MACS1.
Cours d' Analyse numérique

L. Halpern

24 septembre 2010

Table des matières

Bibliographie

[1] G. Allaire, S.M. Kaber, *Algèbre linéaire numérique*. Ellipses, 2002

[2] M. Schatzmann, *Numerical Analysis, A Mathematical Introduction*.Oxford University Press, 2002.

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

[4] E. Hairer : consulter la page `http ://www.unige.ch/ hairer/polycop.html`

Première partie

Introduction générale

Le Calcul Scientifique se propose de mettre un problème issu de la physique, de l'économie, de la chimie, de l'ingénierie, en équations, c'est l'étape de la **modélisation**, et de les résoudre. Ces équations sont souvent très complexes, et font intervenir énormément de paramètres. Elles sont en général impossible à résoudre de façon exacte (comme le serait une équation différentielle du second degré par exemple, modélisant le mouvement d'un pendule de longueur l :

$$x'' + \frac{g}{l} sinx = 0. \tag{1}$$

Le problème linéarisé pour de petits mouvements du pendule s'écrit

$$x'' + \frac{g}{l} x = 0 \tag{2}$$

peut se résoudre sous la forme $x = x_0 cos\sqrt{\frac{g}{l}}t + x'_0 sin\sqrt{\frac{g}{l}}t$. On peut alors calculer $x(t)$ pour tout temps de façon exacte (modulo les erreurs d'arrondi). Par contre on ne connait pas de solution exacte de l'équation (1). On est donc amené à en chercher une solution approchée en un certain nombre de points (cf cours de mise à niveau) : On souhaite calculer x dans l'intervalle $]0, T[$, connaissant $x(0) = x_0$ et $x'(0) = x'_0$. On se donne une suite d'instants $t_n = n\Delta t$, avec $T = N\Delta t$, et on écrit une approximation de la dérivée seconde

$$x''(t_n) = \frac{x(t_{n+1}) - 2x(t_n) + x(t_{n-1})}{\Delta t^2} + \mathcal{O}(\Delta t^2) \tag{3}$$

et on remplace l'équation par

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta t^2} + \frac{g}{l} sin(y_n) = 0. \tag{4}$$

Puisque l'on a une équation de récurrence à 2 niveaux, il faut se donner y_0 et y_1. Nous verrons plus tard comment calculer y_1. L'équation (4) est une *approximation* de l'équation (1). Il est souhaitable que

1. (4) ait une solution unique,
2. $y_n \approx x(t_n)$, c'est la **consistance**,
3. une erreur petite sur les données initiales y_0 et y_1 produise une erreur faible sur y_n : c'est la **stabilité**.

Ce sont les 3 notions de base en Calcul Scientifique. L'équation (4) peut aussi se mettre sous forme condensée, $F(Y) = b$, c'est alors un système non linéaire dont il faut trouver une solution.

L'approximation par différences finies de (2) s'écrit

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta t^2} + \frac{g}{l} y_n = 0. \tag{5}$$

Elle peut se mettre sous forme d'un système linéaire, c'est-à-dire $AY = b$, où

$$A = \begin{pmatrix} 1 & 0 & & & \\ 0 & 1 & 0 & & \\ 0 & -1 & \alpha & -1 & \\ & \ddots & \ddots & \ddots & -1 \\ & & 0 & -1 & \alpha \end{pmatrix}, Y = \begin{pmatrix} y_0 \\ \vdots \\ y_N \end{pmatrix}, b = \begin{pmatrix} y_0 \\ y_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

avec $\alpha = 2 + \frac{g}{l}\Delta t^2$. De manière générale, toute équation issue de la modélisation est ensuite **discrétisée** puis mise sous la forme d'un système, différentiel ou non, linéaire ou non. Tout commence par la résolution des systèmes linéaires, puis des équations non linéaires, puis des équations différentielles. Nous verrons en deuxième année les modèles d'équations aux dérivées partielles.

Deuxième partie

Résolution numérique de systèmes linéaires

Chapitre 1

Généralités

Sommaire

1.1 Rappels sur les matrices

1.1.1 Définitions

Une matrice (m, n) est un tableau à m lignes et n colonnes

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

C'est aussi la matrice d'une application linéaire $\mathcal{A}$ de K^n dans K^m où K est $\mathbf{R}$ ou $\mathbf{C}$: une base $\mathbf{e_1}, \cdots, \mathbf{e_n}$ étant choisie dans K^n, et une base $\mathbf{f_1}, \cdots, \mathbf{f_m}$ dans K^m, $\mathcal{A}$ est défini par

$$1 \leq j \leq n, \mathcal{A}(\mathbf{e_j}) = \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{a_{ij}} \mathbf{f_i}$$

Le j-ème vecteur colonne de A représente donc la décomposition de $\mathcal{A}(\mathbf{e_j})$ dans la base $\mathbf{f_1}, \cdots, \mathbf{f_m}$.

Définition 1.1 *L'application linéaire $\mathcal{A}$ est injective si*

$$\mathcal{A}(\mathbf{x}) = \mathbf{0} \Rightarrow \mathbf{x} = \mathbf{0}$$

Définition 1.2 *L'application linéaire $\mathcal{A}$ est surjective si pour tout $\mathbf{b}$ dans K^m, on peut trouver $\mathbf{x}$ dans K^n tel que $\mathcal{A}(\mathbf{x}) = \mathbf{b}$*

Définition 1.3 *L'application linéaire $\mathcal{A}$ est bijective si elle est à la fois injective et surjective.*

Si $\mathcal{A}$ est bijective , on a $m = n$, la matrice A est carrée.
Opérations sur les matrices

1. Somme : On peut ajouter deux matrices de même dimension (m, n) et
$$(A + B)_{ij} = (A)_{ij} + (B)_{ij}$$
2. Produit par un scalaire : Pour α dans K, on peut faire le produit αA et
$$(\alpha A)_{ij} = \alpha (A)_{ij}$$
3. Produit de 2 matrices : Pour $A(m, n)$ et $B(n, p)$ on peut faire le produit AB, de dimension(m, p) et
$$(AB)_{ij} = \sum_{k=1}^{n} (A)_{ik} (B)_{kj}$$
4. Transposée d'une matrice : Pour $A(m, n)$ la transposée de A est de dimension (n, m) et est définie par $({}^tA)_{ij} = A_{ji}$.
5. Adjointe d'une matrice : Pour $A(m, n)$ l'adjointe de A est de dimension (n, m) et est définie par $(A^*)_{ij} = \bar{A}_{ji}$.
6. Inverse d'une matrice **carrée** : on dit que la matrice carrée A est inversible si il existe une matrice B telle que $AB = BA = I$. La matrice B est appelée l'inverse de A et notée A^{-1}.

1.1.2 Cas particulier de matrices

Toutes les matrice considérées dans ce paragraphe sont carrées.

1. Matrices symétriques : elles vérifient ${}^tA = A$, ou encore $a_{ij} = a_{ji}$.
2. Matrices hermitiennes : elles vérifient $A^* = A$, ou encore $\bar{a}_{ij} = a_{ji}$.
3. Matrices diagonales : elles vérifient $a_{ij} = 0$ pour $i \neq j$.
4. Matrices triangulaires inférieures : elles vérifient $a_{ij} = 0$ pour $j > i$, *i.e.* elles ont la forme

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

5. Matrices triangulaires supérieures : elles vérifient $a_{ij} = 0$ pour $j < i$, *i.e.* elles ont la forme

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

Les matrices triangulaires sont importantes pour la résolution numérique des systèmes car elles ont les propriétés suivantes :

- La transposée d'une matrice triangulaire inférieure est triangulaire supérieure et réciproquement ;
- Le produit de deux matrices triangulaires inférieures est triangulaire inférieure et le produit de deux matrices triangulaires supérieures est triangulaire supérieure.
- L'inverse d'une matrice triangulaire inférieure est triangulaire inférieure et l'inverse d'une matrice triangulaire supérieure est triangulaire supérieure.

1.1.3 Déterminants

Le déterminant d'une **matrice carrée** A se note $det\, A$, ou

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

Il obéit à la règle de calcul de développement par rapport à une ligne ou une colonne et a les propriétés suivantes

1. $det\, I = 1$,
2. $det\ {}^tA = det\, A$,
3. $det\, A^* = \overline{det\, A}$,
4. pour tout scalaire (complexe ou réel) α, $det(\alpha\, A) = \alpha^n det\, A$,
5. $det\, AB = det\, A \times det\, B$,
6. Si A est inversible, $det\, A^{-1} = \dfrac{1}{det\, A}$,
7. Le déterminant d'une matrice triangulaire est égal au produit de ses éléments diagonaux.

1.1.4 Produit de matrices par blocs

On décompose la matrice carrée A de la façon suivante :

$$A = \left(\begin{array}{ccc|ccc} a_{11} & \cdots & a_{1J} & a_{1J+1} & \cdots & a_{1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{I1} & \cdots & a_{IJ} & a_{IJ+1} & \cdots & a_{In} \\ \hline a_{I+11} & \cdots & a_{I+1J} & a_{I+1J+1} & \cdots & a_{I+1n} \\ \vdots & & \vdots & \vdots & & \vdots \\ a_{n1} & \cdots & a_{nJ} & a_{nJ+1} & \cdots & a_{nn} \end{array}\right) = \begin{pmatrix} A^{11}_{(I,J)} & A^{21}_{(I,n-J)} \\ A^{12}_{(n-I,J)} & A^{22}_{(n-I,n-J)} \end{pmatrix}$$

La matrice $A^{11}_{(I,J)} = \begin{pmatrix} a_{11} & \cdots & a_{1J} \\ \vdots & & \vdots \\ a_{I1} & \cdots & a_{IJ} \end{pmatrix}$ est de dimension(I, J),

$A^{21}_{(I,n-J)} = \begin{pmatrix} a_{1J+1} & \cdots & a_{1n} \\ \vdots & & \vdots \\ a_{IJ+1} & \cdots & a_{In} \end{pmatrix}$ est de dimension$(I, n-J)$,

$$A^{12}_{(n-I,J)} = \begin{pmatrix} a_{I+11} & \cdots & a_{I+1J} \\ \vdots & & \vdots \\ a_{n1} & \cdots & a_{nJ} \end{pmatrix} \text{ est de dimension}(n-I, J),$$

$$A^{21}_{(n-I,n-J)} = \begin{pmatrix} a_{I+1J+1} & \cdots & a_{I+1n} \\ \vdots & & \vdots \\ a_{nJ+1} & \cdots & a_{nn} \end{pmatrix} \text{ est de dimension}(n-I, n-J).$$

Si l'on prend une matrice B partitionnée de la façon suivante :

$$B = \begin{pmatrix} B^{11}_{(J,K)} & B^{21}_{(J,n-K)} \\ B^{12}_{(n-J,K)} & B^{22}_{(n-J,n-K)} \end{pmatrix},$$

alors on peut faire le produit AB comme si l'on avait affaire à des matrices 2×2 :

$$AB = \begin{pmatrix} A^{11}B^{11} + A^{12}B^{21} & A^{11}B^{12} + A^{12}B^{22} \\ A^{21}B^{11} + A^{22}B^{21} & A^{21}B^{12} + A^{22}B^{22} \end{pmatrix},$$

1.2 Réduction des matrices

Soit A une matrice carrée $n \times n$, on dit que λ est valeur propre si il existe un $x \neq 0$ tel que $Ax = \lambda x$. On dit alors que x est un vecteur propre associé à la valeur propre λ. Les valeurs propres sont les zéros du polynôme caractéristique $p(x) = det(A - xI)$. L'espace propre associé à la valeur propre λ est $E_\lambda = Ker(A - \lambda I)$. On appelle multiplicité de la valeur propre λ sa multiplicité en tant que zéro de p.

On dit que A est diagonalisable si il existe une base $(f_1, \cdots, f_n)$ de $\mathbb{R}^n$ constituée de vecteurs propres de A associés au valeurs propres $\lambda_1, \cdots, \lambda_n$ (comptées sans la multiplicité). On a alors pour tout i $Af_i = \lambda_i f_i$. On pourra écrire matriciellement $A = P\Lambda P^{-1}$ où Λ est la matrice diagonale des valeurs propres de A, et P la matrice des vecteurs propres.

Théorème 1.1 *La matrice A est diagonalisable sur $\mathbb{R}$ (resp. sur $\mathbb{C}$) si et seulement si*

1. *ses valeurs propres sont dans $\mathbb{R}$ (resp. sur $\mathbb{C}$),*
2. *pour chaque valeur propre la dimension du sous-espace propre est égale à la multiplicité.*

Corollaire 1.1 *Une matrice dont toutes les valeurs propres sont simples est diagonalisable.*

Théorème 1.2 *Une matrice symétrique est diagonalisable en base orthonormée.*

Théorème 1.3 (Théorème de Schur) *Pour toute matrice carrée A, il existe une matrice unitaire U telle que U^*AU est triangulaire. Si de plus A est normale, il existe une matrice une matrice unitaire U telle que U^*AU est diagonale.*

1.3 Algorithme, complexité

Qu'est-ce qu'un algorithme ? C'est une suite d'opérations élémentaires nécessaires pour réaliser une tâche donnée. Qu'est-ce qu'une opération élémentaire ? Dans l'algorithme d'Euclide par exemple pour trouver le pgcd de 2 polynômes a et b, une opération élémentaire est la division euclidienne :

$$\begin{array}{l} a = bq_0 + r_0, r_0 = 0 \text{ ou } d^\circ r_0 < d^\circ a, \\ b = r_0q_1 + r_1, r_1 = 0 \text{ ou } d^\circ r_1 < d^\circ r_0 \\ r_0 = r_1q_2 + r_2, r_2 = 0 \text{ ou } d^\circ r_2 < d^\circ r_1 \end{array}$$

on a alors $a \wedge b = b \wedge r_0 = \cdots = r_{n-1} \wedge r_n$ tant que $r_n \neq 0$. La suite des $d^\circ r_k$ est une suite d'entiers strictement décroissante, il existe donc un n tel que $r_{n+1} = 0$. On a alors $a \wedge b = r_n$. C'est la forme que l'on a apprise à l'école. On peut écrire l'algorithme sous la forme

$$\begin{array}{c} d_1 = d^\circ a; d_2 = d^\circ b\,; \\ \text{si } d_1 < d_2,\ p_1 = b \ \&\ p_2 = a\,; \\ \text{sinon } p_1 = a \ \&\ p_2 = b\,; \\ \text{tant que } p_2 \neq 0, \\ p_1 = q * p_2 + r \\ \text{si } r = 0, pgcd = p_2 \\ \text{sinon } p_1 = p_2; p_2 = r \\ \text{et on normalise.} \end{array}$$

Les opérations élémentaires sont $+, -, *, /$. La complexité d'un algorithme est le nombre d'opérations élémentaires nécessaires à la résolution de l'algorithme. Prenons l'exemple du produit de deux matrices. Soit A une matrice $m \times n$, B une matrice $n \times p$. Pour calculer le produit AB on a l'algorithme avec des boucles

Données A,B
%Initialisation C=zeros(m, p) ;
Pour $i = 1 : m$;
Pour $j = 1 : p$;

Pour $k = 1 : n$;
$C(i,j) = C(i,j) + A(i,k) * B(k,j)$;
Fin ;
Fin ;
Fin ;
Pour calculer chaque $C(i,j)$ on a n multiplications et $n-1$ additions. Ce qui fait nmp multiplications et $(n-1)mp$ additions. Pour des matrices carrées, on obtient en tout, $(2n-1)n^2$. On a longtemps travaillé sur ces questions : cet algorithme est-il optimal, c'est-à-dire existe-t-il des algorithmes de calcul de AB qui nécessitent moins d'opérations ? La réponse est oui : l'algorithme de Strassen qui nécessite moins de $n^{\log_2(7)}$ opérations élémentaires avec $log_2(7) \approx 2.81$. Voir [1].

1.4 Systèmes linéaires, définitions

Résoudre un système linéaire de n équations à m inconnues, c'est trouver m nombres, réels ou complexes, $x_1, \cdots, x_m$, tels que

$$\left\{ \begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m & = & b_2 \\ \vdots & \cdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m & = & b_n \end{array} \right. \tag{1.1}$$

Les données sont les coefficients $a_{ij}, 1 \leq i \leq n, 1 \leq j \leq m$ et $b_j, 1 \leq j \leq n$. On appelle système homogène associé le système obtenu pour $b = (0, \cdots, 0)$. Il est équivalent de se donner la matrice A des coefficients a_{ij} et le vecteur b des b_j :

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{pmatrix}, b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}$$

et de chercher un vecteur

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}$$

tel que

$$Ax = b$$

Il sera souvent utile d'écrire (1.1) sous la forme

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + x_m \begin{pmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{1.2}$$

Soit, en notant $a^{(j)}$ le j-ème vecteur colonne de A, $a^{(j)} = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{pmatrix}$,

$$x_1 a^1 + x_2 a^2 + \cdots x_m a^m = b$$

On en déduit un premier résultat : le système (1.1) admet une solution si et seulement si b appartient au sous-espace vectoriel de $\mathbb{R}^n$ engendré par les m vecteurs colonnes de A : $\mathcal{L}(a^{(1)}, \cdots, a^{(m)})$. On appelle rang du système le rang de la matrice A, c'est la dimension de $\mathcal{L}(a^{(1)}, \cdots, a^{(m)})$. C'est aussi la taille d'un mineur de A d'ordre maximum non nul.

1. On s'interesse d'abord aux systèmes carrés, tels que $m = n$.

   Théorème 1.4 *Supposons $m = n$. Alors les propriétés suivantes sont équivalentes :*
   (i) A est inversible,
   (ii) $\det(A) \neq 0$,
   (iii) pour tout b dans $\mathbb{R}^n$, le système (1.1) admet une solution et une seule,
   (iv) le système homogène associé n'admet que la solution triviale $x = (0, \cdots, 0)$.

   Remarquons que si A n'est pas inversible, son noyau est de dimension $n - r$ d'après le théorème du rang. Soit b est dans ImA, il existe une solution X, et toutes les solutions sont obtenues en ajoutant à X un élément de $KerA$. Si b n'est pas dans ImA, l'ensemble des solutions est vide. Remarquons qu'il n'est pas forcément évident de savoir si A est inversible, cela vient en général avec l'étude de la méthode numérique dont le système est issu.

2. Supposons maintenant que $r = n < m$. Il y a plus d'inconnues que d'équations : le système est sous-déterminé. Supposons, par un changement de numérotation, que le déterminant

$$\begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}$$

soit non nul. On dit que $x_1, \cdots, x_n$ sont les *inconnues principales*, $x_{r+1}, \cdots, x_n$ sont les *inconnues non principales*. On réécrit le système sous la forme

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 - (a_{1r+1}x_{r+1} + \cdots + a_{1m}x_m) \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 - (a_{2r+1}x_{r+1} + \cdots + a_{2m}x_m) \\ \vdots & \ldots \quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n - (a_{nr+1}x_{r+1} + \cdots + a_{nm}x_m) \end{cases}$$

Pour un second membre donné b, pour chaque choix de $(x_{r+1}, \cdots, x_n)$ dans $\mathbb{R}^{n-r}$, il y a une seule solution $x_1, \cdots, x_r$. L'ensemble des solutions est un espace affine de dimension $n - r$. On dira qu'il y a une indétermination d'ordre $m - n$.

3. Cas où $r < n$. Alors on ne peut pas avoir une solution pour tout second membre b, puisque $\mathcal{L}(a^{(1)}, \cdots, a^{(m)}) \subsetneqq \mathbb{R}^n$. On a alors r équations principales, supposons que ce soient les r premières. Nous raisonnons maintenant sur les vecteurs lignes. Notons $l^{(i)}$ les vecteurs lignes de A. Le système se récrit

$$\begin{cases} l^{(1)} \cdot x &= b_1 \\ l^{(2)} \cdot x &= b_2 \\ \vdots & \ldots \quad \vdots \\ l^{(r)} \cdot x &= b_r \\ l^{(r+1)} \cdot x &= b_{r+1} \\ \vdots & \ldots \quad \vdots \\ l^{(n)} \cdot x &= b_n \end{cases}$$

$(l^{(1)}, \cdots, l^{(r)})$ forment un système libre. Le système constitué des r premières équations relèvent donc de l'analyse 2. On peut alors exprimer $l^{(r+1)}, \cdots, l^{(n)}$ en fonction de $(l^{(1)}, \cdots, l^{(r)})$:

$$l^{(j)} = \lambda_{j1} l^{(1)} + \cdots \lambda_{jr} l^{(r)}$$

Si l'on fait une combinaison linéaire des r premières lignes avec les coefficients λ_{jk}, on obtient

$$l^{(j)} \cdot x = \lambda_{j1} b_1 + \cdots \lambda_{jr} b_r$$

d'une part, et b_j d'après le système. On a donc le

Théorème 1.5 *Suppososns que les r premières lignes $(l^{(1)}, \cdots, l^{(r)})$ sont indépendantes, et que les autres lignes vérifient*

$$l^{(j)} = \lambda_{j1} l^{(1)} + \cdots \lambda_{jr} l^{(r)}, r+1 \leq j \leq n \tag{1.3}$$

Alors toute solution du système principal (système des r premières équations) est solution de (1.1) si et seulement si

$$b_j = \lambda_{j1} b_1 + \cdots \lambda_{jr} b_r, r+1 \leq j \leq n \tag{1.4}$$

(1.4) constituent les conditions de compatibilité. Si elles ne sont pas satisfaites, le système est impossible. Si elles sont satisfaites, il y a une indétermination d'ordre m-r comme en 2.

1.5 Norme de vecteurs et de matrices

Définition 1.4 *Une **norme** sur un espace vectoriel V est une application $\|\cdot\| : V \to \mathbb{R}^+$ qui vérifie les propriétés suivantes*

- $\|\mathbf{v}\| = 0 \Longleftrightarrow \mathbf{v} = 0$,
- $\|\alpha \mathbf{v}\| = |\alpha| \, \|\mathbf{v}\|, \forall \alpha \in \mathbb{K}, \, \forall \mathbf{v} \in V$,
- $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|, \, \forall (\mathbf{u}, \mathbf{v}) \in V^2$ *(inégalité triangulaire)*

*Une norme sur V est également appelée **norme vectorielle** . On appelle **espace vectoriel normé** un espace vectoriel muni d'une norme.*

Les trois normes suivantes sont les plus couramment utilisées sur $\mathbb{C}^n$:

$$\|\mathbf{v}\|_1 = \sum_{i=1}^{n} |v_i|$$

$$\|\mathbf{v}\|_2 = \left(\sum_{i=1}^{n} |v_i|^2 \right)^{1/2}$$

$$\|\mathbf{v}\|_\infty = \max_i |v_i| \, .$$

La deuxième norme est la norme euclidienne sur $\mathbb{C}^n$. Elle dérive du produit scalaire $(u, v)_2 = \sum_{i=1}^{n} u_i \bar{v}_i$.

Théorème 1.6 *Soit V un espace de dimension finie. Pour tout nombre réel $p \geq 1$, l'application $v \mapsto \|v\|_p$ définie par*

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{1/p}$$

est une norme.

Rappel 1.1 *Pour $p > 1$ et $\frac{1}{p} + \frac{1}{q} = 1$, l'inégalité*

$$\|\mathbf{uv}\|_1 = \sum_{i=1}^{n} |u_i v_i| \leq \left(\sum_{i=1}^{n} |u_i|^p \right)^{1/p} \left(\sum_{i=1}^{n} |v_i|^q \right)^{1/q} = \|\mathbf{u}\|_p \|\mathbf{v}\|_q$$

*s'appelle l'**inégalité de Hölder**.*

Définition 1.5 *Deux **normes** $\|\cdot\|$ et $\|\cdot\|'$, définies sur un même espace vectoriel V, sont **équivalentes** s'il existe deux constantes C et C' telles que*

$$\|\mathbf{v}\|' \leq C \|\mathbf{v}\| \quad et \quad \|\mathbf{v}\| \leq C' \|\mathbf{v}\|' \quad pour\ tout\ \mathbf{v} \in V.$$

Rappel 1.2 *Sur un espace vectoriel de dimension finie toutes les normes sont équivalentes.*

Définition 1.6 *Soit $\mathcal{M}_n$ l'anneau des matrices d'ordre n, à éléments dans le corps $\mathbb{K}$. Une **norme matricielle** est une application $\|\cdot\| : \mathcal{M}_n \to \mathbb{R}^+$ vérifiant*

1. $\|\mathbb{A}\| = 0 \Longleftrightarrow \mathbb{A} = 0$,
2. $\|\alpha\mathbb{A}\| = |\alpha| \|\mathbb{A}\|, \forall \alpha \in \mathbb{K}, \forall \mathbb{A} \in \mathcal{M}_n$,
3. $\|\mathbb{A}+\mathbb{B}\| \leq \|\mathbb{A}\| + \|\mathbb{B}\|, \forall (\mathbb{A}, \mathbb{B}) \in \mathcal{M}_n^2$ *(inégalité triangulaire)*
4. $\|\mathbb{A}\mathbb{B}\| \leq \|\mathbb{A}\| \|\mathbb{B}\|, \forall (\mathbb{A}, \mathbb{B}) \in \mathcal{M}_n^2$

Rappel 1.3 *Etant donné une norme vectorielle $\|\cdot\|$ sur $\mathbb{K}^n$, l'application $\|\cdot\| : \mathcal{M}_n(\mathbb{K}) \to \mathbb{R}^+$ définie par*

$$\|\mathbb{A}\| = \sup_{\substack{\mathbf{v} \in \mathbb{K}^n \\ \mathbf{v} \neq \mathbf{0}}} \frac{\|\mathbb{A}\mathbf{v}\|}{\|\mathbf{v}\|} = \sup_{\substack{\mathbf{v} \in \mathbb{K}^n \\ \|\mathbf{v}\| \leq \mathbf{1}}} \|\mathbb{A}\mathbf{v}\| = \sup_{\substack{\mathbf{v} \in \mathbb{K}^n \\ \|\mathbf{v}\| = \mathbf{1}}} \|\mathbb{A}\mathbf{v}\|,$$

*est une norme matricielle, appelée **norme matricielle subordonnée** (à la norme vectorielle donnée).*

De plus

$$\|\mathbb{A}\mathbf{v}\| \leq \|\mathbb{A}\| \, \|\mathbf{v}\| \quad \forall \mathbf{v} \in \mathbb{K}^n$$

et la norme $\|\mathbb{A}\|$ *peut se définir aussi par*

$$\|\mathbb{A}\| = \inf \{\alpha \in \mathbb{R} : \|\mathbb{A}\mathbf{v}\| \leq \alpha \, \|\mathbf{v}\| \, , \ \forall \mathbf{v} \in \mathbb{K}^n\} \, .$$

Il existe au moins un vecteur $\mathbf{u}$ *tel que*

$$\mathbf{u} \neq 0 \ \ et \ \ \|\mathbb{A}\mathbf{u}\| = \|\mathbb{A}\| \, \|\mathbf{u}\| \, .$$

Enfin une norme subordonnée vérifie toujours

$$\|\mathbb{I}\| = 1$$

Théorème 1.7 *Soit* $\|\mathbb{A}\| = (a_{ij})$ *une matrice carrée. Alors*

$$\|\mathbb{A}\|_1 \overset{déf.}{=} \sup_{\substack{\mathbf{v} \in \mathbb{C}^n \\ \mathbf{v} \neq \mathbf{0}}} \frac{\|\mathbb{A}\mathbf{v}\|_1}{\|\mathbf{v}\|_1} = \max_j \sum_i |a_{ij}|$$

$$\|\mathbb{A}\|_2 \overset{déf.}{=} \sup_{\substack{\mathbf{v} \in \mathbb{C}^n \\ \mathbf{v} \neq \mathbf{0}}} \frac{\|\mathbb{A}\mathbf{v}\|_2}{\|\mathbf{v}\|_2} = \sqrt{\rho\left(\mathbb{A}^*\mathbb{A}\right)} = \sqrt{\rho\left(\mathbb{A}\mathbb{A}^*\right)} = \|\mathbb{A}^*\|_2$$

$$\|\mathbb{A}\|_\infty \overset{déf.}{=} \sup_{\substack{\mathbf{v} \in \mathbb{C}^n \\ \mathbf{v} \neq \mathbf{0}}} \frac{\|\mathbb{A}\mathbf{v}\|_\infty}{\|\mathbf{v}\|_\infty} = \max_i \sum_j |a_{ij}|$$

La norme $\|\cdot\|_2$ *est invariante par transformation unitaire :*

$$\mathbb{U}\mathbb{U}^* = \mathbb{I} \Longrightarrow \|\mathbb{A}\|_2 = \|\mathbb{A}\mathbb{U}\|_2 = \|\mathbb{U}\mathbb{A}\|_2 = \|\mathbb{U}^*\mathbb{A}\mathbb{U}\|_2 \, .$$

Par ailleurs, si la matrice $\mathbb{A}$ *est normale :*

$$\mathbb{A}\mathbb{A}^* = \mathbb{A}^*\mathbb{A} \Longrightarrow \|\mathbb{A}\|_2 = \rho\left(\mathbb{A}\right) .$$

Remarque 1.1

1. *Si une matrice* $\mathbb{A}$ *est hermitienne, ou symétrique (donc normale), on a* $\|\mathbb{A}\|_2 = \rho\left(\mathbb{A}\right)$ *.*
2. *Si une matrice* $\mathbb{A}$ *est unitaire ou orthogonale (donc normale), on a* $\|\mathbb{A}\|_2 = 1$*.*

Théorème 1.8

1. *Soit* $\mathbb{A}$ *une matrice carrée quelconque et* $\|\cdot\|$ *une norme matricielle subordonnée ou non, quelconque. Alors*

$$\rho\left(\mathbb{A}\right) \leq \|\mathbb{A}\| \, .$$

2. *Etant donné une matrice* $\mathbb{A}$ *et un nombre* $\varepsilon > 0$, *il existe au moins une norme matricielle subordonnée telle que*

$$\|\mathbb{A}\| \leq \rho(\mathbb{A}) + \varepsilon.$$

Théorème 1.9 *L'application* $\|\cdot\|_E : \mathcal{M}_n \to \mathbb{R}^+$ *définie par*

$$\|\mathbb{A}\|_E = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2} = \sqrt{tr(\mathbb{A}^*\mathbb{A})},$$

pour toute matrice $\mathbb{A} = (a_{ij})$ *d'ordre* n, *est une norme matricielle non subordonnée (pour* $n \geq 2$*), invariante par transformation unitaire :*

$$\mathbb{U}\mathbb{U}^* = \mathbb{I} \Longrightarrow \|\mathbb{A}\|_E = \|\mathbb{A}\mathbb{U}\|_E = \|\mathbb{U}\mathbb{A}\|_E = \|\mathbb{U}^*\mathbb{A}\mathbb{U}\|_E$$

et qui vérifie

$$\|\mathbb{A}\|_2 \leq \|\mathbb{A}\|_E \leq \sqrt{n}\,\|\mathbb{A}\|_2\,, \quad \forall \mathbb{A} \in \mathcal{M}_n.$$

De plus $\|\mathbb{I}\|_E = \sqrt{n}$.

1.6 Conditionnement

1.6.1 Erreur d'arrondi

Un nombre réel s'écrit de façon unique $x = \pm a10^b$, où a est la mantisse, b l'exposant, entier. a est un nombre réel tel que $0.1 \leq a < 1$. L'arrondi de x à ℓ termes est noté $arr_\ell(x) = \bar{x}$ et est égal à $\pm\,\bar{a}10^b$, avec $\bar{a} = 0.\underbrace{\cdots}_{\ell}$. par exemple $\pi = 3.141592653\cdots$ s'écrit $\pi = 0.\underbrace{31415926}_{8}53\cdots\,10^1$, et avec $\ell = 8$, on a $\bar{\pi} = 0.\underbrace{31415927}_{8}\,10^1$.

Définition 1.7 *La précision de l'ordinateur est le plus petit eps tel que* $arr_\ell(1 + eps) > 1$.

$$x = 0.\underbrace{10\cdots0}_{\ell}49\cdots\,10^1, \quad arr_\ell(x) = 1,$$
$$x = 0.\underbrace{10\cdots0}_{\ell}50\cdots\,10^1, \quad arr_\ell(x) = 1.\underbrace{10\cdots0}_{}1\,10^1 > 1,$$

On en déduit que $eps = 510^{-\ell}$. Si l'on calcule en base 2, on aura 2^{-l}.

On a pour tout $x \neq 0$, $\left|\frac{x-arr_\ell(x)}{x}\right| < eps$. En effet

$$\frac{x - arr_\ell(x)}{x} = \frac{(a-\bar{a})\,10^b}{a\,10^b} = \frac{(a-\bar{a})}{a} \leq \frac{510^{-\ell-1}}{10^{-1}} = 510^{-\ell} = eps$$

On peut écrire aussi $arr_\ell(x) = x(1+\varepsilon)$, avec $|\varepsilon| < eps$.

1.6.2 Conditionnement d'un problème

Soit P un problème, c'est-à-dire une application de $\mathbb{R}^N$ dans $\mathbb{R}$. Par exemple le produit de 2 nombres s'écrit

$$(x_1, x_2) \mapsto x_1 x_2.$$

Le conditionnement de P mesure l'influence d'une perturbation de x sur la solution du problème $P(x)$:

Définition 1.8 *La condition $\mathcal{K}$ de P est le plus petit nombre tel que*

$$\left|\frac{x-\hat{x}}{x}\right| < eps \Rightarrow \left|\frac{P(x)-P(\hat{x})}{P(x)}\right| < \mathcal{K} \cdot eps$$

Si $\mathcal{K}$ est grand, le problème est mal conditionné, si $\mathcal{K}$ n'est pas trop grand, le problème est bien conditionné.

Exemple 1.1 *: produit de 2 nombres. Soient $\hat{x}_i = (1+\varepsilon_i)x_i$, et $\varepsilon = max(|\varepsilon_i|)$.*

$$\frac{x_1x_2 - \hat{x}_1\hat{x}_2}{x_1x_2} = (1+\varepsilon_1)(1+\varepsilon_2) - 1 = \varepsilon_1 + \varepsilon_2 + \varepsilon_1\varepsilon_2,$$

d'où

$$\left|\frac{x_1x_2 - \hat{x}_1\hat{x}_2}{x_1x_2}\right| \leq \varepsilon^2 + 2\varepsilon$$

Comme ε^2 est négligeable devant ε, on a $\mathcal{K} \approx 2$.

1.6.3 Conditionnement d'une matrice

On veut estimer $x - y$, où x est solution du système linéaire, et y solution du système perturbé

$$\begin{aligned} \mathbb{A}\boldsymbol{x} &= \boldsymbol{b}, \\ (\mathbb{A} + \Delta\mathbb{A})\,\boldsymbol{y} &= (\boldsymbol{b} + \Delta\boldsymbol{b})\,. \end{aligned}$$

Exemple de R.S. Wilson :

$$\mathbb{A} = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}, \quad \boldsymbol{b} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix},$$

$$\mathbb{A} + \Delta\mathbb{A} = \begin{pmatrix} 10 & 7 & 8,1 & 7,2 \\ 7,08 & 5,04 & 6 & 5 \\ 8 & 5,98 & 9,89 & 9 \\ 6,99 & 4,99 & 9 & 9,98 \end{pmatrix}, \quad \boldsymbol{b} + \boldsymbol{\Delta b} = \begin{pmatrix} 32,01 \\ 22,99 \\ 33,01 \\ 30,99 \end{pmatrix},$$

$$\Delta\mathbb{A} = \begin{pmatrix} 0 & 0 & 0,1 & 0,2 \\ 0,08 & 0,04 & 0 & 0 \\ 0 & -0,02 & -0,11 & 0 \\ -0,01 & -0,01 & 0 & -0,02 \end{pmatrix}, \quad \boldsymbol{\Delta b} = \begin{pmatrix} 0,01 \\ -0,01 \\ 0,01 \\ -0,01 \end{pmatrix}.$$

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \Longleftrightarrow \boldsymbol{x} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

$$\mathbb{A}\boldsymbol{u} = \boldsymbol{b} + \boldsymbol{\Delta b} \Longleftrightarrow \boldsymbol{u} = \begin{pmatrix} 1,82 \\ -0,36 \\ 1,35 \\ 0,79 \end{pmatrix}, \Longrightarrow \boldsymbol{\Delta x} = \boldsymbol{u} - \boldsymbol{x} = \begin{pmatrix} 0,82 \\ -1,36 \\ 0,35 \\ -0,21 \end{pmatrix}$$

$$(\mathbb{A} + \Delta\mathbb{A})\,\boldsymbol{v} = \boldsymbol{b} \Longleftrightarrow \boldsymbol{v} = \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}, \Longrightarrow \boldsymbol{\Delta x} = \boldsymbol{v} - \boldsymbol{x} = \begin{pmatrix} -82 \\ 136 \\ -35 \\ 21 \end{pmatrix}$$

Définition 1.9 *Soit $\|.\|$ une norme matricielle subordonnée, le conditionnement d'une matrice régulière $\mathbb{A}$, associé à cette norme, est le nombre*

$$\mathrm{cond}(\mathbb{A}) = \|\mathbb{A}\| \left\|\mathbb{A}^{-1}\right\|.$$

Nous noterons $\mathrm{cond}_p(\mathbb{A}) = \|\mathbb{A}\|_p \left\|\mathbb{A}^{-1}\right\|_p$.

Théorème 1.10 *Soit $\mathbb{A}$ une matrice inversible. Soient $\boldsymbol{x}$ et $\boldsymbol{x} + \boldsymbol{\Delta x}$ les solutions respectives de*

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \text{ et } \mathbb{A}\left(\boldsymbol{x} + \boldsymbol{\Delta x}\right) = \boldsymbol{b} + \boldsymbol{\Delta b}.$$

Supposons $\boldsymbol{b} \neq \boldsymbol{0}$, *alors l'inégalité*

$$\frac{\|\boldsymbol{\Delta x}\|}{\|\boldsymbol{x}\|} \leq \mathrm{cond}(\mathbb{A}) \frac{\|\boldsymbol{\Delta b}\|}{\|\boldsymbol{b}\|}$$

est satisfaite, et c'est la meilleure possible : pour une matrice $\mathbb{A}$ *donnée, on peut trouver des vecteurs* $\boldsymbol{b} \neq \boldsymbol{0}$ *et* $\boldsymbol{\Delta b} \neq \boldsymbol{0}$ *tels qu'elle devienne une égalité.*

Démonstration Il suffit de soustraire les 2 équations. $\boldsymbol{\Delta x}$ est solution du système linéaire

$$\mathbb{A}\boldsymbol{\Delta x} = \boldsymbol{\Delta b}$$

d'où

$$\|\boldsymbol{\Delta x}\| \leq \left\|\mathbb{A}^{-1}\right\| \frac{\|\boldsymbol{\Delta b}\|}{\|\boldsymbol{b}\|} \|\boldsymbol{b}\| \leq \left\|\mathbb{A}^{-1}\right\| \|\mathbb{A}\| \|\boldsymbol{x}\| \frac{\|\boldsymbol{\Delta b}\|}{\|\boldsymbol{b}\|}$$

■

Théorème 1.11 *Soit* $\mathbb{A}$ *une matrice inversible. Soient* $\boldsymbol{x}$ *et* $\boldsymbol{x} + \boldsymbol{\Delta x}$ *les solutions respectives de*

$$\mathbb{A}\boldsymbol{x} = \boldsymbol{b} \text{ et } (\mathbb{A} + \Delta\mathbb{A})(\boldsymbol{x} + \boldsymbol{\Delta x}) = \boldsymbol{b}.$$

Supposons $\boldsymbol{b} \neq \boldsymbol{0}$, *alors l'inégalité*

$$\frac{\|\boldsymbol{\Delta x}\|}{\|\boldsymbol{x} + \boldsymbol{\Delta x}\|} \leq \mathrm{cond}(\mathbb{A}) \frac{\|\Delta\mathbb{A}\|}{\|\mathbb{A}\|}.$$

est satisfaite, et c'est la meilleure possible : pour une matrice $\mathbb{A}$ *donnée, on peut trouver un vecteur* $\boldsymbol{b} \neq \boldsymbol{0}$ *et une matrice* $\Delta\mathbb{A} \neq 0$ *tels qu'elle devienne une égalité.*

Théorème 1.12 *1. Pour toute une matrice inversible* $\mathbb{A}$,

$$\begin{aligned} \mathrm{cond}(\mathbb{A}) &\geq 1, \\ \mathrm{cond}(\mathbb{A}) &= \mathrm{cond}(\mathbb{A}^{-1}), \\ \mathrm{cond}(\alpha\mathbb{A}) &= \mathrm{cond}(\mathbb{A}), \text{ pour tout scalaire } \alpha \neq 0 \end{aligned}$$

2. Pour toute matrice inversible $\mathbb{A}$,

$$\mathrm{cond}_2(\mathbb{A}) = \frac{\mu_{\max}}{\mu_{\min}}$$

où $\mu_{\max}$ *et* $\mu_{\min}$ *sont respectivement la plus grande et la plus petite valeur singulière de* $\mathbb{A}$.

3. *Si* $\mathbb{A}$ *est une matrice normale,*

$$\mathrm{cond}_2(\mathbb{A}) = \frac{\max\limits_i |\lambda_i(\mathbb{A})|}{\min\limits_i |\lambda_i(\mathbb{A})|}$$

où les $\lambda_i(\mathbb{A})$ *sont les valeurs propres de* $\mathbb{A}$.

4. *Le conditionnement* $\mathrm{cond}_2(\mathbb{A})$ *d'une matrice unitaire ou orthogonale vaut* 1.

5. *Le conditionnement* $\mathrm{cond}_2(\mathbb{A})$ *est invariant par transformation unitaire*

$$\mathbb{U}\mathbb{U}^* = \mathbb{I} \Longrightarrow \mathrm{cond}_2(\mathbb{A}) = \mathrm{cond}_2(\mathbb{A}\mathbb{U}) = \mathrm{cond}_2(\mathbb{U}\mathbb{A}) = \mathrm{cond}_2(\mathbb{U}^*\mathbb{A}\mathbb{U}).$$

Rappel 1.4 *Les valeurs singulières d'une matrice rectangulaire* $\mathbb{A}$ *sont les racines carrées positives des valeurs propres de* $\mathbb{A}^*\mathbb{A}$.

1.7 Notion de préconditionnement

Lorsque l'on veut résoudre un système linéaire $Ax = b$ avec une matrice mal conditionnée, il peut être intéressant de de multiplier à gauche par une matrice C telle CA soit mieux conditionnée. L'exemple le plus simple est le *préconditionnement diagonal*, où la matrice C est la matrice diagonale constituée des inverses des éléments diagonaux de A.

Chapitre 2

Méthodes directes

Sommaire

2.1 Méthode de Gauss

2.1.1 Systèmes triangulaires

Considérons un système triangulaire (inférieur) du type :

$$\begin{cases} a_{11}x_1 & = & b_1 \\ a_{21}x_1 + a_{22}x_2 & = & b_2 \\ \ddots & & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = & b_n \end{cases}$$

c'est-à-dire associé a une matrice triangulaire inférieure

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

la résolution est très aisée : on commence par résoudre la première équation :

$$\text{Si } a_{11} \neq 0,\ x_1 = b_1/a_{11}$$

puis on reporte la valeur de x_1 ainsi determinée dans la deuxième équation et on calcule x_2, etc. A l'étape i on a :

$$\text{Si } a_{ii} \neq 0,\ x_i = (b_i - a_{i1}x_1 - a_{i2}x_2 - \cdots - a_{ii-1}x_{i-1})/a_{ii}.$$

ce qui nécessite 3 instructions pour l'implémentation. Regardons la complexité de l'algorithme, c'est-a-dire le nombre d'opérations élémentaires pour la résolution du système. Une opération élémentaire est une des 4 opérations addition($+$), soustraction($-$), multiplication($*$) et division ($/$). En général on les groupe en addition/soustraction et multiplication/division. On appelle N^+ et N^* les nombres d'opérations correspondants.

On peut établir le tableau suivant

ligne	N^+	N^*
1	0	1
$\vdots$	$\vdots$	$\vdots$
i	$i-1$	i
$\vdots$	$\vdots$	$\vdots$
n	$n-1$	n

D'où le nombre total d'opérations en sommant sur i :

$$N^+ = \sum_{i=1,n} (i-1) = \frac{n(n-1)}{2}$$

$$N^* = \sum_{i=1,n} (i) = \frac{n(n+1)}{2}$$

La démarche est la même pour un système triangulaire supérieur, i.e.

$$\left\{\begin{array}{lllll} a_{11}x_1 & + & a_{12}x_2 + \cdots & + & a_{1n}x_n = b_1 \\ & + & a_{22}x_2 \cdots & + & a_{2n}x_n = b_2 \\ & & & \ddots & \vdots \\ & & & & a_{nn}x_n = b_n \end{array}\right.$$

On le résout en remontant :

$$\text{Si } a_{nn} \neq 0,\ x_n = b_n/a_{nn}$$

puis on reporte la valeur de x_n ainsi déterminée dans la deuxième équation et on calcule $x_n - 1$, etc. A l'etape i on a :

$$\text{Si } a_{ii} \neq 0,\ x_i = (b_i - a_{ii+1}x_{i+1} - a_{ii+2}x_{i+2} - \cdots - a_{in}x_n)/a_{ii}.$$

Le nombre d'opérations est le même.

2.1.2 Décomposition LU : un résultat théorique

Le principe de la méthode est de se ramener à deux systèmes triangulaires.

1) On décompose la matrice A en le produit de deux matrices

$$A = LU$$

où U est triangulaire supérieure, et L est triangulaire inférieure avec des 1 sur la diagonale. On a alors à résoudre le système

$$LUx = b,$$

2) On résout le système triangulaire

$$Ly = b$$

d'inconnue y,

3) On résout le système triangulaire

$$Ux = y$$

d'inconnue x.

Reste maintenant à savoir comment faire cette décomposition LU.

Commençons par un résultat théorique

Théorème 2.1 *Soit A une matrice inversible d'ordre n dont les mineurs principaux sont non nuls. Alors il existe une unique matrice L triangulaire inférieure avec des 1 sur la diagonale, et une unique matrice U triangulaire supérieure telles que $A = LU$. De plus $det\,(A) = \prod_{i=1}^{n} u_{ii}$.*

Rappelons que le mineur principal d'ordre i de A est le déterminant des i premières lignes et premières colonnes :

$$(det\,A)_i = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1i} \\ a_{21} & a_{22} & \cdots & a_{2i} \\ \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{ii} \end{vmatrix}$$

La démonstration se fait par récurrence sur n.

Etape 1 : le résultat est évidemment vrai pour $n = 1$.

Etape 2 : on suppose le résultat vrai pour $n-1$. On décompose la matrice A par blocs sous la forme

$$A = \begin{pmatrix} A^{(n-1)} & c \\ b^T & a_{nn} \end{pmatrix}$$

où $A^{(n-1)}$ est la matrice $(n-1) \times (n-1)$ des $(n-1)$ premières lignes et colonnes de A, c et b sont des vecteurs colonnes donnés par

$$c = \begin{pmatrix} a_{1n} \\ \vdots \\ a_{n-1n} \end{pmatrix},\ b = \begin{pmatrix} a_{n1} \\ \vdots \\ a_{nn-1} \end{pmatrix}$$

La matrice $A^{(n-1)}$ a les mêmes mineurs principaux que A, on peut donc lui appliquer l'hypothese de récurrence :il existe deux matrices $L^{(n-1)}$ triangulaire inférieure avec des 1 sur la diagonale, et $U^{(n-1)}$ triangulaire supérieure telles que $A^{(n-1)} = L^{(n-1)}U^{(n-1)}$. Cherchons alors L et U décomposées par blocs sous la forme

$$L = \begin{pmatrix} L^{(n-1)} & 0 \\ {}^t l & 1 \end{pmatrix},\ U = \begin{pmatrix} U^{(n-1)} & u \\ 0 & u_{nn} \end{pmatrix},$$

En effectuant le produit par blocs et en identifiant a la décomposition de A, on obtient le systeme d'équations

$$\begin{aligned} A^{(n-1)} &= L^{(n-1)}U^{(n-1)} \\ {}^t b &= {}^t l U^{(n-1)} \\ c &= L^{(n-1)}u \\ a_{nn} &= {}^t l u + u_{nn} \end{aligned}$$

Ceci se résout immédiatement par

$$\begin{aligned} {}^t l &= {}^t b (U^{(n-1)})^{-1} \\ u &= (L^{(n-1)})^{-1} c \\ u_{nn} &= a_{nn} - {}^t b (A^{(n-1)})^{-1} c \end{aligned}$$

La question de l'unicité se règle de la facon suivante. Supposons qu'il existe 2 couples de matrices $(L_{(1)}, U_{(1)})$ et $(L_{(2)}, U_{(2)})$ tels que

$$A = L_{(1)} U_{(1)} = L_{(2)} U_{(2)}$$

Puisque toutes ces matrices sont inversibles, on en déduit que

$$U_{(1)} (U_{(2)})^{-1} = (L_{(1)})^{-1} L_{(2)}$$

Dans le membre de gauche on a une matrice triangulaire supérieure, dans le membre de droite on a une matrice triangulaire inférieure avec des 1 sur la diagonale. Pour qu'elles coincident, il faut qu'elles soient égales a l'identité.

2.1.3 Décomposition LU : méthode de Gauss

Pour construire les matrices L et U, on applique la méthode de Gauss qui consiste à trigonaliser le système pas a pas. Reprenons le système (1.2), et notons L_i la ième ligne du système. En supposant le premier **pivot** a_{11} non nul, soustrayons à chaque ligne L_i la première ligne L_1 divisée par a_{11} et multipliée par a_{i1} : cette opération annule le coefficient de x_1 dans les lignes 2 à n.

$$\begin{array}{lll} L_1 & a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n & = b_1 \\ L_2 - \frac{a_{21}}{a_{11}} & a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n & = b_2 \\ \vdots & \vdots & \vdots \\ L_i - \frac{a_{i1}}{a_{11}} & a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n & = b_i \\ \vdots & \vdots & \vdots \\ L_n - \frac{a_{n1}}{a_{11}} & a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n & = b_n \end{array}$$

On note $m_{i1} = \frac{a_{i1}}{a_{11}}$ pour $1 \leq i \leq n$. On obtient alors le nouveau système

$$\begin{array}{lllllll} a_{11}x_1 & + & a_{12}x_2 & + \cdots + & a_{1n}x_n & = & b_1 \\ & & (a_{22} - m_{21}a_{12})x_2 & + \cdots + & (a_{2n} - m_{21}a_{1n})x_n & = & b_2 - m_{21}b_1 \\ & & \vdots & & \vdots & & \\ & & (a_{i2} - m_{i1}a_{12})x_2 & + \cdots + & (a_{in} - m_{i1}a_{1n})x_n & = & b_i - m_{i1}b_1 \\ & & \vdots & & \vdots & & \\ & & (a_{n2} - m_{n1}a_{12})x_2 & + \cdots + & (a_{nn} - m_{n1}a_{1n})x_n & = & b_n - m_{n1}b_1 \end{array}$$

ou encore

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}-m_{21}a_{12} & \cdots & a_{2n}-m_{21}a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{i2}-m_{i1}a_{12} & \cdots & a_{in}-m_{i1}a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}-m_{n1}a_{12} & \cdots & a_{nn}-m_{n1}a_{1n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2-m_{21}b_1 \\ \vdots \\ b_i-m_{i1}b_1 \\ \vdots \\ b_n-m_{n1}b_1 \end{pmatrix}$$

La matrice du nouveau système est

$$A^{(2)} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}-m_{21}a_{12} & \cdots & a_{2n}-m_{21}a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{i2}-m_{i1}a_{12} & \cdots & a_{in}-m_{i1}a_{1n} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2}-m_{n1}a_{12} & \cdots & a_{nn}-m_{n1}a_{1n} \end{pmatrix}$$

et le second membre est

$$b^{(2)} = \begin{pmatrix} b_1 \\ b_2-m_{21}b_1 \\ \vdots \\ b_i-m_{i1}b_1 \\ \vdots \\ b_n-m_{n1}b_1 \end{pmatrix}$$

Le nouveau système s'ecrit alors

$$A^{(2)}x = b^{(2)}$$

et il est équivalent au système de départ.

On introduit la matrice

$$M^{(1)} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ -m_{21} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -m_{i1} & 0 & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -m_{n1} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Il est facile de voir que $A^{(2)} = M^{(1)}A$ et $b^{(2)} = M^{(1)}b$:**les manipulations sur les lignes reviennent à multiplier la matrice et le second membre du système par la matrice** $M^{(1)}$. La matrice $A^{(2)}$ contient maintenant uniquement des zeros sous la diagonale dans la première colonne. C'est ce processus que nous allons continuer : a l'étape k nous avons la matrice$A^{(k)}$ qui a la forme suivante

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & \cdots & & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & & & & a_{2n}^{(k)} \\ \vdots & \vdots & & & & \\ 0 & 0 & a_{k-1k-1}^{(k)} & a_{k-1k}^{(k)} & \cdots & a_{kn}^{(k-1)} \\ \vdots & \vdots & 0 & \vdots & & \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

et le système associé s'écrit

$$\begin{array}{ccccccccccc} a_{11}^{(k)}x_1 & + & a_{12}^{(k)}x_2 & + & \cdots & + & \cdots & + & a_{1n}^{(k)}x_n & = & b_1^{(k)} \\ & & a_{22}^{(k)}x_2 & + & \cdots & + & \cdots & + & a_{2n}^{(k)}x_n & = & b_2^{(k)} \\ & & & & a_{33}^{(k)}x_3 & + & \cdots & + & a_{3n}^{(k)}x_n & = & b_3^{(k)} \\ & & & & & & \cdots & & & & \vdots \\ & & & & & & a_{kk}^{(k)}x_k & + & a_{kn}^{(k)}x_n & = & b_k^{(k)} \\ & & & & & & & & & & \vdots \\ & & & & & & a_{nk}^{(k)}x_k & + & a_{nn}^{(k)}x_n & = & b_n^{(k)} \end{array}$$

soit sous forme compacte $A^{(k)}x = b^{(k)}$.

Il faut maintenant faire les manipulations sur les lignes adaptées. Supposons que le **k-ème pivot** $a_{kk}^{(k)}$ est non nul, et notons $L_i^{(k)}$ la i-ème ligne du système.

$$
\begin{array}{lccccccccccc}
L_1^{(k)} & a_{11}^{(k)}x_1 & + & \cdots & + & \cdots & + & \cdots & + & a_{1n}^{(k)}x_n & = & b_1^{(k)} \\
L_2^{(k)} & & & a_{22}^{(k)}x_2 & + & \cdots & + & \cdots & + & a_{2n}^{(k)}x_n & = & b_2^{(k)} \\
\vdots & & & & & \vdots & & \vdots & & & & \vdots \\
L_k^{(k)} & & & & & a_{kk}^{(k)}x_k & + & \cdots & + & a_{kn}^{(k)}x_n & = & b_k^{(k)} \\
L_{k+1}^{(k)} - \frac{a_{k+1k}^{(k)}}{a_{kk}^{(k)}} L_k^{(k)} & & & & & a_{k+1k}^{(k)}x_k & + & \cdots & + & a_{k+1n}^{(k)}x_n & = & b_{k+1}^{(k)} \\
L_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} L_k^{(k)} & & & & & \vdots & & \vdots & & & & \vdots \\
L_n^{(k)} - \frac{a_{nk}^{(k)}}{a_{kk}^{(k)}} L_k^{(k)} & & & & & a_{nk}^{(k)}x_k & + & \cdots & + & a_{nn}^{(k)}x_n & = & b_n^{(k)}
\end{array}
$$

Cette opération annule les coefficients de x_k dans les lignes $k+1$ à n. Nous avons fait un pas de plus vers la trigonalisation de la matrice A. Notons maintenant $m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ pour $k \leq i \leq n$ et introduisons la matrice

$$
M^{(k)} = \begin{pmatrix}
1 & 0 & \cdots & \cdots & & 0 \\
0 & 1 & & & & 0 \\
\vdots & 0 & & & & \\
0 & 0 & 1 & 1 & \cdots & 0 \\
\vdots & \vdots & 0 & -m_{k+1k} & 1 & \\
\vdots & \vdots & \vdots & \vdots & 0 & \vdots \\
0 & 0 & 0 & -m_{nk} & \vdots & 1
\end{pmatrix}
$$

Alors les manipulations sur les lignes reviennent à multiplier la matrice et le second membre du système par la matrice $M^{(k)}$. On obtient donc le système $A^{(k+1)}x = b^{(k+1)}$, avec $A^{(k+1)} = M^{(k)}A^{(k)}$ et $b^{(k+1)} = M^{(k)}b^{(k)}$, et l'on a gagné une nouvelle colonne de zéros. A l'étape n, on obtient une matrice $A^{(n)}$ qui est triangulaire supérieure, et le système

$$A^{(n)}x = b^{(n)}$$

avec $A^{(n)} = M^{(n)} \cdots M^{(1)}A$ et $b^{(n)} = M^{(n)} \cdots M^{(1)}b$.

Posons maintenant $U = A^{(n)}$ et $L = (M^{(n)} \cdots M^{(1)})^{-1}$, alors $A = LU$, U est triangulaire supérieure et il est facile de voir que L est triangulaire inférieure avec des 1 sur la diagonale. Ses coefficients sont les m_{ik}. Nous avons ainsi obtenu pratiquement la décomposition LU.

2.1.4 Méthode de Crout

Pour calculer explicitement les matrices L et U, on a intérêt à procéder par substitution : c'est la méthode de Crout. Ecrivons le produit LU :

$$LU = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ m_{21} & 1 & 0 & \cdots & 0 \\ m_{i1} & m_{i2} & 1 & 0 & \cdots \\ \vdots & \vdots & \vdots & 1 & \vdots \\ m_{n1} & m_{n2} & \vdots & m_{nn-1} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & & u_{1n} \\ 0 & u_{22} & & & & \\ \vdots & 0 & u_{33} & & & \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & 0 & 0 & u_{nn} \end{pmatrix}$$

Ecrivons l'égalité des coefficients ligne par ligne

- Ligne 1
  Pour $j = 1, \cdots, n$, $a_{1j} = u_{1j}$, ce qui permet de calculer

$$j = 1, \cdots, n, \quad u_{1j} = a_{1j}$$

- Ligne 2
  - Colonne 1 $a_{21} = l_{21}u_{11}$, et puisque u_{11} est maintenant connu, on en déduit

$$l_{21} = \frac{a_{21}}{u_{11}}$$

  - Colonne j, pour $j \geq 2$, $a_{2j} = l_{21}u_{1j} + u_{2j}$, et donc

$$j = 2, \cdots, n, \quad u_{2j} = a_{2j} - l_{21}u_{1j}$$

- Ligne i : supposons que nous avons été capable de calculer

$$\begin{matrix} u_{11} & u_{12} & \cdots & & u_{1n} \\ l_{21} & u_{22} & \cdots & & u_{2n} \\ \vdots & & \cdots & & \\ \vdots & & \cdots & & \\ l_{i-11} & \cdots & l_{i-1i-2} & u_{i-1i-1} & \cdots & u_{i-1n} \end{matrix}$$

  - Colonne 1 : $a_{i1} = l_{i1}u_{11}$, on en déduit l_{i1} :

$$l_{i1} = \frac{a_{i1}}{u_{11}}$$

  - Colonne $j < i$: $a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \cdots + l_{ij}u_{jj}$, d'où

$$j = 1, \cdots, j, \quad l_{ij} = \frac{a_{ij} - l_{i1}u_{1j} - \cdots l_{ij-1}u_{j-1j}}{u_{jj}}$$

– Colonne $j \geq i$:$a_{ij} = l_{i1}u_{1j} + l_{i2}u_{2j} + \cdots + l_{ii}u_{ij}$, d'où

$$j = i, \cdots, n, \quad u_{ij} = a_{ij} - l_{i1}u_{1j} - \cdots - l_{ii-1}u_{i-1j}$$

Remarquons qu'à la ligne i nous utilisons les valeurs de A a la ligne i et les valeurs de L et U calculées précédemment. D'un point de vue informatique, on mettra L et U à la place de A ligne par ligne.

Calculons le nombre d'opérations nécessaires à la décomposition LU.

2.1.5 Complexité de l'algorithme

Considérons l'algorithme de Crout. Avec les notations de la section 1.2.1, reprenons notre tableau. La ligne 1 nécessite 0 opérations. A la ligne i, notons N_i^+ et N_i^* le nombre d'opérations élémentaires :

colonne		
$j < i$	$j-1$	j
$\cdots$	$\cdots$	$\cdots$
$j \geq i$	$i-1$	$i-1$
$\cdots$	$\cdots$	$\cdots$
total	N_i^+	N_i^*

On a donc $N_i^* = \sum_{j=1}^{i-1}(j) + \sum_{j=1}^{i-1}(i-1) = \frac{i(i-1)}{2} + (i-1)(n-i+1)$ et $N^* = \sum_{i=1}^{n}(N_i^*) = \frac{n(n^2-1)}{3}$. On fait le même calcul pour N^+ et on a

$$N^* = \frac{n(n^2-1)}{3}, \quad N^+ = \frac{n(n-1)(2n-1)}{6}$$

Exercice 2.1 *Evaluer le coût de la décomposition LU par la méthode d'élimination de Gauss.*

Ce calcul est surtout important lorsque l'on résout des gros systèmes. On a en résumé pour la résolution d'un système linéaire par la méthode de Crout.

Décomposition LU : $\frac{2n^3}{3}$ opérations élémentaires, Résolution des 2 systèmes triangulaires : $2n^2$ opérations élémentaires.

Comparons avec l'utilisation des formules de Cramer : On écrit $x_j = \frac{D_j}{D_0}$ où chaque D représente un déterminant $n \times n$. Or le déterminant d'une matrice $n \times n$ est calculé par

$$det = \sum_{\sigma \text{ permutation de } \{1,\cdots,n\}} \varepsilon(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}$$

Pour chaque permutation, il y a $n-1$ multiplications, et il y a $n!$ permutations. On a donc $N* = (n-1)n!$, et $N \equiv n!$ pour chaque déterminant. Comme il y en a $n+1$ à calculer, on a $N \equiv n^2 n!$. D'après la formule de Stirling, $n! \equiv n^{n+1/2} e^{-n} \sqrt{(2\pi)}$.

Ex (*http ://clusters.top500.org*) le 28° calculateur CEA AlphaServer SC45, 1 GHz (2560 processeurs) est a 3980 GFlops, soit environ 410^{12}Flops. Pour $n = 100$, on a $N \approx 10^{162}$. il faudrait environ $2 \cdot 10^{149}$ années pour le résoudre. Rappelons que l'univers a 15 milliards d'années, *i.e* $15 \cdot 10^9$. Remarquons néanmoins que les formules de Cramer restent très compétitives pour $n = 3$!

Par la méthode LU, il faut environ $7\,10^6$ opérations, soit 1 millionnième de seconde. Pour un système à 10^6 inconnues, il faut 610^{17} opérations, soit 10^5 secondes, soit $\approx 25h$.

Rappelons la définition d'un FLOPS : floating point operations per second. Les nombres sont en général stockés en flottant, c'est-à-dire avec le nombre de chiffres significatifs, le signe, la base.

2.1.6 méthode du pivot partiel

Il peut se passer dans la pratique que l'un des pivots $a_{kk}^{(k)}$ soit nul. D'autre part, examinons le système ci-dessous :

$$\begin{array}{rcccl} 10^{-4} & x & + & y & = 1 \\ & x & + & y & = 2 \end{array}$$

et appliquons la méthode de Gauss avec comme pivot 10^{-4}. On obtient formellement

$$(1 - 1/10^{-4})y = 2 - 10^{-4}$$

Ceci, calculé en virgule flottante avec 3 chiffres significatifs, (cf fichier MAPLE joint) donne $y = 1$, puis $x = 0$, ce qui est notoirement faux.

Echangeons maintenant les deux lignes

$$\begin{array}{rcccl} & x & + & y & = 2 \\ 10^{-4} & x & + & y & = 1 \end{array}$$

et prenons comme pivot 1 . On obtient maintenant

$$(1 - 10^{-4})y = 1 - 2 \cdot 10^{-4}.$$

Ceci, calculé en virgule flottante avec 3 chiffres significatifs, donne $y = 1$, puis $x = 1$.

En fait la raison du problème est que le pivot 10^{-4} est trop petit .

```
> restart;
> Digits:=3;
> a1:=1-1/(10^(-4));
> a2:=2-1/(10^(-4));
> y:=evalf(a2/a1);
```

$$Digits := 3$$
$$a1 := -9999$$
$$a2 := -9998$$
$$y := 1.000$$

```
> x:=(1-y)/(10^(-4));
```

$$x := 0$$

```
>
> restart;
> Digits:=4;
> a1:=1-1/(10^(-4));
> a2:=2-1/(10^(-4));
> y:=evalf(a2/a1);
> x:=(1-y)/(10^(-4));
```

$$Digits := 4$$
$$a1 := -9999$$
$$a2 := -9998$$
$$y := .9999$$
$$x := 1.$$

```
>
```

Page 1

FIG. 2.1 – pivot

Explicitons maintenant la méthode. Pour cela reprenons la matrice $A^{(k)}$

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & \cdots & & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & & & & a_{2n}^{(k)} \\ \vdots & \vdots & a_{k-1k-1}^{(k)} & & & \\ 0 & 0 & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}$$

Si A est inversible, si $a_{kk}^{(k)} = 0$, il existe forcément un indice i supérieur à k tel que $a_{ik}^{(k)}$. En effet A est inversible si et seulement si $A^{(k)}$ l'est, et le déterminant de $A^{(k)}$ est égal à :

$$det\, A^{(k)} = a_{11}^{(k)} \cdots a_{k-1k-1}^{(k)} \begin{vmatrix} a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{vmatrix}$$

Donc si A est inversible, au moins un des éléments de la première colonne de cette dernière matrice est non nul.

Soit i_0 l'indice du nombre le plus grand en module :

$$|a_{i_0k}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|.$$

La **méthode du pivot partiel** consiste à échanger la ligne k et la ligne i_0 du système ; En fait cela revient à multiplier à gauche les deux membres du système matriciel par une **matrice de permutation** : la matrice correspondant à la transposition τ_k de $\{1, \ldots, n\}$ définie par

$$\begin{aligned} &\tau_k(k) = i_0, \\ &\tau_k(i_0) = k, \\ &\tau_k(i) = i \text{ si } i \neq k \text{ et } i \neq i_0. \end{aligned}$$

La matrice correspondante est définie par ses vecteurs colonnes

$$P_{\tau_k}(\boldsymbol{e}_j) = \boldsymbol{e}_{\tau_k(j)},$$

ou encore par ses éléments $(P_{\tau_k})_{ij} = \delta_{i\tau_k(j)}$.

On peut définir plus généralement la matrice de permutation associée à une permutation σ de $\{1, \ldots, n\}$ par

$$P_\sigma(\boldsymbol{e}_j) = \boldsymbol{e}_{\sigma(j)},$$

ou encore par ses éléments $(P_\sigma)_{ij} = \delta_{i\sigma(j)}$.

Ces matrices sont inversibles, leur déterminant est égal à la signature de la permutation, donc ± 1, et on a les résultats suivants :
Multiplier la matrice A à gauche par la matrice P_σ revient à effectuer la permutation σ^{-1} sur les lignes de A,
Multiplier la matrice A à droite par la matrice P_σ revient à effectuer la permutation σ sur les colonnes de A.
Soient σ et τ deux permutations, $P_\sigma P_\tau = P_{\sigma\circ\tau}$.

Donc à l'étape k, on multiplie la matrice $A^{(k)}$ par une matrice de permutation P_{τ_k}, puis on fait la $(k+1)$ ème étape de la réduction de Gauss sur la matrice $P_{\tau_k} A^{(k)}$. On obtient donc ainsi

$$U = M^{(n-1)} P_{\tau_{n-1}} \cdots M^{(1)} P_{\tau_1} A.$$

Théorème 2.2 *Soit A une matrice carrée régulière d'ordre n. Il existe une matrice de permutation P et deux matrices L et U, L étant triangulaire inférieure à diagonale unité et U étant triangulaire supérieure, telles que*

$$PA = LU$$

Démonstration Il suffit de remarquer que pour toute permutation σ de $1, \cdots, n$, pour toute matrice M, la matrice $\tilde{M} = P_\sigma^{-1} M P_\sigma$ est obtenue en effectuant la permuation σ sur les lignes et les colonnes de M. Si M est de type $M^{(k)}$ et σ de type τ_j avec $j \geq k$, alors $\tilde{M}$ a la même forme que M. On peut alors écrire

$$U = \tilde{M}^{(n-1)} \cdots \tilde{M}^{(1)} P_{\tau_{n-1}} \cdots P_{\tau_1} A.$$

Posons $\sigma = \tau_{n-1} \circ \cdots \circ$, alors

$$U = \tilde{M}^{(n-1)} \cdots \tilde{M}^{(1)} P_\sigma A,$$

et l'on conclut comme précédemment avec $L = (\tilde{M}^{(n-1)} \cdots \tilde{M}^{(1)})^{-1}$. ■

Remarque 2.1 *Pour calculer le déterminant d'une matrice, les formules de Cramer sont à prohiber. On utilise la décomposition LU et $D(A) = \prod u_{ii}$.*

Remarque 2.2 *On peut écrire la décomposition LU sous la forme LDV où V est à diagonale unité et D une matrice diagonale.*

2.2 Méthode de Cholewski

D"après la remarque 2.2, si A est une matrice symétrique, par l'unicité de la décomposition, on peut écrire $A = LD\,{}^tL$.

Théorème 2.3 *Soit A une matrice symétrique définie positive. Alors il existe une unique matrice L triangulaire inférieure à diagonale unité, et une unique matrice diagonale D à coefficients strictement positifs, telles que*

$$A = LD\,{}^tL$$

Démonstration On applique la décomposition LU, en vérifiant que si A est symétrique définie positive, les mineurs principaux sont non nuls. ■

Une factorisation de Cholewski de A est une factorisation sous la forme $A = B\,{}^tB$, où B est une matrice triangulaire inférieure.

Théorème 2.4 *Soit A une matrice symétrique définie positive. Alors il existe une unique décomposition de Cholewski de A sous la forme $A = B\,{}^tB$, où B est une matrice triangulaire inférieure à coefficients diagonaux strictement positifs.*

Démonstration D'après le théorème précédent, A s'écrit sous la forme $LD\,{}^tL$. Puisque D est diagonale à éléments strictement positifs, on peut définir la matrice racine carrée de D comme la matrice dont les éléments diagonaux sont $\sqrt{d_{ii}}$. On définit alors $B = L\sqrt{D}$. L'unicité se démontre comme pour la décomposition LU. ■

Chapitre 3

Méthodes itératives

Sommaire

3.1 Suite de vecteurs et de matrices

Définition 3.1 *Soit V un espace vectoriel muni d'une norme $\|.\|$, on dit qu'une suite (v_k) d'éléments de V* ***converge vers un élément*** *$v \in V$, si*

$$lim_{k\to\infty}\|v_k - v\| = 0$$

et on écrit

$$v = lim_{k\to\infty}v_k.$$

Remarque 3.1 *Sur un espace vectoriel de dimension finie, toutes les normes sont équivalentes. Donc v_k tend vers v si et seuelement si $\|v_k - v\|$ tend vers 0 pour une norme.*

Théorème 3.1 *1. Soit $\|\cdot\|$ une norme matricielle subordonnée, et $\mathbb{B}$ une matrice vérifiant*

$$\|\mathbb{B}\| < 1.$$

Alors la matrice $(\mathbb{I}+\mathbb{B})$ est inversible, et

$$\left\|(\mathbb{I}+\mathbb{B})^{-1}\right\| \leq \frac{1}{1-\|\mathbb{B}\|}.$$

2. Si une matrice de la forme $(\mathbb{I}+\mathbb{B})$ est singulière, alors nécessairement

$$\|\mathbb{B}\| \geq 1$$

pour toute norme matricielle, subordonnée ou non.

Théorème 3.2 *Soit $\mathbb{B}$ une matrice carrée. Les conditions suivantes sont équivalentes :*

1. $lim_{k\to\infty}\mathbb{B}^k = 0$,
2. $lim_{k\to\infty}\mathbb{B}^k v = 0$ *pour tout vecteur* v,
3. $\varrho(\mathbb{B}) < 1$,
4. $\|\mathbb{B}\| < 1$ *pour au moins une norme matricielle subordonnée* $\|.\|$.

La démonstration repose sur la série de Neumann $\sum B^n$.

Théorème 3.3 *Soit $\mathbb{B}$ une matrice carrée, et $\|.\|$ une norme matricielle quelconque. Alors*

$$lim_{k\to\infty}\|\mathbb{B}^k\|^{1/k} = \varrho(\mathbb{B}).$$

3.2 Méthode de Jacobi, Gauss-Seidel, S.O.R.

Soit $\mathbb{A} \in \mathcal{M}_n(\mathbb{K})$ une matrice régulière et $\boldsymbol{b} \in \mathbb{K}^n$. Il s'agit de résoudre le système $\mathbb{A}\boldsymbol{x} = \boldsymbol{b}$ par une méthode itérative, c'est-à-dire de créer une suite $\boldsymbol{x}^k$ qui converge vers $\boldsymbol{x}$. On note $\mathbb{D} = \text{diag}(\mathbb{A})$, $\mathbb{E}$ la matrice triangulaire inférieure vérifiant

$$\begin{cases} e_{ij} = 0, & i \leq j \\ e_{ij} = -a_{ij} & i > j \end{cases}$$

et $\mathbb{F}$ la matrice triangulaire supérieure vérifiant

$$\begin{cases} f_{ij} = 0, & i \geq j \\ f_{ij} = -a_{ij} & i > j \end{cases}$$

On a alors

$$\mathbb{A} = \begin{pmatrix} \ddots & & -\mathbb{F} \\ & \mathbb{D} & \\ -\mathbb{E} & & \ddots \end{pmatrix} = \mathbb{D} - \mathbb{E} - \mathbb{F}$$

Méthode de Jacobi

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} \right) \quad \forall i \in \{1, \ldots, n\}$$

Méthode de Gauss-Seidel

$$x_i^{(k+1)}) = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) \quad \forall i \in \{1, \ldots, n\}$$

Méthodes de relaxation

$$x_i^{(k+1)} = \omega \hat{x}_i^{(k+1)} + (1-\omega) x_i^{(k)}$$

où $\hat{x}_i^{(k+1)}$ est obtenu à partir de $x^{(k)}$ par l'une des deux méthodes précédentes.

Avec la méthode de Jacobi

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1, j\neq i}^{n} a_{ij} x_j^{(k)} \right) + (1-\omega) x_i^{(k)} \ \forall i \in \{1, \ldots, n\}.$$

Avec la méthode de Gauss-Seidel

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) + (1-\omega) x_i^{(k)} \ \ \forall i \in \{1, \ldots, n\}$$

Cette méthode de relaxation est appelée méthode S.O.R. (successive over relaxation) Toutes ces méthodes se mettent sous la forme

$$Mx^{k+1} = Nx^k + b$$

avec

Jacobi	$M = D$	$N = E + F$
Gauss-Seidel	$M = D - E$	$N = F$
SOR	$M = \frac{1}{\omega} D - E$	$N = \frac{1-\omega}{\omega} D + F$

Programmation d'une étape de l'algorithme de Jacobi :

```
Pour i=1:N
  S:=B(i)
     Pour j=1:I-1
     S=S-A(i,j)*X(j)
     Pour j=i+1:N
     S=S-A(i,j)*X(j)
  Y(i)=S/A(i,i)
Pour i=1:N
  X(i):=Y(i)
```

Test d'arrêt : on définit le résidu à l'étape k comme $r^{(k)} = b - Ax^{(k)}$. Le test s'écrit : tant que $\|r^{(k)}\| >$ eps, on continue.

Exercice 3.1 *Ecrire une étape de l'algorithme SOR.*

3.3 Résultats généraux de convergence

Soit donc l'algorithme

$$Mx^{k+1} = Nx^k + b \tag{3.1}$$

avec $M - N = A$. Si la suite converge, elle converge vers la solution x de $Ax = b$, et l'erreur $e^{(k)} = x^{(k)} - x$ est solution de $Me^{(k+1)} = Ne^{(k)}$. On note $B = M^{-1}N$. D'après le théorème 3.2, on a

Théorème 3.4 *La suite $x^{(k)}$ converge pour toute donnée initiale x^0 si et seulement si $\varrho(\mathbb{B}) < 1$, si et seulement si $\|\mathbb{B}\| < 1$ pour au moins une norme matricielle subordonnée $\|.\|$.*

Il est d'usage d'affecter les noms suivants aux matrices des méthodes précédentes

Jacobi	$J = D^{-1}(E + F)$
SOR	$\mathcal{L}_\omega = (\frac{1}{\omega}D - E)^{-1}(\frac{1-\omega}{\omega}D + F)$

Lemme 3.1 *Pour tout $\omega \neq 0$, on a $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$.*

On en déduit par le théorème 3.4,

Théorème 3.5 *Si la méthode de relaxation converge pour toute donnée initiale, on a*

$$0 < \omega < 2$$

On définit le taux de convergence asymptotique par

$$R(B) = -\ln \rho(B)$$

Théorème 3.6 *Le nombre d'itérations nécessaires pour réduire l'erreur d'un facteur ε est au moins égal à $K = \dfrac{-\ln \varepsilon}{R(B)}$.*

3.4 Cas des matrices hermitiennes

Théorème 3.7 *Soit A une matrice hermitienne définie positive, $A = M - N$, où M est inversible. Si $M + N^*$ (qui est toujours hermitienne), est définie positive, la méthode itérative converge pour toute donnée initiale.*

Corollaire 3.1 *Soit A une matrice hermitienne définie positive. Si $\omega \in]0, 2[$, la méthode de relaxation converge pour toute donnée initiale.*

3.5 Cas des matrices tridiagonales

Théorème 3.8 *Soit A une matrice tridiagonale. Alors $\rho(\mathcal{L}_1) = (\rho(J))^2$: les méthodes de Jacobi et Gauss-Seidel convergent ou divergent simultanément. Si elles convergent, la méthode de Gauss-Seidel est la plus rapide.*

Théorème 3.9 *Soit A une matrice tridiagonale telles que les valeurs propres de J soient réelles. Alors les méthodes de Jacobi et de relaxation convergent ou divergent simultanément pour $\omega \in]0, 2[$. Si elles convergent, la fonction $\omega \mapsto \rho(\mathcal{L}_\omega)$ a l'allure suivante : avec $\omega^* = \dfrac{2}{1 + \sqrt{1 - (\rho(J))^2}}$.*

Remarque 3.2 *On ne connaît pas précisément ce ω^* si on ne connaît pas $\rho(J)$. Dans ce cas, le graphe ci-dessus montre que qu'il vaut mieux choisir ω trop grand que trop petit.*

3.6 Matrices à diagonale dominante

Théorème 3.10 *Soit A une matrice à diagonale strictement dominante ou irréductible à diagonale dominante. Alors la méthode de Jacobi converge.*

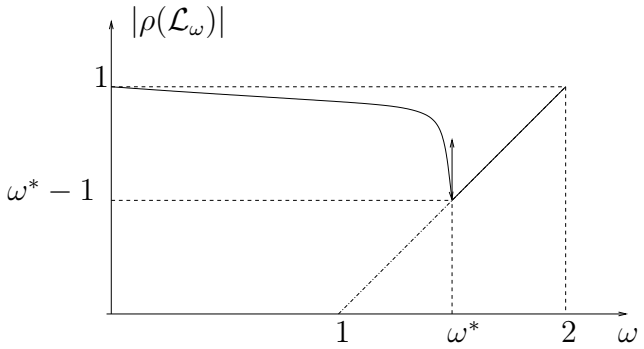

FIG. 3.1 – variations de $\rho(\mathcal{L}_\omega)$ en fonction de ω

Théorème 3.11 *Soit A une matrice à diagonale strictement dominante ou irréductible à diagonale dominante. Si $0 < \omega \leq 1$, la méthode de relaxation converge.*

3.7 La matrice du laplacien

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

On a

$$\rho(J) = 1 - \frac{\pi^2}{2n^2} + \mathcal{O}(n^{-4}), \rho(\mathcal{L}_1) = 1 - \frac{\pi^2}{n^2} + \mathcal{O}(n^{-4}),$$

$$\omega^* = 2(1 - \frac{\pi}{n} + \mathcal{O}(n^{-2})), \rho(\mathcal{L}_{\omega^*}) = \omega^* - 1 == 1 - \frac{2\pi}{n} + \mathcal{O}(n^{-2}).$$

Pour n=100, pour obtenir une erreur de $\varepsilon = 10^{-1}$, on doit faire

- 9342 itérations de l'algorithme de Jacobi,
- 4671 itérations de l'algorithme de Gauss-Seidel,
- 75 itérations de l'algorithme de l'algorithme de relaxation optimale.

3.8 Complexité

Supposons la matrice A pleine. La complexité d'une itération est d'environ $2n^2$. Si l'on fait au moins n itérations, on a donc une complexité totale

de $2n^3$, à comparer aux $2n^3/3$ de la méthode de Gauss.

Pour résoudre un système linéaire, on préferera les méthodes directes dans le cas des matrices pleines, et les méthodes itératives dans le cas des matrices creuses.

Chapitre 4

Calcul des valeurs propres et vecteurs propres

Sommaire

4.1 Généralités, outils matriciels

4.1.1 Matrices de Householder

Pour tout vecteur v de $\mathbb{C}^n - 0$, on introduit la matrice de Householder$H(v)$ définie par

$$H(v) = I - 2\frac{vv^*}{v^*v} \tag{4.1}$$

$H(v)$ est la matrice de la symétrie orthogonale par rapport à l'hyperplan de $\mathbb{C}^n$ orthogonal à v. La matrice $H(v)$ est hermitienne et unitaire. Par abus de langage, on considèrera l'identité comme une matrice de Householder, et l'on écrira $I = H(0)$.

Lemme 4.1 *Pour tout x dans $\mathbb{C}^n$, on a $(x - H(v)x)^*(x + H(v)x) = 0$.*

Lemme 4.2 *Soient x et y deux vecteurs linéairement indépendants. Si v est un vecteur de $\mathbb{C}^n - 0$, et ω un nombre complexe de module 1 tels que $\omega y = H(v)x$, alors il existe un nombre complexe λ tel que*

$$v = \lambda(x - \omega y) \text{ et } \bar{\omega} y^* x = \omega x^* y \tag{4.2}$$

On en déduit :

Proposition 4.1 *pour tout couple (x, y) dans $\mathbb{C}^n$ tel que $\|x\|_2 = \|y\|_2$, il existe une matrice de Householder $H(v)$ et un nombre complexe ω de module 1 tels que*

$$H(v)x = \omega y \tag{4.3}$$

D'après les lemmes on a $v = \lambda(x - \omega y)$ et $\omega = \pm e^{-i\theta}$ où θ est l'argument de $\psi^* x$. v étant défini à une constante multiplicative près, on peut le choisir de sorte que $\|v\|_2 = 1$. De plus le choix pratique du signe dans ω est gouverné par des considérations de conditionnement. On choisira ω tel que $\|x - \omega y\|_2$ est maximal.

4.1.2 Quotients de Rayleigh

Définition 4.1 *Soit A une matrice hermitienne de dimension n. Pour $x \neq 0$, on pose*

$$r_A(x) = \frac{x^* A x}{x^* x}$$

r_A est appelé le quotient de Rayleigh associé à A.

On ordonne les valeurs propres de A par ordre décroissant $\lambda_1 \geq \cdots \geq \lambda_n$.

Théorème 4.1 *On a*

$$\lambda_n = \inf_{x \neq 0} r_A(x), \quad \lambda_1 = \sup_{x \neq 0} r_A(x)$$

$$\lambda_k = \sup_{dim\, V = k} \inf_{x \in V - 0} r_A(x), \lambda_k = \inf_{dim\, W = n-k+1} \sup_{x \in W - 0} r_A(x), \ 1 \leq k \leq n$$

4.1.3 Conditionnement d'un problème de valeurs propres

Théorème 4.2 *Soient A et A' deux matrices hermitiennes, et $E = A' - A$. On note λ_i et λ'_i les valeurs propres de A et A', μ_i les valeurs propres de E, toutes ordonnées dans l'ordre décroissant. On a alors pour $1 \leq k \leq n$,*

$$\lambda_i + \mu_n \leq \lambda'_i \leq \lambda_i + \mu_1,$$
$$|\lambda'_i - \lambda_i| \leq \|E\| \text{ pour toute norme matricielle.}$$

4.2 Décompositions

4.2.1 Décomposition QR

Théorème 4.3 *Soit $A \in \mathcal{M}_n(\mathbb{C})$ (resp. $\mathcal{M}_n(\mathbb{R})$). Alors il existe une matrice Q unitaire (resp. orthogonale) et une matrice R triangulaire supérieure telles que $A = QR$. De plus on peut assurer que $R_{ii} \geq 0$. Si A est inversible, la décomposition avec $R_{ii} > 0$ est unique.*

Lien avec la décomposition de Gram-Schmidt Notons a^j les colonnes de A, q^j les colonnes de Q. Q est unitaire si et seulement si les q^j forment une base orthonormée, et

$$A = QR \iff \forall j, a^j = \sum_{1 \leq \ell \leq j} R_{\ell j} q^j$$

ce qui se réécrit

$$\begin{aligned}
a^1 &= R_{1,1} q^1 \\
&\vdots \\
a^j &= R_{j,j} q^j + R_{j-1,j} q^{j-1} + \cdots + R_{1,j} q^1 \\
&\vdots \\
a^n &= R_{n,n} q^n + R_{n-1,n} q^{n-1} + \cdots + R_{1,n} q^1
\end{aligned}$$

Si A est inversible, le système de ses vecteurs colonnes est un système libre, et on sait qu'on peut construire un système orthonormal par le procédé de Gram-Schmidt : supposons connus $q^1, \cdots, q^{j-1}$, et les coefficients $R_{k,i}$ pour $1 \leq i \leq j-1$ et $k \leq i$. On calcule alors à la ligne j les coefficients $R_{k,j}$ par

$$R_{j,j} q^j = a^j - R_{j-1,j} q^{j-1} - \cdots - R_{1,j} q^1$$

On écrit $(q^j, q^k) = 0$, ce qui donne $(a^j, q^k) - R_{k,j} = 0$ pour $k < j$, puis $(q^j, q^j) = 1$ ce qui donne $R_{j,j} = (a^j, q^j)$ ou encore

$$R_{j,j} = \|a^j\|_2^2 - \sum_{k<j} (a^j, q^k)^2$$

On peut compter le nombre d'opérations nécessité par ce procédé. On a $2n^3$ opérations élémentaires + n extractions de racines carrées. De plus ce procédé est peu stable numériquement. On préfere utiliser les matrices de Householder.

D'après la proposition 4.1, il existe une matrice de Householder $H(v^{(1)})$ et un nombre complexe ω_1 de module 1 tels que

$$H(v^{(1)})\, a^1 = \omega_1 \|a^1\| e_1 \tag{4.4}$$

On note $H^{(1)} = H(v^{(1)})$. La première colonne de la matrice $A^{(2)} = H^{(1)}A$ est donc de la forme ${}^t(r_{1,1}, 0, \cdots, 0)$. Par récurrence, on construit une suite de matrices $A^{(k)}$ de la forme

$$A^{(k)} = \begin{pmatrix} r_{1,1} & \cdots & r_{1,k-1} & a^{(k)}_{1,k} & \cdots & a^{(k)}_{1,n} \\ & 0_L & \ddots & \vdots & & \vdots \\ & & r_{k-1,k-1} & a^{(k)}_{k-1,k} & & a^{(k)}_{k-1,n} \\ & & & a^{(k)}_{k,k} & \cdots & a^{(k)}_{k,n} \\ & & 0 & \vdots & & \vdots \\ & & & a^{(k)}_{n,k} & & a^{(k)}_{n,n} \end{pmatrix}$$

et une suite de matrices de Householder$H^{(k)} = H(v^{(k)})$, telles que

$$A^{(k+1)} = H^{(k)}A^{(k)}, k = 1, \cdots, n. \tag{4.5}$$

On cherche $v^{(k)}$ sous la forme ${}^t v^{(k)} = (0, {}^t\tilde{v}^{(k)})$, et l'on vérifie que

$$H^{(k)} = \begin{pmatrix} I_k & 0 \\ 0 & \tilde{H}^{(k)} \end{pmatrix}, \text{ avec } \tilde{H}^{(k)} = I_{n-k+1} - 2\tilde{v}^{(k)}\ {}^t\tilde{v}^{(k)}$$

On a donc

$$A^{(n)} = H^{(n-1)} \cdots H^{(1)} A$$

et la matrice $A^{(n)}$ est triangulaire supérieure. Si l'on pose ${}^tQ = H^{(n-1)} \cdots H^{(1)}$, Q est une matrice orthogonale et $A = {}^tQ^{-1}A^{(n)} = QA^{(n)}$. On a ainsi construit les matrices Q et R.

Remarque 4.1

1. *Si A est réelle, les $H^{(k)}$ sont réelles, avec $\omega_k = \pm 1$, Q et R sont réelles, et Q est orthogonale.*
2. *Le nombre d'opérations nécessaires pour calculer Q et R est de l'ordre de $\frac{4}{3}n^3 + n$ racines carrées. De plus cette méthode est beaucoup plus stable que le procédé de Gram-Schmidt.*
3. *Par contre ce n'est pas une méthode compétitive pour résoudre un système linéaire.*

4.2.2 Tridiagonalisation d'une matrice symétrique

Théorème 4.4 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique. Alors il existe une matrice P orthogonale telle que ${}^tPAP$ soit tridiagonale.*

La démonstration est du même type que pour la méthode QR. Ici on multiplie à droite et à gauche par une matrice de Householder.

4.3 Algorithmes pour le calcul de toutes les valeurs propres d'une matrice

4.3.1 Méthode de Jacobi

Soit A une matrice réelle symétrique. On sait qu'il existe une matrice orthogonale O telle que ${}^tOAO$ soit diagonale.

$$ {}^tOAO = \Lambda = \begin{pmatrix} \lambda_1 & & & \\ & \ddots & 0 & \\ 0 & & \ddots & \\ & & & \lambda_n \end{pmatrix} $$

La méthode de Jacobi consiste à construire une suite de matrices de rotation $O^{(k)}$ telles que la suite définie par $A^{(k+1)} = {}^tO^{(k)}A^{(k)}O^{(k)}$ converge vers Λ. Soit R_θ^{pq} la matrice de rotation dans le plan (e_p, e_q) d'angle θ :

$$ R_{p,q}(\theta) = \begin{array}{c} \\ \\ \\ \mathbf{p} \\ \\ \\ \\ \mathbf{q} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} & & & & & & \mathbf{p} & & & & \mathbf{q} & & \end{array} \\ \begin{bmatrix} 1 & & & & 0 & 0 & \cdots & \cdots & 0 & & & \\ & \ddots & & & \vdots & & & & \vdots & & & \\ & & 1 & & 0 & & & & 0 & & & \\ 0 & \cdots & 0 & & \cos(\theta) & 0 & \cdots & 0 & \sin(\theta) & 0 & \ldots & 0 \\ & & & & 0 & 1 & & & 0 & & & \\ & & & & \vdots & & \ddots & & \vdots & & & \\ & & & & 0 & & & 1 & 0 & & & \\ & & & & -\sin(\theta) & 0 & \cdots & 0 & \cos(\theta) & & & \\ & & & & & & & & & 1 & & \\ & & & & & & & & & & \ddots & \\ & & & & & & & & & & & 1 \end{bmatrix} \end{array} $$

On pose $B = {}^tR_{p,q}(\theta)AR_{p,q}(\theta)$. Les $b_{i,j}$ sont alors donnés par

$$\begin{array}{lcll} b_{i,j} & = & a_{i,j}, & i \neq p,q, j \neq p,q \\ b_{p,j} & = & a_{p,j}\cos(\theta) - a_{q,j}\sin(\theta), & j \neq p,q \\ b_{q,j} & = & a_{p,j}\sin(\theta) + a_{q,j}\cos(\theta), & j \neq p,q \\ b_{p,p} & = & a_{p,p}\cos^2(\theta) + a_{q,q}\sin^2(\theta) - a_{p,q}\sin(2\theta) & \\ b_{q,q} & = & a_{p,p}\sin^2(\theta) + a_{q,q}\cos^2(\theta) + a_{p,q}\sin(2\theta) & \\ b_{p,q} & = & a_{p,q}\cos(2\theta) + \dfrac{a_{p,p} - a_{q,q}}{2}\sin(2\theta) & \\ b_{p,q} & = & b_{q,p}. & \end{array}$$

Théorème 4.5 *Si $a_{p,q} \neq 0$, il existe un unique θ dans $]-\frac{\pi}{4}, 0[\cup]0, \frac{\pi}{4}[$ tel que $b_{p,q} = 0$. C'est l'unique racine de l'équation*

$$cotg\, 2\theta = \frac{a_{q,q} - a_{p,p}}{2a_{p,q}}$$

Etape 1. On choisit p_1 et q_1 tels que $|a_{p_1,q_1}| = \max_{i\neq j}|a_{i,j}|$.On choisit θ_1 tel que $A^{(1)} = {}^tR_{p_1,q_1}(\theta_1)AR_{p_1,q_1}(\theta_1)$ vérifie $a^{(1)}_{p_1,q_1} = 0$.

Etape 2. On choisit p_2 et q_2 tels que $|a^{(1)}_{p_2,q_2}| = \max_{i\neq j}|a^{(1)}_{i,j}|$.On choisit θ_2 tel que $A^{(2)} = {}^tR_{p_2,q_2}(\theta_2)A^{(1)}R_{p_2,q_2}(\theta_2)$ vérifie $a^{(2)}_{p_2,q_2} = 0$. Puisque $p_2 \neq p_1, q_1$, $q_2 \neq p_1, q_1$, on a aussi $a^{(2)}_{p_1,q_1} = 0$.

Etape k. On choisit p_k et q_k tels que $|a^{(k-1)}_{p_k,q_k}| = \max_{i\neq j}|a^{(k-1)}_{i,j}|$.On choisit θ_k tel que $A^{(k)} = {}^tR_{p_k,q_k}(\theta_k)A^{(k-1)}R_{p_k,q_k}(\theta_k)$ vérifie $a^{(k)}_{p_k,q_k} = 0$. On a $a^{(k)}_{p_1,q_1} = \cdots = a^{(k)}_{p_k,q_k} = 0$.

On vide ainsi la matrice de tous ses éléments extradiagonaux.

Théorème 4.6 *Chaque élément diagonal $a^{(k)}_{i,i}$ converge vers une valeur propre de A quand k tend vers $+\infty$.*

On a à l'étape k, $A^{(k)} = {}^tR_{p_k,q_k}(\theta_k)\cdots{}^tR_{p_1,q_1}(\theta_1)AR_{p_1,q_1}\cdots R_{p_k,q_k}(\theta_k) = {}^tO^{(k)}AO^{(k)}$, où $O^{(k)}$ est une matrice orthogonale. Lorsque k tend vers l'infini, $O^{(k)}$ tend donc vers la matrice des vecteurs propres de A. Pour calculer les vecteurs propres de A, il suffit donc de calculer les matrices $O^{(k)}$, ce qui et néanmoins assez coûteux.

4.3.2 Méthode de Givens ou bisection

Soit A une matrice symétrique réelle. La méthode de bisection permet de calculer toutes les valeurs propres de A. Le principe est le suivant.

Etape 1. On se ramène à une matrice symétrique tridiagonale réelle par la méthode de Householder.La matrice

$$B = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \cdots & 0 & b_n & a_n \end{pmatrix}$$

a les mêmes valeurs propres que A.

Etape 2. On calcule les valeurs propres de B.

L'étape 1 a déjà été décrite, passons à l'étape 2. On suppose d'abord tous les c_i non nuls, sinon on décompose B par blocs qui ont les mêmes valeurs propres. On note p_i le polynôme caractéristique de la matrice A_i définie pour $i \geq 1$ par

$$A_1 = (a_1),\ A_2 = \begin{pmatrix} a_1 & b_2 \\ b_2 & a_2 \end{pmatrix}, \cdots, A_i = \begin{pmatrix} a_1 & b_2 & 0 & \cdots & 0 \\ b_2 & a_2 & b_3 & \ddots & \vdots \\ 0 & b_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_i \\ 0 & \cdots & 0 & b_i & a_i \end{pmatrix}, \cdots$$

On posera par convention $p_0 \equiv 1$. On a la relation de récurrence

$$p_i(\lambda) = (a_i - \lambda)p_{i-1}(\lambda) - b_i^2 p_{i-2}(\lambda)$$

Lemme 4.3 *Les polynômes p_i ont les propriétés suivantes :*

1. $\lim_{\lambda \to -\infty} = +\infty$, $1 \leq i \leq n$.
2. $p_i(\lambda_0) = 0 \Rightarrow p_{i-1}(\lambda_0)p_{i+1}(\lambda_0) < 0$, $1 \leq i \leq n-1$.
3. *Le polynôme p_i possède i racines réelles distinctes, qui séparent les (i+1) racines du polynôme p_{i+1}, $1 \leq i \leq n-1$.*

Théorème 4.7 *Soit $\omega(\lambda)$ le nombre de changements de signe de l'ensemble $\{p_0(\lambda), \cdots, p_n(\lambda)\}$. Alors p_n possède $\omega(b) - \omega(a)$ racines dans l'intervalle $[a, b[$.*

La méthode consiste alors en deux étapes.

Etape 1. On cherche un intervalle $[a, b]$ qui contient toutes les valeurs propres (par exemple l'union des disques de Gerschgörin $D(a_k, |b_k| + |b_{k+1}|)$. On a alors $\omega(a) = 0$, $\omega(b) = n$.

Etape 2. On applique une méthode de dichotomie. On calcule $\omega(\frac{a+b}{2})$, ce qui détermine le nombre de racines dans les intervalles $[a, \frac{a+b}{2}[$ et $[b, \frac{a+b}{2}, [$. On itère.

4.4 Méthode de la puissance itérée

Elle permet le calcul de la valeur propre de plus grand module et d'un vecteur propre associé.

On choisit $q^{(0)} \in \mathbb{C}^n$ tel que $\|q^{(0)}\| = 1$.

Pour $k = 1, 2, \ldots$ on calcule :

$$\begin{cases} x^{(k)} &= Aq^{(k-1)} \\ \lambda_j^{(k)} &= \dfrac{x_j^{(k)}}{q_j^{(k-1)}} \quad j = 1, \ldots, n \\ q^{(k)} &= \frac{x^{(k)}}{\|x^{(k)}\|} \end{cases}$$

On fera l'hypothèse suivante :

(H) la valeur propre de plus grand module est unique.

On suppose que A est diagonalisable, et on note V l'espace propre associé à λ_1.

Théorème 4.8 *On suppose que A est diagonalisable et que l'hypothèse (H) est vérifiée. On suppose de plus que q_0 n'est pas orthogonal à V. Alors on a*

1. $\lim_{k\to\infty} \|Aq_k\|_2 = |\lambda_1|$,
2. $\lim_{k\to\infty} \lambda_j^{(k)} = \lambda_1 \quad 1 \leq j \leq n$, *si* $q_j^{(k)} \neq 0$,
3. $\lim_{k\to\infty} \left(\dfrac{|\ \lambda_1\ |}{\lambda_1}\right)^k q^{(k)}$ *est un vecteur propre associé à* λ_1.

On remarque que q_k est également défini par $q_k = \dfrac{A^k q_0}{\|A^k q_0\|_2}$.

La méthode de la puissance inverse permet de calculer la plus petite valeur propre en module de A en appliquant la méthode de la puissance à A^{-1}.