geq 0$ . D'où :

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} \leq \max_i \lambda_i = \rho(A^t A)$$

et

$$\|A\|_2^2 \leq \rho(A^t A)$$

Comme pour  $x = (\delta_{i,k})$  où  $k$  est l'indice tel que  $|\lambda_k| = \rho(A^t A)$  on a

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \rho(A^t A)$$

on obtient bien :

$$\|A\|_2 = \sqrt{\rho(A^t A)}$$

■

**Theorem 9** Si  $A$  symétrique  $\|A\|_2 = \rho(A)$

**Démonstration**

$$\|A\|_2^2 = \rho(A^2) = \rho(A)^2$$

■

**Theorem 10** Pour  $A$  symétrique définie positive on a :

$$\text{cond}_2(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$$

### 3.2 Décomposition de Cholesky

**Theorem 11** Pour  $A$  symétrique les propositions suivantes sont équivalentes :

- (1)  $\forall x \neq 0, (Ax, x) > 0$
- (2) les matrices principales  $\tilde{A}_k$  ont toutes un déterminant  $> 0$
- (3) tous les pivots de LU sont  $> 0$
- (4)  $\exists B$  triangulaire inf. inversible tq  $A = B^t B$  (Cholesky)

**Démonstration**

(1)  $\Rightarrow$  (2)

Soit

$$x = \begin{pmatrix} x_k \\ 0 \end{pmatrix}$$

et

$$A = \begin{pmatrix} \tilde{A}_k & C_k \\ {}^t C_k & B_k \end{pmatrix}$$

$$(Ax, x) = (\tilde{A}_k x_k, x_k) > 0$$

d'où  $\tilde{A}_k$  symétrique définie positive, d'où ses valeurs propres sont  $> 0$ , d'où  $\det A > 0$

(2)  $\Rightarrow$  (3)

Si on a :

$$\det \tilde{A}_k > 0$$

à chaque étape de Gauss, on a

$$\tilde{A}_k = \tilde{L}_k \tilde{U}_k$$

avec

$$\det \tilde{L}_k = 1$$

d'où

$$\det \tilde{A}_k = \det \tilde{U}_k$$

or les pivots sont les termes de la diagonale de  $U_k$ . Par récurrence, tous les pivots sont  $> 0$ .

(3)  $\Rightarrow$  (4)

Les pivots de  $LU$  sont  $> 0$ . Donc Gauss ne s'arrête pas. On applique donc la méthode de Gauss sans pivot :

$$A = LU$$

On n'a que des 1 dans  $\text{diag}(L)$ . La diagonale de  $U$  est  $> 0$  (elle contient les pivots). Posons :

$$\Lambda = \sqrt{\text{diag}(U)} = \begin{pmatrix} \sqrt{a_{1,1}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{a_{n,n}} \end{pmatrix}$$

$$A = (L\Lambda)(\Lambda^{-1}U)$$

Posons :

$$B = L\Lambda \text{ et } C = \Lambda^{-1}U$$

et montrons que

$$C = B^t$$

On a :

$$A = BC$$

$A$  symétrique donc

$$C^t B^t = BC$$

et donc

$$B^{-1}C^t = CB^{-t}$$

$B^{-1}C^t$  est triangulaire inférieure et  $CB^{-t}$  est triangulaire supérieure donc  $B^{-1}C^t$  et  $CB^{-t}$  sont diagonales et

$$\text{diag}(B^{-t}) = \begin{pmatrix} \frac{1}{\sqrt{a_{1,1}}} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\sqrt{a_{n,n}}} \end{pmatrix} \text{ et } \text{diag}(C) = \begin{pmatrix} \sqrt{a_{1,1}} & & 0 \\ & \ddots & \\ 0 & & \sqrt{a_{n,n}} \end{pmatrix}$$

donc

$$CB^{-t} = I \text{ et } C = B^t \text{ d'où } B = C$$

(4)  $\Rightarrow$  (1)

Si

$$A = BB^t$$

pour  $x \neq 0$  on a :

$$(Ax, x) = (BB^t x, x) = (B^t x, B^t x) = \|B^t x\|_2^2 > 0 \text{ (si } x \neq 0)$$

■

**Corollary 12** Pour une matrice symétrique définie positive, la méthode de Gauss ne s'arrête pas .

**Definition 6** La décomposition d'une matrice symétrique définie positive sous la forme  $BB^t$  avec  $B$  triangulaire inférieure et  $\text{diag}(B) > 0$  s'appelle décomposition de Cholesky de  $A$ .

**Theorem 13** La décomposition de Cholesky est unique.

**Démonstration**

Remarque préliminaire

Si on a

$$A = B^t B$$

On pose

$$\Delta = \text{diag} B$$

$$A = \underbrace{(B\Delta^{-1})}_L \underbrace{(\Delta^t B)}_U$$

Comme  $L$  n'a que des 1 sur la diagonale, l'expression précédente est "la" décomposition  $LU$  de  $A$ .

Supposons :

$$A = B_1 B_1^t = B_2 B_2^t$$

avec

$$\Delta_1 = \text{diag}(B_1) > 0 \text{ et } \Delta_2 = \text{diag}(B_2) > 0$$

On écrit

$$A = (B_1 \Delta_1^{-1})(\Delta_1 B_1^t) = (B_2 \Delta_2^{-1})(\Delta_2 B_2^t)$$

d'après la remarque préliminaire :

$$\begin{cases} B_1 \Delta_1^{-1} &= B_2 \Delta_2^{-1} \\ \Delta_1 B_1^t &= \Delta_2 B_2^t \end{cases}$$

et donc

$$(\Delta_1 B_1^t)_{i,i} = (B_1)_{i,i}^2 = (B_2)_{i,i}^2$$

or

$$(B_1)_{i,i} > 0 \text{ et } (B_2)_{i,i} > 0$$

d'où

$$(B_1)_{i,i} = (B_2)_{i,i}$$

et donc

$$\Delta_1 = \Delta_2$$

d'où enfin

$$B_1 = B_2$$

■

### 3.3 Calcul de la décomposition de Cholesky

On peut calculer autrement la décomposition de Cholesky d'une matrice  $A$  symétrique définie positive que par la méthode de Gauss.

Soit  $A = BB^t$ , avec  $B$  triangulaire inférieure et  $\text{diag} B > 0$ .  $B = (b_{i,j})$ . On a :

$$a_{i,j} = \sum_{k=1}^n b_{i,k} b_{j,k}$$

On va calculer les  $b_{ij}$  par colonne.

**Colonne 1**

$$\begin{aligned} a_{1,1} &= b_{1,1}^2 \\ a_{i,1} &= b_{i,1} b_{1,1} \end{aligned}$$

d'où

$$\begin{aligned} b_{11} &= \sqrt{a_{1,1}} \\ b_{i1} &= \frac{a_{i,1}}{b_{1,1}} \end{aligned}$$

Supposons qu'on ait calculé les  $j-1$  premières colonnes de  $B$ .

**Colonne  $j$**

$$a_{i,j} = \sum_{k=1}^j b_{i,k} b_{j,k}$$

D'où pour  $i = j$  :

$$a_{j,j} = \sum_{k=1}^j b_{j,k}^2$$

et

$$b_{j,j} = \sqrt{a_{j,j} - \sum_{k=1}^{j-1} b_{j,k}^2}$$

et les autres termes s'obtiennent par :

$$b_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} b_{i,k} b_{j,k}}{b_{j,j}}$$

**Exercice 2** Démontrez que le nombre d'opérations de la décomposition de Cholesky est en  $n^2$  (il faut donc utiliser ces formules pour calculer la décomposition de Cholesky et non la méthode de Gauss qui est en  $n^3$ )

### 3.4 Inégalité de Kantorovitch

**Theorem 14** Pour  $A$  symétrique définie positive, on a pour tout  $x \neq 0$  :

$$1 \leq \frac{(Ax, x)(A^{-1}x, x)}{(x, x)^2} \leq \frac{(1 + \text{cond}_2(A))^2}{4\text{cond}_2 A}$$

et les bornes sont optimales (elles sont atteintes)

**Démonstration**

Soit

$$0 < \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n \text{ les valeurs propres de } A$$

Soit

$$p(\lambda) = \lambda^2 - (\lambda_1 + \lambda_n)\lambda + \lambda_1\lambda_n$$

Comme  $p(\lambda_1) = p(\lambda_n) = 0$ , alors

$$\forall i, p(\lambda_i) \leq 0$$

Soit

$$C = A^2 - (\lambda_1 + \lambda_n)A + \lambda_1\lambda_n Id$$

(remarque :  $vp(A + I) = vp(A) + 1$ ) Les valeurs propres de  $C$  sont les  $p(\lambda_i)$  qui sont donc négatifs. Les valeurs propres de

$$B = A^{-1}C = A - (\lambda_1 + \lambda_n)Id + \lambda_1\lambda_n A^{-1}$$

sont donc aussi  $\leq 0$  ( $C$  et  $A^{-1}$  sont diagonalisables dans la même base et commutent, le calcul des valeurs propres de  $B$  en fonction de celle de  $A$  est immédiat). Donc on a :

$$(Bx, x) = (Ax, x) - (\lambda_1 + \lambda_n)(x, x) + \lambda_1\lambda_n(A^{-1}x, x) \leq 0$$

Pour  $x$  fixé non nul, soit la fonction  $f$  définie par :

$$f(\lambda) = \lambda^2(Ax, x) - (\lambda_1 + \lambda_n)(x, x)\lambda + \lambda_1\lambda_n(A^{-1}x, x)$$

On a

$$f(0) = \lambda_1\lambda_n(A^{-1}x, x) > 0$$

et

$$f(1) = (Bx, x) \leq 0$$

donc le trinôme  $f$  a au moins une racine et donc  $\Delta \geq 0$

$$\Delta = (\lambda_1 + \lambda_n)^2(x, x)^2 - 4(Ax, x)\lambda_1\lambda_n(A^{-1}x, x) \geq 0$$

d'où

$$\frac{(Ax, x)(A^{-1}x, x)}{(x, x)^2} \leq \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n} = \frac{(1 + \text{cond}_2(A))^2}{4\text{cond}_2(A)}$$

Montrons qu'il peut y avoir égalité dans cette inégalité. Soit

$$x = V_1 + V_n$$

avec  $V_1$  et  $V_n$  2 vecteurs propres de norme 1 associés à  $\lambda_1$  et  $\lambda_n$ . On sait que

$$(V_1, V_n) = 0$$

donc

$$(Ax, x) = (\lambda_1 V_1 + \lambda_n V_n, V_1 + V_n) = \lambda_1 + \lambda_n$$

$$(A^{-1}x, x) = \frac{1}{\lambda_1} + \frac{1}{\lambda_n}$$

$$(x, x) = (V_1 + V_n, V_1 + V_n) = 2$$

d'où

$$\begin{aligned} \frac{(Ax, x)(A^{-1}x, x)}{(x, x)^2} &= \frac{(\lambda_1 + \lambda_n)\left(\frac{1}{\lambda_1} + \frac{1}{\lambda_n}\right)}{4} \\ &= \frac{(\lambda_1 + \lambda_n)^2}{4\lambda_1\lambda_n} \end{aligned}$$

### Deuxième partie de l'inégalité

On se place dans la base orthonormale de diagonalisation de  $A$ .

$$\frac{(Ax, x)(A^{-1}x, x)}{(x, x)^2} = \frac{(\sum_i \lambda_i x_i^2)(\sum_i \frac{1}{\lambda_i} x_i^2)}{(\sum_i x_i^2)^2}$$

$$\left(\sum_i x_i^2\right)^2 = \left(\sum_i \left(\frac{x_i}{\sqrt{\lambda_i}}\right)(x_i\sqrt{\lambda_i})\right)^2 \leq \left(\sum_i \frac{x_i^2}{\lambda_i}\right)\left(\sum_i x_i^2\lambda_i\right)$$

par Cauchy Schwarz, d'où l'inégalité. On sait enfin qu'il peut y avoir égalité dans cette inégalité en prenant les vecteurs quand les vecteurs  $\left(\frac{x_i}{\sqrt{\lambda_i}}\right)$  et  $(x_i\sqrt{\lambda_i})$  sont colinéaires.

■

## Chapitre 4

# Irréductibilité

**Definition 7** *A matrice carrée est réductible ssi*

$$\exists S \subset [1, n], S \neq \emptyset, S \neq [1, n]$$

tel que

$$\forall i \in S, \forall j \in [1, n] \setminus S, a_{ij} = 0$$

**Definition 8** *A est irréductible ssi elle n'est pas réductible.*

**Theorem 15** *A est réductible ssi  $\exists P$  matrice de permutation tel que*

$$PAP^t = \begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{pmatrix}$$

**Démonstration**

$\Rightarrow$

A réductible. Alors il existe  $S$  et  $T$  tels que :

$$\begin{aligned} T &= \{i_1, i_2, \dots, i_r\}, S = \{i_{r+1}, \dots, i_n\}, S \cup T = [1, n] \\ \forall i \in S, \forall j \in T, a_{ij} &= 0 \end{aligned}$$

Soit  $(e_i)$  la base canonique de  $R^n$ . Soit  $P$  la matrice de permutation dont les vecteurs colonnes sont  $(e_{i_1}, e_{i_2}, \dots, e_{i_n})$ . Calculons  $M = PAP^t$ .

$$(AP^t)_{ij} = \sum_{k=1}^n a_{ik} P_{jk}$$

or

$$P_{jk} = 0 \text{ sauf si } k = i_j$$

d'où

$$(AP^t)_{ij} = a_{i i_j}$$

$$M_{ij} = (PA^tP)_{ij} = \sum_{k=1}^n P_{ik}(A^tP)_{k,j} = \sum_{k=1}^n P_{ik}a_{ki_j}$$

$$M_{ij} = a_{i i_j}$$

Or si  $i \in [r + 1, n]$  et  $j \in [1, r]$  alors  $i_i \in T$  et  $i_j \in S$  et donc  $M_{ij} = 0$ , ce qui correspond bien à la forme annoncée de  $PAP^t$

←

Si on a

$$\begin{aligned} A &= P^t M P \\ M &= \begin{pmatrix} A_{1,1} & A_{1,2} \\ 0 & A_{2,2} \end{pmatrix} \end{aligned}$$

avec  $P_{ij} = \delta_{s(i)j} = \delta_{is^{-1}(j)}$  où  $s$  est la permutation de  $[1, n]$  associée à  $P$ . On a alors :

$$\begin{aligned} (MP)_{ij} &= \sum_{k=1}^n M_{ik} \delta_{ks^{-1}(j)} = M_{is^{-1}(j)} \\ a_{ij} &= (P^t M P)_{ij} = \sum_{k=1}^n \delta_{ks^{-1}(i)} M_{ks^{-1}(j)} = M_{s^{-1}(i)s^{-1}(j)} \end{aligned}$$

Comme pour  $i = r + 1, \dots, n$  et  $j = 1, \dots, r$ ,  $M_{ij} = 0$ , si on pose  $S = s\{r + 1, \dots, n\}$  et  $T = s\{1, \dots, r\}$ , on a  $S \cup T = [1, n]$ ,  $S \cap T = \emptyset$  et  $\forall i \in S, \forall j \in T, s^{-1}(i) \in \{r + 1, \dots, n\}$  et  $s^{-1}(j) \in \{1, \dots, r\}$  et donc  $a_{ij} = 0$ . Et donc  $A$  est réductible.

■

**Remark 2** *Tout cela n'est pas "utilisable" dans la pratique pour des matrices de taille importante.*

## 4.1 Graphe orienté

A matrice carrée d'ordre  $n$ .  $P_1, \dots, P_n$   $n$  points distincts du plan. On joint  $P_i$  à  $P_j$  avec une flèche si  $a_{ij} \neq 0$

**Definition 9** *On appelle graphe orienté (ou direct) de  $A$  l'ensemble de flèches ainsi obtenues (sans tracer pas  $\overrightarrow{P_i P_i}$ )*

**Definition 10** *On dit qu'un graphe est fortement connexe si pour deux points quelconques du graphe, il existe un chemin orienté qui va de l'un à l'autre.*

**Theorem 16 (Varga)**  *$A$  irréductible  $\Leftrightarrow$  le graphe de  $A$  est fortement connexe.*

### Démonstration

⇒

$A$  irréductible,  $i_0$  donné, on considère l'ensemble :

$$T = \{j \neq i_0, P_{i_0} \text{ connecté à } P_j\}$$

On veut montrer que  $T$  est l'ensemble de tous les indices.

Si  $T$  était vide,  $P_{i_0}$  ne serait connecté à rien et donc  $\forall j \neq i_0, a_{i_0,j} = 0$ , ce qui contredit l'irréductibilité de  $A$  en prenant  $S = \{i_0\}$  et  $T$ . Donc  $T$  non vide.

Soit  $S = N_n \setminus T$

Supposons  $S \neq \emptyset$ . On sait déjà que  $T$  est aussi non vide. Soit  $j \in T, k \in S$ . Supposons  $a_{jk} \neq 0$ . Comme  $j \in T, P_{i_0}$  est connecté à  $P_j$ . Comme  $a_{jk} \neq 0$ , on a aussi  $P_j$  connecté à  $P_k$  et donc  $P_{i_0}$  est connecté à  $P_k$ , ce qui signifie que  $k \in T$ , ce qui est impossible. Donc  $S = \emptyset$ , et donc  $T = [1, n]$  ce qui signifie bien que le

graphe est connexe.

⇐

Si le graphe est connexe.

Si  $A$  est réductible,

$$\exists S \neq \emptyset, T \neq \emptyset, S \cap T = \emptyset, \forall j \in S, \forall k \in T, a_{j,k} = 0$$

Soit  $j \in S, k \in T$  fixés. Comme le graphe est supposé connexe, il y a un chemin de  $P_j$  à  $P_k$  :

$$\overrightarrow{P_{j,i_1}} \dots \overrightarrow{P_{i_{r-1},k}}$$

$j \in S$  donc  $i_1 \in S$ . De même,  $i_2 \in S$  et de proche en proche  $k \in S$  ce qui est absurde, donc  $A$  est irréductible.

■

**Example 2**

$$A = \begin{pmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{pmatrix}$$

$$\begin{array}{ccccccc} P_1 & & P_2 & & & & P_n \\ \otimes & \rightleftharpoons & \otimes & \rightleftharpoons & \dots & \rightleftharpoons & \otimes \end{array}$$

Graphe connexe  $\Rightarrow A$  irréductible.

**Remark 3** Si  $A$  est symétrique, on ne met plus deux flèches mais un trait non orienté.

# Chapitre 5

## Les valeurs propres

**Theorem 17** *Gershgorin*

$$C_i = \{z \in \mathbb{C}, |z - a_{i,i}| \leq \sum_{j \neq i} |a_{i,j}|\}$$

$$C'_i = \{z \in \mathbb{C}, |z - a_{i,i}| \leq \sum_{i \neq j} |a_{i,j}|\}$$

$$\text{alors } Sp(A) \subset \left( \bigcup_i C_i \right) \cap \left( \bigcup_i C'_i \right)$$

### Démonstration

$\lambda$  valeur propre,  $x$  vecteur propre  $\neq 0$

$$Ax = \lambda x$$

$$\forall i, \sum_{j=1}^n a_{i,j} x_j = \lambda_i x_i$$

$$\forall i, (\lambda_i - a_{i,i}) x_i = \sum_{j \neq i} a_{i,j} x_j$$

$$\forall i, |\lambda_i - a_{i,i}| |x_i| \leq \sum_{j \neq i} |a_{i,j}| |x_j|$$

$\exists i_0$  tq

$$|x_{i_0}| = \max_i |x_i| \neq 0$$

Pour  $i = i_0$

$$|\lambda_{i_0} - a_{i_0,i_0}| \leq \sum_{j \neq i_0} |a_{i_0,j_0}|$$

$$\Rightarrow \lambda \in C_{i_0} \subset \bigcup_i C_i$$

$\lambda \in \bigcup_i C'_i$  puisque les valeurs propres de  $A^t$  sont les mêmes que celles de  $A$ .

■

## 5.1 Application à la localisation des zéros d'un polynôme

Soit  $P$  le polynôme :

$$P(X) = X^n + a_1 X^{n-1} + \dots + a_n$$

On lui associe la matrice compagnon

$$A = \begin{pmatrix} 0 & & & -a_n \\ 1 & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \ddots & 0 \\ 0 & & & 1 & -a_1 \end{pmatrix}$$

dont on peut vérifier que le polynôme caractéristique est :

$$\text{Det}(XI - A) = \begin{vmatrix} X & & & a_n \\ -1 & \ddots & & \vdots \\ & \ddots & \ddots & \vdots \\ & & \ddots & X & a_2 \\ 0 & & & -1 & X + a_1 \end{vmatrix} = P(X)$$

Les racines de  $P$  sont donc les valeurs propres de  $A$ . On applique le théorème de Gershorin à  $A$  : Soient  $z$  une racine de  $P$

- Sur les lignes :

$$|z| \leq |a_n|$$

ou

$$|z| \leq 1 + \max_{i \in \{2, \dots, n-1\}} |a_i|$$

ou

$$|z + a_1| \leq 1$$

et

- Sur les colonnes :

$$|z| \leq 1$$

ou

$$|z + a_1| \leq \sum_{i \neq 1} |a_i|$$

### Exemple 3

$$z^3 + 2z^2 - 3z - 1 = 0$$

$$|z| \leq 1$$

ou

$$|z| \leq 1 + 3 = 4$$

ou

$$|z + 2| \leq 1$$

ou

$$|z| \leq 1$$

ou

$$|z + 2| \leq 4$$

# Chapitre 6

## Méthodes itératives

### 6.1 Généralités

On ne considère que des matrices carrées.

**Theorem 18** *Les normes induites sont des normes matricielles. Il n'y a pas de réciproque, la norme de Frobenius est matricielle mais pas induite :*

$$\sqrt{\sum |a_{i,j}|^2} = \|A\|_F \text{ (norme de Frobenius)}$$

**Theorem 19** *A carrée,  $\|\cdot\|$  matricielle.*

- (i)  $\rho(A) \leq \|A\|$
- (ii)  $\forall \varepsilon > 0, \exists \|\cdot\|$  induite tel que  $\|A\| \leq \rho(A) + \varepsilon$

#### Démonstration

(i) Soit  $\lambda$  valeur propre de  $A$  tel que

$$\rho(A) = |\lambda|$$

$$\exists u \neq 0, Au = \lambda u$$

alors

$$\exists v \in \mathbf{R}^n / uv^t \neq 0$$

par exemple  $v = u$  convient. On peut écrire :

$$\|Auv^t\| \leq \|A\| \|uv^t\|$$

puisque la norme est matricielle. Mais on a aussi :

$$\|Auv^t\| = \|\lambda uv^t\| = |\lambda| \|uv^t\|$$

et donc

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

(ii) Soit  $\varepsilon > 0$  fixé. Toute matrice complexe étant trigonalisable en base orthonormée  $\exists U$  orthogonale telle que

$$U^{-1}AU = \begin{pmatrix} \lambda_1 & & t_{i,j} \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix} = T$$

Soit  $\delta > 0$  pas encore fixé et soit :

$$D_\delta = \text{Diag}(1, \delta, \dots, \delta^{n-1})$$

On a :

$$D_\delta^{-1}TD_\delta = \begin{pmatrix} \lambda_1 & \delta t_{1,2} & \delta^2 t_{1,3} & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & \delta^2 t_{n-2,n} \\ & & & \ddots & \delta t_{n-1,n} \\ 0 & & & & \lambda_n \end{pmatrix}$$

plus précisément :

$$(D_\delta^{-1}TD_\delta)_{i,j} = \delta^{j-i} t_{i,j} = (D_\delta^{-1}U^{-1}AUD_\delta)_{i,j}$$

On peut choisir  $\delta$  suffisamment petit pour que

$$\forall i \sum_{j=i+1}^n |\delta^{j-i} t_{i,j}| \leq \varepsilon$$

On fixe à présent  $\delta$  vérifiant l'inégalité ci dessus. Soit la norme  $N$  telle que

$$N(X) = \|(UD_\delta)^{-1}X(UD_\delta)\|_\infty$$

On peut montrer aisément que  $N$  est une norme et que  $N$  est induite par la norme sur  $R^n$  :

$$n(v) = \|(UD_\delta)^{-1}v\|_\infty$$

Or on sait que :

$$\|Z\|_\infty = \max_i \sum_{j=1}^n |Z_{ij}|$$

et donc :

$$N(A) = \max_i \left( \sum_{j=i+1}^n |\delta^{j-i} t_{i,j}| + |\lambda_i| \right) \leq \varepsilon + \rho(A)$$

■

**Theorem 20**  $B$  matrice carrée. Les propriétés suivantes sont équivalentes :

- (i)  $B^k \rightarrow 0$
- (ii)  $\forall v \in \mathbf{R}^n, B^k v \rightarrow 0$
- (iii)  $\rho(B) < 1$
- (iv)  $\exists \|\cdot\|$  induite tel que  $\|B\| < 1$

**Démonstration**

(i)  $\Rightarrow$  (ii)

si

$$B^k \rightarrow 0$$

alors pour tout  $v$  :

$$\|B^k v\|_2 \leq \|B^k\|_2 \cdot \|v\|_2$$

et donc  $B^k v$  tend vers 0.

(ii)  $\Rightarrow$  (iii)

Si

$$B^k v \rightarrow 0$$

Supposons  $\rho(B) \geq 1$  et soit  $|\lambda| = \rho(B)$ . Soit  $u$  vecteur propre associé à  $\lambda$  :  $Bu = \lambda u$ . Alors :

$$\begin{aligned} B^k u &= \lambda^k u \\ \|\lambda^k u\| &= \|\lambda\|^k \|u\| \end{aligned}$$

et  $B^k u$  ne peut pas tendre vers 0.

(iii)  $\Rightarrow$  (iv)

si

$$\rho(B) < 1$$

alors pour tout  $\varepsilon > 0 \exists \|\cdot\|$  induite tel que

$$\|B\| \leq \rho(B) + \varepsilon$$

On considère cette norme pour

$$\varepsilon = \frac{1 - \rho(B)}{2}$$

On a alors

$$\|B\| < 1$$

(iv)  $\Rightarrow$  (i)

si  $\exists \|\cdot\|$  induite telle que

$$\|B\| < 1$$

Comme la norme est matricielle

$$\|B^k\| \leq \|B\|^k < 1$$

D'où la convergence de  $B^k$  vers 0.



## 6.2 Méthodes itératives par splitting

### 6.2.1 Méthodes itératives

On considère le système linéaire  $Ax = b$ . Ce système peut se mettre sous la forme :

$$x = Bx + c$$

avec  $B = A + I$  et  $c = -b$ . On va étudier les suites de la forme

$$x_{k+1} = Bx_k + c$$

Si la suite converge, ce sera vers la solution du système linéaire  $Ax = b$  ( $x = Bx + c$ ) par continuité des applications en jeu.

**Theorem 21** *La suite  $x_{k+1} = Bx_k + c$  converge (indépendamment de  $x_0$ ) si et seulement si  $\rho(B) < 1$*

**Démonstration**

$\Rightarrow$

On suppose que la suite  $x_{k+1} = Bx_k + c$  converge vers  $x$  qui vérifie donc :  $x = Bx + c$ . Soit  $e_k = x - x_k$ .  $e_k$  converge vers 0. Alors en soustrayant les deux égalités précédentes, on a :

$$e_{k+1} = Be_k$$

or  $e_k = B^k e_0$ , donc d'après le théorème précédent  $\rho(B) < 1$ .

$\Leftarrow$

Supposons que  $\rho(B) < 1$ . Alors  $x = Bx + c$  a une solution. Supposons que cela ne soit pas le cas.  $x = Bx + c$  est équivalent à  $(B - I)x = -c$ . Alors la matrice  $B - I$  n'est pas inversible, donc elle a 0 comme valeur propre, donc  $B$  a 1 comme valeur propre, ce qui est absurde puisque  $\rho(B) < 1$ .

Soit donc  $x$  tel que  $x = Bx + c$ . On a donc  $x_{k+1} = Bx_k + c$  et  $x = Bx + c$ . Par soustraction on obtient comme précédemment :

$$e_{k+1} = Be_k$$

comme  $\rho(B) < 1$ ,  $\lim e_k = 0$  et donc

$$\lim x_k = x$$

■

## 6.2.2 Splitting

**Definition 11** *Splitting régulier de  $A$  :  $A = M - N$  avec  $M$  inversible*

Dans la suite on ne considèrera que des splittings réguliers. On a donc :

$$Ax = b \Leftrightarrow Mx - Nx = b$$

On considère les suites de la forme

$$Mx_{k+1} - Nx_k = b$$

Si cette suite converge, c'est vers  $x$  tel que

$$Mx - Nx = b$$

c'est à dire vers la solution du système linéaire de départ. Ce que l'on espère c'est que la suite converge, mais avant tout que les  $x_k$  soient "calculables" puisque pour chacun d'eux il est nécessaire de résoudre un système linéaire de matrice

$M$ . Il ne faudrait pas que cette résolution soit plus complexe que le problème de départ !

Comme  $M$  est inversible, on peut écrire la suite sous la forme :

$$x_{k+1} = M^{-1}Nx_k + M^{-1}b$$

qui est de la forme :

$$x_{k+1} = Bx_k + c$$

**Remark 4** Il faut faire très attention que bien qu'on écrive  $M^{-1}$  on ne calcule évidemment pas cette matrice, on résout les systèmes linéaires associés.

**Theorem 22** La suite associée au splitting converge ssi  $\rho(M^{-1}N) < 1$

### 6.2.3 Méthode de Jacobi

On utilise le splitting où la matrice  $M$  est le plus simple possible, à savoir diagonale. On pose donc :

$$A = D + (A - D)$$

avec  $D = \text{diag}(A)$ . Pour que le splitting soit régulier, il faut que  $\text{diag}A$  soit inversible.

**Remark 5** Attention!  $A$  inversible  $\nRightarrow \text{diag}A$  inversible

La méthode de Jacobi s'écrit donc :

$$Dx_{k+1} = (D - A)x_k + b$$

soit encore

$$x_{k+1} = D^{-1}(D - A)x_k + D^{-1}b$$

**Theorem 23** La méthode de Jacobi converge ssi  $\rho(D^{-1}(D - A)) < 1$

### Calcul des coordonnées des itérés de la méthode de Jacobi

$$\begin{cases} a_{1,1}x_1^{k+1} + a_{1,2}x_2^k + \dots + a_{1,n}x_n^k & = b_1 \\ a_{2,1}x_1^k + a_{2,2}x_2^{k+1} + \dots + a_{2,n}x_n^k & = b_2 \\ \vdots & \vdots \\ a_{n,1}x_1^k + a_{n,2}x_2^k + \dots + a_{n,n}x_n^{k+1} & = b_n \end{cases} \quad (6.1)$$

On remarque qu'il n'y a pas de système à résoudre : on calcule toutes les nouvelles coordonnées en fonction des anciennes, donc pas de problèmes de mise en oeuvre.

### 6.2.4 Méthode de Gauss-Seidel

Quand on a  $x_1^{k+1}$ , on pourrait remplacer dans la deuxième ligne  $x_1^k$  par  $x_1^{k+1}$  qui vient d'être calculé, et ainsi de suite. Cela donne :

$$\begin{cases} a_{1,1}x_1^{k+1} + a_{1,2}x_2^k + \dots + a_{1,n}x_n^k & = b_1 \\ a_{2,1}x_1^{k+1} + a_{2,2}x_2^{k+1} + \dots + a_{2,n}x_n^k & = b_2 \\ \vdots & \vdots \\ a_{n,1}x_1^{k+1} + a_{n,2}x_2^{k+1} + \dots + a_{n,n}x_n^{k+1} & = b_n \end{cases}$$

En faisant ceci, on résout un système triangulaire qui est la partie triangulaire inférieure de  $A$ .

$$A = D + L + U$$

avec  $D$  diagonale de  $A$ ,  $L$  partie inférieure de  $A$  et  $U$  partie supérieure de  $A$  :

$$L = \begin{pmatrix} 0 & \cdots & 0 \\ & \ddots & \vdots \\ a_{ij} & & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0 & & a_{ij} \\ \vdots & \ddots & \\ 0 & \cdots & 0 \end{pmatrix}$$

d'où le splitting :

$$A = (D + L) - (-U)$$

**Theorem 24** *La méthode de Gauss Seidel converge ssi  $\rho((D + L)^{-1}U) < 1$*

### 6.2.5 Méthode de relaxation

On peut essayer d'améliorer la méthode de Gauss Seidel (diminuer le rayon spectral de la matrice d'itération) en modulant la "part" de  $D$  que l'on prend. Plus précisément, soit  $\omega \neq 0$ . On pose :

$$M = \frac{D}{\omega} + L$$

d'où

$$A = \left(\frac{D}{\omega} + L\right) + \left(\frac{\omega - 1}{\omega}D + U\right)$$

**Remark 6** *Relaxer une valeur signifie "prendre la dernière valeur connue" : Jacobi, Gauss Seidel sont des méthodes de relaxation, dans Jacobi, on prend le terme de la diagonale et on relaxe les autres termes à l'étape suivante.*

**Theorem 25** *La méthode de relaxation converge ssi  $\rho\left(\left(\frac{D}{\omega} + L\right)^{-1}\left(\frac{\omega - 1}{\omega}D + U\right)\right) < 1$*

**Remark 7** *Trouver le  $\omega$  optimal (celui qui minimise le rayon spectral précédent est très difficile.*

**Theorem 26**  *$A$  symétrique définie positive;  $A = M - N$  splitting régulier. Si la matrice symétrique  ${}^tM + N$  est définie positive, alors*

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

#### Démonstration

Tout d'abord vérifions que la matrice  $M^t + N$  est symétrique. Comme  $A = M - N$  est symétrique, alors :

$$\begin{aligned} M^t + N &= (A + N)^t + N \\ &= A^t + N^t + N = A + N + N^t \end{aligned}$$

ce qui prouve bien la symétrie de la matrice.

On suppose donc que  $M^t + N$  est définie positive.  $A$  symétrique définie positive définit un produit scalaire et une norme :

$$\|x\|_A^2 = (Ax, x) = x^t Ax$$

On exprime  $M^{-1}N$  sous la forme :

$$M^{-1}N = M^{-1}(M - A) = I - M^{-1}A$$

et on va voir que pour la norme induite sur les matrices par  $\|\cdot\|_A$ , la norme de  $M^{-1}N$  est plus petite que 1, ce qui prouvera que son rayon spectral est aussi plus petit que 1. On a :

$$\|I - M^{-1}A\| = \sup_{\|x\|_A=1} \|x - M^{-1}Ax\|$$

Soit  $x$  tel que  $\|x\|_A = 1$  ( $x^tAx = 1$ ). Soit  $y = M^{-1}Ax$ . On a :

$$\begin{aligned} \|x - M^{-1}Ax\|_A^2 &= \|x - y\|_A^2 = (x - y)^t A (x - y) \\ &= x^t Ax - y^t Ax - x^t Ay + y^t Ay \\ &= 1 - y^t My - y^t M^t A^{-t} Ay + y^t Ay \\ &= 1 - y^t My - y^t M^t y + y^t Ay \\ &= 1 - y^t (M + M^t - A)y = 1 - y^t (M^t + N)y \end{aligned}$$

Or  $y \neq 0$  car  $M^{-1}A$  est inversible. Donc pour tout  $x$  de norme 1, on a :

$$\|x - M^{-1}Ax\|_A^2 < 1$$

La borne supérieure de cette quantité est-elle aussi strictement inférieure à 1 ? Mais  $S = \{x \in R^n / \|x\|_A = 1\}$  est compact. L'application de  $S$  dans  $R$  qui à  $x$  associe  $\|x - M^{-1}Ax\|$  est continue sur le compact  $S$ . L'image de ce compact est donc un compact (théorème de Heine), la borne supérieure de cette application est donc atteinte et donc :

$$\|M^{-1}N\| = \sup_{\|x\|_A=1} \|x - M^{-1}Ax\|_A < 1$$

■

**Theorem 27** Si  $A$  symétrique définie positive, la méthode de relaxation converge pour  $0 < \omega < 2$ .

**Démonstration**

$$\begin{aligned} A &= M - N = \left(\frac{D}{\omega} + L\right) - \left(\frac{1-\omega}{\omega}D - U\right) \\ M^t + N &= \left(\frac{D}{\omega} + L^t\right) + \frac{1-\omega}{\omega}D - U \end{aligned}$$

$A$  symétrique entraîne  $L^t = U$ . La diagonale d'une matrice définie positive est définie positive.

$$M^t + N = \frac{2-\omega}{\omega}D$$

Si  $0 < \omega < 2$ ,  $\rho(M^{-1}N) < 1$  d'après le théorème précédent.

**Theorem 28** La méthode de relaxation ne converge pas pour  $\omega \notin ]0, 2[$  et plus précisément, si on pose

$$M_\omega = \left(\frac{D}{\omega} + L\right)^{-1} \left(\frac{\omega-1}{\omega}D + U\right)$$

alors

$$\rho(M_\omega) \geq |\omega - 1|$$

**Démonstration**

$$\begin{aligned} \det M_\omega &= \frac{\det \left( \frac{\omega-1}{\omega} D + U \right)}{\det \left( \frac{D}{\omega} + L \right)} = \frac{\left( \frac{\omega-1}{\omega} \right)^n \prod_i di}{\left( \frac{1}{\omega} \right)^n \prod_i di} \\ &= (\omega - 1)^n \end{aligned}$$

mais

$$(\rho(M_\omega))^n \geq \prod_i |\lambda_i| = |\omega - 1|^n$$

d'où

$$\rho(M_\omega) \geq |\omega - 1|$$

### 6.3 Méthodes itératives liées à l'optimisation

**Definition 12** *L'optimisation est l'étude de la minimisation de fonctionnelles sur des sous domaines de  $\mathbf{R}^n$ .*

A partir de maintenant,  $A$  symétrique définie positive.

#### 6.3.1 Equivalence système linéaire et optimisation

Soit  $v \in \mathbf{R}^n$ . On considère

$$\begin{aligned} J(v) &= \frac{1}{2}(Av, v) - (b, v) \\ \mathbf{R}^n &\rightarrow \mathbf{R} \end{aligned}$$

**Theorem 29** *Minimiser  $J$  sur  $\mathbf{R}^n$  est équivalent à résoudre  $Ax = b$ . Autrement dit,  $J$  admet un unique minimum en un point qui est la solution du système linéaire  $Ax = b$ .*

**Remark 8** *Rappel  $J$  différentiable en  $u$  :*

$$\begin{aligned} \exists l &: \mathbf{R}^n \rightarrow \mathbf{R} \text{ linéaire (continue)} \\ \forall h \in \mathbf{R}^n, & J(u+h) = J(u) + l(h) + \|h\|\varepsilon(h) \end{aligned}$$

avec

$$\lim_{h \rightarrow 0} \varepsilon(h) = 0$$

on note

$$\begin{aligned} l(h) &= DJ_u(h) = J'(u).h = \nabla J(u).h \\ l &= DJ_u = J'(u) = \nabla J(u) \end{aligned}$$

**Démonstration**

$$\begin{aligned} J(u+h) &= \frac{1}{2}(A(u+h), u+h) - (b, u+h) \\ &= \frac{1}{2}(Au, u) - (b, u) + (Au - b, h) + \frac{1}{2}(Ah, h) \\ &= J(u) + (Au - b, h) + \frac{1}{2}(Ah, h) \end{aligned}$$

d'où  $J$  est différentiable car  $|(Ah, h)| \leq \|Ah\| \cdot \|h\| \leq \|A\| \cdot \|h\|^2 = \|h\|\varepsilon(h)$  et  $\nabla J(u) = Au - b$ .

Donc, si  $J$  est minimum en  $u$ , alors  $\nabla J(u) = 0$  et  $u$  est solution de  $Ax = b$ . Réciproquement si  $Au = b$ , alors pour tout  $h \neq 0$  :

$$J(u + h) = J(u) + \frac{1}{2}(Ah, h) > J(u)$$

car  $A$  est définie positive, et donc  $J$  admet un minimum en  $u$ .

■

On va donc essayer de construire des méthodes pour minimiser  $J$ .

### 6.3.2 Méthodes de descente

**Definition 13** On appelle méthode de descente une suite de la forme

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

où  $\alpha_k \in R$  et  $p_k \in \mathbf{R}^n$

**Definition 14** On appelle  $p_k$  la direction de descente.

**Remark 9** Les différents choix de  $\alpha_k$  et  $p_k$  donnent naissance à différentes méthodes.

On voudrait que  $x_k$  converge vers  $x$  solution de  $Ax = b$ .

**Notations**

$$\begin{aligned} A\bar{x} &= b \\ J(x) &= \frac{1}{2}(Ax, x) - (b, x) \\ \text{erreur } e(x) &= x - \bar{x} \\ \text{résidu } r(x) &= b - Ax \\ E(x) &= (Ae(x), e(x)) \\ \nabla J(x) &= -r(x) \end{aligned}$$

**Theorem 30** On a  $E(x) = (r(x), A^{-1}r(x))$  et minimiser  $J$  est équivalent à minimiser  $E$

**Démonstration**

$$\begin{aligned} A^{-1}r(x) &= A^{-1}b - x \\ &= \bar{x} - x = -e(x) \end{aligned}$$

$$\begin{aligned} (r(x), A^{-1}r(x)) &= (r(x), -e(x)) \\ Ae(x) = Ax - A\bar{x} &= Ax - b = -r(x) \end{aligned}$$

d'où

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

et

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

et  $E$  et  $J$  ne diffèrent que d'une constante.

■

**Theorem 31** On a

$$r_{k+1} = r_k - \alpha_k Ap_k$$

**Démonstration**

Comme  $x_{k+1} = x_k + \alpha_k p_k$ , en multipliant cette expression par  $-A$  et en ajoutant  $b$ , on obtient le résultat annoncé.

■

**6.3.3 Méthode de descente optimale**

**Definition 15**  $p_k$  étant choisi non nul, on appelle descente optimale le choix de  $\alpha_k$  qui minimise  $J$  sur l'ensemble des vecteurs de la forme  $x_k + \alpha p_k$  (qui constituent un droite).

$\alpha_k$  minimise  $J$  sur la droite  $x_k + \alpha P$  :

$$\forall \alpha \in R, J(x_k + \alpha_k p_k) \leq J(x_k + \alpha p_k)$$

Est-ce possible ?

$$J(x_k + \alpha p_k) = \frac{1}{2}(A(x_k + \alpha p_k), x_k + \alpha p_k) - (b, x_k + \alpha p_k)$$

Cette expression est un trinôme du second degré en  $\alpha$ .

$$J(x_k + \alpha p_k) = \frac{1}{2}\alpha^2 (Ap_k, p_k) + \alpha((Ap_k, x_k) - (b, p_k)) + \frac{1}{2}(Ax_k, x_k) - (b, x_k)$$

Pour que  $J$  ait un minimum, il faut que

$$\frac{1}{2}(Ap_k, p_k) > 0$$

ce qui est vrai ici puisque  $A$  est définie positive. Le minimum est atteint là où la dérivée est nulle ce qui donne :

$$\alpha_k = \frac{-(Ap_k, x_k) + (b, p_k)}{(Ap_k, p_k)}$$

On peut "arranger" un peu cette expression :

$$\begin{aligned} -(Ap_k, x_k) + (b, p_k) &= -(p_k, Ax_k) + (b, p_k) = (p_k, -Ax_k + b) \\ &= (p_k, r_k) \end{aligned}$$

d'où

$$\alpha_k = \frac{(p_k, r_k)}{(Ap_k, p_k)}$$

**Theorem 32**

$$(p_k, r_{k+1}) = 0$$

**Démonstration**

$$(p_k, r_{k+1}) = (p_k, r_k - \alpha_k Ap_k) = (p_k, r_k) - \alpha_k (p_k, Ap_k) = 0$$

■

**Theorem 33** On a

$$E(x_{k+1}) = E(x_k)(1 - \gamma_k)$$

avec

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

**Démonstration**

On a

$$\begin{aligned} E(x_{k+1}) &= E(x_k) - 2\alpha_k(r_k, p_k) + \alpha_k^2(Ap_k, p_k) \\ &= E(x_k) - \frac{(r_k, p_k)^2}{(Ap_k, p_k)} = E(x_k) \left( 1 - \frac{(r_k, p_k)^2}{E(x_k)(Ap_k, p_k)} \right) \\ &= E(x_k) \left( 1 - \frac{(r_k, p_k)^2}{(A^{-1}r_k, r_k)(Ap_k, p_k)} \right) \end{aligned}$$

■

On élimine dans toute la suite le cas  $r_k = 0$  ( $Ax_k = b$  : on a trouvé la solution).

**Theorem 34** Si  $p_k \neq 0$  alors

$$\gamma_k \geq \frac{1}{\text{cond}_2(A)} \left( \frac{r_k}{\|r_k\|}, \frac{p_k}{\|p_k\|} \right)^2$$

**Démonstration**

On a

$$\text{cond}_2 A = \frac{\lambda_{\max}}{\lambda_{\min}}$$

$$\begin{aligned} (Ap_k, p_k) &\leq \lambda_{\max} \|p_k\|^2 \\ (A^{-1}r_k, r_k) &\leq \frac{1}{\lambda_{\min}} \|r_k\|^2 \end{aligned}$$

d'où

$$\gamma_k \geq \frac{\lambda_{\min}(r_k, p_k)^2}{\lambda_{\max} \|p_k\|^2 \|r_k\|^2} = \frac{1}{\text{cond}_2(A)} \left( \frac{r_k}{\|r_k\|}, \frac{p_k}{\|p_k\|} \right)^2$$

■

**Theorem 35** Si

$$\exists \mu > 0, \forall k \left( \frac{r_k}{\|r_k\|}, \frac{p_k}{\|p_k\|} \right)^2 \geq \mu > 0$$

alors

$$x_k \rightarrow \bar{x}$$

**Démonstration**

$$E(x_{k+1}) \leq E(x_k) \left( 1 - \frac{\mu}{\text{cond}_2(A)} \right)$$

On a aussi

$$\begin{aligned} \mu &\geq 0 \\ \text{cond}_2 A &\geq 1 \end{aligned}$$

et d'après Cauchy-Schwarz  $\mu \leq 1$ . D'où

$$0 \leq E(x_k) \leq \left(1 - \frac{\mu}{\text{cond}_2(A)}\right)^k E(x_0) \rightarrow 0$$

donc

$$E(x_k) \rightarrow 0$$

or

$$E(x_k) = (A(x_k - \bar{x}), x_k - \bar{x}) \geq \lambda_{\min} \|x_k - \bar{x}\|^2$$

et donc

$$\|x_k - \bar{x}\|^2 \leq \frac{E(x_k)}{\lambda_{\min}}$$

et

$$x_k \rightarrow \bar{x}$$

■

### 6.3.4 Méthode de gradient à pas optimal

Dans la méthode de descente à pas optimal, on a encore le choix de  $p_k$ .

**Theorem 36** *La méthode de gradient à pas optimal est la méthode de descente optimale où l'on a choisit  $p_k = r_k$ .*

**Remark 10** *Le nom de la méthode bien du fait que  $r_k = -\nabla J(x_k)$ .*

**Algorithme**

$x_0$  donné, on peut calculer

$$\begin{aligned} k &= 0 \\ x_0 &= \text{donnée} \\ r_0 &= b - Ax_0 \\ &\text{itérez jusqu'à convergence} \\ \alpha_k &= \frac{(r_k, r_k)}{(Ar_k, r_k)} \\ x_{k+1} &= x_k + \alpha_k r_k \\ r_{k+1} &= r_k - \alpha_k Ar_k \end{aligned}$$

**Theorem 37** *La méthode de gradient à pas optimal est convergente.*

**Démonstration**

On a

$$\forall k, \left( \frac{r_k}{\|r_k\|}, \frac{p_k}{\|p_k\|} \right)^2 = 1 > 0$$

donc la méthode converge.

■

**Theorem 38** *La méthode de gradient à pas optimal vérifie :*

$$\|x_k - \bar{x}\| \leq \sqrt{\frac{E(x_0)}{\lambda_{\min}}} \left( \frac{K(A) - 1}{K(A) + 1} \right)^k$$

où  $K(A) = \text{cond}_2(A)$ .

**Démonstration**

Comme  $p_k = r_k$ , on a

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

d'où d'après Kantorovitch

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

et donc comme

$$\lambda_{\min} \|x_k - \bar{x}\|^2 \leq E(x_k)$$

on obtient le résultat annoncé.



**6.3.5 Méthode de descente à pas fixe**

On repart de la méthode précédente, mais on décide de fixer  $\alpha$  une fois pour toute.

**Definition 16** On appelle méthode de descente à pas fixe la méthode définie par :

$$x_{k+1} = x_k + \alpha r_k$$

avec  $\alpha > 0$ .

**Theorem 39** La descente à pas fixe converge ssi  $0 < \alpha < \frac{2}{\lambda_{\max}}$

**Démonstration**

On a  $x_{k+1} = x_k + \alpha(b - Ax_k)$  et la limite éventuelle de  $x_k$  est bien la solution de  $Ax = b$ . On peut d'autre part écrire  $x_{k+1} = (I - \alpha A)x_k + \alpha b$  et cette suite converge ssi  $\rho(I - \alpha A) < 1$ , ce qui est équivalent à  $\alpha \in ]0, \frac{2}{\lambda_{\max}}[$ .



Existe-t-il un  $\alpha$  optimal? : y-a-t-il un  $\alpha$  tel que  $\rho(I - \alpha A)$  soit le plus petit possible?

$$f(\alpha) = \rho(I - \alpha A) = \max_i |1 - \alpha \lambda_i| = \max \{ |1 - \alpha \lambda_1|, |1 - \alpha \lambda_n| \}$$

Cette fonction  $f$  a un minimum. Ce point a pour abscisse  $\alpha$  tel que :

$$\begin{aligned} |1 - \alpha \lambda_1| &= |1 - \alpha \lambda_n| \\ 1 - \alpha \lambda_1 &= \alpha \lambda_n - 1 \\ \alpha &= \frac{2}{\lambda_1 + \lambda_n} \end{aligned}$$

l'ordonnée  $f(\alpha)$  fournit le rayon spectral de  $I - \alpha A$  :

$$\rho(I - \alpha A) = \left| \frac{\lambda_1 + \lambda_n - 2\lambda_1}{\lambda_1 + \lambda_n} \right| = \left| \frac{-\lambda_1 + \lambda_n}{\lambda_1 + \lambda_n} \right| = \frac{K(A) - 1}{K(A) + 1}$$

pour cet  $\alpha$  est optimal :

$$\|x_k - \bar{x}\| \leq \left( \frac{K(A) - 1}{K(A) + 1} \right)^k \|x_0 - \bar{x}\|$$

Cette méthode n'est pas utilisable "pratiquement", car pour avoir le  $\alpha$  optimal, il faut calculer la plus grande et la plus petite valeurs propres de  $A$ ...

### 6.3.6 Méthode du gradient conjugué

On utilise une méthode de descente optimale, c'est à dire :

$$\begin{aligned} x_{k+1} &= x_k + \alpha_k p_k \\ \alpha_k &= \frac{(r_k, p_k)}{(Ap_k, p_k)} \end{aligned}$$

Il reste à choisir  $p_k$ . Pour la méthode de gradient à pas optimal, on avait choisi  $p_k = r_k$ . On va faire un autre choix ici. Tout d'abord, on prend  $p_0 = r_0$ . Puis : on prend  $p_k$  dans le plan défini par  $r_k$  et  $p_{k-1}$  :

$$p_k = r_k + \beta_k p_{k-1}$$

avec  $\beta_k \in \mathbb{R}$  tel que la minimisation de  $E(x_{k+1})$  par rapport à  $E(x_k)$  soit maximale. C'est à dire que comme :

$$E(x_{k+1}) = E(x_k)(1 - \gamma_k)$$

on prend  $\beta_k$  qui maximise  $\gamma_k$  :

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

Il faut que  $\gamma_k$  soit le plus grand possible (le plus proche de 1). Examinons chacun des termes figurant dans l'expression précédente :

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

Or  $(r_k, p_{k-1}) = 0$  d'où  $(r_k, p_k) = \|r_k\|^2$  ne dépend pas de  $p_k$ .

$(A^{-1}r_k, r_k)$  ne dépend pas non plus de  $p_k$ .

Il reste donc à minimiser  $(Ap_k, p_k)$

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

Il s'agit d'un trinôme du second degré de coefficient principal positif (si  $p_{k-1} \neq 0$ , ce que l'on suppose) qui a un minimum en  $\beta_k$  annulant la dérivée :

$$\begin{aligned} 2\beta_k (Ap_{k-1}, p_{k-1}) &= -2(Ap_{k-1}, r_k) \\ \beta_k &= -\frac{(Ap_{k-1}, r_k)}{(Ap_{k-1}, p_{k-1})} \end{aligned}$$

**Theorem 40** *Deux directions de descentes successives sont orthogonales pour le produit scalaire défini par A (A-conjuguées).*

$$(Ap_{k-1}, p_k) = 0$$

#### Démonstration

$$\begin{aligned} (Ap_{k-1}, p_k) &= (Ap_{k-1}, r_k + \beta_k p_{k-1}) = (Ap_{k-1}, r_k) + \beta_k (Ap_{k-1}, p_{k-1}) \\ &= (Ap_{k-1}, r_k) - \frac{(Ap_{k-1}, r_k)}{(Ap_{k-1}, p_{k-1})} (Ap_{k-1}, p_{k-1}) = 0 \end{aligned}$$

■

L'algorithme s'écrit donc :

$$\begin{aligned}
 x_0 \text{ donné, } r_0 &= b - Ax_0, p_0 = r_0, k = 0 \\
 \alpha_k &= \frac{(r_k, p_k)}{(Ap_k, p_k)} \\
 x_{k+1} &= x_k + \alpha_k p_k \\
 r_{k+1} &= r_k - \alpha_k Ap_k \\
 \beta_{k+1} &= -\frac{(Ap_k, r_{k+1})}{(Ap_k, p_k)} \\
 p_{k+1} &= r_{k+1} + \beta_{k+1} p_k
 \end{aligned}$$

En fait, comme on va le voir un peu plus loin, on peut écrire l'algorithme un peu plus simplement (avec un produit scalaire de moins).

**Theorem 41**

$$\begin{aligned}
 \text{Si } \forall i \leq k, r_i &\neq 0 \\
 \forall k &\geq 0, (r_{k+1}, r_k) = 0 \\
 \forall k &\geq 1, \beta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}
 \end{aligned}$$

**Démonstration**

$$\begin{aligned}
 (r_{k+1}, r_k) &= (r_k - \alpha_k Ap_k, r_k) = \|r_k\|^2 - \alpha_k (Ap_k, p_k - \beta_k p_{k-1}) \\
 &= \underbrace{\|r_k\|^2 - \alpha_k (Ap_k, p_k)}_{=0} + \alpha_k \beta_k \underbrace{(Ap_k, p_{k-1})}_{=0} = 0
 \end{aligned}$$

Et pour

$$\beta_k = -\frac{(Ap_{k-1}, r_k)}{(Ap_{k-1}, p_{k-1})}$$

on écrit :

$$\begin{aligned}
 Ap_{k-1} &= \frac{1}{\alpha_{k-1}} (r_{k-1} - r_k) \\
 (Ap_{k-1}, r_k) &= -\frac{1}{\alpha_{k-1}} \|r_k\|^2
 \end{aligned}$$

et

$$\begin{aligned}
 (Ap_{k-1}, p_{k-1}) &= \left(\frac{1}{\alpha_{k-1}}\right) \left[ \underbrace{(r_{k-1}, p_{k-1})}_{=0} - \underbrace{(r_k, p_{k-1})}_{=0} \right] \\
 &= \frac{\|r_{k-1}\|^2}{\alpha_{k-1}}
 \end{aligned}$$

car  $(r_k, p_{k-1}) = 0$  et  $(r_{k-1}, p_{k-1}) = \|r_{k-1}\|^2$ . D'où :

$$\beta_k = \frac{\|r_k\|^2}{\|r_{k-1}\|^2}$$

■  
On peut donc écrire l'algorithme sous la forme :

$$\begin{aligned}\alpha_k &= \frac{\|r_k\|^2}{(Ap_k, p_k)} \\ x_{k+1} &= x_k + \alpha_k p_k \\ r_{k+1} &= r_k - \alpha_k Ap_k \\ \beta_{k+1} &= \frac{\|r_{k+1}\|^2}{\|r_k\|^2} \\ p_{k+1} &= r_{k+1} + \beta_{k+1} p_k\end{aligned}$$

**Theorem 42**

$$\begin{aligned} \forall i \leq k, r_i &\neq 0 \\ \forall i \leq k-1, (r_k, p_i) &= 0 \\ H_k &= \text{Vect}(r_0, \dots, r_k) = \text{Vect}(r_0, Ar_0, A^2r_0, \dots, A^k r_0) \\ H_k &= \text{Vect}(p_0, \dots, p_k) = \text{Vect}(p_0, Ap_0, A^2p_0, \dots, A^k p_0) \\ \forall i \leq k-1, (Ap_k, p_i) &= 0 \end{aligned}$$

**Démonstration** (laissée en exercice)

Par récurrence sur  $k$  pour les 4 propriétés en même temps.

Comme

$$\forall i \leq k-1, (Ap_k, p_i) = 0$$

dans un espace de dimension  $n$ , il ne peut y avoir que  $n$  vecteurs orthogonaux 2 à 2. Au pire à la  $n+1^{\text{ième}}$  itération :

$$r_n \perp p_0, \dots, p_{n-1} \text{ et donc } r_n = 0$$

et on a le :

**Theorem 43** *La méthode du gradient conjugué converge en  $n$  itérations au plus.*

On a donc en fait une méthode "directe" !! Mais à cause des erreurs d'arrondis, le système de vecteurs n'est pas rigoureusement orthogonal, et la méthode est utilisée comme une méthode itérative !

De plus, la méthode du gradient conjugué à une propriété d'optimalité :

**Theorem 44**  *$E$  atteint son minimum sur  $x_0 + H_k$  en  $x_k$*

$$\forall x \in x_0 + H_k, E(x_k) \leq E(x)$$

En faisant  $n$  minimisations dans  $\mathbf{R}$ , on a fait une minimisation dans  $\mathbf{R}^n$ , ce qui est remarquable et peu courant !