Université des Sciences et Technologies de Lille
U.F.R. de Mathématiques Pures et Appliquées

M307 : Algèbre matricielle numérique

Notes de cours par Clément Boulonne

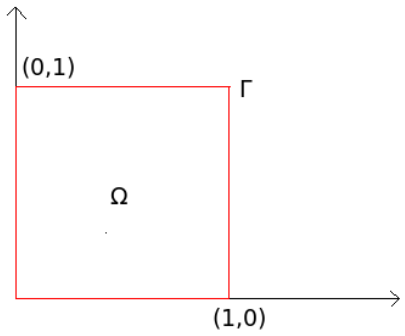
L3 Mathématiques **2008 - 2009**

Table des matières

Introduction à l'algèbre matricielle numérique

0.1 Un exemple

Exemple 0.1.1. Imaginons une plaque dont on connait la température aux bords :



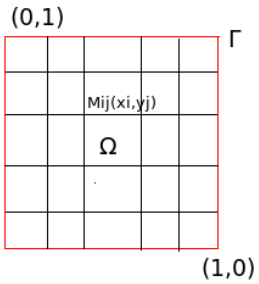
On cherche à connaître la température en tout point $M(x, y)$ de Ω, sachant qu'on la connait sur la frontière Γ. On note $u(x, y)$ cette température en $M(x, y)$. On modélise le phénomène physique :

$$\begin{cases} -\Delta u = 0 \text{ dans } \Omega \\ u(x, y) = T(x, y) \text{ sur } \Gamma \end{cases}$$

Rappel (Laplacien).

$$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}$$

On peut supposer que ce problème admet une unique solution (dans un sens précis). On va approcher la solution u en cherchant une approximation numérique. On va mailler le domaine Ω par des points $M_{ij}(x_i, y_j)$:



avec :

$$\begin{cases} x_i = \frac{i}{n} & (0 \leq i \leq n) \\ y_j = \frac{j}{n} & (0 \leq j \leq n) \end{cases}$$

et $h = \frac{1}{n}$ est le pas du maillage.

On prend : $1 \leq i \leq n-1$ et $1 \leq j \leq n-1$. On a alors :

$$u(x_{i+1}, y_j) = u(x_i, y_j) + h\frac{\partial u}{\partial x}(x_i, y_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) + \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i, y_j) + \mathcal{O}(h^4) \qquad (1)$$

$$u(x_{i-1}, y_j) = u(x_i, y_j) - h\frac{\partial u}{\partial x}(x_i, y_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial x^2}(x_i, y_j) - \frac{h^3}{6}\frac{\partial^3 u}{\partial x^3}(x_i, y_j) + \mathcal{O}(h^4) \qquad (2)$$

$$u(x_i, y_{j+1}) = u(x_i, y_j) + h\frac{\partial u}{\partial y}(x_i, y_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial y^2}(x_i, y_j) + \frac{h^3}{6}\frac{\partial^3 u}{\partial y^3}(x_i, y_j) + \mathcal{O}(h^4) \qquad (3)$$

$$u(x_i, y_{j-1}) = u(x_i, y_j) - h\frac{\partial u}{\partial y}(x_i, y_j) + \frac{h^2}{2}\frac{\partial^2 u}{\partial y^2}(x_i, y_j) - \frac{h^3}{6}\frac{\partial^3 u}{\partial y^3}(x_i, y_j) + \mathcal{O}(h^4) \qquad (4)$$

On va former des combinaisons linéaires. En utilisant (1) et (2) :

$$-\frac{\partial^2 u}{\partial x^2}(x_i, y_j) = \frac{-u(x_{i+1}, y_j) + 2u(x_i, y_j) - u(x_{i-1}, y_j)}{h^2} + \mathcal{O}(h^2)$$

De même avec (3) et (4) :

$$-\frac{\partial^2 u}{\partial y^2}(x_i, y_j) = \frac{-u(x_i, y_{j+1}) + 2u(x_i, y_j) - u(x_i, y_{j-1})}{h^2} + \mathcal{O}(h^2)$$

Remarque. On suppose ici que u étant aussi régulière qu'il le fallait.

Idée. On va noter u_{ij} une approximation de $u(x_i, y_j)$ et l'équation approchée au point M_{ij} va s'écrire pour $1 \leq i \leq n-1$ et $1 \leq j \leq n-1$:

$$\frac{-u_{i+1,j} + 2u_{ij} - u_{i-1,j}}{h^2} + \frac{-u_{i,j+1} + 2u_{ij} - u_{i,j-1}}{h^2} = 0$$

$$\Leftrightarrow \frac{4u_{ij} - u_{i-1,j} - u_{i+1,j} - u_{i,j-1} - u_{i,j+1}}{h^2} = 0 \qquad (*)$$

Soit :

$$U = \begin{pmatrix} u_{i1} \\ u_{i2} \\ \cdots \\ u_{i,j-1} \\ u_{ij} \\ u_{i,j+1} \\ \cdots \\ u_{i,n-1} \end{pmatrix}$$

On peut montrer que $(*)$ est équivalent à :

$$AU = B$$

avec

$$A = \begin{pmatrix} A_{11} & A_{12} & & \\ A_{21} & A_{22} & \ddots & \\ & \ddots & \ddots & A_{n-2,n-1} \\ 0 & & A_{n-1,n-2} & A_{n-1,n-1} \end{pmatrix}_{(n-1)^2 \times (n-1)^2}$$

tel que :

$$A_{ii} = \begin{pmatrix} 4 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 4 \end{pmatrix}_{(n-1)(n-1)}$$

et :

$$A_{i,i+1} = \begin{pmatrix} -1 & & \\ & \ddots & \\ & & -1 \end{pmatrix}_{(n-1)(n-1)}$$

B est un vecteur qui contient l'information des conditions aux limites.

0.2 Remarques

1) La méthode ici exposée est la méthode des différences limites. Il resterait à prouver sa convergence. On retient qu'on a à résoudre un grand système linéaire.
2) Si on regarde la matrice A :
   → elle a au plus 5 coefficients non nuls par ligne. Le taux de remplissage est de $\frac{5(n-1)}{(n-1)^2(n-1)^2} \sim \frac{5}{n^2}$.
   → elle est symétrique.
   → elle est tri-diagonale par bloc.
   → elle est symétrie définie possitive (admis).
   ⇒ il faut en tenir compte pour savoir comment on résout le système.
3) On pourrait utiliser d'autres méthodes numériques
   → éléments finies (matrices creuses [1] mais non tri-diagonale par bloc).
   → Méthodes intégrales (matrice pleine)
   → Méthodes spectrales, probabilistes...

0.3 D'autres problèmes menant à la résolution d'un système linéaire

→ Electrostatisque (où mettre les relais ds téléphones portables ?)
→ Mécanique des fludies (quel temps fera demain ? quel dimension doit avoir le A380 ?)
→ Structure (Pont)
→ Thermique
→ Représentation de surfaces (Programmation des jeux vidéo)
→ Finances (Bourse ?)
→ On peut aussi avoir besoin d'informations sur un système linéaire sans pour autant devoir le résoudre.
  → Valeurs propres ? (Mécanique vibratoire, étude de la dynamique des populations, connexion interurbaines, compression d'images)

0.4 Objectifs

Résoudre de grands systèmes linéaires, trouver une méthode la plus efficace.

[1] c'est-à -dire ayant un taux de remplissage faible

0.5 Bibliographie pour le cours

(1) *Analyse numérique appliqueé à l'art de l'Ingénieurie*, tome 1, méthodes directes, Lascaux, Théodore (Manson)

(2) *Introduction à l'analyse numérique*, Rappaz, Picasso

(3) *Introduction à l'analyse matricielle et à l'optimisation*, (MANSON), P.G.

Chapitre 1

Méthodes directes de résolution de système linéaire

Problème. *On veut résoudre $AX = B$ avec :*
- *A matrice à coefficients réels de taille $n \times n$ et supposée inversible.*
- *B vecteur de $\mathbb{R}^n$*
- *X inconnue à chercher de $\mathbb{R}^n$*

1.1 Méthode du pivot de Gauss

1.1.1 Un exemple

Exemple 1.1.1. Soit :

$$A^{(1)} = \begin{pmatrix} 4 & 8 & 12 \\ 3 & 8 & 13 \\ 2 & 9 & 18 \end{pmatrix}$$

et

$$B^{(1)} = \begin{pmatrix} 4 \\ 5 \\ 11 \end{pmatrix}$$

On a à résoudre ce système linéaire :

$$\begin{cases} 4x_1 + 8x_2 + 12x_3 = 4 & (1) \\ 3x_1 + 8x_2 + 13x_3 = 5 & (2) \\ 2x_1 + 9x_2 + 18x_3 = 18 & (3) \end{cases}$$

Etape 1 : on choisit $a_{11}^{(1)} = 4$ comme pivot. On remplace l'équation (2) par $(2) - \frac{a_{21}^{(1)}}{a_{11}^{(1)}}(1)$. On remplace l'équation (3) par $(3) - \frac{a_{31}^{(1)}}{a_{11}^{(1)}}(1)$. On a ainsi :

$$A^{(2)} = \begin{pmatrix} 4 & 8 & 12 \\ 0 & 2 & 4 \\ 0 & 5 & 12 \end{pmatrix}$$

$$B^{(2)} = \begin{pmatrix} 4 & (1)' \\ 2 & (2)' \\ 9 & (3)' \end{pmatrix}$$

Etape 2 : On remplace $(3)'$ par $(3)' - \frac{a_{32}^{(2)}}{a_{22}^{(2)}}(2)'$. On a ainsi :

$$A^{(3)} = \begin{pmatrix} 4 & 8 & 12 \\ 0 & 2 & 4 \\ 0 & 0 & 2 \end{pmatrix}$$

$$B^{(3)} = \begin{pmatrix} 4 \\ 2 \\ 4 \end{pmatrix}$$

Ce qui est équivalent à ce système linéaire :

$$\begin{cases} 4x_1 + 8x_2 + 12x_3 = 4 \\ \quad\quad 2x_2 + 4x_3 = 2 \\ \quad\quad\quad\quad 2x_3 = 4 \end{cases}$$

Après calculs, on trouve :

$$\begin{cases} x_3 = 2 \\ x_2 = -3 \\ x_1 = 1 \end{cases}$$

1.1.2 Algorithme d'élimination

Définition 1.1.1. On appelle $A^{(i)}$ et $B^{(i)}$ la matrice et le second membre obtenus avant la $i$ème étape. On a :

$$A^{(i)} = \begin{pmatrix} a_{11}^{(i)} & a_{12}^{(i)} & \cdots & \cdots & \cdots & a_{1n}^{(i)} \\ 0 & a_{22}^{(i)} & \cdots & \cdots & \cdots & a_{2n}^{(i)} \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & & \boxed{a_{ii}^{(i)}} & \cdots & a_{in}^{(i)} \\ 0 & 0 & \cdots & 0 & \cdots & a_{i+1,n}^{(i)} \\ \vdots & \vdots & & & & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & a_{nn}^{(i)} \end{pmatrix}$$

Pour passer de $A^{(i)}$ à $A^{(i+1)}$ et de $B^{(i)}$ à $B^{(i+1)}$, on procède comme suit :

Pour k variant de $i+1$ jusque n et pour j variant de $i+1$ à n :

$$a_{kj}^{(i+1)} \leftarrow a_{kj}^{(i)} - \frac{a_{ki}^{(i)}}{a_{ii}^{(i)}} a_{ij}^{(i)}$$

de même pour le second membre :

$$b_k^{(i+1} \leftarrow b_k^{(i)} - \frac{a_{ki}^{(i)}}{a_{ii}^{(i)}} b_i^{(i)},\ i+1 \leq k \leq n$$

Algorithme 1.1.1.
PivotdeGauss(A, B, n)
1 **for** $i \leftarrow 1$ **to** $n-1$
2 **do for** $k \leftarrow i+1$ **to** n

3 **do** $c \leftarrow \frac{a_{ki}}{a_{ii}}$
4 **for** $j \leftarrow i+1$ **to** n
5 **do** $a_{kj}^{(i+1)} \leftarrow a_{kj} - c * a_{ij}$
6 $b_k^{(i+1)} \leftarrow b_k - c * b_i$
7
8
9
10 **for** $i \leftarrow n-1$ **to** 1
11 **do** $x_i \leftarrow \dfrac{\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right)}{a_{ii}}$

Remarque.
1. Lorsque $a_{kj}^{(i+1)}$ a été calculée, on n'aura plus jamais besoin dans l'algorithme de $a_{kj}^{(i)}$ donc on écrase le coefficient $a_{kj}^{(i)}$. Même remarque pour $b_h^{(i)}$.
2. On peut stocker $\frac{a_{ki}}{a_{ii}}$ dans une variable c.
3. Cet algorithme ne fonctionne pas toujours.

Exemple 1.1.2. Soit :

$$A^{(1)} = \begin{pmatrix} 2 & 1 & 0 & 4 \\ -4 & 2 & 3 & -7 \\ 4 & 1 & -2 & 8 \\ 0 & -3 & -12 & -1 \end{pmatrix}$$

$$B^{(1)} = \begin{pmatrix} 2 \\ -9 \\ 2 \\ 2 \end{pmatrix}$$

On a :

$$A^{(2)} = \begin{pmatrix} 2 & 1 & 0 & 4 \\ 0 & \boxed{0} & 3 & 1 \\ 0 & -1 & -2 & 0 \\ 0 & -3 & -12 & -1 \end{pmatrix}$$

$$B^{(2)} = \begin{pmatrix} 2 \\ -5 \\ -2 \\ 2 \end{pmatrix}$$

Mais on ne peut pas avoir $A^{(3)}$ car on a un pivot nul. Pour s'en sortir, il faut permuter des lignes. On peut par exemple, permuter la ligne 2 et 4 de la matrice $A^{(2)}$ et de la matrice $B^{(2)}$. On aura :

$$A^{(2)'} = \begin{pmatrix} 2 & 1 & 0 & 4 \\ 0 & -3 & -12 & -1 \\ 0 & -1 & -2 & 0 \\ 0 & 0 & 3 & 1 \end{pmatrix}$$

$$B^{(2)'} = \begin{pmatrix} 2 \\ 2 \\ -2 \\ -5 \end{pmatrix}$$

Définition 1.1.2. Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$ et $1 \leq k \leq n$. A_k est la sous-matrice principale d'ordre k de A si A_k est la matrice de dimension $k \times k$ composée des coefficients de $(a_{ij})_{1 \leq i,j \leq k}$ de la matrice A.

Theorème 1.1.1. *Soit $A \in \mathcal{M}_{n,n}(\mathbb{R})$.*

- *Si toutes les sous-matrices principales A_k de la matrice de départ A sont régulières* [1]*, alors les pivots obtenus successivement dans l'algorithme de l'élimination de Gauss sont tous non nuls.*
- *Si tous les pivots obtenus dans l'algorithme de l'élimination de Gauss sont non nuls alors toutes les matrices principales A_k de la matrice de départ A sont régulières.*

Démonstration. Soit $A_k^{(i)}$ la sous-matrice principale d'ordre k de $A^{(i)}$. On a $A_i^{(i)}$ qui est triangulaire supérieure. On a :

$$\det(A_i^{(i)}) = \prod_{k=1}^{i} a_{kk}^{(i)}$$

Détail :

$$\det A_1 = a_{11}$$
$$\det A_2^{(1)} = \det A_2^{(2)} = a_{11}^{(2)} a_{22}^{(2)} = a_{22}^{(1)} a_{11}^{(1)}$$
$$\vdots \qquad \vdots$$
$$\det A_i^{(1)} = \cdots = \det A_i^{(i)} = \prod_{k=1}^{i} a_{kk}^{(i)} = \prod_{k=1}^{i} a_{kk}^{(k)}$$

($\Rightarrow$) • $\det(A_1) \neq 0 \Rightarrow a_{11}^{(1)} \neq 0$.
- Si on a : $\det A_2 \neq 0$ et $a_{11}^{(1)} \neq 0$ alors $a_{22}^{(2)} \neq 0$
- ...
- Si on a : $\det A_i \neq 0$ et $a_{11}^{(i)}$, $a_{22}^{(i)}$,...,$a_{i-1,i-1}^{(i-1)}$ sont non nuls alors $a_{kk}^{(k)}$ sont non nuls

($\Leftarrow$) Evident.

□

1.1.3 Complexité algorithmique

On va calculer le nombre d'opérations (c'est-à -dire opérations élémentaires (/, +, ×, −)) necéssaires à la factorisation de la matrice. On considère que chaque opération, prend le même temps de calcul.

On note N, le nombre d'opérations :

$$N = \sum_{i=1}^{n-1} (n-1)[1 + 2(n-1) + 2] = 3 \sum_{i=1}^{n-1} (n-i) + 2 \sum_{i=1}^{n-1} (n-i)^2$$

$$= 2 \sum_{i=1}^{n-1} i + 2 \sum_{i=1}^{n-1} i^2 = 3\frac{(n-1)n}{2} + 2\left(\frac{n^3}{3} - \frac{n^2}{2} + \frac{n}{2}\right) = \mathcal{O}\left(\frac{2n^3}{3}\right)$$

Conclusion : quand n est multiplié par 2, le nombre d'opérations est multiplié par 8.

En ce qui concerne l'algorithme de remontée, il est en $\mathcal{O}(n^2)$. Il a un coût négligeable devant celui de la factorisation.

Remarque (Choix du pivot). Dans l'algorithme, on n'a pas, pour le moment, utilisé de changement de pivot. On peut utiliser différentes stratégies :

[1] c'est-à -dire inversibles

1) Du pivotage partiel : on échange deux lignes pour avoir comme pivot, le nombre de valeur absolue la plus grande possible. La complexité de cette opération est en $\mathcal{O}(n^2)$ (négligeable devant $\mathcal{O}\left(\frac{2n^3}{3}\right)$).
2) Du pivotage total (ou complet) : on échange non seulement les lignes mais aussi les colonnes (on change l'ordre des inconnues). La complexité de cette opération est en $\mathcal{O}(n^3)$ (très couteux, il ne faut jamais l'utiliser).

1.2 Décomposition LU

On suppose qu'on veuille résoudre $AX_1 = B_1$ et $AX_2 = B_2$.

Idée. On va factoriser la matrice A.

1.2.1 Matrice de permutation

Définition 1.2.1. Une permutation σ est une application bijective de l'ensemble $\{1, ..., n\}$ dans lui-même. L'application de σ à $\{1, ..., n\}$ revient donc à réordonner les n nombres.

Définition 1.2.2. On associe à σ l'application linéaire g tel que : $g(e_i) = e_{\sigma(i)}$ pour $i = 1, ..., n$ où $(e_i)_{1\leq i\leq n}$ est une base canonique de $\mathbb{R}^n$. La matrice P qui représente g dans la base $(e_i)_{1\leq i\leq n}$ est appelé matrice de permutation.

Exemple 1.2.1.

i	1	2	3	4
$\sigma(i)$	3	2	4	1

Alors la matrice P de permutation pour σ est :

$$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$$

Remarque. Plus généralement, $(P)_{ij} = \delta_{i,\sigma(j)}$.

Proposition 1.2.1. *Pour toute matrice de transposition, on a :*

$$P^{-1} = {}^tP$$

Définition 1.2.3. On appelle matrice de permutation élémentaire, une permutation σ définie par :

$$\sigma(i) = j,\ \sigma(j) = i,\ \sigma(k) = k \text{ pour tout } k \neq i, j$$

On remarque, dans ce cas :

$$P_{ij} = {}^tP_{ij}$$

et donc :

$$P_{ij}^{-1} = P_{ij}$$

Quelques propriétés

Propriété 1.2.2. *Toute permutation peut s'obtenir comme composition d'au plus $n-1$ permutations élémentaires.*

Exemple 1.2.2. $(1,2,3,4) \to (3,2,4,1)$ peut être vue comme :

$$(1,2,3,4) \xrightarrow{PE} (1,2,4,3) \xrightarrow{PE} (3,2,4,1)$$

Propriété 1.2.3.
- *Si on multiplie à gauche une matrice A par une matrice de permutation P, on permute les lignes de A : la i-ème ligne devient la $\sigma(i)$-ème ligne.*
- *Si on multiplie à droite une matrice A par une matrice de permutation P, on permute les colonnes de A : la j-ème colonne devient à la $\sigma^{-1}(j)$-ème colonne.*

1.2.2 Mise en œuvre de la factorisation

Le principe est d'aboutir à une factorisation de type $PA = LU$ avec :
- P : matrice de permutation.
- A : matrice de départ qu'on suppose inversible.
- L : matrice triangulaire inférieure à diagonale unité.
- U : matrice triangulaire supérieure

Pourquoi ? On veut résoudre :

$$AX = B$$

Ce qui est équivalent à résoudre :

$$PAX = PB$$

Mais avec la factorisation LU, on a à résoudre :

$$LUX = PB$$

Ce qui est équivalent à résoudre :

$$\begin{cases} LY = PB \text{ (descente } \mathcal{O}(n^2)) \\ UX = Y \text{(remontée } \mathcal{O}(n^2)) \end{cases}$$

Quel lien y-a-t-il entre $A^{(i)}$ et $A^{(i+1)}$? Soit $1 \leq i \leq n-1$

1) D'abord, on choisit le $i$ème pivot en prenant l'indice k tel que $|a_{ki}^{(i)}| = \max\limits_{1 \leq l \leq n} |a_{li}^{(i)}|$.

2) On va alors permuter les lignes i et k de la matrice

$$\tilde{A}^{(i)} = P_{ik} A^{(i)}$$

Pour simplifier, on note $P^{(i)} = P_{ik}$. On a :

$$P^{(i)} = \left(P^{(i)}\right)^{-1}$$

Donc :

$$A^{(i)} = P^{(i)} \tilde{A}^{(i)}$$

3) Maintenant, pour obtenir $A^{(i+1)}$ à partir de $\tilde{A}^{(i)}$, on retranche à la $k$ième ligne de $\tilde{A}^{(i)}$, $\frac{\tilde{a}_{kj}^{(i)}}{\tilde{a}_{ii}^{(i)}}$ à la ligne i de $\tilde{A}^{(i)}$ $(1 \leq i \leq k \leq n)$.

Pour obtenir $\tilde{A}^{(i)}$ à partir de $A^{(i+1)}$, on ajoute à la $k$ième ligne de $A^{(i+1)}$, $\frac{\tilde{a}_{kj}^{(i)}}{\tilde{a}_{ii}^{(i)}}$ fois la ligne i de $A^{(i+1)}$. On a ainsi :

$$\tilde{A}^{(i)} = L^{(i)} A^{(i+1)}$$

avec :

$$L^{(i)} = \begin{pmatrix} & & i & \\ 1 & & \vdots & \\ & \ddots & \vdots & 0 \\ & & l_i(k) & \\ & 0 & l_i(k+1) & \ddots \\ & & \vdots & 1 \end{pmatrix}$$

où les $l_i(k) = \frac{\tilde{a}_{kj}^{(i)}}{\tilde{a}_{ii}^{(i)}}$ avec $1 \leq k \leq n$. On a ainsi :

$$A^{(i)} = P^{(i)} \tilde{A}(i) = P^{(i)} L^{(i)} A^{(i+1)}$$

d'où

$$A = A^{(1)} = P^{(1)} L^{(1)} A^{(2)} = ... = P^{(1)} L^{(1)} P^{(2)} l^{(2)} ... P^{(n-1)} L^{(n-1)} \underbrace{A^{(n)}}_{U} \qquad (*)$$

Lemme 1.2.4. *Soit $P^{(p)}$ matrice de permutation élémentaire entre les indices p et $q \geq p$. Alors pour $k < p$, on a :*

$$L^{(k)} P^{(p)} = P^{(p)} L'^{(k)}$$

ou encore :

$$P^{(p)} L^{(k)} P^{(p)} = L'^{(k)}$$

où $L'^{(k)}$ se déduit de $L^{(k)}$ par permutation des lignes p et q dans la $k$ième colonne.

Démonstration. On considère la base canonique de $\mathbb{R}^n$. Donc :

$$e_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \leftarrow k\text{-ième ligne}$$

Alors pour $p > k$, on a : $P^{(p)}e_k = e_k$. On définit l_k par $L^{(k)} = I + l_k e_k^t$ [2]. Donc :

$$l_k = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ l_k(k+1) \\ \vdots \\ l_k(n) \end{pmatrix}$$

Donc :

$$\begin{aligned} L'^{(k)} = P^{(p)}L^{(k)}P^{(p)} &= P^{(p)}(I + l_k e_k^t)P^{(p)} = (P^{(p)})^2 + P^{(p)}l_k e_k^t P^{(p)} \\ &= I + P^{(p)}l_k\,{}^t(P^{(p)}e_k) \text{ car } {}^tP^{(p)} = P^{(p)} \\ &= I + P^{(p)}l_k e_k^t \end{aligned}$$

□

Lemme 1.2.5. *Soit $(l_k)_{1\leq k\leq n-1}$ une suite de $(n-1)$ vecteurs de dimension n tel que les k premières composantes de l_k soient nulles. Alors :*

$$\prod_{k=0}^{n-1}(I + l_k e_k^t) = I + \sum_{k=1}^{n-1} l_k e_k^t$$

Démonstration. On développe le terme de gauche et on remarque que les termes contenant un facteur de type $l_i e_i^t l_j e_j^t$ avec ij sont nuls car $e_i^t l_j = l_j(i) = 0$ pour $j > i$. □

Retour à la décomposition

On retrouve l'expression et on utilise le **Lemme 1.2.4.** pour avoir :

$$A = P^{(1)}P^{(2)}...P^{(n-1)}L''^{(1)}L''^{(2)}...L''^{n-1}U$$

avec :

$$L''^{(k)} = P^{(n-1)}P^{(n-2)}...P^{(k+1)}L^{(k)}P^{(k+1)}P^{(k+2)}...P^{(n-2)}P^{(n-1)}$$

On a ainsi :

$$L''^{(1)}L''^{(2)}...L''^{(n-1)} = L$$

L est triangulaire inférieur à diagonale unité (d'après le **Lemme 1.2.5.**). On pose :

$$Q = P^{(1)}P^{(2)}...P^{(n-1)}$$

On a ainsi :

$$A = QLU$$

et si on pose $Q^{-1} = P$, on a :

$$\boxed{PA = LU}$$

Theorème 1.2.6. *1) Soit A une matrice carrée régulière d'ordre n. Il existe une matrice de permutations P et deux matrices L et U respectivement triangulaire inférieure à diagonale unité et triangulaire supérieure tel que $PA = LU$.*

[2] il faut lire $I + l_k$ transposée de e_k

2) Si A vérifie les hypothèses du ***Théorème 1.1.1.*** *alors on peut prendre $P = I$ et la décomposition $A = LU$ est unique.*

Démonstration de l'unicité de la décomposition $A = LU$. On suppose que $A = L_1U_1$ et $A = L_2U_2$. Alors :

$$L_1U_1 = L_2U_2 \Leftrightarrow L_2^{-1}L_1U_1 = U_2 \Leftrightarrow L_2^{-1}L_1 = U_2U_1^{-1} = X$$

Comme $X = L_2^{-1}L_1$, X triangulaire inferieur à diagonal unité. On a aussi : $X = U_2U_1^{-1}$ alors X tiangulaire supérieure. Alors $X = I$. Donc $L_2 = L_1$ et $U_2 = U_1$.

□

1.2.3 Complexité algorithmique

Factorisation de la matrice $A = LU$

Soit

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

$$L = \begin{pmatrix} 1 & & & \\ l_{21} & \ddots & 0 & \\ l_{31} & l_{32} & \ddots & \\ \vdots & \vdots & \ddots & \\ l_{n1} & & l_{n,n-1} & 1 \end{pmatrix}$$

et

$$U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1n} \\ & u_{21} & \cdots & u_{2n} \\ 0 & & \ddots & \\ & & & u_{nn} \end{pmatrix}$$

On a donc :

$$(1) : \begin{cases} a_{11} = u_{11} \\ a_{22} = u_{22} \\ \vdots \\ a_{nn} = u_{nn} \end{cases} \qquad (2) : \begin{cases} a_{21} = l_{21}u_{11} \\ a_{31} = l_{31}u_{11} \\ \vdots \\ a_{n1} = l_{n1}u_{11} \end{cases}$$

(1) : on a la première ligne de U. (2) : on a la première ligne de L.

$$(3) : \begin{cases} a_{22} = l_{21}u_{12} + u_{22} \\ \vdots \\ a_{2n} = l_{21}u_{1n} + u_{2n} \end{cases} \qquad (4) : \begin{cases} a_{32} = l_{31}u_{12} + l_{32}u_{22} \\ a_{42} = l_{41}u_{12} + l_{42}u_{22} \\ \vdots \\ a_{n2} = l_{n1}u_{12} + l_{n2}u_{22} \end{cases}$$

Algorithme 1.2.1.
CONSTRUCTION DE LA PREMIÈRE LIGNE DE $L()$
1 **for** $i \leftarrow 2$ **to** n
2 **do** $a_{i1} \leftarrow \frac{a_{i1}}{a_{11}}$

CONSTRUCTION DE LA kIÈME LIGNE DE U ET DE LA kIÈME COLONNE DE $L()$

for $k \leftarrow 2$ **to** $n-1$
  do $a_{kk} \leftarrow a_{kk} - \sum_{j=1}^{k-1} a_{kj}a_{jk}$
    for $i \leftarrow k+1$ **to** n
      do $a_{ki} \leftarrow a_{ki} - \sum_{j=1}^{k-1} a_{kj}a_{ji}$
        $a_{ik} \leftarrow \frac{1}{a_{kk}}\left(a_{ij} - \sum_{j=1}^{k-1} a_{ij}a_{jk}\right)$

CONSTRUCTION DE LA DERNIÈRE LIGNE DE $U()$

$a_{nn} \leftarrow a_{nn} - \sum_{j=1}^{n-1} a_{nj}a_{jn}$

Résolution du système linéaire

$LUX = B$

1) Résolution de $LY = B$, T stocké dans X

$LY = B()$

for $i \leftarrow 1$ **to** n
do $c \leftarrow 0$
  for $j \leftarrow 1$ **to** $i-1$
  do $c \leftarrow c + a_{ij}x_j$
    $x_i \leftarrow b_i - c$

2) Résolution de $UX = Y$, X contenant Y en entrée et "le vrai X" en sortie.

$UX = Y()$

for $i \leftarrow n$ **to** 1
do $c \leftarrow 0$
  for $j \leftarrow i+1$ **to** n
  do $c \leftarrow c + a_{ij} \times x_j x_i = \frac{x_i - c}{a_{ii}}$

Remarque. $\rightarrow$ Complexité de la factorisation $A = LU$ est de $\mathcal{O}\left(\frac{2n^3}{3}\right)$

1.2.4 Le cas des matrices bandes

Définition 1.2.4. Soit A une matrice de dimension $n \times n$ et de coefficients a_{ij}, $1 \leq i, j \leq n$, et soit $l \in \mathbb{N}^*$ tel que $l \leq n$. On dit que A est une matrce de bande de demi-largeur l si pour tout i, j satisfaisant $1 \leq i, j \leq n$ et $|i-j| \geq l$, on ait $a_{ij} = 0$.

Exemple 1.2.3.

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

Matrice bande de demi-largeur $l = 2$ car soit $1 \leq i, j \leq n$:

$$|i-j| \geq 2 \Rightarrow a_{ij} = 0$$

Theorème 1.2.7. *Soit A une matrice de dimension $n \times n$ tel que A est une matrice bande de demi-largeur l, dont toutes les sous-matrices principales sont régulières (existence de la factorisation $A = LU$). Alors la décomposition LU de A donne lieu à des matrices triangulaires L et U qui sont aussi de demi-largeur de bande l. La complexité algorithmique de la factorisation est alors de l'ordre de $\mathcal{O}(nl^2)$ quand n est grand.*

1.3 Décomposition de Cholesky

On va voir ici quelques cas particuliers de matrices.

1.3.1 Les matrices symétriques

Theorème 1.3.1. *Soit A une matrice symétrique, régulière, possédant une décomposition $A = LU$. Alors on peut factoriser A sous la forme :*

$$A = LD^t$$

avec L matrice triangulaire inférieure à diagonale unité et D une matrice diagonale.

Démonstration. $A = LU$, U triangulaire supérieure. On peut donc écrire $U = DR$ avec D matrice diagonale et R triangulaire à diagonale unité. Cette décomposition est unique :

$$A = LDR \text{ et } {}^tA = {}^tRD^tL$$

Comme $A = {}^tA$ alors :

$$LDR = {}^tRD^tL$$

avec :

- L = triangulaire inférieure à diagonale unité.
- DR = triangulaire supérieure
- ${}^tR$ = triangulaire inférieure à diagonale unité.
- D^tL = triangulaire supérieure.

L'unicité de la décomposition $LU \Rightarrow L = {}^tR$. Donc : $A = LD^tL$. □

1.3.2 Les matrices symétriques définies positives

Définition 1.3.1. Une matrice $A \in \mathbb{R}^{n\times n}$ est dite symétrique définie positive si :

(i) $A = {}^tA$

(ii) ${}^tyAy \geq 0$, $\forall y \in \mathbb{R}^n$

(iii) ${}^tyAy = 0 \Leftrightarrow y = 0$.

Theorème 1.3.2. *Si $A \in \mathbb{R}^{n\times n}$ est symétrique définie positive alors toutes ses sous-matrices principales sont symétriques définies positives.*

Démonstration. (i) Soit A_k la sous-matrice principale d'ordre k ($1 \leq k \leq n$). On a ${}^tA_k = A_k$.

(ii) Soit z un vecteur de dimension k ($z \in \mathbb{R}^k$) et $y \in \mathbb{R}^n$ tel que :

$$y = \begin{bmatrix} \vdots \\ z \\ \vdots \\ 0 \\ \vdots \end{bmatrix}$$

On aura :

$$^tzA_kz = {}^tyAy$$

Comme A est symétrique définie positive, les propositions (ii) et (iii) sont vérifiés pour A_k.

□

Corollaire. *Soit $A \in \mathbb{R}^{n\times n}$ symétrique. A est définie positive si et seulement si toutes ses valeurs propres sont définies positives.*

Theorème 1.3.3. *Soit A une matrice symétrique définie positive alors il existe une unique matrice B triangulaire inférieure à valeurs diagonales positives telle que :*

$$A = B^tB$$

Démonstration. Si A_k, L_k et U_k des sous-matrices principales d'ordre k de A, L et U, on vérifie que :

$$A_k = L_kU_k$$

Comme A_k est symétrique définie positive, d'après le **Théorème 1.3.2**, on a :

$$\det A_k = \prod_{i=1}^{k} u_{ii} > 0$$

En prenant successivement $k = 1, 2, ..., n$, on a :

$$u_{jj} > 0 \text{ tel que } j = 1, 2, .., n$$

on définit donc D par :

$$D = \begin{pmatrix} \sqrt{u_{11}} & & & \\ & \sqrt{u_{22}} & 0 & \\ & 0 & \ddots & \\ & & & \sqrt{u_{nn}} \end{pmatrix}$$

et E par :

$$E = \begin{pmatrix} u_{11} & & & \\ & u_{22} & 0 & \\ & 0 & \ddots & \\ & & & u_{nn} \end{pmatrix}$$

$\tilde{U} = E^{-1}U$ triangulaire supérieure à diagonale unité.

$$A = LE\tilde{U} \underbrace{=}_{A \text{ symétrique}} {}^t\tilde{U}E^tL$$

(i) L triangulaire inférieure à diagonale unité

(ii) ${}^t\tilde{U}$ triangulaire inférieure à diagonale unité

(iii) Décomposition LU unique

(i), (ii) et (iii) implique que ${}^tL = \tilde{U}$.

$$A = {}^t\tilde{U}E\tilde{U} = {}^t\tilde{U}DD\tilde{U}$$

On pose $B = {}^t\tilde{U}D$, B triangulaire inférieure à termes diagonaux positifs. On a ainsi :

$$A = B\,{}^tB$$

Unicité : On suppose :

$$\begin{cases} A = B_1\,{}^tB_1 \\ A = B_2\,{}^tB_2 \end{cases}$$

Donc :

$$B_1\,{}^tB_1 = B_2\,{}^tB_2 \Leftrightarrow {}^tB_2^{-1}B_1\,{}^tB_1 = {}^tB_2 \Leftrightarrow \underbrace{{}^tB_2^{-1}B_1}_{\text{triang inf}} = \underbrace{{}^tB_2\,{}^tB_2^{-1}}_{\text{triang sup}} \qquad (*)$$

donc $B_2^{-1}B_1$ par diagonale :

$$(*) \Leftrightarrow \frac{(b_2)_{jj}}{(b_1)_{jj}} = \frac{(b_1)_{jj}}{(b_2)_{jj}} \qquad 1 \leq j \leq n$$

Donc $B_2^{-1}B_1 = I$ car $(b_1)_{jj}$ et $(b_2)_{jj}$ sont > 0. Donc $B_1 = B_2$. □

Proposition d'algorithme :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ & a_{22} & & \vdots \\ & sym & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix} = \begin{pmatrix} b_{11} & & 0 & \\ b_{21} & b_{22} & & \\ \vdots & & \ddots & \\ b_{n1} & \cdots & \cdots & b_{nn} \end{pmatrix} \times \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ & b_{22} & & \vdots \\ & & \ddots & \vdots \\ & 0 & & b_{nn} \end{pmatrix}$$

Construction de $B()$

```
1  a_11 ← √a_11
2  # Première colonne de B
3  for i ← 2 to n
4      do a_i1 ← a_i1 / a_11
5
6  # Colonne de B, 2 ≤ k ≤ n − 1
7  for k ← 2 to n − 1
8      do a_kk ← (a_kk − Σ_{j=1}^{k−1} a_kj^2)^{1/2}
9          for i ← k + 1 to n
10             do a_ik ← (1/a_kk) (a_kk − Σ_{j=1}^{k−1} a_ij a_kj)
11
12 # Dernière colonne de B
13 a_nn ← a_nn − (Σ_{j=1}^{n−1} a_nj^2)^{1/2}
```

Chapitre 2

Normes et conditionnement

2.0 Quelques rappels d'algèbre matricielle

On considère le corps $\mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$). On considère une matrice $A \in \mathbb{K}^{m\times n}$ avec $m \in \mathbb{N}^*$ et $n \in \mathbb{N}^*$. On note $A = (a_{ij})_{1\leq i\leq m, 1\leq j\leq n}$ les coefficients de A.

Définition 2.0.1. On appelle matrice adjointe de $A \in \mathbb{C}^{m\times n}$ et on note A^* la matrice définie par : $A^* = (\overline{a_{ji}})_{1\leq i\leq m, 1\leq j\leq n}$

Définition 2.0.2. On appelle matrice transposée de $A \in \mathbb{R}^{m\times n}$ et on note ${}^tA$, la matrice définie par ${}^tA = (a_{ji})_{1\leq i\leq m, 1\leq j\leq n}$.

Remarque. (i) $(AB)^{-1} = A^{-1}B^{-1}$ si A et B inversibles.

(ii) $({}^tA)^{-1} = {}^t(A^{-1})$

(iii) $(A^*)^{-1} = (A^{-1})^*$

(iv) ${}^t(AB) = {}^tB{}^tA$

(v) $(AB)^* = B^*A^*$

Définition 2.0.3. Une matrice A est dite :
- symétrique si $A \in \mathbb{R}^{m\times n}$ et ${}^tA = A$
- hermitienne si $A \in \mathbb{C}^{m\times n}$ et $A^* = A$
- orthogonale si $A \in \mathbb{R}^{m\times n}$ et $A{}^tA = {}^tAA = I$
- unitaire si $A \in \mathbb{C}^{m\times n}$ et $AA^* = A^*A = I$
- normale si $A \in \mathbb{C}^{m\times n}$ et $AA^* = A^*A$

Définition 2.0.4. On appelle vecteur propre u associé à une valeur propre λ de A :

$$Au = \lambda u,\ \lambda \in \mathbb{K}$$

(λ, u) est l'élément propre.

Définition 2.0.5. Le spectre de A est l'ensemble des valeurs propres $\mathrm{Sp}(A) = \bigcup_{i=1}^{n}(u_i, \lambda_i)$.

Définition 2.0.6. Le rayon spectral de A est défini par :

$$\rho(A) = \max_{1\leq i\leq n} \lambda_i(A), \lambda_i \in \mathrm{Sp}(A), 1 \leq i \leq n$$

Définition 2.0.7. Sur $\mathbb{R}^n$, on définit le produit scalaire euclidien :

$$(u,v) = {}^tvu = \sum_{i=1}^{n} u_i v_i, u \in \mathbb{R}^n, v \in \mathbb{R}^n$$

Sur $\mathbb{C}^n$, on définit le produit scalaire hermitien :

$$(u,v) = v^*u = \overline{u^*v} = \sum_{i=1}^{n} u_i \overline{v_i}$$

Deux vecteurs u et v sont dits orthogonaux si $(u,v) = 0$.

Propriété 2.0.1. *Soit $u \in \mathbb{K}^n$ et $v \in \mathbb{K}^m$, $A \in \mathbb{K}^{m\times n}$.*
- *Sur $\mathbb{R}^n$, $(Au,v) = (u, {}^tAv)$*
- *Sur $\mathbb{C}^n$, $(Au,v) = (u, A^*v)$*

2.1 Normes matricielles

2.1.1 Normes vectorielles

Définition 2.1.1. Soit E un espace vectoriel sur le corps $\mathbb{K}$. Une norme sur E est une application de E dans $\mathbb{R}^+$ qui a tout élément x de E, fait correspondre le nombre réel positif non nul noté $\|x\|$, appelé norme de x et possédant les propriétés suivantes :

(1) $\|x\| = 0 \Leftrightarrow x = 0, \forall x \in E$

(2) $\|\lambda x\| = |\lambda|\|x\|, \forall x \in E, \forall \lambda \in \mathbb{K}$

(3) $\|x+y\| \leq \|x\| + \|y\|, \forall (x,y) \in E^2$

Exemple 2.1.1. $x \in \mathbb{C}^n$

A) $x \to \|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$

B) $x \to \|x\|_1 = \sum_{i=1}^{n} |x_i|$

C) $x \to \|x\|_\infty = \max_{1\leq i\leq n} |x_i|$

Démonstration de (3) *pour la norme A*). On utilise le fait que $\alpha \geq 0$, $\beta \geq 0$

$$\Rightarrow \alpha\beta \leq \frac{\alpha^2}{2} + \frac{\beta^2}{2}$$

car $(\alpha - \beta)^2 \geq 0$. Soient $u \in \mathbb{C}^n$ et $v \in \mathbb{C}^n$ et $1 \leq i \leq n$.

$$\frac{|u_i v_i|}{\|u\|_2\|v\|_2} = \frac{|u_i|}{\|u\|_2}\frac{|v_i|}{\|v\|_2} \leq \frac{1}{2}\frac{|u_i|^2}{\|u\|_2} + \frac{1}{2}\frac{|v_i|^2}{\|v\|_2}$$

donc :

$$\sum_{i=1}^{n} |u_i v_i| \leq \|u\|_2\|v\|_2$$

$$\|x-y\|_2 = \sum_{i=1}^{n} |x_i + y_i|^2 \leq \sum_{i=1}^{n} (|x_i|^2 + |y_i|^2 - 2|x_i||y_i|)$$

$$\leq \sum_{i=1}^{n} |x_i|^2 + \sum_{i=1}^{n} |y_i|^2 + 2\left(\sum_{i=1}^{n} |x_i|\right)\left(\sum_{i=1}^{n} |y_i|\right) = (\|x\|_2 + \|y\|_2)^2$$

□

Démonstration. On peut montrer aussi l'équivalence des normes, c'est-à -dire :

$$\frac{1}{n}\|x\|_1 \leq \|x\|_\infty \leq \|x\|_1$$

$$\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2$$

$$\|x\|_2 \leq \|x\|_1 \leq n\|x\|_2$$

□

2.1.2 Normes matricielles

Définition 2.1.2. $A \in \mathbb{K}^{m\times n}$. Soient deux normes vectorielles définies sur $\mathbb{K}^m$ et sur $\mathbb{K}^n$ et toutes les deux notées $\|\cdot\|$ (par exemple, on prend comme première norme, la norme $\|\cdot\|_2$ sur $\mathbb{K}^m$ et comme deuxième norme, la norme $\|\cdot\|_2$ sur $\mathbb{K}^n$). On appelle norme matricielle subordonée le nombre :

$$\|A\| = \max_{\|x\|=1, x\in\mathbb{R}^n} \|Ax\|$$

Lemme 2.1.1. *Pour toute norme matricielle subordonée :*

$$\|Ax\| \leq \|A\|\|x\|, \forall x \in \mathbb{K}^n$$

Démonstration.
- Vrai pour $x = 0$.
- Soit $x \neq 0$, $\frac{x}{\|x\|}$ est un vecteur unitaire :

$$\left\|A\frac{x}{\|x\|}\right\| \leq \max_{\|x\|=1} \|Au\|$$

On a donc :

$$\|Ax\| \leq \|A\|\|x\|$$

□

Theorème 2.1.2. *Toute norme matricielle subordonée vérifie les propriétés suivantes :*

(i) $\forall A \in \mathbb{K}^{m\times n}$, $\|A\| = 0 \Leftrightarrow A = 0$

(ii) $\forall A \in \mathbb{K}^{m\times n}$, $\forall \lambda \in \mathbb{K}$, $\|\lambda A\| = |\lambda|\|A\|$

(iii) $\forall A, B \in \mathbb{K}^{m\times n}$, $\|A + B\| \leq \|A\| + \|B\|$.

(iv) $\forall A \in \mathbb{K}^{m\times n}$, $\forall B \in \mathbb{K}^{n\times p}$, $\|AB\| \leq \|A\|\|B\|$.

Démonstration. (i) ($\Leftarrow$) $A = 0 \Rightarrow \forall x \in \mathbb{R}^n$, $Ax = 0 \Rightarrow \|A\| = \max_{\|x\|=1} \|Ax\| = 0$
($\Rightarrow$) $\|A\| = \max_{\|x\|=1} \|Ax\| = 0 \Rightarrow \|Ax\| = 0$, $\forall x \in \mathbb{K}^n \Rightarrow A = 0$.

(ii) évident

(iii)

$$\|(A + B)x\| = \|Ax + Bx\| \leq \|Ax\| + \|Bx\|$$

$$\max_{\|x\|=1} \|(A \times B)x\| \leq \max_{\|x\|=1} \|Ax\| + \max_{\|x\|=1} \|Bx\|$$

$$\|A + B\| \leq \|A\| + \|B\|$$

(iv)

$$\|ABx\| = \|A(Bx)\| \leq \|A\|\|Bx\| \text{ d'après le Lemme 2.1.1.}$$
$$\max_{\|x\|=1} \|ABx\| \leq \|A\| \max_{\|x\|=1} \|Bx\| = \|A\|\|B\|$$

□

Remarque. Pour toute norme matricielle subordonnée, on a $\|I\| = 1$ car :

$$\|I\| = \max_{\|x\|=1} \|Ix\| = \max_{\|x\|=1} \|x\| = 1$$

Définition 2.1.3. On appelle norme de Schur de $A \in \mathbb{K}^{n\times m}$ le nombre :

$$\|A\|_S = \left(\sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}$$

Theorème 2.1.3. *La norme de Schur est une norme matricielle au sens où elle vérifie les* 3 *premières propriétés du* ***Théorème 2.1.2.****. Mais ce n'est pas une norme subordonée. De plus, elle vérifie la dernière propriété du* ***Théorème 2.1.2.***

Démonstration. Démonstration de la propriété (iv) : Pour $A \in \mathbb{R}^{n\times n}$ et $B \in \mathbb{R}^{n\times n}$

$$\|AB\|_S^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}\left|\sum_{k=1}^{n} a_{ik}b_{kj}\right|^2 \leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}||b_{kj}|\right)^2$$
$$\leq \sum_{i=1}^{n}\sum_{j=1}^{n}\left(\sum_{k=1}^{n}|a_{ik}^2|\right)\left(\sum_{k=1}^{n}|b_{kj}|^2\right)$$
$$\leq \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ik}|^2\right)\left(\sum_{j=1}^{n}\sum_{k=1}^{n}|b_{kj}|^2\right) = \|A\|_S^2\|B\|_S^2$$

□

Remarque. On a : $\|I\|^2 = n \neq 1$. Ce n'est pas une norme induite.

Définition 2.1.4. On appelle norme matricielle, une application $\mathbb{K}^{m\times n} \to \mathbb{R}^+$ qui vérifie les 3 premières propriétés du **Théorème 2.1.2.**

2.1.3 Valeurs singulières

Définition 2.1.5. On appelle valeurs singulières d'une matrice $A \in \mathbb{C}^{m,n}$, les racines carrées positives ou nulles des valeurs propres de la matrice $A^*A \in \mathbb{C}^{n,n}$.

Remarque. Ces valeurs singulières sont bien définies. En effet, A^*A est une matrice hermitienne. Donc, il existe une matrice unitaire θ et une matrice diagonale D tel que :

$$A^*A = \theta^* D\theta$$

et les valeurs propres sont réelles. Les valeurs propres sont forcément positives ou nulles car :

$$\text{Pour } x \neq 0,\ A^*Ax = \lambda x \Rightarrow (A^*Ax, x) = \lambda(x,x) \Rightarrow (Ax, Ax) = \lambda(x,x) \Leftrightarrow \underbrace{\|A_x\|_2^2}_{\geq 0} = \lambda \underbrace{\|x\|_2^2}_{\geq 0}$$

Donc : $\lambda \geq 0$.

Theorème 2.1.4 (Décomposition en valeurs singulières)**.** *Soit $A \in \mathbb{C}^{m,n}$. Il existe deux matrices unitaires et carrées U et V avec $U \in \mathbb{C}^{m,m}$ et $V \in \mathbb{C}^{n,n}$ tel que :*

$$U^*AV = \Sigma$$

où Σ est une matrice rectangulaire de format $m \times n$ et :

$$\Sigma = \begin{pmatrix} \mu_1 & & & & \\ & \mu_2 & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \mu_n \\ \cdots & \cdots & 0 & \cdots & \cdots \end{pmatrix}^{(m>n)}$$

où les $(\mu_i)_{1\leq i\leq n}$ sont les valeurs singulières de la matrice A.

Démonstration. A^*A est hermitienne donc il existe une matrice V unitaire d'ordre n tel que :

$$V^*(A^*A)V = \begin{pmatrix} \mu_1^2 & & & \\ & \mu_2^2 & 0 & \\ & 0 & \ddots & \\ & & & \mu_n^2 \end{pmatrix}$$

Soit C_j le vecteur correspondant à la $j$ième colonne de la matrice $C = AV$.

$$C^*C = \mathrm{diag}(\mu_i^2)$$

Donc :

$$C_i^*C_j = \mu_i^2\delta_{ij}$$

Soit $\{\mu_1, ..., \mu_r\}$ l'ensemble des valeurs singulières non nulles et soit $\{\mu_{r+1}, ..., \mu_n\}$ l'ensemble des valeurs singulières nulles. On a donc :

$$C_j = 0 \text{ pour } r+1 \leq j \leq n \text{ car } \|C_j\| = 0$$

On pose $u_j = \frac{c_j}{\mu_j}$ pour $1 \leq j \leq n$. Par construction :

$$u_i^*u_j = \delta_{ij} \text{ pour } 1 \leq j \leq n$$

On complète alors les r vecteurs u_j (pour $1 \leq j \leq n$) pour ainsi obtenir une base orthonormée de $\mathbb{C}^n$.

$$\mathbb{C}^n = \mathrm{Sp}\{u_1, u_2, ..., u_r, u_{r+1}, ..., u_n\} \qquad \text{(Gram-Schmidt)}$$

On a maintenant :

$$u_i^*u_j = \delta_{ij} \text{ pour } 1 \leq i, j \leq m$$

et :

$$C_j = \mu_j u_j \text{ pour } 1 \leq i \leq n$$

Soit U la matrice carrée de dimension $m \times m$ dont la $j$ième colonne est formée du vecteur u_j. On a ainsi une matrice unitaire. On a $(U^*AV)_{ij} = u_i^*C_j$ pour $1 \leq i \leq m$, $1 \leq j \leq n$. Ce qui s'écrit :

$$U^*AV = \Sigma = \begin{pmatrix} \mu_1 & & \\ & \ddots & \\ & & \mu_n \\ \cdots & 0 & \cdots \end{pmatrix}$$

□

Remarque. Dans le cas où $A \in \mathbb{R}^{m\times n}$ alors $A^* = {}^tA$. On remplace alors les * par des ^{t}. On parle de matrice orthogonale (à la place de matrice unitaire) et de matrice symétrique (à la place de hermitienne).

2.1.4 Calcul de normes matricielles

Theorème 2.1.5. *Soit $A \in \mathbb{C}^{m,n}$. Alors*

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \mu_1$$

la plus grande valeur singulière de A. On rappelle que :

$$\rho(A) = \max_{\lambda\in \mathrm{Sp}(A)} |\lambda_i|$$

Démonstration. Il existe une base orthonormée de $\mathbb{C}^n$, notée $\{e_i\}_{1\leq i\leq n}$, constituée de vecteurs propres de A^*A associés aux valeurs propres μ_i^2. Donc :

$$\forall 1 \leq i \leq n,\ A^*Ae_i = \mu_i^2 e_i$$

Soit $x \in \mathbb{C}^n$. Alors :

$$x = \sum_{i=1}^{n} \tilde{x_i} e_i$$

$$\|x\|_2^2 = \sum_{i=1}^{n} |\tilde{x_i}|^2$$

Donc :

$$\|Ax\|_2^2 = (Ax, Ax) = (A^*Ax, x) = \left(AA^* \sum_{i=1}^{n} \tilde{x_i} e_i, \sum_{j=1}^{n} \tilde{x_j} e_j \right)$$

$$= \sum_{i=1}^{n}\sum_{j=1}^{n} (\tilde{x_i}\mu_i^2 e_i, \tilde{x_j} e_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{x_i}\tilde{x_j}\mu_i^2 \underbrace{(e_i, e_j)}_{\delta_{ij}} = \sum_{i=1}^{n} \mu_i^2 |\tilde{x_i}|^2$$

On a ainsi :

$$\|Ax\|_2^2 \leq \mu_1^2 \sum_{i=1}^{n} |\tilde{x_i}^2| = \mu_1^2 \|x\|_2^2$$

Donc, on a que, pour $\|x\|_2 = 1$:

$$\|Ax\|_2^2 \leq \mu_i^2$$

et donc :

$$\underbrace{\max_{\|x\|_2=1} \|Ax\|_2^2}_{\|A\|_2} \leq \mu_1$$

On va choisir maintenant un x particulier pour lequel l'égalité est vérifiée. On choisit $x = e_1$ et on aura :

$$\|Ax\|_2^2 = \mu_1$$

□

Corollaire. *Soit $A \in \mathbb{C}^{m,n}$,*

(i) $\|A\|_2 = \|A^*\|_2$

(ii) Soit A matrice hermitienne. Alors $\|A\|_2 = \rho(A)$.

(iii) Soit $A \in \mathbb{C}^{m,n}$, $U \in \mathbb{C}^{n,n}$ et $V \in \mathbb{C}^{m,m}$ avec U et V unitaires alors $\|VAU\|_2 = \|A\|_2$.

Démonstration. (i) Soit $\lambda \in \mathbb{R}$, $\lambda \neq 0$ tel que :

$$A^*Ax = \lambda x,\ x \neq 0$$

donc :

$$A(A^*Ax) = \lambda(Ax) \Leftrightarrow AA^*(Ax) = \lambda(Ax)$$

donc λ est valeur propre de AA^* associé au vecteur propre $Ax \neq 0$. Les valeurs propres non nulles de A^*A et de AA^* sont les mêmes.

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

(ii) Soit A une matrice hermitienne, on a :

$$\|A\|_2 = \sqrt{\rho(A^*A)} = \sqrt{\rho(A^2)} = \sqrt{\rho(A)^2} = \rho(A)$$

(iii) Soit $A \in \mathbb{C}^{m,n}$, $U \in \mathbb{C}^{n,n}$ et $V \in \mathbb{C}^{m,m}$ avec U et V unitaires, on a :

$$\|VAU\|_2 = \max_{\|x\|_2=1} \|VAU(x)\|_2 = \max_{\|y\|_2=1} \|VAy\|_2$$

car en posant $y = U(x)$, on a :

$$\|y\|_2 = \|x\|_2 \ [1]$$

or :

$$\max_{\|y\|_2=1} \|VAy\|_2^2 = \max_{\|y\|_2=1} (VAy, VAy) = \max_{\|y\|_2=1} (\underbrace{V^*V}_{I} Ay, Ay) = \max_{\|y\|_2=1} \|Ay\|_2^2 = \|A\|_2$$

donc : $\|VAU\|_2 = \|A\|_2$

□

Theorème 2.1.6. *Soit $A \in \mathbb{C}^{m,n}$ alors :*

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$$

Démonstration. Soit $x \in \mathbb{C}^n$ tel que $\|x\|_1 = 1$ alors :

$$\|Ax\|_1 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij}x_j \right| \leq \sum_{j=1}^{n} \sum_{i=1}^{m} |a_{ij}||x_j| = \underbrace{\sum_{j=1}^{n} |x_j|}_{=1} \sum_{i=1}^{m} |a_{ij}| \leq \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|$$

Donc :

$$\max_{\|x\|_1=1} \|Ax\|_1 \leq \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|$$

On note k tel que :

$$\max_{1\leq j\leq n} \left(\sum_{i=1}^{m} |a_{ij}| \right) = \sum_{i=1}^{m} |a_{ij}|$$

[1] $\|y\|^2 = (y, y) = (Ux, Ux) = (\underbrace{U^*U}_{I} xx, x) = \|x\|_2^2$

Soit le vecteur x de composante x_j suivante :

$$x_j = \begin{cases} 0 \text{ si } j \neq k \\ 1 \text{ si } j = k \end{cases} = \delta_{jk} \qquad \|x\|_1 = 1$$

On a ainsi :

$$\|Ax\|_1 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| = \sum_{i=1}^{m} |a_{ik}| = \max_{1 \leq j \leq n} \left(\sum_{i=1}^{m} |a_{ij}| \right)$$

□

Theorème 2.1.7. *Soit $A \in \mathbb{C}^{m,n}$ alors :*

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$$

Démonstration. Soit x tel que $\|x\|_\infty = 1$ alors :

$$\|Ax\|_\infty = \max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}||x_j| \leq \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$$

Soit k tel que :

$$\max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{kj}|$$

Soit x défini par :

$$x_j = \begin{cases} \frac{a_{kj}}{|a_{kj}|} \text{ si } a_{kj} \neq 0 \\ 1 \text{ si } a_{kj} = 0 \end{cases} \qquad \|x\|_\infty = 1$$

– Si $i \neq k$:

$$\left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \sum_{j=1}^{n} |a_{ij}| \leq \sum_{j=1}^{n} |a_{kj}|$$

– Si $i = k$:

$$\left| \sum_{j=1}^{n} a_{ij} x_j \right| = \sum_{j=1}^{n} |a_{kj}|$$

donc :

$$\max_{1 \leq i \leq m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| = \sum_{j=1}^{m} |a_{kj}| = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|$$

□

Remarque. Pour une matrice $A \in \mathbb{R}^{m,n}$:

$$\|A\|_\infty = \|{}^t A\|_1$$

2.1.5 Quelques propriétés

Définition 2.1.6. Une norme matricielle $\|\cdot\|_M$, une norme vectorielle $\|\cdot\|_V$ sont dites compatibles si :

$$\|Ax\|_V \leq \|A\|_M \|x\|_V \text{ pour tout vecteur } x \text{ de dimension compatible}$$

Exemple 2.1.2.
- Une norme vectorielle et la matrice matricielle induite (ou subordonnée) par cette norme vectorielle sont compatible (par définition).
- La norme de Schur est compatible avec la norme hermitienne. En effet :

$$\|A\|_S^2 = \text{tr}(A^*A) = \sum_{i=1}^{n} \mu_i^2$$

Comme on sait par ailleurs que $\|A\|_2^2 = \mu_1^2$, on a que :

$$\|A\|_2^2 \leq \|A\|_S^2 \leq n\|A\|_2^2$$

On a ainsi :

$$\|Ax\|_2 \leq \|A\|_2\|x\|_2 \leq \|A\|_S\|x\|_2$$

Theorème 2.1.8. *A toute norme matricielle (voir la **Définition 2.1.3.**) qui de plus vérifie la quatrième propriété du **Théorème 2.1.2.**, alors on peut associer une norme vectorielle qui lui soit compatibles.*

Démonstration. Soit $\|.\|$ la norme matricielle, x un vecteur. On va introduire la matrice X tel que :

$$X = \begin{pmatrix} x & 0 & \cdots & 0 & 0 \\ & & & & \end{pmatrix}$$

Or par définition : $\|x\| = \|X\|$. Il est clair que $x \to \|x\|$ est une norme vectorielle :

$$\|Ax\| = \|[Ax, 0, 0, \cdots, 0]\| = \|AX\| \leq \|A\|\|X\| = \|A\|\|x\|$$

□

Theorème 2.1.9. *Soit $\|\cdot\|$ une norme matricielle telle que pour $A \in \mathbb{C}^{n,n}$ et $B \in \mathbb{C}^{n,n}$,$\|AB\| \leq \|A\|\|B\|$. Alors pour $A \in \mathbb{C}^{n,n}$, $\rho(A) \leq \|A\|$.*

Démonstration. On a vu qu'à cette norme matricielle, on pouvait leur associer une norme vectorielle compatible, que l'on note également $\|\cdot\|$.

Soit (λ, u) un élément propre de A, on a ainsi : $Au = \lambda u$.

$$\|\lambda u\| = |\lambda|\|u\| = \|Au\| \leq \|A\|\|u\|$$

donc clairement, on a : $|\lambda| \leq \|A\|$. Ce qui implique $\rho(A) \leq \|A\|$. □

Theorème 2.1.10. *Soit $A \in \mathbb{C}^{n,n}$ et $\varepsilon > 0$. Il existe au moins une norme matricielle subordonnée telle que $\|A\| \leq \rho(A) + \varepsilon$.*

Démonstration. Il exsite une matrice inversible U telle que $U^{-1}AU$ soit triangulaire supérieure.

$$U^{-1}AU = \begin{pmatrix} \lambda_1 & t_{12} & t_{13} & \cdots & t_{1n} \\ & \lambda_2 & t_{21} & \cdots & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \lambda_n \end{pmatrix}$$

avec $\{\lambda_i\}_{1\leq i\leq n}$ valeurs prores de A. A tout scalaire $\delta \neq 0$, on va considérer la matrice :

$$D_\delta = \begin{pmatrix} 1 & & & & \\ & \delta & & 0 & \\ & & \delta^2 & & \\ & 0 & & \ddots & \\ & & & & \delta^{n-1} \end{pmatrix}$$

$$(UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}\underbrace{U^{-1}AU}_{T}D_\delta = \begin{pmatrix} \lambda_1 & \delta t_{12} & \delta t_{13} & \cdots & \delta^{n-1}t_{1n} \\ & \lambda_2 & \delta t_{23} & \cdots & \delta^{n-2}t_{2n} \\ & & \ddots & & \vdots \\ & & & \lambda_{n-1} & \delta t_{n-1,n} \\ & & & & \lambda_n \end{pmatrix}$$

Soit $\varepsilon > 0$, on fixe $\delta > 0$ tel que :

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

L'application $\|\cdot\| : B \in \mathbb{C}^{n\times n} \to \|B\| = \|(UD_\delta)^{-1}B(UD_\delta)\|_\infty$ répond à la question. En effet :

$$\|A\| = \|(UD_\delta)^{-1}A(UD_\delta)\|_\infty \leq \rho(A) + \varepsilon$$

D'autre part, $\|\cdot\|$ est bien une norme matricielle car c'est une norme induite par la norme vectorielle qui a tout $v \in \mathbb{C}^n \to \|(UD_\delta)^{-1}v\|_\infty$ (**Théorème 2.1.2.**). □

Theorème 2.1.11. *Soit $B \in \mathbb{C}^{n,n}$. Les propriétés suivantes sont équivalentes :*

(i) $\lim_{k\to\infty} B^k = 0$

(ii) $\lim_{k\to\infty} B^k v = 0$, $\forall v \in \mathbb{C}^n$

(iii) $\rho(B) < 1$

(iv) $\|B\| < 1$ *pour au moins une norme matricielle subordonnée* $\|\cdot\|$.

Démonstration. *(i) ⇒ (ii)* Soit $\|\cdot\|$ une norme vectorielle, et $\|\cdot\|$ la norme matricielle induite (ou subordonnée) correspondante.

$$\|B^k v\| \leq \|B^k\|\|v\| \text{ donc } \lim_{k\to+\infty} B^k v = 0$$

(ii) ⇒ (iii) Par l'absurde : On suppose que $\rho(B) \geq 1$. On peut trouver $p \in \mathbb{C}^n$, $p \neq 0$ tel que $Bp = \lambda p$, $|\lambda| \geq 1$. Comme $B^k p = \lambda^k p$ alors il est impossible que $\lim_{k\to+\infty} B^k p = 0$.

(iii) ⇒ (iv) Voir **Théorème 2.1.10**, il suffit de choisir ε suffisamment petit.

$\underline{(iv) \Rightarrow (i)}$ On applique l'inégalité $\|B^k\| \leq \|B\|^k$ pour la norme en question.

□

Theorème 2.1.12. *(1) $B \in \mathbb{C}^{n,n}$ et $\|\cdot\|$ une norme matricielle vérifiant $\|AB\| \leq \|A\|\|B\|$ pour $(A, B) \in (\mathbb{C}^{n,n})^2$ alors :*

$$\lim_{k\to+\infty} \|B^k\|^{1/k} = \rho(B)$$

(2) La série $\sum_{k=0}^{\infty} B^k$ converge vers $(I-B)^{-1}$ si et seulement si $\rho(B) < 1$.

Démonstration. (1) D'après le **Théorème 2.1.9.**, on a : $\rho(B) \leq \|B\|$. D'autre part, $\rho(B^k) = (\rho(B))^k$ donc $\rho(B) = (\rho(B^k))^{1/k}$ donc : $\rho(B) \leq \|B^k\|^{1/k}$, $\forall k \in \mathbb{N}^*$.

On va montrer que :

$$\forall \varepsilon > 0,\ \exists l_\varepsilon \text{ tel que } k \geq l_\varepsilon \Rightarrow \|B^k\|^{1/k} \leq \rho(B) + \varepsilon$$

Soit $\varepsilon > 0$, on définit :

$$B_\varepsilon = \frac{B}{\rho(B) + \varepsilon}$$

On sait que $\rho(B_\varepsilon) < 1$. Donc d'après le **Théorème 2.1.11.**, on a :

$$\lim_{k\to+\infty} B_\varepsilon^k = 0$$

donc :

$$\exists l_\varepsilon \text{ tel que } k \geq l_\varepsilon \Rightarrow \|B_\varepsilon^k\| = \frac{\|B^k\|}{(\rho(B)+\varepsilon)^k} \leq 1$$

c'est-à -dire :

$$\|B^k\|^{1/k} \leq \rho(B) + \varepsilon$$

(2) $(\Leftarrow)$ $\rho(B) < 1 \Rightarrow 1 \notin \mathrm{Sp}(B)$ donc $(I-B)$ inversible.

$$\begin{cases} A^k = I + B + ... + B^k & (L_1) \\ BA^k = B + B^2 + ... + B^{k+1} & (L_2) \end{cases} \overset{(L_2)-(L_1)}{\Rightarrow} (I-B)A_k = I - B^{k+1}$$

$$A_k = (I-B)^{-1}(I - B^{k+1})$$

donc :

$$\|A_k - (I-B)^{-1}\| \leq \|(I-B)^{-1}\|\|B^{k+1}\| \xrightarrow[k\to+\infty]{} 0$$

donc : $\lim_{k\to\infty} A_k = (I-B)^{-1}$.

$(\Rightarrow)$ On sait que $\lim_{k\to\infty} A_k$ existe donc $\lim_{k\to+\infty} B^k = 0$ donc $\rho(B) < 1$.

□

2.2 Conditionnement

2.2.1 Position du problème

Supposons que l'on veuille résoudre le système $Ax = B$. En réalité, cette résolution n'est jamais exacte, elle est entachée d'erreur qui peuvent provenir :

→ de mesures ou de calculs pour déterminer les coefficients de A et B.

$\rightarrow$ d'erreur dûes à la représentation des chiffres de l'ordinateur.
Donc : on ne résout pas véritablement $Ax = B$ mais plutôt $(A + \Delta A)y = (B + \Delta B)$.

Question : Que dire sur $y - x$?

Exemple 2.2.1 (Wilson)**.**

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \; B = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

$$B + \Delta B = \begin{pmatrix} 32.1 \\ 22.9 \\ 33.1 \\ 30.9 \end{pmatrix}, \; A + \Delta A = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.06 & 6 & 5 \\ 8 & 5.98 & 9.89 & 9 \\ 6.99 & 4.99 & 9 & 9.96 \end{pmatrix}$$

$$AX = B \Rightarrow X = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

$$AY = B + \Delta B \Rightarrow Y = \begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ -11 \end{pmatrix}$$

$$(A + \Delta A)Z = B \Rightarrow Z = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}$$

2.2.2 Résultats principaux

Theorème 2.2.1. *Soit A matrice carrée inversible. Soient x et $x + \Delta x$, les solutions de :*

$$\begin{cases} Ax = B \\ A(x + \Delta x) = B + \Delta B \end{cases} \quad \text{on suppose que } B \neq 0$$

(1) Alors :

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\| \frac{\|\Delta B\|}{\|B\|}$$

où $\|\cdot\|$ est la norme induite par la norme vectorielle correspondante.

(2) Cette inégalité est optimale : pour une matrice A donnée, on peut trouver $B \neq 0$ et $\Delta B \neq 0$ tel qu'elle deviennent une égalité.

Démonstration. (1)

$$Ax + A\Delta = B + \Delta B$$

$\Delta x = A^{-1}\Delta B$

$$\|\Delta x\| \leq \|A^{-1}\|\|\Delta B\| \qquad (*)$$

d'autre part $Ax = B$ donc :

$$\|B\| \leq \|A\|\|x\| \qquad (**)$$

On a ainsi :

$$(*) + (**) \Rightarrow \frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta B\|}{\|B\|}$$

(2) Soit A fixée donc il existe ΔB tel que $\|A^{-1}\Delta B\| = \|A^{-1}\|\|\Delta B\|$. On pose $\Delta x = A^{-1}\Delta B$.

$$\|\Delta x\| = \|A^{-1}\Delta B\| = \|A^{-1}\|\|\Delta B\|$$

Maintenant il existe x tel que $\|Ax\| = \|A\|\|x\|$. On pose : $B = Ax$, $\|B\| = \|A\|\|x\|$. On a donc : $A(x + \Delta x) = B + \Delta B$ et :

$$\frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1}\|\|\Delta B\|\|A\|}{\|x\|\|A\|} = \|A\|\|A^{-1}\|\frac{\|\Delta B\|}{\|B\|}$$

□

Theorème 2.2.2. *A matrice carrée inversible. Soient x et $x + \Delta x$ les solutions de :*

$$\begin{cases} Ax = B \\ (A + \Delta A)(x + \Delta x) = B \end{cases} \qquad (B \neq 0)$$

1)

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta A\|}{\|A\|}$$

où $\|\cdot\|$ est la norme induite par la norme vectorielle correspondants.

2) Cette majoration est optimale : pour une matrice A donnée, on peut trouver $B \neq 0$ et $\Delta A \neq 0$ tels qu'elle dénomme une égalité.

Démonstration. 1)

$$\Delta x = -A^{-1}\Delta A(x + \Delta x) \Rightarrow \|\Delta x\| \leq \|A^{-1}\|\|\Delta A\|\|x + \Delta x\|$$

$$\frac{\|\Delta x\|}{\|x + \Delta x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta A\|}{\|A\|}$$

Remarque. Ici on ne doit pas supposer que $(A + \Delta A)$ est inversible, mais seulement que le système linéaire $(A + \Delta A)(x + \Delta x) = B$ ait au moins une solution, notée $x + \Delta x$.

2) admis

□

Définition 2.2.1. Si $\|\cdot\|$ est une norme matricielle, on appelle conditionnement d'une matrice $A \in \mathbb{C}^{n,n}$ le nombre :

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

Si $\|\cdot\| = \|\cdot\|_p$, on note $\text{cond}_p(A) = \|A\|_p\|A^{-1}\|_p$.

Exemple 2.2.2 (Retour sur l'**Exemple 2.2.1** de Wilson)**.**

$$\|A\|_2 = \max_{1 \leq i \leq 4} |\lambda_i| \approx 30$$

$$\|A^{-1}\|_2 = \max_{1 \leq i \leq 4} \left|\frac{1}{\lambda_i}\right| \approx 100$$

$$\text{cond}_2(A) \approx 3000$$

Exemple 2.2.3 (Interprétation sur un exemple simple)**.** Soit à résoudre le système linéaire suivant :

$$\begin{cases} 4.218613x_1 + 6.327917x_2 = 10.546530 \\ 3.141592x_1 + 4.712390x_2 = 7.853982 \end{cases}$$

de solution : $x_1 = x_2 = 1$ et :

$$\begin{cases} 4.218611x_1 + 6.327917x_2 = 10.546530 \\ 3.141594x_1 + 4.712390x_2 = 7.853980 \end{cases}$$

de solution : $x_1 = -5$, $x_2 = 5$

2.2.3 Propriétés

Soit $A \in \mathbb{C}^{n,n}$ inversible, $\|\cdot\|$ une norme matricielle induite, $\text{cond}(A) = \|A\|\|A^{-1}\|$

Propriété 2.2.3. *(1)* $\forall \alpha \in \mathbb{R}^*$, $\text{cond}(\alpha A) = \text{cond}(A)$.

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

(3) $\text{cond}(A) \geq 1$

(4) $\text{cond}_2(A) = \frac{\mu_n(A)}{\mu_1(A)}$ *où* $\mu_1(A) > 0$ *et* $\mu_n(A) > 0$ *désignent respectivement la plus petite et la plus grande valeur singulière de* A.

(5) Si A *est symétrique,* $\text{cond}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}$ *où* $\{\lambda_i\}_{1 \leq i \leq n}$ *sont les valeurs propres de* A.

(6) Le conditionnement en norme $\|\cdot\|_2$ *(*$\text{cond}_2(A)$*) est invariant par transformation unitaire (ou orthogonale)* $UU^* = I$ *alors* $\text{cond}_2(A) = \text{cond}_2(AU) = \text{cond}_2(UA) = \text{cond}_2(U^*AU)$.

Démonstration. (1) $\text{cond}(\alpha A) = \|\alpha A\|\|(\alpha A)^{-1}\| = \frac{|\alpha|}{|\alpha|}\text{cond}(A)$

(2) évident

(3) $I = AA^{-1}$ alors :

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

(4) $\text{cond}_2(A) = \|A\|_2\|A^{-1}\|_2 = \mu_{\max}\tilde{\mu}_{\max}$ où $\mu_{\max}$ est la plus grande valeur singulière de A et $\tilde{\mu}_{\max}$ est la plus grand valeur singulière de A^{-1} mais :

$$\tilde{\mu}_{\max} = \frac{1}{\mu_{\min}}$$

avec $\mu_{\min}$ plus petite valeur singulière de A.

(5) Evident avec (4).

(6) Si U unitaire alors :

$$\|U\| = \sqrt{\rho(U^*U)} = \sqrt{\rho(I)} = 1$$

$$\|AU\|_2 = \max_{\|x\|_2=1} \frac{\|AUx\|_2}{\|x\|_2} = \max_{\|y\|_2=1} = \frac{\|Ay\|_2}{\|y\|_2} = \|A\|_2$$

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

idem pour le reste.

□

2.2.4 Quelques remarques

Remarque. 1) On dit qu'une matrice A est "bien conditionnée" si son conditionnement n'est pas beaucoup plus grand que 1.

2) Les matrices unitaires (ou orthogonales dans $\mathbb{R}$) sont les mieux conditionnées $\Rightarrow$ Utilisation fréquente en analyse numérique.

3) il y a aucun lien avec le conditionnement et le déterminant.

Exemple 2.2.4.

$$A = \begin{pmatrix} 1 & & & & \\ & 0.1 & & & \\ & & 0.1 & & \\ & & & \ddots & \\ & & & & 0.1 \end{pmatrix}_{100\times 100}$$

On a : $\det(A) = (0.1)^{99} \leq \text{cond}_2(A) = 10$.

$$B = \begin{pmatrix} 1 & 2 & & \\ & \ddots & \ddots & \\ & & \ddots & 2 \\ & & & 1 \end{pmatrix}_{n\times n}$$

On a : $\det(B) = 1$ On peut montrer que :

$$A^{-1} = \begin{pmatrix} 1 & -2 & 4 & -8 & \cdots & (-2)^{j-1} & \cdots & (-2)^{n-1} \\ & \ddots & -2 & & \cdots & (-2)^{j-2} & \cdots & (-2)^{n-2} \\ & & \ddots & & & & & \\ & & & \ddots & & & & \\ & & & & \ddots & & & \\ & & & & & \ddots & & \\ & & & & & & \ddots & \\ & & & & & & & 1 \end{pmatrix}_{n\times n}$$

On a : $\|B\|_1 = \|B\|_\infty = 3$, $\|B^{-1}\|_\infty = \|B^{-1}\|_1 = 2^n - 1$ et $\text{cond}_\infty(B) = \text{cond}_1(B) = 3(2^n - 1) >> 1$ pour n assez grand.

Chapitre 3

Moindres carrés

3.1 Moindres carrés : le cas continu

On notera $\mathcal{P}_n$ l'ensemble des polynômes de degré au plus égal à n.

Motivation : On se donne une fonction f, et on va chercher un polynôme de degré n, qui est "le plus proche possible" de f au sens d'une norme à préciser. Plus précisement, si V est un espace vectoriel de fonctions $[a,b] \to \mathbb{R}$ qui est normé pour la norme $\|\cdot\|$ et que $\mathcal{P}_n \subset V$, $\forall n \in \mathbb{N}$, alors le prolbème auquel on s'interesse est le suivant :

$$\text{"Trouver } u_n \in \mathcal{P}_n \text{ tel que } \|u-u_n\| \leq \|u-v\|, \forall v \in \mathcal{P}_n :$$
$$\|u-u_n\| = \inf_{v\in\mathcal{P}_n} \|u-v\|\text{"}$$

3.1.1 Existence

Theorème 3.1.1. *$\forall u \in V$, il existe au moins $u_n \in \mathcal{P}_n$ tel que $\|u-u_n\| = \inf_{v\in\mathcal{P}_n} \|u-v\|$.*

Démonstration. Soit $\{v_k\}_{k\in\mathbb{N}}$ une suite minimisante, c'est-à -dire telle que $v_k \in \mathcal{P}_n$ et

$$\|u-v_k\| \xrightarrow[k\to+\infty]{} \omega = \inf_{v\in\mathcal{P}_n} \|u-v\|$$

On a que :

$$\|v_k\| \leq \|v_k - u\| + \|u\| \leq C \qquad \forall k \in \mathbb{N}$$

donc la suite $\{\|v_k\|\}_{k\in\mathbb{N}}$ est bornée. Comme $\mathcal{P}_n$ est de dimension finie, on peut extraire une sous-suite convregente $v_{k_l} \to u_n$ dans V. Comme $\mathcal{P}_n$ est fermé dans V et que $v_k \in \mathcal{P}_n$, on en déduit que $u_n \in \mathcal{P}_n$. Comme :

$$\Big|\|u-v_{k_l}\| - \|u-u_n\|\Big| \leq \|u-v_{k_l}-u+u_n\| = \|u_n - v_{k_l}\| \xrightarrow[l\to+\infty]{} 0$$

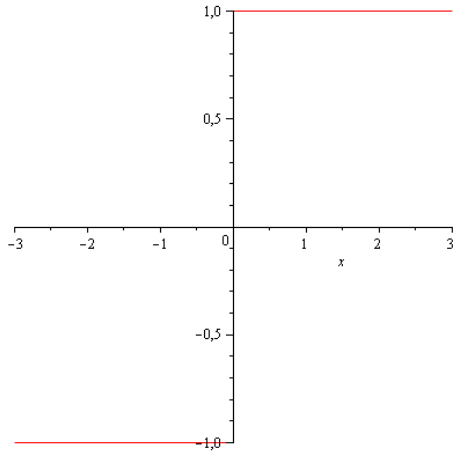
donc $\|u-v_n\| = \omega$. □

Remarque. L'existence est donc garantie mais pas l'unicité.

Exemple 3.1.1. $V = L^1_{\text{Rie}}([-1,1])$, c'est-à -dire l'ensemble des fonctions définies sur $[-1,1]$ et telle que $\int_{-1}^{1} |u(x)|dx < \infty$ et on prend pour norme :

$$\|u\|_{L^1_{\text{Rie}}} = \int_{-1}^{1} |u(x)|dx$$

Soit $u(x) = \text{sgn}(x)$:

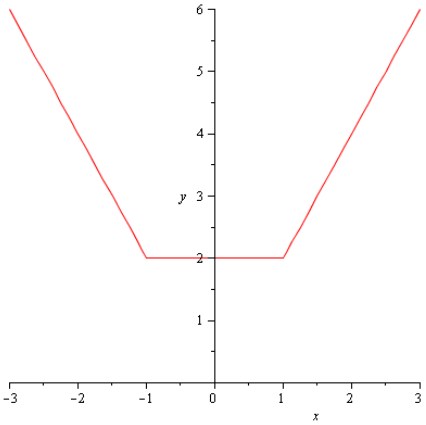


On cherche le polynôme de meilleure approximation dans $\mathcal{P}_0 : p(x) = \alpha$ tel que :

$$\begin{aligned}
\inf_{\alpha\in\mathbb{R}} \|u(x)-\alpha\|_{L^1_{\text{Rie}}} &= \inf_{\alpha\in\mathbb{R}} \int_{-1}^{1} |\operatorname{sgn}(x)-\alpha| dx \\
&= \inf_{\alpha\in\mathbb{R}} \left[\int_0^1 |1-\alpha| dx + \int_{-1}^{0} |-1-\alpha| dx\right] \\
&= \inf_{\alpha\in\mathbb{R}} |1-\alpha| + |1+\alpha| \\
&= \inf_{\alpha\in\mathbb{R}} F(\alpha)
\end{aligned}$$

avec :

$$F(\alpha) = \begin{cases} -2\alpha \text{ si } \alpha \leq -1 \\ 2 \text{ si } -1 \leq \alpha \leq 1 \\ 2\alpha \text{ si } \alpha > 1 \end{cases}$$



$$\inf_{\alpha\in\mathbb{R}} F(\alpha) = 2 = F(\beta) \text{ pour } -1 \leq \beta \leq 1$$

donc il y a une infinité de solutions minimisantes.

3.1.2 Quelques exemples classiques

Exemple 3.1.2. a) La meilleure approximation en norme $\|\cdot\|_\infty$ sur $V = \mathcal{C}^0([a,b])$:

$$\|u\|_\infty = \max_{a\leq x\leq b} |u(x)|$$

(approximation au sens de Tchebycheff), unicité de meilleur approximant.

b) Cas hilbertien : on va utiliser une norme $\|\cdot\|$, déduite d'un produit scalaire :

$$\|u\|^2 = (u,u) \qquad \forall u \in V$$

$(V,(.,.))$ est un espace préhilbertien. C'est dans ce cadre qu'on se place à partir de maintenant.

3.1.3 Polynômes orthogonaux

Theorème 3.1.2. *Soit $(V,(.,.))$ un espace préhilbertien tel que $\mathcal{P}_n \subset V$. Alors il existe une unique suite de polynômes p_n tel que :*

1) $\deg(p_n) = n$

2) Le coefficient directeur (ou coefficient dominant) de p_n est égal à 1.

3) $(p_n, q) = 0$, $\forall q \in \mathcal{P}_{n-1}$.

Cette suite est appelée suite de polynômes orthogonaux.

Démonstration. Gram-Schmidt appliquée à la base canonique $1, x, x^2, ..., x^n$. On pose :

$$\begin{aligned}
&p_0(x) = 1\\
&p_1(x) = x - \alpha p_0 \text{ où } \alpha \text{ est choisi tel que } (p_1,p_0) = 0 \Leftrightarrow \alpha = \tfrac{(1,x)}{(1,1)}\\
&\vdots\\
&p_n(x) = x^n - \sum_{i=0}^{n-1} \lambda_{i,n} p_i(x)
\end{aligned}$$

où les $\lambda_{i,n}$ sont choisis de telle sorte que $(p_n, p_j) = 0$, $\forall j = 0, ..., n-1$, c'est-à -dire :

$$(x^n, p_j) = \lambda_{j,n}(p_j,p_j) \Leftrightarrow \lambda_{j,n} = \frac{(x^n,p_j)}{(p_j,p_j)}$$

□

Exemple 3.1.3. $\omega :]a,b[\to \mathbb{R}$ avec $a, b \in \overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ tel que :

a) $\omega(x) > 0$, $\forall x \in]a,b[$

b) $\int_a^b x^n \omega(x) dx < \infty$, $\forall n \in \mathbb{N}$

On considère :

$$V = L^2_\omega(a,b) = \{v \in]a,b[\to \mathbb{R}\,;\, v^2\omega \text{ est intégrable sur }]a,b[\}$$

et on prend comme produit scalaire :

$$(u,v) = \int_a^b uv\omega dx$$

(i) $a = -1$, $b = 1$ et $\omega(x) = 1$: ce sont les polynômes de Legendre.

(ii) $a = -1$, $b = 1$ et $\omega(x) = (1-x^2)$: ce sont les polynômes de Tchebycheff.

(iii) $a = 0$, $b = +\infty$, $\omega(x) = e^{-x}$: ce sont les polynômes de Laguerre.

(iv) $a = -\infty$, $b = +\infty$, $\omega(x) = e^{-x^2}$: ce sont les polynômes d'Hermite.

3.1.4 Approximation au sens des moindres carrés

Theorème 3.1.3. *Soit $(V,(.,.))$ un espace préhilbertien. Alors $\forall u \in V$, $\exists! u_n \in \mathcal{P}_n$ tel que :*

$$\|u - v_n\| = \inf_{v \in \mathcal{P}_n} \|u - v\|$$

Cette meilleure approximation est caractérisé par les propriétés suivantes :

(1) $(u_n - u, v) = 0$, $\forall v \in \mathcal{P}_n$.

(2) De plus, on a [1] *:*

$$u_n = \sum_{i=0}^{n} \frac{(u, p_i)}{\|p_i\|^2} p_i$$

(3) $\|u_n\| \leq \|u\|$

Démonstration. 1) On va montrer que $u_n \in \mathcal{P}_n$ qui satisfait (1) existe, est unique et elle est donnée par (2). Comme les $\{p_i\}_{0 \leq i \leq n}$ forment une base de $\mathcal{P}_n$ alors :

$$u_n = \sum_{i=0}^{n} \alpha_i p_i$$

(1) s'écrit alors :

$$\underbrace{\left(\sum_{i=0}^{n} \alpha_i p_i - u, v\right) = 0}_{\forall v \in \mathcal{P}_n} \Leftrightarrow \underbrace{\left(\sum_{i=0}^{n} \alpha_i p_i - u, p_j\right) = 0}_{\forall j = 0,..,n} \Leftrightarrow \|p_j\|^2 \alpha_j = \left(\sum_{i=0}^{n} \alpha_i p_i, p_j\right) = (u, p_j) \, \forall j$$

$$\Leftrightarrow \alpha_j = \frac{(u, p_j)}{\|p_j\|^2} \qquad \forall j$$

d'où existence et unicité de u_n défini par (1). On a aussi que :

$$(u_n - u, v) = 0 \Leftrightarrow \|u_n\|^2 = (u, u_n) \overset{CS}{\leq} \|u\| \|u_n\|$$

donc $\|u_n\| \leq \|u\|$

2) Reste donc à montrer que $u_n \in \mathcal{P}_n$ satisfaissant (1) est la meilleure approximation de u au sens des moindres carrés. Soit $v \in \mathcal{P}_n$:

$$\|u-v\|^2 = \|u-u_n+u_n-v\|^2 = (u-u_n+u_n-v, u-u_n+u_n+v) = \|u-u_n\|^2+\|u_n-v\|^2+2\underbrace{(u-u_n, u_n-v)}_{0}$$

Donc :

$$\|u - v\|^2 = \|u - u_n\|^2 + \|u_n - v\|^2 \geq \|u - u_n\|^2$$

Ce qui prouve que :

$$\|u - v\| \geq \|u - u_n\| \qquad \forall v \in \mathcal{P}_n$$

De plus, si $v \neq u_n$ alors $\|u - v\| > \|u - u_n\|$ donc :

$$\forall v \in \mathcal{P}_n \text{ avec } v \neq u_n \text{ alors } \|u - v\| > \|u_n - v\|$$

Ce qui montre que u_n est l'unique polynôme qui minimise $\|u - v\|$, $v \in \mathcal{P}_n$.

□

[1] $\{p_i\}_{0 \leq i \leq n}$ famille de polynômes orthogonaux de $\mathcal{P}_n$ associée au $(.,.)$

Corollaire. *Soit $V = \mathcal{C}^0([a,b])$ muni de la norme suivante :*

$$\|u\|^2_{L^2([a,b])} = \int_a^b (u(x))^2 dx$$

Alors $\forall u \in V$, la meilleure approximation au sens des moindres carrés $u_n \in \mathcal{P}_n$ tend vers u quand $n \to +\infty$, c'est-à -dire :

$$\|u - u_n\|_{L^2([a,b])} \xrightarrow[n\to+\infty]{} 0$$

Démonstration. Soit $u \in \mathcal{C}^0([a,b])$. Par le théorème de Weierstrass (admis, démontré dans le module **M315**), il existe une suite de polynômes $f_n \in \mathcal{P}_n$, $n \in \mathbb{N}$, tel que :

$$\|f_n - u\|_{L^\infty([a,b])} = \max_{x\in[a,b]} |f_n(x) - u(x)| \xrightarrow[n\to+\infty]{} 0$$

On calcule :

$$\|f_n - u\|^2_{L^2([a,b])} = \int_a^b |f_n(x) - u(x)|^2 dx \leq (b-a)\|f_n - u\|_\infty \xrightarrow[n\to+\infty]{} 0$$

On a forcément :

$$\|u - u_n\|_{L^2([a,b])} \leq \|u - f_n\|_{L^2([a,b])}$$

donc $\|u - u_n\|_{L^2([a,b])} \xrightarrow[n\to+\infty]{} 0$. □

Remarque. La démonstration fonctionne encore pour :

$$\|u\| = \int_a^b |u(x)|^2 \omega(x) dx$$

avec $\omega : [a,b] \to \mathbb{R}$ intégrable tel que :

$$0 \leq \omega(x) \leq M < \infty, \forall x \in [a,b]$$

Lien avec les séries de Fourier

On pourrait maintenant reprendre cette théorie de l'approximation au sens des moindres carrées en considérant : $V =$ espaces des fonctions f avec :

$$f : [-\pi, \pi] \to \mathbb{R} \text{ tel que } \int_{-\pi}^{\pi} |f(x)|^2 dx$$

Cette norme est induite par le produit scalaire :

$$(f,g) = \int_{-\pi}^{\pi} f(x)g(x)dx$$

Plutôt que de prendre $\mathcal{P}_n$ comme l'espace des polynômes de degré au plus n, on va prendre :

$$\mathcal{I}_n = \text{span}(1, \cos(kx), \sin(kx))_{k=1,\dots,n}$$

où span veut dire "l'espace vectoriel engendré par". $\mathcal{I}_n$ a les propriétés suivantes :

1) $\mathcal{I}_n \subset V$

2) $\dim \mathcal{I}_n < \infty$

On vérifie de plus que $\{1, \cos(kx), \sin(kx)\}_{k=0}^{n}$ est une base orthogonale de $\mathcal{I}_n$. Le **Théorème 3.1.3.** reste valable et on a le théorème suivant :

Theorème 3.1.4. *$\forall u \in V$, $\exists! u_n \in \mathcal{I}_n$ tel que :*

$$\|u - u_n\| = \inf_{v \in \mathcal{I}_n} \|u - v\|$$

il vérifie $(u - u_n, v) = 0$, $\forall v \in \mathcal{I}_n$ et $\|u_n \leq \|u\|$ et il s'écrit :

$$u_n(x) + a_0 + \sum_{k=1}^{n} (a_k \cos(kx) + b_k \sin(kx))$$

où :

$$a_0 = \frac{(u,1)}{2\pi} = \frac{1}{2\pi} \int_{-\pi}^{\pi} u(x)dx$$

$$a_k = \frac{1}{\pi}(u, \cos(k)) = \frac{1}{\pi} \int_{-\pi}^{\pi} u(x)\cos(kx)dx$$

$$b_k = \frac{1}{\pi}(u, \sin(k)) = \frac{1}{\pi} \int_{-\pi}^{\pi} u(x)\sin(kx)dx$$

à cause des propriétés d'orthogonalité :

$$\|u_n\|^2 = 2\pi a_0^2 + \pi \sum_{k=1}^{n} (a_k^2 + b_k^2)$$

il reste à montrer que $u_n \xrightarrow[n \to +\infty]{} u$ dans V (admis).

3.2 Moindres carrés discrets

Position du problème

Supposons que l'on cherche à identifier une fonction y qui dépend de n paramètres x_j $(1 \leq j \leq n)$

$$y(t) = \sum_{j=1}^{n} x_j f_j(t)$$

où les f_j sont linéairement indépendants. On va prendre pour plusieurs t_i $(1 \leq i \leq m)$, des valeurs mesurées y_i et on suppose $m \geq n$. On cherche donc x_j $(1 \leq j \leq n)$ tel que :

$$\sum_{j=1}^{n} x_j f_j(t_i) = y_i \qquad (1 \leq i \leq m)$$

où les y_i est la valeur mesurée approche y_i. C'est un système linéaire : n inconnues et m équations. A priori, $\nexists$ solutions (système surdimensionné car $m \geq n$). On va chercher à satisfaire au mieux les équations, en minimisant la quantité :

$$\sum_{i=1}^{m} \left(\sum_{j=1}^{n} x_j f_j(t_i) - y_i \right)^2$$

C'est une approximation au sens des moindres carrés.

3.2.1 Existence et unicité

Soient $n \in \mathbb{N}^*$, $m \in \mathbb{N}^*$ avec $m > n$, $b \in \mathbb{R}^m$, $A \in \mathbb{R}^{m,n}$ (dans l'exemple précédent, on avait $a_{ij} = f_j(t_i)$ et on cherche à résoudre au sens des moindres carrés $AX = b$, $X \in \mathbb{R}^n$. On pose $\phi(X) = \|AX - b\|_2$ et on cherche $X \in \mathbb{R}^n$ qui minimise $E(X) = (\phi(X))^2$.

Theorème 3.2.1. *(i) α réalise le minimum de E sur $\mathbb{R}^n$*

$$\Leftrightarrow {}^tAA\alpha = {}^tAb \qquad \textit{(équations normales)}$$

(ii) $\mathrm{rg}(A) = n \Leftrightarrow \mathrm{rg}({}^tAA) = n$ *et dans ce cas, le problème de minoration admet une unique solution.*

(iii) Si α_1 et α_2 minimisant tous les deux E alors :

$$\alpha_1 - \alpha_2 \in \ker A$$

(iv) Le problème de minimisation possède toujours au moins une solution.

Démonstration. (i) ($\Leftarrow$) Soit $\alpha \in \mathbb{R}^n$ tel que ${}^tAA\alpha = {}^tAb$, $y \in \mathbb{R}^n$:

$$\|Ay - b\|_2^2 = (Ay - b, Ay - b) = (A(y - \alpha) + A\alpha - b, A(y - \alpha) + A\alpha - b)$$

$$= \|A(y - \alpha)\|_2^2 + \|A\alpha - b\|_2^2 + 2(A(y - \alpha), A\alpha - b)$$

$$= \|A\alpha - b\|_2^2 + \|A(y - \alpha)\|_2^2 + 2(y - \alpha, {}^tA(A\alpha - b)) \geq \|A\alpha - b\|_2^2$$

($\Rightarrow$) Soit α est solution du problème de minimisation. Soit $y \in \mathbb{R}^n$, $t \in \mathbb{R}$, on a donc :

$$\|A(\alpha + ty) - b\|_2^2 \geq \|Ax - b\|_2^2$$

d'où :

$$2t(Ay, A\alpha - b) + t^2\|Ay\|_2^2 \geq 0$$

1^{er} cas : $t > 0$. Dans ce cas :

$$2(Ay, A\alpha - b) + t\|Ay\|_2^2 \geq 0$$

et si on fait $t \to 0^+$, on obtient $(Ay, A\alpha - b) \geq 0$.
$2^{\text{ème}}$ cas : $t > 0$ alors :

$$2(Ay, A\alpha - b) + t\|Ay\|_2^2 \leq 0$$

et $t \to 0^-$, on a : $(Ay, Ax - b) \leq 0$.
D'après les deux cas, si $t \to 0$:

$$(Ay, A\alpha - b) = 0 \Leftrightarrow (y, {}^tA(A\alpha - x) = 0$$

donc : ${}^tAAx = Ab$

(ii) On montre que $\ker(A) = \ker({}^tAA)$:
($\subset$) $Ax = 0 \Rightarrow {}^tAAx = 0 \Rightarrow x \in \ker({}^tAA)$
($\supset$)

$${}^tAAx = 0 \Leftrightarrow ({}^tAAx, x) = 0 \Leftrightarrow (Ax, Ax) = 0 \Leftrightarrow \|Ax\|_2^2 = 0$$

$$\Leftrightarrow Ax = 0 \Leftrightarrow x \in \ker(A)$$

$$\mathrm{rg}(A) = \dim(\mathrm{Im}\, A) = n - \dim(\ker A) = n - \dim(\ker {}^tAA) = \mathrm{rg}({}^tAA)$$

Si $\mathrm{rg}(A) = n$ alors $\mathrm{rg}({}^tAA) = n$ donc ${}^tAA$ inversible donc :

$${}^tAA\alpha = {}^tAb \Leftrightarrow \alpha = ({}^tAA)^{-1}({}^tA)b$$

Unicité de x.

(iii)

$$\left.\begin{array}{l} {}^tAA\alpha_1 = {}^t Ab \\ {}^tAA\alpha_2 = {}^tAb \end{array}\right\} \Leftrightarrow {}^tAA(\alpha_1 - \alpha_2) = 0 \Leftrightarrow \alpha_1 - \alpha_2 \in \ker({}^tAA) = \ker(A)$$

(iv) $A : \mathbb{R}^n \to \mathbb{R}^m$, $m > n$:

$$\mathbb{R}^m = \mathrm{Im}\, A + (\mathrm{Im}\, A)^\perp = \mathrm{Im}\, A + \ker {}^tA$$

$b \in \mathbb{R}^m$, $b = s + r$, $s \in \mathrm{Im}\, A$, $r \in \ker {}^tA$, $\exists \alpha_0 \in \mathbb{R}^n$ tel que $A\alpha_0 = s$.

$$b = A\alpha_0 + r \qquad {}^tAb = {}^tA(A\alpha_0 + r) = {}^tAA\alpha_0$$

Interprétation géométrique : Π la projection orthogonale sur $\mathrm{Im}\, A$:

$$\mathbb{R}^m = \mathrm{Im}\, A + (\mathrm{Im}\, A)^\perp$$

$$b = \underbrace{\Pi b}_{\in \mathrm{Im}\, A} + \underbrace{(I - \Pi)b}_{\in (\mathrm{Im}\, A)^\perp}$$

Admettons que X^* soit une solution au sens des moindres carrés.

$$\|AX^* - b\|_2^2 = \|AX^* - \Pi b + (\Pi - I)b\|_2^2 = (AX^* - \Pi + (\Pi - I)b, AX^* - \Pi b + (\Pi - I)b) = \|AX^* - \Pi b\|_2^2 = |$$

X^* est solution de $AX^* = \Pi b$, car ce système à toujours au moins une solution grâce à la surjectivité de A sur $\mathrm{Im}\, A$ ($\Pi b \in \mathrm{Im}\, A$). L'ensemble des solutions est constitué par :

$$S = \{X \in \mathbb{R}^n,\ AX = \Pi b\}$$

Si de plus, $\mathrm{rg}(A) = n$ alors :

$${}^tAAX^* = {}^tAb \Leftrightarrow X^* = ({}^tAA)^{-1}({}^tA)b \Rightarrow AX^* = \underbrace{A({}^tAA)^{-1}}_{\Pi} b$$

□

3.2.2 Lien avec la décomposition en valeurs singulières

On suppose ici que A n'est pas de rang maximale et que X^* est solution de :

$$E(X^*) = \min_{X \in \mathbb{R}^n} E(X) \qquad E(X) = \|AX - b\|_2^2$$

On a vu que le vecteur $X^* + Z$ avec $Z \in \ker(A)$ est aussi solution. On va alors chercher à minimiser la norme euclidienne de la solution elle-même. On cherche à trouver $X^* \in \mathbb{R}^n$, de norme euclidienne minimale tel que :

$$\|AX^* - b\|_2^2 \leq \max_{X \in \mathbb{R}^n} \|AX - b\|_2^2 \qquad (**)$$

Définition 3.2.1. Soit $A \in \mathbb{C}^{m,n}$ de rang r tel que elle admette une décomposition en valeurs singulières de type $U^*AV = \Sigma$. La matrice $A^\dagger = V\Sigma^\dagger U^*$ est appelée pseudo-inverse (de Moore-Penrose) où $\Sigma^\dagger = \text{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, ..., \frac{1}{\sigma_k}, 0, ..., 0\right)$ avec $\Sigma = \text{diag}(\sigma_1, ..., \sigma_n, 0, ..., 0)$.

Theorème 3.2.2. *Soit $A \in \mathbb{R}^{m,n}$ dont la décomposition en valeurs singulières et donnée par $A = U\Sigma^t V$. Alors l'unique solution de $(**)$ est donnée par $X^* = A^\dagger b$ où $A^\dagger$ est la pseudo-inverse de A.*

Démonstration. $(**) \Leftrightarrow$ Trouver $\omega = V^t x$ tel que ω est une norme euclidienne minimale et :

$$\|\Sigma\omega - U^t b\|_2^2 = \|\Sigma y - U^t b\|_2^2 \qquad \forall y \in \mathbb{R}^n$$

Si r le nombre de valeurs singulières non nulles de A.

$$\|\Sigma\omega - U^t b\|_2^2 = \sum_{i=1}^{r}(\sigma_i\omega_i - (U^t b)_i) + \sum_{i=r+1}^{n}(U^t b)_i^2$$

avec $(U^t b)_i$ est la i-ième composante du vecteur $U^t b$. Ceci est minimal par $\omega_i = \frac{(U^t b)_i}{\sigma_i}$ pour $1 \leq i \leq r$

Parmi les vecteurs dont les r premières composantes sont fixées, celui de norme euclidienne la plus petite seet celui dont les $(n-r)$ composantes restantes sont nulles. Donc on choisit $\omega_i = 0$, $r+1 \leq i \leq n$, donc :

$$\omega = \Sigma^\dagger U^t b \qquad X^* = V\omega = V\Sigma^\dagger U^t b = A^\dagger b$$

□

3.2.3 Retour vers les équations normales

Une idée pour résoudre le problème des moindres carrés consisterait à résoudre les équations normales :

$$^tAAX = {}^tAB$$

(on suppose ici que $\text{rg}(A) = n$). On peut résoudre ce système par factorisation de Cholesky car ^tAA est symétrique définie positive.

Problème. *1)*

$$A = \begin{pmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{pmatrix} \qquad \varepsilon \in \mathbb{R}^* \qquad \text{rg}(A) = 3$$

Donc :

$$^tAA = \begin{pmatrix} 1+\varepsilon^2 & 1 & 1 \\ 1 & 1+\varepsilon^2 & 1 \\ 1 & 1 & 1+\varepsilon^2 \end{pmatrix}$$

Soit ε_n la précision machine. Supposons que[2] *$\varepsilon^2 < \varepsilon_n < \varepsilon$. On a que :* $\text{rg}(^tAA) = 3$ *mais pour la machine,* $\text{rg}(^tAA) = 1$. *De façon plus générale, les perturbations de la solution dues aux erreurs d'arrondis quand on utilise les équations normales peuvent dégénérées fortement (quand A carrée, elles deviennent proportionnelles à* $\text{cond}(A)^2$*).*

[2] manque de chance !

2) si A est très creuse alors ${}^tAA$ peut l'être beaucoup moins. Exemple :

$$A = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ & 0 & 0 & \end{pmatrix} \qquad {}^tAA = \begin{pmatrix} a_1^2 & a_1a_2 & \cdots & a_1a_n \\ \vdots & & & \vdots \\ a_1a_n & \cdots & \cdots & a_n^2 \end{pmatrix}$$

3.3 Factorisation QR

3.3.1 Intérêt de la factorisation

On suppose $\mathrm{rg}(A) = n$. On va montrer qu'il est possible de factoriser A vers la forme $A = QR$ avec :

1) Q matrice orthogonale de dimension $m \times m$ (${}^tQQ = Q{}^tQ = I$)

2) R matrice $m \times n$ "triangulaire supérieure"

$$R = \begin{pmatrix} \ddots & & * \\ & \ddots & \\ 0 & & \ddots_n \\ -- & -- & -- \\ & 0 & \end{pmatrix}_{m \times n} \qquad \mathrm{rg}(A) = n$$

On cherche $X \in \mathbb{R}^n$ tel que $\|AX - b\|_2^2$ est minimale :

$$\|AX - b\|_2^2 = \|QRX - b\|_2^2 = \|{}^tQ(QRX - b)\|_2^2 = \|RX - {}^tQb\|_2^2 \qquad (*)$$

Soit :

$$C = {}^tQB = \begin{pmatrix} c_1 \\ --_n \\ c_2 \end{pmatrix} \qquad \overline{R} = \begin{pmatrix} \ddots & & * \\ & \ddots & \\ 0 & & \ddots_n \end{pmatrix}$$

$$(*) = \|\overline{R}X - c_1\|_2^2 + \|c_2\|_2^2$$

et donc la solution est donnée par X^* vérifiant :

$$\overline{R}X^* = c_1$$

3.3.2 Existence de la factorisation QR, unicité

Soit $A \in \mathbb{R}^{n,n}$, dont les colonnes a_i $(1 \leq i \leq n)$ sont linéairement indépendants. On va montrer qu'on peut construire une matrice Q orthogonale et une matrice R triangulaire supérieure tel que $A = QR$. On va d'abord définir Q par le procédé d'orthonormalisation de Gram-Schmidt.

$$A = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ & & & \end{pmatrix}$$

$$\begin{aligned} q_1 &= \frac{a_1}{\|a_1\|_2} \\ q_2 &= \frac{a_2-({}^tq_1a_2)q_1}{\|a_2-({}^tq_1a_2)q_1\|_2} \\ \vdots & \quad \vdots \\ q_i &= \frac{a_i-\sum_{j=1}^{i-1}({}^tq_ja_i)q_j}{\|a_i-\sum_{j=1}^{i-1}({}^tq_ja_i)q_j\|} \end{aligned}$$

$$Q = \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix}$$

Maintenant $A = QR \Leftrightarrow {}^tQA = R$.

$$\begin{array}{ccc} & \times & \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \\ \begin{pmatrix} {}^tq_1 \\ {}^tq_2 \\ \vdots \\ {}^tq_n \end{pmatrix} & & = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & a_n \end{pmatrix} \end{array}$$

Cela permet donc de définir les coefficients de R par :

$$r_{ij} = {}^tq_ia_{ji} \qquad 1 \leq i \leq j \leq n$$

Cela montre donc l'existence de la décomposition QR.

Remarque. $A = QR$:

$$\begin{array}{ccc} & \times & \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & a_n \end{pmatrix} \\ \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} & = & \begin{pmatrix} q_1 & q_2 & \cdots & q_n \end{pmatrix} \end{array}$$

$$\begin{cases} a_1 = r_{11}q_1 \\ a_2 = r_{12}q_1 + r_{22}q_2 \\ \vdots \\ a_n = r_{1n}q_1 + r_{2n}q_2 + ... + r_{nn}q_n \end{cases}$$

Chaque colonne q_i est obtenue en soustrayant de a_i ses projecteurs orthogonales sur les $(i-1)$ premiers vecteurs de la base orthonormée déjà calculés puis en normant le résultat.

Algorithme 3.3.1.
GRAM-SCHMIDT()
1 **for** $i \leftarrow 1$ **to** n # *Calcul de la ième colonne de Q et de R*
2 **do for** $j \leftarrow 1$ **to** $i-1$

3 **do** $r_{ji} \leftarrow {}^t q_j a_i$
4
5 $b_i \leftarrow a_i - \sum_{j=1}^{i-1} r_{ji} q_j$
6 $r_{ii} \leftarrow \|b_i\|_2$
7 $q_i \leftarrow \frac{b_i}{r_{ii}}$

Theorème 3.3.1. *Soit $A \in \mathbb{R}^{n,n}$ inversible. Il existe un unique couple de matrices (Q, R) avec Q orthogonale et R triangulaire supérieure à diagonale strictement positive tel que $A = QR$.*

Démonstration. (1) Existence : déjà fait, par construction.

(2) Unicité : supposons $A = Q_1R_1 = Q_2R_2$. Soit $T = R_2R_1^{-1} = {}^tQ_2Q_1$. T est une matrice triangulaire supérieur qui vérifie :

$$ {}^tTT = {}^t({}^tQ_2Q_1){}^tQ_2Q_1 = I$$

D'autre part $T_{ii} > 0$, $1 \leq i \leq n$ donc T est la matrice de la factorisation de Cholesky de la matrice I. Mais il y a unicité de cette décomposition donc necessairement $T = I$ donc $R_2R_1^{-1} = I \Rightarrow R_2 = R_1$ et $Q_1 = Q_2$.

□

3.3.3 Remarques "théoriques"

1) Si on n'impose plus le caractère strictement positive de la diagonale de R, on perd l'unicité. Dans ce cas-là , deux factorisations différentes seraient liées par la relation $R_1 = ER_2$ où E est une matrice diagonale avec des coefficients égaux à 1 ou -1.

Démonstration. En effet : $T = R_2R_1^{-1}$ triangulaire supérieure.

$$T{}^tT = I \qquad \begin{pmatrix} t_{11} & & & \\ t_{21} & t_{22} & & \\ & & \ddots & \\ t_{n1} & & & t_{nn} \end{pmatrix} \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ & t_{22} & & \\ & & \ddots & \\ & & & t_{nn} \end{pmatrix}$$

donc $t_{ij} = 0$ et $t_{ii}^2 = 1$ donc $|t_{ii}| = 1$. Soit E matrice diagonale telle que $e_{ii} = \operatorname{sgn}(t_{ii})$.

$$R_2R_1^{-1} = E \Leftrightarrow R_2 = ER_1$$

□

2) Si A est non inversible : on perd l'unicité et l'algorithme de Gram-Schmidt ne fonctionne plus.

Exemple 3.3.1.

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 2 & 4 \\ 1 & 2 & 4 \end{pmatrix}$$

$$b_1 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \qquad r_{11} = \sqrt{3} \qquad q_1 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

$$r_{12} = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} . \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} = \frac{6}{\sqrt{3}}$$

$$b_2 = \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} - \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \frac{6}{\sqrt{3}} = 0$$

Mais on peut compléter q_1 pour avoir une base orthonormale de $\mathbb{R}^3$.

Exemple 3.3.2.

$$\underbrace{\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}}_{q_1} \underbrace{\begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}}_{\text{choix arbitraire}} \xrightarrow{\text{GRAM-SCHMIDT}} \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \frac{1}{\sqrt{6}} \begin{pmatrix} 2 \\ -1 \\ -1 \end{pmatrix} \frac{1}{\sqrt{2}} \begin{pmatrix} 0 \\ 1 \\ -1 \end{pmatrix}$$

$$A = \begin{pmatrix} 1/\sqrt{3} & 2/\sqrt{6} & 0 \\ 1/\sqrt{3} & -1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & -1/\sqrt{6} & -1/\sqrt{2} \end{pmatrix} \begin{pmatrix} \sqrt{3} & -6/\sqrt{3} & 12/\sqrt{3} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

3) Si $A \in \mathbb{R}^{m,n}$, $m > n$.

- Si $\text{rg}(A) = n$, on applique le même procédé. Une fois déterminer les $q_1, q_2, ..., q_n$, on complète par $q_{n+1}, ..., q_m$ pour avoir une base orthonormale de $\mathbb{R}^n$. On a : $\tilde{Q} \in \mathbb{R}^{m,m}$ et $\tilde{R} = \mathbb{R}^{m,n}$. On a ainsi :

$$A = \tilde{Q}\tilde{R}$$

avec :

$$\tilde{Q} = \begin{pmatrix} q_1 & q_2 & \cdots & q_n & q_{n+1} & \cdots & q_m \end{pmatrix}$$

$$\tilde{R} = \begin{pmatrix} & R & \\ & & \\ -- & -- & --_n \\ & 0 & \end{pmatrix}_{m \times n} \qquad R = \begin{pmatrix} \ddots & * \\ 0 & \ddots \end{pmatrix}$$

il n'y a plus unicité de la décomposition car cela dépend comment on choisit $q_{n+1}, q_{n+2}, ..., q_n$.

– Si $\text{rg}(A) = r < n$, on peut montrer en permuttant les colonnes de A de façon convenable que $\exists \tilde{Q}, \tilde{R}$, P matrice de permutations tel que :

$$AP = \tilde{Q}\tilde{R} \qquad \tilde{R} = \begin{pmatrix} \ddots & & * \\ & \ddots & \\ 0 & & \ddots \\ -- & -- & --_r \\ & 0 & \\ -- & -- & --_n \\ & 0 & \\ -- & -- & --_m \end{pmatrix}$$

3.3.4 Remarques "numériques"

→ Gram-Schmidt coûte cher numériquement

→ C'est un procédé instable : si les différents vecteurs a_i ont des normes comparables, on a necessairement $\|b_i\|$ décroit quand i croit et on divise par $\|b_i\|_2$.

Exemple 3.3.3.

$$A = \begin{pmatrix} 1 & 1-\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix},\ \varepsilon > 0$$

Gram-Schmidt $\Rightarrow r_{22} \sim |\varepsilon|$ mais si $\varepsilon = 10^{-20}$, $\varepsilon = 0$ pour l'ordinateur. ON verra dans la **Section 3.4.** comment procéder.

3.3.5 Autre démonstration en lien avec les équations normales

Pour montrer que $A \in \mathbb{R}^{m,n}$ inversible admet une décomposition QR, on peut aussi utiliser ${}^tAA$ qui est symétrique définie positive car ${}^tAA$ admet une factorisation de Cholesky. ${}^tAA = {}^tBB$, B triangulaie supérieure à diagonale > 0. On pose ${}^tQ = {}^tB^{-1}\,{}^tA$ (Q orthogonale).

$${}^tQQ = {}^tB^{-1}\,{}^tAAB^{-1} = I$$

$$A = {}^tA^{-1}\,{}^tBB = {}^tA({}^tB^{-1})^{-1}B = ({}^tB^{-1}\,{}^tA)^{-1}B = ({}^tQ)^{-1}B = QR$$

3.4 Méthode de Householder

3.4.1 Les matrices de Householder

Définition 3.4.1. On appelle matrice élémentaire de Householder une matrice H de la forme :

$$H = I_n - 2u^tu \text{ où } u \in \mathbb{R}^n \text{ et } \|u\|_2 = 1$$

Theorème 3.4.1. *Toute matrice de Householder est :*

(i) symétrique

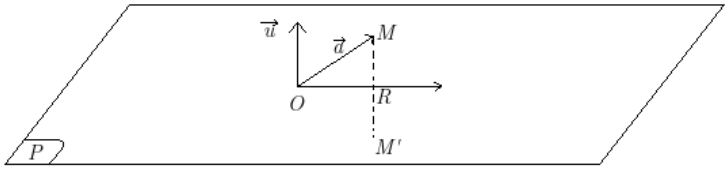
(ii) orthogonale

Démonstration. (i) $H = I - 2u^tu$ et ${}^tH = {}^tI - 2^t(u^tu) = I - 2u^tu = H$

(ii) ${}^tHH = (I - 2u^tu)(I - 2u^tu) = I - 4(u^tu) + 4(u\underbrace{|^tuu|}_{=1}\,{}^tu) = I$

□

Interprétation géométrique : H est la matrice qui caractérise la symétrie par rapport à un plan P orthogonal à $\overrightarrow{u}$.



$$\overrightarrow{OM'} = \overrightarrow{OR} + \overrightarrow{RM'} = \overrightarrow{OM} - 2\overrightarrow{RM}$$

mais $\overrightarrow{RM} = \alpha\,\overrightarrow{u}$ avec $\alpha = \|\overrightarrow{RM}\|$

$$\|\overrightarrow{RM}\| = (\overrightarrow{OM}, \overrightarrow{u}) = {}^t\overrightarrow{u}\,\overrightarrow{a} \text{ car } \|\overrightarrow{u}\|_2 = 1$$

$$\overrightarrow{RM} = \|\overrightarrow{RM}\|\,\overrightarrow{u} = ({}^t\overrightarrow{u}\,\overrightarrow{a})\,\overrightarrow{u}$$

Donc :

$$\overrightarrow{OM'} = \overrightarrow{a} - 2({}^t\overrightarrow{u}\,\overrightarrow{a})\,\overrightarrow{u} = (I - 2\overrightarrow{u}\,{}^t\overrightarrow{u})\,\overrightarrow{a} = H\,\overrightarrow{a}$$

Theorème 3.4.2. *Soient a et b deux vecteurs de $\mathbb{R}^n$ non colinéaires tel que $\|b\|_2 = 1$. Alors il existe $u \in \mathbb{R}^n$ avec $\|u\|_2 = 1$ et une réel $\alpha \in \mathbb{R}$ tel que si $H = I - 2u\,{}^t u$ (noté (1)) alors $Ha = \alpha b$ (noté (2)).*

Démonstration. Pour que (2) ait lieu, il faut :

$$\|a\|_2 = \|Ha\|_2 = |\alpha|\|b\| = |\alpha|$$

donc il y a deux candidats pour α :

- $\alpha = -\|a\|_2$
- $\alpha = \|a\|_2$

$$(2) \Leftrightarrow (I - 2u\,{}^t u)a = \alpha b \Leftrightarrow a - 2u\,\underbrace{{}^t ua}_{\text{scalaire}} = \alpha b$$

$$\Leftrightarrow a - 2({}^t ua)u = \alpha b \Leftrightarrow a - \alpha b = 2\lambda u$$

où in pose $\lambda = {}^t ua$. Pour trouver u, il suffit d'avoir λ. On va faire :

$${}^t a(a - \alpha b) = 2\lambda\,{}^t au$$

$${}^t aa - \alpha\,{}^t ab = 2\lambda\,{}^t au = 2\lambda^2$$

$$\lambda^2 = \frac{1}{2}({}^t aa - \alpha\,{}^t ab)$$

Une condition nécessaire pour que $\lambda > 0$ est que ${}^t aa - \alpha\,{}^t ab > 0$

$$|(a,b)| = |{}^t ab| \overset{CS}{\lneq} \underbrace{\|a\|_2}_{=|\alpha|}\underbrace{\|b\|_2}_{=1} \qquad (*)$$

$\lneq$ car a et b non colinéaires, ce qui est vrai car $(*)$ donc finalement on peut trouver $\lambda \neq 0$ et on pose $u = \frac{1}{2\lambda}(a - \alpha b)$

□

Remarque. (1) H ne dépend pas du signe de λ choisi ($H = I - 2u\,{}^t u$)

(2) Pour éviter d'avoir à calculer une racine carrée, on écrit $H = I - \frac{1}{\beta}v\,{}^t v$ avec :

$$v = 2\lambda u = a - \alpha b$$

$$\beta = 2\lambda^2 = \alpha^2 - \alpha({}^t ab)$$

(3) Il y a deux solutions pour α ($\alpha > 0$ et $\alpha < 0$). Comme β intervient au dénominateur, pour des raisons de stabilités numérique, on choisit celui qui mène à la plus grande valeur de β en prenant celui de signe opposé à ${}^t ab$.

(4) $u = 2\lambda u$.

$$^tvv = {}^t(2\lambda u)(2\lambda u) = 4\lambda^2 \text{ car } \|u\|_2 = 1$$

donc $\beta = 2\lambda^2 = \frac{\|v\|_2}{2}$. Autrement dit, on peut eventuellement ne pas stocker la valeur de β car on la retrouve immédiatement avec celle de v.

(5) Cas particulier : $b = e_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

$$\alpha^2 = \|a\|_2^2 = \sum_{i=1}^{n} a_i^2$$

$$\operatorname{sgn}(\alpha) = \operatorname{sgn}({}^tae_1) = -\operatorname{sgn}(a_1)$$

$$\beta = \alpha^2 - \alpha a_1 = \alpha^2 + |\alpha e_1|$$

$$v = a - \alpha e_1 = \begin{pmatrix} a_1 - \alpha \\ a_2 \\ \vdots \\ a_n \end{pmatrix}$$

3.4.2 Application à la factorisation QR

Soit $A \in \mathbb{R}^{m,n}$, $m \geq n$ et $\operatorname{rg}(A) = n$.

Etape 1 : Soit $a_1 \in \mathbb{R}^m$ la première colonne de $A = A_0$. On détermine :

$$H_1 = I - \frac{1}{\beta_1} v_1 {}^t v_1 \text{ tel que } H_1 a_1 = \alpha_1 e_1$$

donc : $A_1 = H_1 A_0$ a sa première colonne égale à $\alpha_1 e_1$:

$$A_1 = \begin{pmatrix} * \\ 0 \\ \cdots \\ 0 \end{pmatrix}$$

la $i$ème colonne de A_1 est égale à $H_1 a_1$ avec a_i : $i$ème colonne de A_0 $(2 \leq i \leq n)$.

Etape 2 : Même principe appliqué à la matrice constituée des $(m-1)$ dernières lignes et $(n-1)$ dernières colonnes de A_1.

$$A_1 = \begin{pmatrix} * & * & \cdots & * \\ 0 & & & \\ \vdots & & (\tilde{A}_1)_{m-1,n-1} & \\ 0 & & & \end{pmatrix}$$

$$H_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & (\tilde{H}_2)_{m-1,n-1} & \\ 0 & & & \end{pmatrix}$$

On determiner $\tilde{H}_2$ tel que :

$$\tilde{H}_2\tilde{a}_1 = \alpha_2(e_1)_{m-1}$$

avec :

- $\tilde{a}_1$: première colonne de $\tilde{A}_1$
- $(e_1)_{m-1}$: vecteur de la base canonique :

$$\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix}_{m-1}$$

$$\tilde{H}_2 = I_{m-1} - \frac{1}{\beta_2}(\tilde{v}_2)_{m-1}({}^t\tilde{v}_1)_{m-1}$$

$$H_2 = I_m - \frac{1}{\beta_2}(\tilde{v}_2)_m({}^t\tilde{v}_2)_m$$

$$A_2 = H_2A$$

avec :

$$A_2 = \begin{pmatrix} * & * & & & & \\ 0 & * & & & & \\ \vdots & 0 & \ddots & & & \\ \vdots & \vdots & & * & & \\ 0 & 0 & \cdots & \cdots & \cdots & 0 \end{pmatrix}$$

Etape successif : $A_i = H_iA_{i-1}$ $(1 \leq i \leq n-1)$ avec $A_{n-1} = R$ où :

$$H_i = H_i^{-1} = I - \frac{1}{\beta_i}v_i{}^tv_i$$

$$H_i = \begin{pmatrix} 1 & & & & \\ & \ddots & & \boxed{0} & \\ & & 1 & & \\ & & & \overset{i-1\times i-1}{} & \\ & & & \boxed{*} & \\ & & & \underset{(m-i+1)\times(m-i+1)}{} & \end{pmatrix}$$

$$A_{i-1} = H_i^{-1}A_i = H_iA_i$$

donc :

$$A_0 = \underbrace{H_1H_2...H_{n-1}}_{Q}\underbrace{A_{n-1}}_{R}$$

On a donc bien Q matrice orthogonale comme produit de matrices orthogonales. R a des zéros sous sa diagonale principale (si $m-n$, R triangulaire supérieure).

Remarque. Si $m = n$: pour résoudre $AX = B$, on peut le remplacer par $QRX = B \Leftrightarrow RX = {}^tQB$. L'intérêt de la décomposition QR dans ce cas $A = QR$:

$$\text{cond}_2(A) = \text{cond}_2(QR) = \text{cond}_2(R)$$

Pour la résolution d'un système linéaire, la méthode QR est plus stable que la méthode LU (elle est aussi plus chère...).

3.4.3 Algorithme

Algorithme 3.4.1.
QR()

```
1  # Donnée : A ∈ ℝ^{m,n}
2  H ← I_m
3  for k ← 1 to n − 1
4      do α ← √(Σ_{i=k}^{m} a_ik^2) × (sgn(a_kk))
5          β ← α^2 − α × a_kk
6
7  # Construction de v_k
8  v_k ← a_kk − α
9  for i ← k to n
10     do v_i ← a_ik
11
12 # Mise à jour de A
13 for j ← k to n
14     do c ← (1/β) Σ_{i=1}^{k} v_i a_ij
15         for i ← k to m
16             do a_ij ← a_ij − c v_i
17
18 # Mise à jour de H = H_k H_{k−1} ... H_1
19 for j ← 1 to m
20     do c ← (1/β) Σ_{i=k}^{m} v_i H_ij
21         for i ← k to m
22             do H_ij ← H_ij − c v_i
23 Q = ^t H
24 # et dans A, il y a R
```

Complexité algorithmique : 2 fois plus élevé qu'une décomposition LU.

3.5 Méthode de Givens

3.5.1 Les matrices de rotations élémentaires

Définition 3.5.1. Soient $\theta \in \mathbb{R}$ et $n \in \mathbb{N}^*$. On appelle matrice élémentaire de Givens, la matrice $\vartheta^r_{pq}(\theta)$ définie par :

$$\vartheta_{pq}(\theta) \in \mathbb{R}^{n,n} \qquad \vartheta_{pq}(\theta) = \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & \cos\theta & & & \sin\theta & & \\ & & & & 1 & & & & \\ & & & & & \ddots & & & \\ & & & & & & 1 & & \\ & & & -\sin\theta & & & \cos\theta & & \\ & & & & & & & 1 & \\ & & & & & & & & \ddots \\ & & & & & & & & & 1 \end{pmatrix}_{n\times n}$$

où :

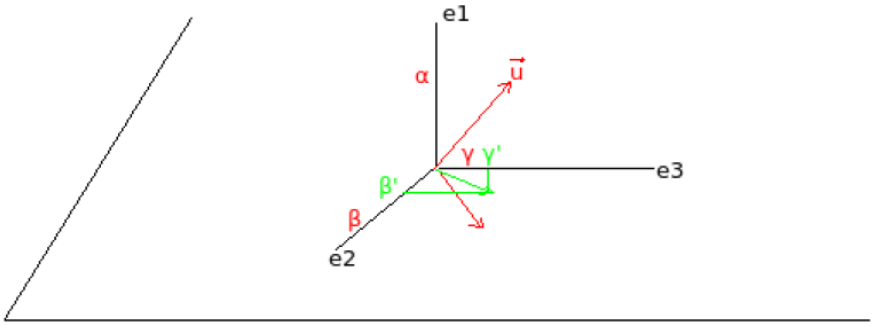
- $\cos\theta$ est à la $p$ième ligne, $p$ième colonne et à la $q$ième ligne, $q$ième colonne.
- $\sin\theta$ est à la $p$ième ligne, $q$ième colonne.
- $-\sin\theta$ est à la $q$ième ligne, $p$ième colonne.

Proposition 3.5.1. *1) Si on multiplie un vecteur $u \in \mathbb{R}^n$ à gauche par $\vartheta_{pq}(\theta)$, on opère une rotation d'angle θ das le plan $\{e_p, e_q\}$.*

Exemple : *en dimension 3 :*

$$u = \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$

$$\vartheta_{23}(\theta).u = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} = \begin{pmatrix} \alpha \\ \beta\cos\theta + \gamma\sin\theta = \beta' \\ -\beta\sin\theta + \gamma\cos\theta = \gamma' \end{pmatrix}$$



2) Si on multiplie à gauche une matrice par $\vartheta_{pq}(\theta)$, on effectue cette rotation sur chacun de ses vecteurs colonnes.

3) $\vartheta_{pq}(\theta)$ est orthogonale ($^t\vartheta_{pq}(\theta)\vartheta_{pq}(\theta) = I$)

3.5.2 Annulation d'un coefficient d'une matrice $A \in \mathbb{R}^{m,n}$

Soit $B \in \mathbb{R}^{n,n}$ tel que $B = \vartheta_{pq}(\theta)A$ $(1 \leq p, q \leq n,\ p \neq q)$. Seules les lignes p et q de la matrice B sont différentes de celles de A :

$$\begin{cases} b_{pk} = \cos\theta a_{pk} + \sin\theta a_{qk} & 1 \leq k \leq n \\ b_{qk} = -\sin\theta a_{pk} + \cos\theta a_{qk} & 1 \leq k \leq n \end{cases}$$

Supposons que l'on cherche à ce que $b_{qp} = 0$, il faut choisir θ tel que $0 = b_{qp} = -\sin\theta a_{pp} + \cos\theta a_{qq}$. On sait de plus que $\cos^2\theta + \sin^2\theta = 1$ donc on a deux possibilités pour choisir θ :

$$\begin{cases} \cos\theta = \dfrac{a_{pp}}{\sqrt{a_{pp}^2 + a_{qp}^2}} \overset{\text{def}}{=} \alpha \\ \sin\theta = \dfrac{a_{qp}}{\sqrt{a_{pp}^2 + a_{qp}^2}} \overset{\text{def}}{=} \beta \end{cases} \quad \text{ou} \quad \begin{cases} \cos\theta = -\alpha \\ sin\theta = -\beta \end{cases}$$

Remarque. Si $a_{pp} \neq 0$ alors :

$$\tan\theta = \frac{\beta}{\alpha} = \frac{a_{qp}}{a_{pp}}$$

Seul la possibilité choisie, on aura :

$$b_{pp} = \pm\alpha a_{pp} \times \pm\beta a_{qp} = \pm\sqrt{a_{pp}^2 + a_{qp}^2}$$

et donc, en choisiant la première possibilité, on peut par exemple s'assurer que $b_{pp} \geq 0$.

3.5.3 Application à la factorisation QR

$A = (a_{ij})_{1 \leq i,j \leq n}$:

1) On écrit $A^{(2,1)} = \vartheta_{21}(\theta)A$ avec le coefficient $(2,1)$ de $A^{(2,1)}$ qui est nul et le coefficient $(1,1)$ de $A^{(2,1)}$ qui est positif.

$$\begin{pmatrix} \cos\theta & \sin\theta & & 0 \\ -\sin\theta & \cos\theta & & \\ & & 1 & \\ & 0 & & \ddots \\ & & & & 1 \end{pmatrix} A = \begin{pmatrix} * & \\ 0 & \\ * & \\ \vdots & \\ * & \end{pmatrix}$$

si $a_{21} = 0$, on ne fait rien !

2) On continue en annulant succesivement tous les coefficients de la première colonne :

$$A^{(k,1)} = \vartheta_{1k}(\theta)A^{(k-1,1)} \qquad 3 \leq k \leq n$$

On a donc :

$$A^{(1)} = A^{(n,1)} = \vartheta_{1n}A^{(n-1),1)} = \vartheta_{1,n}\vartheta_{1,n-1}...\vartheta_{12}A = Q_1$$

Dans $A^{(1)}$, tous les coefficients de la première colonne sont nuls sauf le premier que est $*$.

3) On recommence la même stratégie pour annuler successivement les coefficients $A^{(1)}(3,2)$, $A^{(1)}(4,2)$, ..., $A^{(1)}(n,2)$. On note :

$$\varphi_2 = \vartheta_{2n}\vartheta_{2,n-1}...\vartheta_{23}$$

et on a :

$$A^{(2)} = Q_2 A^{(1)} = Q_2 Q_1 A$$

on a finalement :

$$A^{(n-1)} = \underbrace{Q_{n-1} Q_{n-2} ... Q_1 A}_{\tilde{Q}}$$

On a : $R = \tilde{Q}A \Leftrightarrow A = QR$ avec $Q = {}^t\tilde{Q}$.

Remarque. $A \in \mathbb{R}^{m,n}$, $m \geq n$, $\text{rg}(A) = m$. On applique exactement la même stratégie :

$$\begin{pmatrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{pmatrix} \to \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & * & * \\ 0 & * & * \end{pmatrix} \to \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & * \end{pmatrix} \to \begin{pmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \end{pmatrix}$$

Complexité : Nombre d'opérations est le double du nombre d'opérations pour Householder. Mais cette algorithme est bien adapté pour les matrices creuses.

Il existe une variante appelée "Givens rapide" qui a la même complexité algorithmique que Householder.

3.6 Application à la recherche de valeurs propres

Il existe une méthode dite "QR" pour evaleur les valeurs propres d'une matrice $A \in \mathbb{R}^{n,n}$. On définit :

$$A_1 = A$$

$$A_1 = Q_1 R_1 \text{ et on définit } A_2 = R_1 Q_1$$

$$A_2 = Q_2 R_2 \text{ et on définit } A_3 = R_2 Q_2$$

$$\vdots \qquad \vdots$$

On va obtenir une suite de matrices semblables à A :

$$\begin{aligned} A_{k+1} &= R_k Q_k \\ &= {}^t Q_k A_k Q_k \\ &= ... \\ &= {}^t(Q_1 Q_2 Q_3 ... Q_k) A (Q_1 Q_2 Q_3 ... Q_k) \end{aligned}$$

Theorème 3.6.1. *Soit A inversible et dont les valeurs propres sont toutes de modules différentes. Il existe donc une matrice P inversible tel que $A = PAP^{-1}$ avec :*

$$A = \text{diag}(\lambda_1, \lambda_2, \lambda_n) \text{ et } |\lambda_1| > |\lambda_2| > ... > |\lambda_n| > 0$$

On suppose que P^{-1} admet une décomposition LU. Alors la suite de matrices $(A_k)_{k \geq 1}$ et telle que :

$$\begin{cases} \lim\limits_{k \to +\infty} (A_k)_{ii} = \lambda_1, \ 1 \leq i \leq n \\ \lim\limits_{k \to +\infty} (A_k)_{ij} = 0, \ 1 \leq j < i \leq n \end{cases}$$

Remarque. Il existe d'autres méthodes de recherche d'éléments propores (méthode de la puissance, Jacobi)...

Chapitre 4

Méthodes itératives de résolution de systèmes linéaires

4.1 Principe de la méthode

$A \in \mathbb{R}^{n,m}$ inversible, $n \in \mathbb{N}^*$. On veut résoudre le système linéaire $Ax = b$. L'idée consiste à construire une suite de vecteurs $(x^{(k)})_{0 \leq k}$, qui a pour objectif de converger vers x, la solution du système linéaire. On se donne $x^{(0)} \in \mathbb{R}^n$:

$$x^{(k+1)} = Bx^{(k)} + c \qquad (*)$$

$c \in \mathbb{R}^n$, $B \in \mathbb{R}^{n,n}$ est appelé matrice d'itération.

Remarque. 1) Si on pose le problème de la convergence de la méthode, c'est-à -dire la propriété $\forall x^{(0)} \in \mathbb{R}^n$, $\lim\limits_{k \to 0} x^{(k)} = x$. Se pose aussi le problème de la vitesse de convergence, pour pouvoir comparer plusieurs méthodes entre elles.

2) Pour la mise en oeuvre d'une telle méthode, le coût d'une itération est de l'ordre d'un produit matrice/vecteur qui peut être faible si B est creuse.

Remarque (Conditions nécessaire de convergence). Pour que la méthode $(*)$ converge vers x, solution de $Ax = b$, il faut que :

$$c = (I - B)A^{-1}b \qquad (P)$$

En effet, si la méthode converge vers $\overline{x}$, alors nécessairement :

$$\overline{x} = B\overline{x} + c$$

Comme $x = A^{-1}b$, alors $c = (I - B)A^{-1}b$.

Theorème 4.1.1. *On suppose (P) vérifiée. La méthode itérative est convergente $\Leftrightarrow \rho(B) < 1$* [1]

Démonstration.

$$x^{(k+1)} = Bx^{(k)} + c \qquad x^{(k+1)} - x = Bx^{(k)} - Bx$$

$$\forall 0 \leq n,\ x(n) - x = B(x^{(n-1)} - x) = B^n(x^{(0)} - x) \qquad (**)$$

[1] $\rho(B)$ est le rayon spectral

($\Rightarrow$) Supposons $\rho(\beta) \geq 1$, si $\rho(B) \geq 1$, $\exists y \in \mathbb{R}^n$ tel que $B^n y \nrightarrow 0$ pour $n \to +\infty$. En choissant $x^{(0)} = x + y = A^{-1}b + y$, $(**)$ devient :

$$x^{(n)} - x = B^n y \xrightarrow[n\to\infty]{} 0$$

donc la méthode n'est pas convergeante. Absurde. Donc $\rho(B) < 1$.

($\Leftarrow$) Si $\rho(B) < 1$ alors

$$x^{(n)} - x = B^n(x^{(0)} - x) \xrightarrow[n\to+\infty]{} 0$$

donc $x^{(n)} \xrightarrow[n\to+\infty]{} x = A^{-1}b$. La méthode converge.

□

Nouvelle factorisation. $A \in \mathbb{R}^{n,n}$ inversible, $b \in \mathbb{R}^n$. On va écrire $A = M - N$ avec M inversible (et facile à inverser !). On va définir la méthode itérative :

$$\begin{cases} x^{(0)} \text{ vecteur arbitraire} \\ Mx^{(k+1)} = Nx^{(k)} + b \end{cases} \qquad (**)$$

Remarque. Si cette méthode itérative converge alors elle converge vers la solution du système linéaire. En effet, si $x^{(k)} \xrightarrow[k\to+\infty]{} y \in \mathbb{R}^n$ alors $My = Ny + b \Leftrightarrow (M - N)y = b \Leftrightarrow Ay = b$.

Theorème 4.1.2. *(i) La méthode itérative $(**)$ converge $\Leftrightarrow \rho(M^{-1}N) < 1$*

(ii) La méthode itérative $(**)$ convrege $\Leftrightarrow$ il existe une norme induite tel que $\|M^{-1}N\| < 1$.

Démonstration.

$$Mx^{(k+1)} = Nx^{(k)} + b \Leftrightarrow x^{(k+1)} = \underbrace{M^{-1}N}_{B} x^{(k)} + M^{-1}b$$

Voir le théorème précédent.

($\Leftarrow$) Si il existe une norme induite tel que $\|M^{-1}N\| < 1$ alors $\rho(M^{-1}N) < 1$. Donc la méthode converge grâce à (i).

($\Rightarrow$) Si la méthode converge alors $\rho(M^{-1}N) < 1$ grâce à (i). Donc $\exists \eta > 0$ tel que $\rho(M^{-1}N) = 1 - \eta$. On prend $\varepsilon = \frac{\eta}{2}$, il existe une norme induité $\|\cdot\|$ tel que $\|M^{-1}N\| \leq \rho(M^{-1}N) + \varepsilon < 1$.

□

Theorème 4.1.3 (Condition suffisante de convergence)**.** *$A \in \mathbb{R}^{n,n}$ symétrique définie positive. $A = M - N$, M inversible. Si ${}^tM + N$ est symétrique définie positive alors $\rho(M^{-1}N) < 1$ et donc la méthode $(**)$ converge.*

Démonstration. Si $B \in \mathbb{R}^{n,n}$ et que $\|\cdot\|$ est une norme induite sur $\mathbb{R}^n$, on a $\rho(B) \leq \|B\|$. On va chercher une norme sur $\mathbb{R}^n$, notée $\|\cdot\|_*$ tel que :

$$\|M^{-1}N\|_* = \max_{\|x\|_*=1} \|M^{-1}Nx\|_* < 1$$

Soit $\|\cdot\|_*$ définie par $\|x\|_*^2 = (Ax, x)$, $\forall x \in \mathbb{R}^n$.

Soit $x \in \mathbb{R}^n$, $x \neq 0$:

$$\|MN^{-1}x\|_*^2 = (AM^{-1}Nx, M^{-1}Nx)$$

Soit $y = M^{-1}Ax$, $y \neq 0$ car $x \neq 0$ et $M^{-1}A$ inversible :

$$M^{-1}Nx = M^{-1}(M - A)x = x - y$$

$$\begin{aligned}\|M^{-1}Nx\|_*^2 &= (A(x-y),x-y)\\ &= \underbrace{(Ax,x)-2(Ax,y)+(Ay,y)}_{A \text{ symétrique}}\\ &= \|x\|_*^2-2(Ax,y)+(Ay,y)\end{aligned}$$

Pour montrer que $\|M^{-1}Nx\|_*^2 < \|x_*^2$, $\forall x \in \mathbb{R}^n$, $x \neq 0$. Il suffit de montrer que $-2(Ax,y)+A(y,y)<0$. Comme $My = Ax$:

$$-2(Ax,y)+(Ay,y) = -2(My,y)+(Ay,y)$$

Mais :

$$(My,y) = (y,{}^tMy) = ({}^tMy,y)$$
$$-2(Ax,y)+(Ay,y) = ((-M-{}^tM+A)y,y)$$

Comme $A = M-N$:

$$-2(Ax,y)+(Ay,y) = (-({}^tM+N)y,y) \underbrace{<}_{\text{car } y\neq 0} 0$$

Par hypothèse : $({}^tM+N)$ symétrique définie positive. □

Remarque. Si on note $e^{(k)} = x^{(k)}-x$, avec B matrice d'itération alors :

$$e^{(k)} = x^{(k)}-x = B(x^{(k-1)}-x) = Be^{(k-1)} = B^{(k)}e^{(0)}$$
$$\|e^{(k)}\|_2 = \|B^ke^{(0)}\|_2 \leq \|B^k\|_2\|e^{(0)}\|_2 \underbrace{\leq}_{B \text{ normale}} (\rho(B))^k\|e^{(k)}\|_2$$

Bilan. Dans ce cas, on voit que plus $\rho(B)$ est petit, plus la convergence est rapide.

Dans le cas général, la conclusion est identique asymptotiquement, $\|e^{(k)}\|$ converge vers 0 comme $(\rho(B))^k$.

4.2 Quelques méthodes usuelles

Dans toute cette partie, on prend $A \in \mathbb{R}^{n,n}$, inversible. On cherche à résoudre $Ax = b$ par la méthode itérative.

4.2.1 La méthode de Jacobi

$A = D-E-F$

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n}\\ \vdots & & \vdots\\ a_{n1} & \cdots & a_{nn}\end{pmatrix} \qquad D = \begin{pmatrix} a_{11} & & & 0\\ & a_{22} & & \\ & & \ddots & \\ 0 & & & a_{nn}\end{pmatrix}$$

$$-E = \begin{pmatrix} 0 & \cdots & \cdots & 0\\ a_{21} & \ddots & & \vdots\\ \vdots & & \ddots & \vdots\\ a_{n1} & \cdots & a_{n,n-1} & 0\end{pmatrix} \qquad -F = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n}\\ \vdots & \ddots & & \vdots\\ \vdots & \ddots & a_{n-1,n} & \\ 0 & \cdots & \cdots & 0\end{pmatrix}$$

On choisit :

$$\begin{cases} M = F\\ N = (E+F)\end{cases} \qquad A = M-N$$

Remarque. Cette méthode peut être définie si et seulement si $a_{ii} \neq 0$, $\forall 1 \leq i \leq n$.

$$\begin{array}{rcrcl} Mx^{(k+1)} = Nx^{(k)} + b & \Leftrightarrow & Dx^{(k+1)} & = & (E+F)x^{(k)} + b \\ & \Leftrightarrow & x^{(k+1)} & = & D^{-1}(E+F)x^{(k)} + D^{-1}b \\ & & & = & D^{-1}(D-D+E+F)x^{(k)} + D^{-1}b \\ & & & = & D^{-1}(D-A)x^{(k)} + D^{-1}b \\ & & x^{(k+1)} & = & (F - D^{-1}A)x^{(k)} + D^{-1}b \end{array}$$

Composante par composante :

$$\begin{array}{rcl} x^{(k+1)} & = & x^{(k)} + D^{-1}(b - Ax^{(k)}) \\ \downarrow & & \\ x_i^{(k+1)} & = & x_i^{(k)} + \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^{(k)} \right), \qquad 1 \leq i \leq n \end{array}$$

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j \neq i}^{n} a_{ij} x_j^{(k)} \right), \qquad 1 \leq i \leq n$$

Remarque. $x_i^{(k+1)} - x_i^{(k)} = \dfrac{r_i^{(k)}}{a_{ii}}$ avec $r^{(k)} = b - Ax^{(k)}$ (résidu).

Exemple 4.2.1. 1)

$$A = \begin{pmatrix} 10 & 1 \\ 2 & 10 \end{pmatrix} \qquad b = \begin{pmatrix} 11 \\ 12 \end{pmatrix} \qquad Ax = b \Leftrightarrow x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$x^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad x^{(1)} = \begin{pmatrix} 11/10 \\ 12/10 \end{pmatrix} \qquad x^{(2)} = \begin{pmatrix} 98/100 \\ 99/100 \end{pmatrix} \qquad x^{(3)} = \begin{pmatrix} 1002/1000 \\ 1004/1000 \end{pmatrix}$$

On conjecture la convergence.

2)

$$A = \begin{pmatrix} 1 & 10 \\ 10 & 2 \end{pmatrix} \qquad b = \begin{pmatrix} 11 \\ 12 \end{pmatrix} \qquad Ax = b \Leftrightarrow x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$x^{(0)} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \qquad x^{(1)} = \begin{pmatrix} 11 \\ 6 \end{pmatrix} \qquad x^{(2)} = \begin{pmatrix} -49 \\ -49 \end{pmatrix} \qquad x^{(3)} = \begin{pmatrix} 501 \\ 251 \end{pmatrix} \qquad x^{(4)} = \begin{pmatrix} -2499 \\ -2499 \end{pmatrix}$$

Divergence.

4.2.2 Méthode de Gauss-Seidel

Avec les mêmes définitions des matrices D, E et F que dans la **Section 4.2.1**, on définit :

$$\begin{cases} M = D - E \\ N = F \end{cases}$$

$$(D - E)x^{(k+1)} = Fx^{(k)} + b$$

avec :

$$D - E = \begin{pmatrix} a_{11} & \cdots & 0 \\ \vdots & \ddots & \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \qquad F = \begin{pmatrix} 0 & a_{12} & \cdots & a_{1n} \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ 0 & \cdots & \cdots & 0 \end{pmatrix}$$

Composante par composante :

$$\sum_{j=1}^{i} a_{ij} x_j^{(k-1)} = -\sum_{j=i+1}^{n} a_{ij} x_j^{(k)} + b_i \qquad (1 \leq i \leq n)$$

Supposons qu'on est déjà obtenu $x_1^{(k+1)}, x_2^{(k+1)}, ..., x_{i-1}^{(k+1)}$:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right)$$

Remarque. Comme pour Jacobi, une condition nécessaire et suffisante pour que l'algorithme soit bien définie et que $a_{ii} \neq 0$, $\forall 1 \leq i \leq n$.

En général, Gauss-Seidel converge plus vite que Jacobi.

4.2.3 Méthode de relaxation

Pour Gauss-Seidel, on avait décomposé :

$$A = (D - E) - F$$

Maintenant, on choisit $\omega \in \mathbb{R}^*$ et on prend :

$$A = \underbrace{\left(\frac{D}{\omega} - E\right)}_{M} - \underbrace{\left(\frac{1-\omega}{\omega} D + F\right)}_{N}$$

- $\omega = 1 \Rightarrow$ Gauss-Seidel.
- On a pu aussi "faire passer" une partie de la matrice D dans la matrice N. La matrice d'itération est alors :

$$B_\omega = \left(\frac{D}{\omega} - E\right)^{-1} \left(\frac{1-\omega}{\omega} D + F\right) = (D - \omega E)^{-1}((1-\omega)D + \omega F)$$

On peut ainsi espérer jouer sur ω pour avoir $\rho(B_\omega)$ aussi petit que possible.

Définition 4.2.1.
- $\omega > 1$: sur-relaxation.
- $\omega < 1$: sous-relaxation.

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) + (1-\omega) x_i^{(k)}$$

L'algorithme n'est pas plus couteux que Gauss-Seidel à ceci près qu'il faut tenir compte du temps nécessaire de la détermination de ω. En notant :

$$\tilde{r}_i^{(k)} = b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij} x_j^{(k)}$$

On a :

$$x_i^{(k+1)} - x_i^{(k)} = \frac{\omega}{a_{ii}} \tilde{r}_i^{(k)}$$

4.2.4 Méthode par blocs

On peut généraliser toutes ces méthodes en utilisant des décompositions par bloc de matrices.

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1p} \\ A_{21} & A_{22} & \cdots & A_{2p} \\ \vdots & & & \vdots \\ A_{1p} & \cdots & \cdots & A_{pp} \end{pmatrix} \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_p \end{pmatrix}$$

où :
- $A_{ij} \in \mathbb{R}^{n,n}$, $1 \leq i, j \leq p$,
- $x_i \in \mathbb{R}^n$,
- $b_i \in \mathbb{R}^n$.

Alors la méthode de relaxation peut s'écrire :

$$A_{ii}x_i^{(k+1)} = \omega \left(b_i - \sum_{j=1}^{i-1} A_{ij}x_j^{(k-1)} \right) + (1-\omega)A_{ii}x_i^{(k)} - \omega \sum_{j=i-1}^{p} A_{ij}x_j^{(k)}$$

→ il faut donc résoudre pour obtenir $x_i^{(k+1)}$ un système linéaire de matrice A_{ii}.
→ il faut donc que A_{ii} soit inversibles $(1 \leq i \leq p)$.
→ on souhaite que le système linéaire soit rapide à résoudre, cela avantage la méthode. On peut penser à une décomposition LU de A_{ii} (par exemple) calculée une fois pour toute au début de l'algorithme.

4.2.5 Test d'arrêt

On utilise habituellement $\frac{\|r^{(k)}\|}{\|b\|} \leq \varepsilon$, avec ε donné. A priori, on peut montrer dans ce cas que :

$$\|e^{(k)}\| \leq \text{cond}(A)\varepsilon\|x\|$$

Remarque. Si on utilise Jacobi, on a directement $r^{(k)}$. Si on utilise Gauss-Seidel, on utilise en fait $\tilde{r}^{(k)}$, ce qui évite des calculs supplémentaires ($\frac{\|\tilde{r}^{(k)}\|}{\|b\|} \leq \varepsilon$).

Autre possibilité. B matrice d'itération, B symétrique définie positive. On peut utiliser :

$$\|x^{(k+1)} - x^{(k)}\| \leq \varepsilon\|B\|$$

$$\|e^{(k)}\|_2 = \|e^{(k+1)} - (x^{(k+1)} - x^{(k)})\|_2 \leq \rho(B)\|e^{(k)}\|_2 + \|x^{(k+1)} - x^{(k)}\|_2$$

On en déduit donc :

$$\|e^{(k)}\|_2 \leq \frac{1}{1-\rho(B)} \underbrace{\|x^{(k+1)} - x^{(k)}\|}_{\leq \varepsilon\|b\|_2}$$

- Intéressant pour $\rho(B)$ petit. La constante peut exploser si $\rho(B) \xrightarrow[x \to +\infty]{} 1$.
- Si $\rho(B)$ est petit, ça reste une bonne stratégie pour des matrices non nécessairement symétique définie positive.

Attention :. Le test d'arrêt pourrait être vérifie si $x^{(k)}$ est loin de x.

4.3 Etude de convergence

$A \in \mathbb{R}^{n,n}$, $n \in \mathbb{N}^*$. On veut résoudre $Ax = b$. On pose $A = M - N$:

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$$

Comme $(M - N)x = b$, si la méthode converge, elle converge forcément vers x. On note $B = M^{-1}N$ la matrice d'itération. On a déjà vu une condition nécessaire et suffisante de convergence (**Théorème 4.1.2**). Reprenons les exemples de la **Section 4.2**.

Exemple 1.

$$A = \begin{pmatrix} 10 & 1 \\ 2 & 10 \end{pmatrix}$$

La matrice de Jacobi est :

$$J = I - D^{-1}A = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} 1/10 & 0 \\ 0 & 1/10 \end{pmatrix} \begin{pmatrix} 10 & 1 \\ 2 & 10 \end{pmatrix} = \begin{pmatrix} 0 & -1/10 \\ -2/10 & 0 \end{pmatrix}$$

On a : $\mathrm{Sp}(J) = \left\{-\frac{\sqrt{2}}{10}, \frac{\sqrt{2}}{10}\right\}$, $\rho(J) = \frac{\sqrt{2}}{10} < 1$ converge.

Exemple 2.

$$A = \begin{pmatrix} 1 & 10 \\ 10 & 2 \end{pmatrix} \qquad J = \begin{pmatrix} 0 & -10 \\ 5 & 0 \end{pmatrix}$$

On a : $\mathrm{Sp}(J) = \{-\sqrt{50}, \sqrt{50}\}$, $\rho(J) = \sqrt{50} > 1$ (divergence).

4.3.1 Méthode de relaxation

Theorème 4.3.1. *Soit B_ω la matrice d'itération de la méthode de relaxtion. Alors :*
1) $\rho(B_\omega) < 1 \Rightarrow 0 < \omega < 2$.
2) Si, de plus, A est symétrique définie positive alors $(\rho(B_\omega) < 1 \Leftrightarrow 0 < \omega < 2)$.

Démonstration. 1) On calcule $\det(B_\omega)$, $B_\omega = M^{-1}N$ avec :

$$\begin{cases} M = \frac{1}{\omega}D - E \\ N = \frac{1-\omega}{\omega}D + F \end{cases}$$

$$\det(B_\omega) = (\det(M))^{-1} \det(N)$$

Mais M et N sont triangulaires : le déterminant est le produit des termes diagonaux.

$$\det(B_\omega) = \frac{\left(\frac{1-\omega}{\omega}\right)^n \det(D)}{\left(\frac{1}{\omega}\right)^n \det(D)} = (1-\omega)^n$$

Mais le déterminant est le produit des valeurs propres de la matrice, dont chaque module est inférieur ou égal au rayon spectral. Donc :

$$|\det(B_\omega)| = |(1-\omega)^n| \leq (\rho(B_\omega))^n$$

Donc :

$$\rho(B_\omega) < 1 \Rightarrow 0 < \omega < 2$$

2) On suppose que A est symétrique définie positive et $0 < \omega < 2$. On montre que $\rho(B_\omega) < 1$. On utilise le **Théorème 4.1.3** :

$$ {}^tM = {}^t\left(\frac{D}{\omega} - E\right) = \frac{D}{\omega} - F $$

$$ {}^tM + N = \frac{D}{\omega} - F + F + \frac{1-\omega}{\omega}D = \frac{2-\omega}{\omega}D $$

On a que D est symétrique définie positive car A est définie positive. Donc d'après le **Théorème 4.1.3**, $\rho(B_\omega) < 1$.

□

Remarque. Ce théorème nous dit que A symétrique définie positive ⇒ Gauss-Seidel converge. Mais attention, ce n'est forcément le cas pour Jacobi.

Exemple 4.3.1.

$$ A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix} \qquad -\frac{1}{2} < a < 1 $$

A est symétrique définie positive mais pourtant Jacobi converge $\Leftrightarrow -\frac{1}{2} < a < \frac{1}{2}$.

4.3.2 Comparaison Jacobi / Gauss-Seidel pour les matrices tridiagonales

Theorème 4.3.2. *Soit A matrice tridiagonale. Si les méthodes de Jacobi et de Gauss-Seidel sont définies alors $\rho(B_1) = (\rho(J))^2$. Donc ces méthodes convergent ou divergent simultanément. Si elles convergent, Gauss-Seidel converge plus rapidement.*

Démonstration. 1) Soit $\mu \neq 0$. Soit :

$$ A(\mu) = \begin{pmatrix} b_1 & \mu^{-1}c_1 & & & 0 \\ \mu a_2 & b_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ 0 & & \ddots & \ddots & \mu^{-1}c_{n-1} \\ & & & \mu a_n & b_n \end{pmatrix} $$

On a : $\det(A(\mu)) = \det(A(1))$, $\forall \mu \neq 0$. En effet :

$$ Q(\mu) = \begin{pmatrix} \mu & & & \\ & \mu^2 & 0 & \\ & 0 & \ddots & \\ & & & \mu^n \end{pmatrix} $$

On vérifie que $A(\mu) = Q(\mu)A(1)(Q(\mu))^{-1}$. Donc :

$$ \det(A(\mu)) = \det(Q(\mu))\det(A(1))\det(Q(\mu))^{-1} = \det(A(1)) $$

2) – Les valeurs propres de $J = D^{-1}(E+F)$ sont les racines du polynôme caractéristique :

$$ p_J(\lambda) = \det(D^{-1}(E+F) - \lambda I) $$

On définit q_J par :

$$ q_J(\lambda) = \det(-D)p_J(\lambda) = \det(\lambda D - E + F) $$

Les racines de $p_J(\lambda)$ sont aussi racines de $q_J(\lambda)$.

– De même, les valeurs propres de $B_1 = (D-E)^{-1}F$ sont les racines du polynôme caractéristique :

$$p_{B_1}(\lambda) = \det((D-E)^{-1}F - \lambda I)$$

On définit q_{B_1} par :

$$q_{B_1}(\lambda) = \det(E-D)p_{B_1}(\lambda) = \det(\lambda D - \lambda E - F)$$

Les racines de $p_{B_1}(\lambda)$ sont aussi racines de $q_{B_1}(\lambda)$.

3)

$$\begin{aligned} q_{B_1}(\lambda^2) &= \det(\lambda^2 D - \lambda^2 E - F) \\ &\quad \downarrow 1) + A \text{ tridiagonale} \\ &= \det(\lambda^2 D - \lambda E - \lambda F) = \lambda^n \det(\lambda D - E - F) = \lambda^n q_J(\lambda) \end{aligned}$$

Remarque. Vraie pour $\lambda = 0$ aussi.

– Soit $\beta \in \mathrm{Sp}\{B_1\}$

$$\Rightarrow \rho_{B_1}(\beta) = 0 \Rightarrow q_{B_1}(\beta) = 0 \overset{\beta \neq 0}{\Rightarrow} q_J(\beta^{1/2}) = q_J(-\beta^{1/2}) = 0$$

où $\beta^{1/2}$ est l'une des deux valeurs (complexes) de β.

$$\Rightarrow p_J(\beta^{1/2}) = p_J(-\beta^{1/2}) = 0 \Rightarrow \{-\beta^{1/2}, \beta^{1/2}\} \in \mathrm{Sp}(J)$$

– Réciproquement, $\beta \in \mathrm{Sp}(J)$ et $\beta \neq 0 \Rightarrow \beta^2 \in \mathrm{Sp}(B)$. On en déduit que $\rho(B_1) = (\rho(J))^2$.

□

Remarque. Même résultat quand A est triagonale par blocs :

$$\begin{pmatrix} B & -I & & \\ -I & B & \ddots & \\ & \ddots & \ddots & \ddots \\ & & \ddots & \ddots & -I \\ & & & -I & B \end{pmatrix}$$

avec :

$$I = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}_{m \times m} \qquad B = \begin{pmatrix} 4 & 1 & & \\ -1 & \ddots & & \\ & & \ddots & 1 \\ & & -1 & 4 \end{pmatrix}_{m \times m}$$

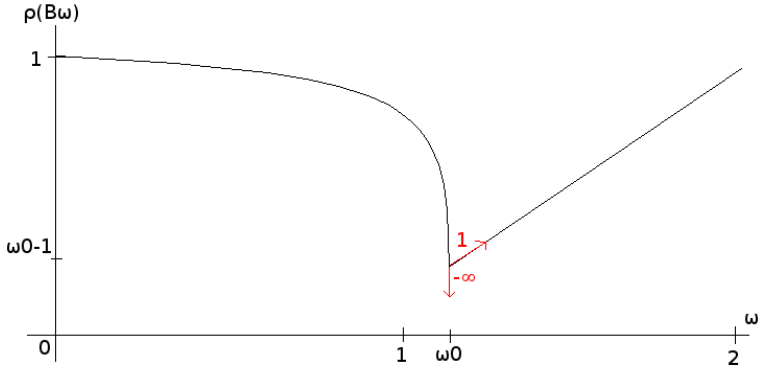
4.3.3 Paramétre de relaxation optimal pour le cas tridiagonal

Theorème 4.3.3. *Soit A matrice tridiagonale avec $a_{ii} \neq 0$. Si toutes les valeurs propres de la matrice de Jacobi sont réelles alors la méthode de Jacobi et la méthode de relaxation ($0 < \omega < 2$) convergent ou divergent simultanément. En cas de convergence, le paramètre ω_0 tel que $\rho(B_{\omega_0}) = \min\{\rho(B_\omega), \omega \in]0,2[\}$ s'exprime en fonction de $\rho(J)$ par la formule :*

$$\omega_0 = \frac{2}{1 + \sqrt{1 + (\rho(J))^2}}$$

et on a : $\rho(B_{\omega_0}) = \omega_0 - 1$.

Remarque. – Cet énoncé est aussi valable pour les matrice tridiagonales par blocs.
– Evolution de $\rho(B_\omega)$ en fonction de ω.



$$\begin{cases} \lim\limits_{\omega \to \omega_0^-} f'(\omega) = -\infty \\ \lim\limits_{\omega \to \omega_0^+} f'(\omega) = 1 \end{cases}$$

Si on a une approximation de ω_0, on privilégie une valeur par excès (sur-relaxation par rapport à la valeur optimale).

4.3.4 Matrices à diagonale dominante

Démonstration. $A = (a_{ij})_{1 \leq i,j \leq n}$

1) On dit que A est à diagonale dominante :

$$\forall i \in \{1, ..., n\} \qquad |a_{ii}| \geq \sum_{j=1, j \neq i}^{n} |a_{ij}|$$

2) On dit que A est à diagonale strictement dominante si :

$$\forall i \in \{1, ..., n\} \qquad |a_{ii}| > \sum_{j=1, j \neq 1}^{n} |a_{ij}|$$

3) On dit que A est à diagonale fortement dominante si :
   a) A est à diagonale dominante,
   b) $\exists k \in \{1, ..., n\}$ tel que :

$$|a_{kk}| > \sum_{j=1, j \neq k}^{n} |a_{kj}|$$

□

Définition 4.3.1. $n \geq 2$, $n \in \mathbb{N}^*$
– A est dite réductible si il existe une matrice de permutation P telle que $B = {}^tPAP$ sont de la forme :

$$\begin{pmatrix} \underbrace{B_{11}}_{p \times p} & \underbrace{B_{12}}_{(n-p) \times p} \\ \underbrace{0}_{p \times (n-p)} & \underbrace{B_{22}}_{(n-p) \times (n-p)} \end{pmatrix}$$

avec B_{11} et B_{22} des matrices carrés d'ordre p et $n - p$ respectivement ($p \in \mathbb{N}^*$).
– A est irréductible si elle n'est pas réductible.

Theorème 4.3.4 (Premier théorème de Gershgorin-Hadamard)**.** *Si λ est valeur propre de A alors :*

$$\lambda \in \bigcup_{k=1}^{n} D_k$$

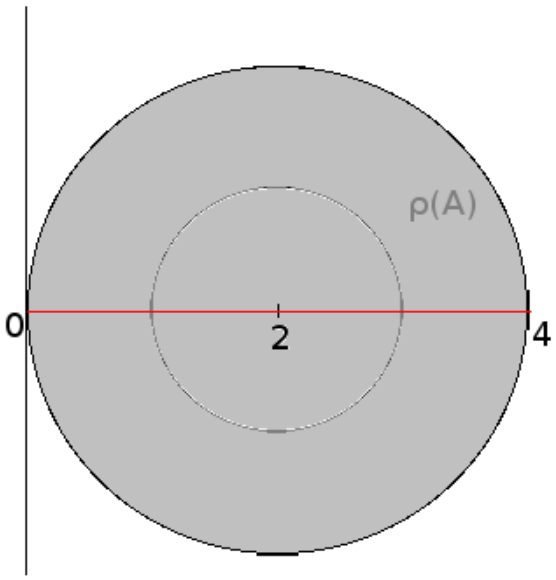
avec :

$$D_k = (a_{kk}, \Lambda_k) \text{ et } \Lambda_k = \sum_{j=1, j\neq i}^{n} |a_{kj}|$$

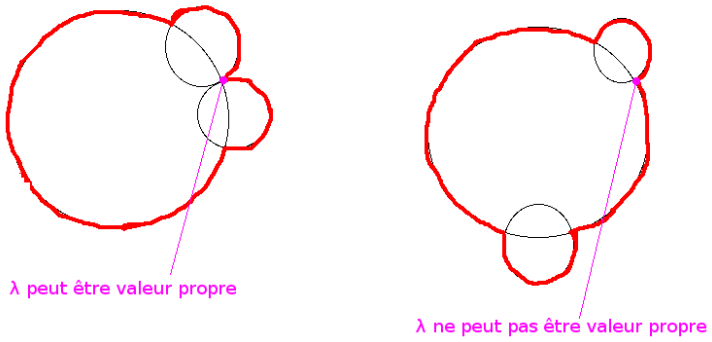
D_k est appelé le kième disque de Greshgorin.

Exemple 4.3.2.

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



Theorème 4.3.5 (Deuxième théorème de Gershgorin-Hadamard)**.** *Soit A irréductible. Si λ est une valeur propre de A située sur la frontière de la réunion des disques de Gershgorin alors tous les cercles de Gershgorin passent par λ.*



Exemple 4.3.3. D'après le deuxième théorème de Greshgorin-Hadamard, 0 n'est pas valeur propre de :

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

Démonstration. 1) Soit (λ, u) un élément propre de A et u choisit tel que $\|u\|_\infty = 1$. λ n'est à l'intérieur d'aucun disque D_k : $\forall k \in \{1, .., n\}$, $|\lambda - a_{kk}| \geq \Lambda_k$.

2) Clairement, il existe $l \in \{1, ..., n\}$ tel que $|u_l| = \max\limits_{1 \leq k \leq n} |u_k| = 1$. D'après le premier théorème de Gershgorin-Hadamard, $|\lambda - a_{ll}| \leq \Lambda_l$. Donc, nécessairement 1) + 2) $\Rightarrow |\lambda - a_{ll}| = \Lambda_l$.

3) Soit $I = \{i \in \{1, ..., n\} \,|\, |u_i| = 1\} \neq \emptyset$.
   – A est irréductible si elle n'est pas réductible :

$$\sum_{j=1, j\neq i}^{n} |a_{ij}||u_j| \geq \left| \sum_{j=1, j\neq i}^{n} a_{ij} u_j \right| = |(\lambda - a_{ii})u_i|$$

$$= \lambda - a_{ii} = \Lambda_i = \sum_{j=1, j\neq i}^{n} |a_{ij}|$$

On vient donc de montrer que :

$$\sum_{j=1, j\neq i}^{n} |a_{ij}|(1 - |u_j|) \leq 0$$

Comme tous les termes de cette somme ≥ 0, on a :

$$|a_{ij}|(1 - |u_j|) = 0 \qquad \forall j$$

Si $j \notin I \Rightarrow a_{ij} = 0$. Soit $J = I^C$ le complémentaire de I par rapport à $\{1, ..., n\}$. $J \neq \emptyset \Rightarrow$ la partition I, J serait telle que $\forall i \in I, \forall j \in J$, $a_{ij} = 0$. On pourrait donc en utilisant une matrice de permutation trouver que :

$$B = {}^tPAP = \begin{pmatrix} \underbrace{*}_{I\times I} & \underbrace{*}_{J\times I} \\ \underbrace{0}_{I\times J} & \underbrace{*}_{J\times J} \end{pmatrix}$$

Cela signifie que A est réductible (contraire à l'hypothèse). Donc nécessairement $J = \emptyset$. Donc $\forall k$, $k \in I$ et $|u_k| = 1$. Donc :

$$|\lambda - a_{kk}| = |(\lambda - a_{kk})u_k| = \left| \sum_{k\neq j} a_{kj} u_j \right| \leq \sum_{j\neq k} |a_{kj}| = \Lambda_k \qquad \forall j \leq k \leq n$$

Donc : λ appartient à tous les disques D_k s'il appartient à la frontière de ces disques (par hypothèse). Donc il appartient à la frontière de tous les D_k.

□

Theorème 4.3.6. *A matrice soit à diagonale strictement dominante, soit irréductible et à diagonale fortement dominante. Alors la méthode de Jacobi est convergente.*

Démonstration. 1) On montre quue A est inversible.

→ Si A est à diagonale strictement dominante. Soit $i \in \{1, ..., n\}$:

$$|0 - a_{ii}| = |a_{ii}| > \sum_{j=1, j\neq i}^{n} |a_{ij}| = \Lambda_i$$

Donc : 0 n'appartient à aucun des disques de Gershgorin. En appliquant le premier théorème de Gershgorin-Hadamard, on peut en déduire que 0 ne peut être valeur propre de A donc A est inversible.

→ Si A est à diagonale fortement dominante et irréductible. Avec le même raisonnement, on a que $\forall i \in \{1, ..., n\}$:

$$|0 - a_{ii}| \geq \Lambda_i \qquad \text{(car à diagonale dominante)}$$

Donc : 0 n'est à l'intérieur d'aucun des disques. Donc si 0 est valeur propre, il est sur la frontière de l'union des disques. A irérréductible donc, d'après le deuxième théorème de Gershgorin-Hadamard, tous les cercles passent par 0. A diagonale fortement dominante donc il existe k tel que $|0 - a_{kk}| > \Lambda_k$. Donc D_k ne peut pas passer par 0. Donc : 0 n'est pas valeur propre $\Rightarrow$ A est inversible.

2) La méthode de Jacobi est bien définie. En effet, supposons que $\exists i \in \{1, ..., n\}$ tel que $a_{ii} = 0$:

$$a_{ii} = 0 \Rightarrow \sum_{j=1, j\neq i}^{n} |a_{ij}| = 0 \qquad \text{(matrice à diagonale dominante)}$$

Donc la $i$ème ligne de A est constituée de 0. Donc : A non inversible $\Rightarrow$ contradiction avec 1). Donc $\forall i \in [\![1, n]\!]$, $a_{ii} \neq 0$ et la méthode de Jacobi est bien définie.

3) On étudie le rayon de convergence de $J = D^{-1}(E + F)$:

$$\begin{cases} J_{ii} = 0 & (1 \leq i \leq n) \\ J_{ij} = -\frac{a_{ij}}{a_{ii}} & (1 \leq i, j \leq n,\ i \neq j) \end{cases}$$

$$\sum_{j=1}^{n} |J_{ij}| = \sum_{j=1, j\neq i}^{n} \left|\frac{a_{ij}}{a_{ii}}\right|$$

a) Si A est a diagonale dominante :

$$\forall i,\ 1 \leq i \leq n, \qquad \sum_{j=1}^{n} |J_{ij}| < 1$$

$$\max_{1\leq i\leq n} \sum_{j=1}^{n} |J_{ij}| = \|J\|_\infty < 1$$

$\rho(J) \leq \|J\|_\infty < 1$: la méthode converge.

b) Si A est à diagonale fortement dominante et irréductible. Si on fait le même raisonnement alors on aura $\rho(J) \leq \|J\|_\infty$. Supposons que $|\lambda| = 1$ soit valeur propre de J alors, d'après le deuxième théorème de Gershgorin-Hadamard, tous les cercle de Gershgorin passerait par λ. Cela est impossible car il existe au moins un indice (fortement dominant) par lequel cette inégalité est strict donc $|\lambda| = 1$ n'est pas valeur propre de J donc $\rho(J) < 1$: la méthode converge.

□

Theorème 4.3.7. *A matrice soit à diagonale strictement dominante, soit à diagonale fortement dominante et irréductible.* $0 < \omega \leq 1 \Rightarrow$ *la méthode de relaxation converge.*

Remarque. Il s'agit d'une condition suffisante de convergence.

Exemple 4.3.4.

$$A = \begin{pmatrix} 2 & 1 \\ 0 & 5 \end{pmatrix} \qquad B_\omega = \begin{pmatrix} 1-\omega & -\frac{\omega}{2} \\ 0 & 1-\omega \end{pmatrix}$$

La méthode converge $\Leftrightarrow |1-\omega| < 1$, c'est-à-dire $0 < \omega < 2$.

4.3.5 Vitesse de convergence d'une méthode itérative

On considère une méthode itérative de matrice d'itération B.

$$e^{(k)} = x^{(k)} - x \qquad e^{(k)} = B^k e^{(0)}$$

$$\begin{cases} Ax = b \\ x(k+1) = Bx^{(k)} + c \end{cases}$$

$$\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \leq \|B^k\| = \max_{e^{(k)} \neq 0} \frac{\|B^k e^{(0)}\|}{\|e^{(0)}\|} = \max_{e^{(0)} \neq 0} \frac{\|e^{(k)}\|}{\|e^{(0)}\|}$$

Définition 4.3.2. On appelle :

$$\sigma = \left(\frac{\|e^{(k)}\|}{\|e^{(0)}\|} \right)^{1/k}$$

le facteur moyen de réduction de l'erreur par itération.

$$\sigma \leq \|B^k\|^{1/k} = \exp(-R_k(B))$$

où :

$$R_k(B) = -\frac{1}{k}(\ln \|B^k\|)$$

Si on définit $N_k = (R_k(B))^{-1}$:

$$\sigma^{N_k} = \exp(N_k \ln(\sigma)) \leq \left(\frac{1}{R_k(B)}(-R_k(B)) \right) = -\frac{1}{e}$$

N_k mesure donc le nombre d'itérations nécessaires pour réduire la norme de l'erreur initiale d'un facteur e.

Si on considère par exemple :

$$\begin{cases} x^{(k+1)} = Bx^{(k)} + c & x^{(0)} \text{ donné} \quad (i) \\ \tilde{x}^{(k+1)} = \tilde{B}\tilde{x}^{(k)} + c & x^{(0)} \text{ donné} \quad (ii) \end{cases}$$

Si $R_k(B) < R_k(\tilde{B})$ alors $N_k > \tilde{N}_k$ et donc la méthode (ii) va être plus rapide que (i).

On va définir :

$$R_\infty(B) = \lim_{k \to +\infty} -\ln \|B^k\|^{1/k} \overset{\text{Ch.2}}{=} -\ln \rho(B)$$

$R_\infty(B)$ est le taux asymptotique de convergence de la méthode.

Si on pose $N_\infty(B) = (R_\infty(B))^{-1}$, on a : $\rho(\tilde{B}) < \rho(B) < 1 \Rightarrow R_\infty(\tilde{B}) > R_\infty(B) \Rightarrow N_\infty(\tilde{B}) < N_\infty(B)$.

Pour comparer deux méthode entre elles, on regardera : $M(B) \times N_\infty(B)$ où $M(B)$ est le nombre d'opérations nécessaires pour effectuer une itération.

4.4 Introduction aux méthodes de gradient

Dans cette partie, on suppose A symétrique définie positive, de valeurs prores $0 < \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n$.

4.4.1 Méthode du gradient à pas constant (Richardson)

Soit $A = M - N$, $\alpha > 0$, $M = \frac{1}{\alpha} I$ et $N = \frac{I}{\alpha} - A$.

$$Mx^{(k+1)} = Nx^{(k)} + b \Leftrightarrow \frac{1}{\alpha} x^{(k+1)} = \left(\frac{1}{\alpha} I - A\right) x^{(k)} + b$$

$$\Leftrightarrow x^{(k+1)} = x^{(k)} - \alpha(Ax^{(k)} - b) \Leftrightarrow x^{(k+1)} = x^{(k)} + \alpha(\underbrace{b - Ax^{(k)}}_{\text{résidu}})$$

Theorème 4.4.1. *On rappelle que A est symétrique définie positive.*

1) La méthode de Richazrdson converge $\Leftrightarrow \alpha \in \left]0, \frac{2}{\lambda_n}\right[$.

2) Le paramètre optimal α_{opt} qui minimalise $\rho(M^{-1}N)$ est donné par :

$$\alpha_{opt} = \frac{2}{\lambda_1 + \lambda_n}$$

pour lequel :

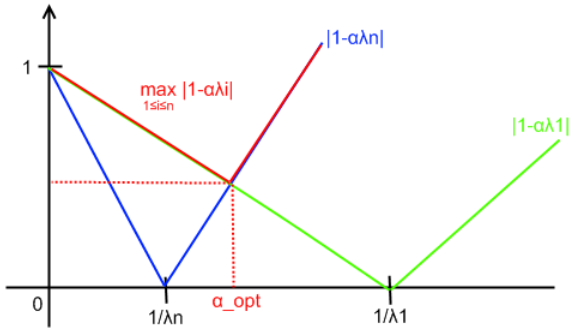
$$\rho(M^{-1}N) = \frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}$$

Démonstration. 1) Pour $\alpha \in \mathbb{R}^{*+}$, on pose $G_\alpha = M^{-1}N$:

$$\rho(G_\alpha) = \rho(I - \alpha A) = \max_{1 \leq i \leq n} |1 - \alpha\lambda_i|$$

$$\rho(G_\alpha) < 1 \Leftrightarrow \forall i \in [\![1, n]\!],\ |1 - \alpha\lambda_i| < 1 \Leftrightarrow \forall i \in [\![1, n]\!],\ 0 < \alpha\lambda_i < 2 \Leftrightarrow 0 < \alpha < \frac{2}{\lambda_n}$$

2) Maintenant, on calcule α_{opt} tel que $\rho(G_{\alpha_{\text{opt}}}) = \min_{0 < \alpha < \frac{2}{\lambda_n}} \rho(G_\alpha)$



On voit que α_{opt} vérifie :

$$1 - \alpha_{\text{opt}}\lambda_1 = \alpha_{\text{opt}}\lambda_n - 1 \Leftrightarrow \alpha_{\text{opt}} = \frac{2}{\lambda_1 + \lambda_n}$$

3)

$$\rho(G_{\alpha_{\text{opt}}}) = 1 - \frac{2\lambda_1}{\lambda_1 + \lambda_n} = \frac{\frac{\lambda_n}{\lambda_1} + 1}{\frac{\lambda_n}{\lambda_1} - 1} = \frac{\text{cond}_2(A) - 1}{\text{cond}_2(A) + 1}$$

□

4.4.2 Interprétation fonctionnelle

A symétrique définie positive, $A \in \mathbb{R}^{n,n}$ et $b \in \mathbb{R}^n$. On considère la fonction :

$$\begin{array}{rccc} J & : & \mathbb{R}^n & \to & \mathbb{R} \\ & & x & \mapsto & J(x) = \frac{1}{2}(Ax.x) - (b.x) \end{array}$$

Cette fonction est appelée fonctionnelle quadratique sur $\mathbb{R}^n$ dû au caractère symétrique définie positive de A.

On peut montrer que J est strictement convexe et qu'elle atteint son minimum en un unique point $x^* \in \mathbb{R}^n$ caractérisé par $\nabla J(x^*) = 0$. Soit $h \in \mathbb{R}^n$:

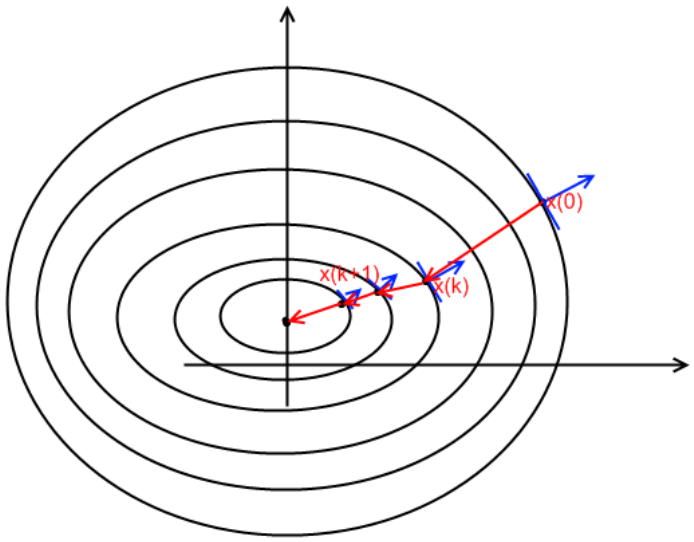
$$(\nabla J(x^*), h) = \frac{1}{2}(Ax^*, h) + \frac{1}{2}(Ah, x^*) - (b, h) \overset{A \text{ sym.}}{=} (Ax^* - b, h)$$

$$\nabla J(x) = 0 \Leftrightarrow Ax^* - b = 0 \Leftrightarrow Ax^* = b$$

$$x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)}) = x^{(k)} - \alpha(Ax^{(k)} - b) = x^{(k)} - \alpha \underbrace{\nabla J(x^{(k)})}_{\text{Méthode du gradient}}$$

Exemple 4.4.1 (en 2D)**.** $x \in \mathbb{R}^2$, $A \in \mathbb{R}^{2,2}$ symétrique définie positive.

Lignes de niveau : $J(x) = \text{cste}$.



Peut-on améliorer les choses ?

1) On choisit α_k à la place de α : méthode du gradiant à pas optimal local.
2) Ne peut-on pas choisir comme direction de descente autre chose que $\nabla J(x^{(k)})$? Méthode plus sophistiquée : méthode de gradient conjugué. Accélération très forte de la convergence.