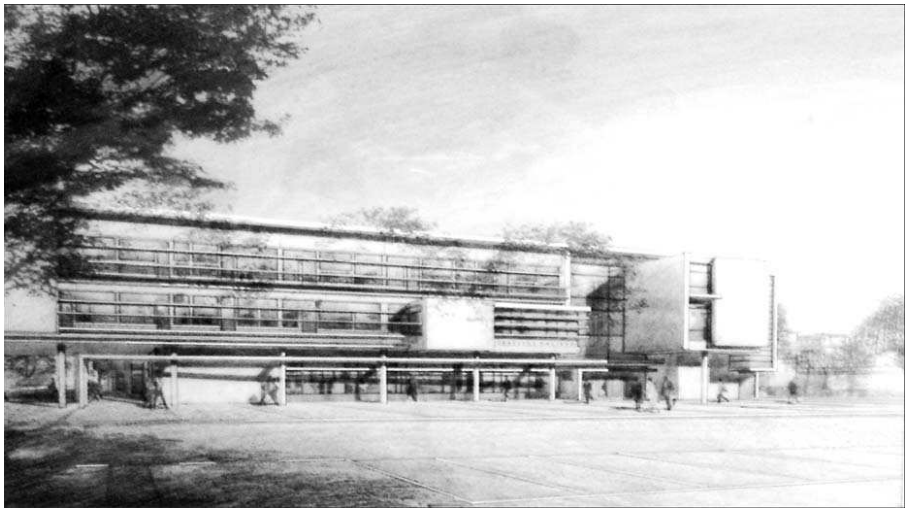
SuP Galilée
L'école d'ingénieurs de
l'Institut Galilée



Spécialité MACS 1ère année

ANALYSE NUMERIQUE I

L. Halpern

© INSTITUT GALILEE, 99 avenue Jean-Baptiste-Clément 93430 VILLETANEUSE 2007/2008

Liste des chapitres

Table des matières

Chapitre 1

Introduction

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 :

$$x'' + \frac{g}{l} sinx = 0. \tag{1.1}$$

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

$$x'' + \frac{g}{l} x = 0 \tag{1.2}$$

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 (1.1). 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

$$x''(t_n) = \frac{x(t_{n+1}) - 2x(t_n) + x(t_{n-1})}{\Delta t^2} + \mathcal{O}(\Delta t^2) \tag{1.3}$$

et on remplace l'équation par

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta t^2} + \frac{g}{l} sin(y_n) = 0. \tag{1.4}$$

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 (1.4) est une *approximation* de l'équation (1.1). Il est souhaitable que

1. (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 (1.4) 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 (1.2) s'écrit

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\Delta t^2} + \frac{g}{l} y_n = 0. \tag{1.5}$$

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.

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`

Chapitre 2

Arithmétique en précision finie

Voir http ://www.math.univ-paris13.fr/ japhet/Doc/Handouts/RoundOffErrors.pdf

Chapitre 3

Généralités

3.1 Rappels sur les matrices

3.1.1 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 3.1 *L'application linéaire $\mathcal{A}$ est injective si*

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

Définition 3.2 *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 3.3 *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}.

3.1.2 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.

3.1.3 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.

3.1.4 Produit de matrices par blocs

On décompose la matrice carrée A de la façon suivante :

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

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

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

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

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

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

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

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

$$AB = \left(\begin{array}{cc} A^{11}B^{11} + A^{12}B^{21} & A^{11}B^{12} + A^{12}B^{22} \\ A^{21}B^{11} + A^{22}B^{21} & A^{21}B^{12} + A^{22}B^{22} \end{array}\right),$$

3.2 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 assocé 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 3.1 *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 3.1 *Une matrice dont toutes les valeurs propres sont simples est diagonalisable.*

Théorème 3.2 *Une matrice symétrique est diagonalisable en base orthonormée.*

Théorème 3.3 (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.*

3.3 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 :

$$\begin{aligned} &a = bq_0 + r_0, r_0 = 0 \text{ ou } d^\circ r_0 < d^\circ a, \\ &b = r_0 q_1 + r_1, r_1 = 0 \text{ ou } d^\circ r_1 < d^\circ r_0 \\ &r_0 = r_1 q_2 + r_2, r_2 = 0 \text{ ou } d^\circ r_2 < d^\circ r_1 \end{aligned}$$

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

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

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].

3.4 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

$$\left\{\begin{array}{lcl} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1m}x_m & = & b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2m}x_m & = & b_2 \\ \vdots & \cdots & \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nm}x_m & = & b_n \end{array}\right. \tag{3.1}$$

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 (3.1) sous la forme

$$x_1 \begin{pmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{pmatrix} + x_2 \begin{pmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{pmatrix} + \cdots + x_m \begin{pmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{3.2}$$

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 (3.1) 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 3.4 *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 (3.1) 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

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

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 $\mathcal{L}(a^{(1)}, \cdots, a^{(m)}) \subsetneq \mathbb{R}^n$. 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

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

$(l^{(1)}, \cdots, l^{(r)})$ 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 3.5 *Suppososns que les r premières lignes $(l^{(1)}, \cdots, l^{(r)})$ sont indépendantes, et que les autres lignes vérifient*

$$l^{(j)} = \lambda_{j1} l^{(1)} + \cdots \lambda_{jr} l^{(r)}, r+1 \leq j \leq n \tag{3.3}$$

Alors toute solution du système principal (système des r premières équations) est solution de (3.1) si et seulement si

$$b_j = \lambda_{j1} b_1 + \cdots \lambda_{jr} b_r, r+1 \leq j \leq n \tag{3.4}$$

(3.4) 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.

3.5 Norme de vecteurs et de matrices

Définition 3.4 *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 3.6 *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 3.1 *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 3.5 *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 3.2 *Sur un espace vectoriel de dimension finie toutes les normes sont équivalentes.*

Définition 3.6 *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 3.3 *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 3.7 *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 :

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

Remarque 3.1 *1. Si une matrice $\mathbb{A}$ est hermitienne, ou symétrique (donc normale), on a $\|\mathbb{A}\|_2 = \rho(\mathbb{A})$.*

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

Théorème 3.8 *1. Soit $\mathbb{A}$ une matrice carrée quelconque et $\|\cdot\|$ une norme matricielle subordonnée ou non, quelconque. Alors*

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

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 3.9 *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 $\|\mathbb{I}\|_E = \sqrt{n}$.

3.6 Conditionnement

3.6.1 Erreur d'arrondi

Un nombre réel s'écrit de façon unique $x = \pm a 10^b$, 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 3.7 *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 $|\varepsilon| < eps$.

3.6.2 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 3.8 *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 3.1 *: 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* $\mathcal{K} \approx 2$*.*

3.6.3 Conditionnement d'une matrice

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

$$A\boldsymbol{x} = \boldsymbol{b},$$
$$(A + \Delta A)\,\boldsymbol{y} = (\boldsymbol{b} + \Delta\boldsymbol{b})\,.$$

Exemple de R.S. Wilson :

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

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

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

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

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

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

Définition 3.9 *Soit $\|.\|$ une norme matricielle subordonnée, le conditionnement d'une matrice régulière A, associé à cette norme, est le nombre*

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

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

Théorème 3.10 *Soit A une matrice inversible. Soient $\boldsymbol{x}$ et $\boldsymbol{x}+\boldsymbol{\Delta x}$ les solutions respectives de*

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

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

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

est satisfaite, et c'est la meilleure possible : pour une matrice 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

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

d'où

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

■

Théorème 3.11 *Soit A une matrice inversible. Soient $\boldsymbol{x}$ et $\boldsymbol{x}+\boldsymbol{\Delta x}$ les solutions respectives de*

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

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

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

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

Théorème 3.12 *1. Pour toute une matrice inversible A,*

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

2. Pour toute matrice inversible A,

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

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

3. Si A est une matrice normale,

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

où les $\lambda_i(A)$ sont les valeurs propres de A.

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

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

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

Rappel 3.4 *Les valeurs singulières d'une matrice rectangulaire A sont les racines carrées positives des valeurs propres de A^*A.*

3.7 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 4

Résolution numérique de systèmes linéaires par méthodes directes

4.1 Méthode de Gauss

4.1.1 Systèmes triangulaires

Considérons un système triangulaire (inférieur) du type :

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

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.

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

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.

4.1.2 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 4.1 *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

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

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.

4.1.3 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 (3.2), 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

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

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.

4.1.4 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

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

– 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.

4.1.5 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 4.1 *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.

4.1.6 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 .

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 :

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

```
> 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

FIGURE 4.1 – pivot

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 :

$$|a_{i_0 k}^{(k)}| = \max_{k \leq i \leq n} |a_{ik}^{(k)}|.$$

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

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

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 4.2 *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 4.1 *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 4.2 *On peut écrire la décomposition LU sous la forme LDV où V est à diagonale unité et D une matrice diagonale.*

4.2 Méthode de Cholewski

D"après la remarque 4.2, si A est une matrice symétrique, par l'unicité de la décomposition, on peut écrire $A = LD\,{}^tL$.

Théorème 4.3 *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 4.4 *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 5

Résolution numérique de systèmes linéaires par méthode itérative

5.1 Introduction

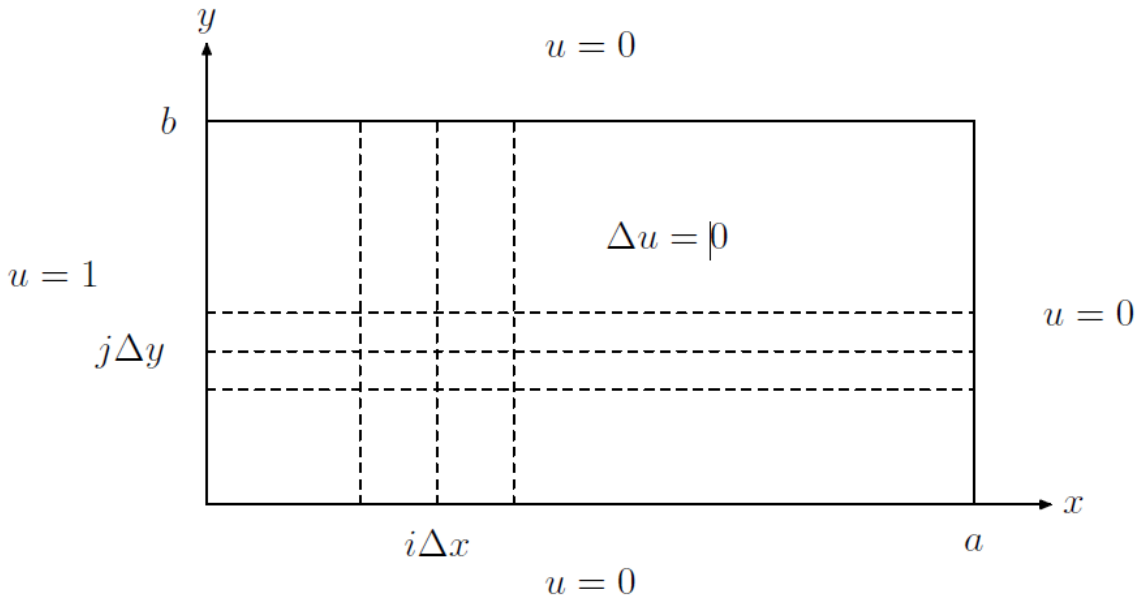
Exemple de base en analyse numérique : équation de la chaleur 2D (et même 3D). On cherche $u(x, y, t)$, solution de

$$\rho \partial_t u - \Delta u = f \text{ dans } \Omega; \quad \Delta u = \partial_{xx} u + \partial_{yy} u + \cdots$$

avec des conditions initiales et aux limites sur $\partial\Omega$.

Equilibre : $u(x, y)$

$$-\Delta u = f \text{ dans } \Omega.$$



On choisit des pas d'espace

$$\Delta x = \frac{a}{m+1}, \quad \Delta y = \frac{b}{m+1}$$

L'opérateur de Laplace en deux dimensions est discrétisé avec des *différences finies*

$$\partial_{xx} u \sim \frac{u(x+\Delta x, y) - 2u(x,y) + u(x-\Delta x, y)}{\Delta x^2};$$
$$\partial_{yy} u \sim \frac{u(x, y+\Delta y) - 2u(x,y) + u(x, y-\Delta y)}{\Delta y^2};$$

On obtient pour $u_{i,j} \sim u(i\Delta x, j\Delta y)$,

$$\frac{1}{\Delta x^2}\left(-u_{i-1,j} + 2u_{i,j} - u_{i+1,j}\right) \quad + \frac{1}{\Delta y^2}\left(-u_{i,j-1} + 2, u_{i,j} - u_{i,j+1}\right) = f(x_i, y_j)$$

On introduit les vecteurs

$$\boldsymbol{u}_j = \begin{pmatrix} u_{1,j} \\ \vdots \\ u_{m,j} \end{pmatrix}$$

pour $j = 1, \cdots, n$. On multiplie l'équation par Δy^2, et on obtient

$$-\boldsymbol{u}_{j-1} + T\boldsymbol{u}_j - \boldsymbol{u}_{j+1} = \delta u_{0,j}\boldsymbol{e}_1 + \boldsymbol{f}_j \Delta y^2$$

$$T = \begin{pmatrix} 2(1+\delta) & -\delta & 0 & & \\ -\delta & 2(1+\delta) & -\delta & & \\ & \ddots & \ddots & \ddots & \\ & & -\delta & 2(1+\delta) & -\delta \\ & & 0 & -\delta & 2(1+\delta) \end{pmatrix} \quad \text{matrice } m \times m$$

On introduit maintenant le vecteur de toutes les inconnues

$$\boldsymbol{U} = \begin{pmatrix} \boldsymbol{u}_1 \\ \vdots \\ \boldsymbol{u}_n \end{pmatrix}$$

On obtient un large système de $m \times n$ équations à $m \times n$ inconnues,

$$A\boldsymbol{U} = \boldsymbol{F}$$

avec

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

La matrice A est CREUSE, TRIDIAGONALE PAR BLOCS. Dans chaque bloc-ligne de A, seuls 3 blocs sur les n sont non-identiquement nuls, avec dans chaque ligne de ces blocs au plus 3 éléments non-nuls. Ca fait dans chaque ligne de A 5 élément non nuls sur $m \times n$ (faire $m = 100, n = 50$). La décomposition de Gauss risque de remplir la matrice. Les méthodes itératives n'ont pas besoin de construire la matrice, mais plutôt de savoir faire le produit matrice-vecteur.

Pour construire une méthode itérative on écrit

$$A = M - N; A\boldsymbol{u} = b \iff M\boldsymbol{u} - N\boldsymbol{u} = b$$

$$\boldsymbol{u} = M^{-1}(N\boldsymbol{u} + b)$$

On utilise le point fixe

$$M\boldsymbol{u}_{k+1} = N\boldsymbol{u}_k + b$$

Compromis entre

- M est une bonne approximation de A (le meilleur M est A)
- le système à résoudre à chaque itération est simple et peu onéreux.

5.2 Norme de vecteurs et de matrices

Définition 5.1 *Une* ***norme*** *sur un espace vectoriel V est une application $\|\cdot\| : V \to \mathbb{R}^+$ qui vérifie les propriétés suivantes*

- $\|\boldsymbol{v}\| = 0 \Longleftrightarrow \boldsymbol{v} = 0$,
- $\|\alpha\boldsymbol{v}\| = |\alpha|\,\|\boldsymbol{v}\|\ \forall\alpha \in \mathbb{K}, \forall\boldsymbol{v} \in V$,
- $\|\boldsymbol{u}+\boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|\ \ \forall\,(\boldsymbol{u},\boldsymbol{v}) \in V^2$ *(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 $\mathbb{C}^n$:

$$\begin{aligned}\|\boldsymbol{v}\|_1 &= \sum_{i=1}^{n} |v_i| \\ \|\boldsymbol{v}\|_2 &= \left(\sum_{i=1}^{n} |v_i|^2\right)^{1/2} \\ \|\boldsymbol{v}\|_\infty &= \max_{1\leq i\leq n} |v_i|.\end{aligned}$$

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

Théorème 5.1 *Soit V un espace de dimension finie. Pour tout nombre réel $p \geq 1$, l'application $v \mapsto \|v\|_p$ définie par*

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

est une norme.

Rappel 5.1 *Pour $p > 1$ et $\frac{1}{p} + \frac{1}{q} = 1$, l'inégalité*

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

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

Définition 5.2 *Deux* ***normes*** *$\|\cdot\|$ et $\|\cdot\|'$, définies sur un même espace vectoriel V, sont* ***équivalentes*** *s'il existe deux constantes C et C' telles que*

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

Rappel 5.2 *Sur un espace vectoriel de dimension finie toutes les normes sont équivalentes.*

Définition 5.3 *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. $\|A\| = 0 \Longleftrightarrow A = 0$,
2. $\|\alpha A\| = |\alpha|\,\|A\|, \forall \alpha \in \mathbb{K},\ \forall A \in \mathcal{M}_n$,
3. $\|A+B\| \leq \|A\| + \|B\|,\ \forall\,(A,B) \in \mathcal{M}_n^2$ *(inégalité triangulaire)*
4. $\|AB\| \leq \|A\|\|B\|,\ \forall\,(A,B) \in \mathcal{M}_n^2$

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

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

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

De plus

$$\|A\boldsymbol{v}\| \leq \|A\|\|\boldsymbol{v}\| \ \ \forall \boldsymbol{v} \in \mathbb{K}^n$$

et la norme $\|A\|$ peut se définir aussi par

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

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

$$\boldsymbol{u} \neq 0 \ \text{ et } \ \|A\boldsymbol{u}\| = \|A\|\|\boldsymbol{u}\|.$$

Enfin une norme subordonnée vérifie toujours

$$\|I\| = 1$$

Rappelons la définition du rayon spectral d'une matrice. Notons $\lambda_i(A)$ les n valeurs propres de la matrice $n \times n$ A. Le rayon spectral de A est

$$\rho(A) = \max_i |\lambda_i(A)|$$

Définition 5.4 *Soit $A = (a_{ij})$ une matrice carrée. Alors*

$$\|A\|_1 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq 0}} \frac{\|A\boldsymbol{v}\|_1}{\|\boldsymbol{v}\|_1}$$
$$\|A\|_2 = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq 0}} \frac{\|A\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2}$$
$$\|A\| = \sup_{\substack{\boldsymbol{v}\in\mathbb{C}^n \\ \boldsymbol{v}\neq 0}} \frac{\|A\boldsymbol{v}\|_\infty}{\|\boldsymbol{v}\|_\infty}$$

Théorème 5.2 *Soit $A = (a_{ij})$ une matrice carrée. Alors*

$$\|A\|_1 = \max_j \sum_i |a_{ij}|$$
$$\|A\|_2 = \sqrt{\rho\left(A^*A\right)} = \sqrt{\rho\left(AA^*\right)} = \|A^*\|_2$$
$$\|A\|_\infty = \max_i \sum_j |a_{ij}|$$

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

$$UU^* = I \Longrightarrow \|A\|_2 = \|AU\|_2 = \|UA\|_2 = \|U^*AU\|_2.$$

Par ailleurs, si la matrice A est normale, c'est-à-dire si $AA^* = A^*A$, alors

$$\|A\|_2 = \rho\left(A\right).$$

Remarque 5.1

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

Théorème 5.3

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

$$\rho\left(A\right) \leq \|A\|.$$

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

$$\|A\| \leq \rho\left(A\right) + \varepsilon.$$

Théorème 5.4 *L'application* $\|\cdot\|_E : \mathcal{M}_n \to \mathbb{R}^+$ *définie par*

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

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

$$UU^* = I \Longrightarrow \|A\|_E = \|AU\|_E = \|UA\|_E = \|U^*AU\|_E$$

et qui vérifie

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

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

5.3 Conditionnement

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

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

Exemple de R.S. Wilson :

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

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

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

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

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

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

Définition 5.5 *Soit $\|.\|$ une norme matricielle subordonnée, le conditionnement d'une matrice régulière A, associé à cette norme, est le nombre*

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

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

Théorème 5.5 *Soit A une matrice inversible. Soient $\boldsymbol{x}$ et $\boldsymbol{x} + \boldsymbol{\Delta x}$ les solutions respectives de*

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

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

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

est satisfaite, et c'est la meilleure possible : pour une matrice 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

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

d'où

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

■

Théorème 5.6 *Soit A une matrice inversible. Soient $\boldsymbol{x}$ et $\boldsymbol{x} + \boldsymbol{\Delta x}$ les solutions respectives de*

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

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

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

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

Théorème 5.7 *1. Pour toute une matrice inversible* A,

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

2. Pour toute matrice inversible A,

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

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

3. Si A *est une matrice normale,*

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

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

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

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

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

Rappel 5.3 *Les valeurs singulières d'une matrice rectangulaire* A *sont les racines carrées positives des valeurs propres de* A^*A.

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 : c'est l'algorithme de Richardson que nous verrons plus loin.

5.4 Suite de vecteurs et de matrices

Définition 5.6 *Soit V un espace vectoriel muni d'une norme $\|.\|$, on dit qu'une suite $(\boldsymbol{v}_k)$ d'éléments de V converge vers un élément $\boldsymbol{v} \in V$, si*

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

et on écrit

$$\boldsymbol{v} = \lim_{k\to\infty} \boldsymbol{v}_k.$$

Remarque 5.1 *Sur un espace vectoriel de dimension finie, toutes les normes sont équivalentes. Donc $\boldsymbol{v}_k$ tend vers $\boldsymbol{v}$ si et seulement si $\|\boldsymbol{v}_k - \boldsymbol{v}\|$ tend vers 0 pour une norme.*

Théorème 5.8 *1. Soit $\|\cdot\|$ une norme matricielle subordonnée, et B une matrice vérifiant*

$$\|B\| < 1.$$

Alors la matrice $(I + B)$ est inversible, et

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

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

$$\|B\| \geq 1$$

pour toute norme matricielle, subordonnée ou non.

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

Théorème 5.9 *Soit B une matrice carrée. Les conditions suivantes sont équivalentes :*

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

Théorème 5.10 *Soit B une matrice carrée, et $\|.\|$ une norme matricielle quelconque. Alors*

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

Soit donc une suite $\boldsymbol{v}_k$ définie par $\boldsymbol{v}_{k+1} = B\boldsymbol{v}_k$. On a $\boldsymbol{v}_k = B^k\boldsymbol{v}_0$.

$$\frac{\|\boldsymbol{v}_k\|}{\|\boldsymbol{v}_0\|} \leq \|B^k\|$$

Si l'on veut une erreur inférieure à ε

$$\frac{\|\boldsymbol{v}_k\|}{\|\boldsymbol{v}_0\|} \leq \|B^k\| \leq \varepsilon$$

$$\|B^k\|^{\frac{1}{k}} \leq \varepsilon^{\frac{1}{k}}$$

Définition 5.7 *On définit le facteur de convergence local de la méthode itérative dont la matrice d'itération est B est $\rho_k(B) = \|B^k\|^{\frac{1}{k}}$. Le facteur asymptotique de convergence asymptotique est $\rho(B)$. Le taux de convergence moyen est $R_k(B) = -\ln \rho_k(B)$, le taux de convergence asymptotique est $R(B) = -\ln \rho(B)$.*

Théorème 5.11 *Le nombre d'itérations nécessaires pour réduire l'erreur d'un facteur ε est au moins égal à $K = \dfrac{-\ln \varepsilon}{R(B)}$.*

5.5 Résultats généraux de convergence

Soit donc l'algorithme

$$M\boldsymbol{u}_{k+1} = N\boldsymbol{u}_k + b \tag{5.1}$$

avec $M - N = A$. Si la suite converge, elle converge vers la solution $\boldsymbol{u}$ de $A\boldsymbol{u} = \boldsymbol{b}$, et l'erreur $\boldsymbol{e}_k = \boldsymbol{u}_k - \boldsymbol{u}$ vérifie $M\boldsymbol{e}_{k+1} = N\boldsymbol{e}_k$. On note $B = M^{-1}N$, c'est la *matrice de l'itération*. On note aussi $\boldsymbol{r}_k := \boldsymbol{b} - A\boldsymbol{u}_k = A(\boldsymbol{u} - \boldsymbol{u}_k) = A\boldsymbol{e}_k$ le *résidu* à l'étape k. D'après le théorème 5.9, on a

Théorème 5.12 *La suite $\boldsymbol{u}_k$ converge pour toute donnée initiale $\boldsymbol{u}_0$ si et seulement si $\varrho(B) < 1$, si et seulement si $\|B\| < 1$ pour au moins une norme matricielle subordonnée $\|.\|$.*

5.5.1 Cas des M-matrices

Définition 5.8 (Matrice non-négative) *Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est dite non-négative (resp. non-positive) si pour tout $i, j \in \{1, \cdots, n\}$, $a_{ij} \geq 0$ (resp. $a_{ij} \leq 0$).*

Théorème 5.13 (Perron-Frobenius) *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice non-négative. Alors A a une valeur propre non-négative égale au rayon spectral de A, et un vecteur propre correspondant qui est aussi non-négatif.*

Définition 5.9 (Décomposition régulière ou regular splitting) *. Une décomposition $A = M - N$ est dite régulière si M est inversible et si M^{-1} et N sont toutes deux non-négatives.*

Théorème 5.14 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice, $A = M - N$ une décomposition régulière. Alors la méthode itérative converge pour toute donnée initiale si et seulement si A est inversible et A^{-1} est non-négative.*

Définition 5.10 (M-matrice) *. Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est une M-matrice si elle possède les quatre propriétés suivantes : 1. $a_{ii} > 0$ pour tout i, 2. $a_{ij} \leq 0$ pour tout $(i, j), i \neq j$. 3. A est inversible. 4. A^{-1} est non négative.*

Corollaire 5.1 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ une M-matrice, $A = M - N$ une décomposition régulière. Alors la méthode itérative $M\boldsymbol{u}_{k+1} = N\boldsymbol{u}_k + \boldsymbol{b}$ converge pour toute donnée initiale vers la solution de $A\boldsymbol{u} = \boldsymbol{b}$.*

Définition 5.11 (Matrice à diagonale strictement dominante) *La matrice est dite à diagonale strictement dominante si*

$$\forall i, 1 \leq i \leq n, |a_{ii}| > \sum_{j \neq i} |a_{ij}|.$$

5.5.2 Cas des matrices hermitiennes

Théorème 5.15 (Householder-John) *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.*

5.6 Méthodes classiques

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

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

et 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

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

Méthode de Jacobi

On choisit $M = D, N = E + F$, et une étape de l'algorithme s'écrit

$$D\boldsymbol{u}_{k+1} = (E + F)\boldsymbol{u}_k + \boldsymbol{b}.$$

Chaque composante $(u_{k+1})_i$ peut être calculée explicitement par

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

Exemple de 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 : tant que $\|\boldsymbol{r}_k\| >$ eps, on continue.

Théorème 5.16 *Si A est à diagonale strictement dominante, l'algorithme de Jacobi converge.*

Théorème 5.17 *Soit $A \in \mathcal{M}(\mathbb{R})$ une matrice symétrique inversible, décomposée en $A = D - E - E^T$, où D est diagonale définie positive, et E strictement triangulaire inférieure. Alors l'algorithme de Jacobi converge si et seulement si A et $2D - A$ sont définies positives.*

Méthode de Gauss-Seidel

Elle correspond à la décomposition $M = D - E, N = F$.

$$(D - E)\boldsymbol{u}_{k+1} = F\boldsymbol{u}_k + \boldsymbol{b}.$$

Le calcul de $\boldsymbol{u}_{k+1}$ en fonction de $\boldsymbol{u}_k$ nécessite la résolution d'un système triangulaire inférieur de matrice $D - E$.

Méthode de relaxation

Elle est apppelée aussi méthode S.O.R. (successive over relaxation). Elle correspond à la décomposition $M = \frac{1}{\omega}D - E$, $N = \frac{1-\omega}{\omega}D + F$, ce qui s'écrit

$$(\frac{1}{\omega}D - E)\boldsymbol{u}_{k+1} = (\frac{1-\omega}{\omega}D + F)\boldsymbol{u}_k + \boldsymbol{b}.$$

De nouveau, une étape de l'algorithme nécessite la résolution d'un système triangulaire. La méthode de Gauss-Seidel correspond à $\omega = 1$.

Tableau résumé

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

Exercice 5.1 *Ecrire une étape de l'algorithme SOR.*

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

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

Théorème 5.18 *Soit A une matrice à diagonale strictement dominante. Si $0 < \omega \leq 1$, la méthode de relaxation converge.*

Théorème 5.19 *Si la méthode de relaxation converge pour toute donnée initiale, on a*

$$0 < \omega < 2$$

La preuve repose sur le lemme suivant, et le théorème 5.12.

Lemme 5.1 *Pour tout $\omega \neq 0$, on a $\rho(\mathcal{L}_\omega) \geq |\omega - 1|$.*

Théorème 5.20 *Soit A une matrice hermitienne définie positive. Si $\omega \in]0, 2[$, la méthode de relaxation converge pour toute donnée initiale.*

C'est une conséquence du théorème 5.15.

5.7 Cas des matrices tridiagonales, comparaison des méthodes

Théorème 5.21 *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 5.22 *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 $\omega^* = \dfrac{2}{1 + \sqrt{1 - (\rho(J))^2}}$. Le facteur de*

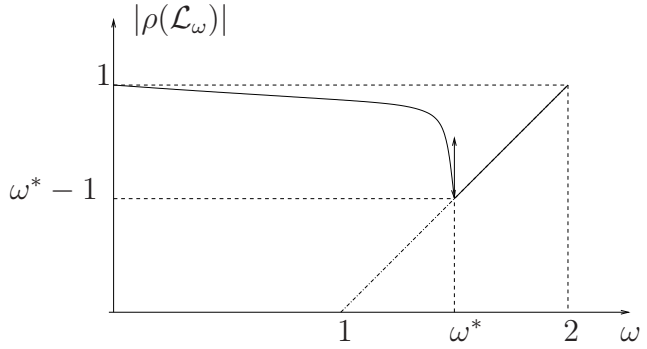


FIGURE 5.1 – variations de $\rho(\mathcal{L}_\omega)$ en fonction de ω

convergence optimal est $\rho(\mathcal{L}_{\omega^*}) = \omega^* - 1$.

Remarque 5.2 *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.*

5.8 Méthode de Richardson

On réécrit l'itération sous la forme

$$\boldsymbol{u}_{k+1} = \boldsymbol{u}_k + M^{-1}\boldsymbol{r}_k.$$

Si on choisit $M^{-1} = \alpha I$, on obtient la méthode de Richardson

$$\boldsymbol{u}_{k+1} = (I - \alpha A)\boldsymbol{u}_k + \alpha \boldsymbol{b}.$$

Si A est une matrice symétrique et définie positive, ses valeurs propres $\lambda_i(A)$ sont strictement positives, nous les ordonnons de façon croissante.

Théorème 5.23 *Soit A une matrice symétrique et définie positive d'ordre n. La méthode de Richardson converge si et seulement si $\alpha \in (0, 2/\rho(A))$. La convergence est optimale pour $\alpha_{opt} = \frac{2}{\lambda_1+\lambda_n}$. On a alors*

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

5.9 La matrice du laplacien en dimension 1

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

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.

5.10 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 préfèrera les méthodes directes dans le cas des matrices pleines, et les méthodes itératives dans le cas des matrices creuses.

5.11 Méthodes par blocs

Reprenons l'exemple des différences finies en dimension 2. La matrice A est tridiagonale par blocs. On fait la décomposition

$$D = \begin{pmatrix} T & & & \\ & T & & \\ & & \ddots & \\ & & & T \end{pmatrix}, \quad E = \begin{pmatrix} 0 & & & \\ I & 0 & & \\ & \ddots & \ddots & \\ & & I & 0 \end{pmatrix}, \quad F = E^T$$

ET on découpera aussi le vecteur $\boldsymbol{u}_k$ par blocs. Tous les théorèmes précédents ont un analogue. Voir [4].

Chapitre 6

Interpolation polynômiale et extrapolation

Deux problèmes classiques

1) Interpolation : on considère une aile d'avion, qu'on soumet à des vents de 10, 50, 100, 200 km/h, et dont on calcule les déformations pour ces valeurs. On veut savoir comment elle résistera à un vent de 150km/h.

2) Extrapolation : on connaît la population française de 1800 à 2010 et on veut en déduire une estimation de la population française dans les 10 prochaines années.

Une solution est de déterminer un polynôme dont la courbe s'approche le plus possible (ou passe par) ces points, et de prendre sa valeur aux nouveaux points. C'est le but de ce chapitre.

Le problème mathématique est le suivant : on se donne $n+1$ mesures $f_0, \cdots, f_n$ en $n+1$ points distincts $x_0, \cdots, x_n$ et on cherche à calculer un polynôme q de degré inférieur ou égal à m, avec $m \leq n$, qui "approche" les mesures $f_0, \cdots, f_n$. La première approche est quand $m = n$: c'est le polynôme d'interpolation.

6.1 Interpolation de Lagrange

Théorème 6.1 *1) Il existe un unique polynôme $p_n \in \mathbf{P}_n$ (espace vectoriel des polynômes de degré inférieur ou égal à n) tel que*

$$\forall i,\ 0 \leq i \leq n, \quad p_n(x_i) = f_i. \tag{6.1}$$

2) Il s'écrit sous la forme

$$p_n(x) = \sum_{i=0}^{n} f_i \ell_i(x), \quad avec \quad \ell_i(x) = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}. \tag{6.2}$$

Les ℓ_i sont les polynômes d'interpolation de Lagrange. p_n est le polynôme d'interpolation aux points x_i pour les mesures f_i.

Démonstration 1)Notons $p_n(x) = \sum_{k=0}^{n} a_k x^k$, $x \in \mathbb{R}$. Résoudre (6.1) est équivalent à résoudre un système linéaire dont les inconnues sont les coefficients a_k :

$$Ay = b \text{ avec}$$

$$A = \begin{pmatrix} 1 & x_0 & \cdots & x_0^n \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & \cdots & x_n^n \end{pmatrix}, \quad y = \begin{pmatrix} a_0 \\ \vdots \\ a_n \end{pmatrix}, \quad b = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix},$$

A est une matrice de Vandermonde. Elle est inversible ce qui conclut la partie 1).

2) ℓ_i est un polynôme de $\mathbf{P}_n$, et vérifie $\ell_i(x_j) = \delta_{ij}$. On vérifie que ce polynôme convient. ■

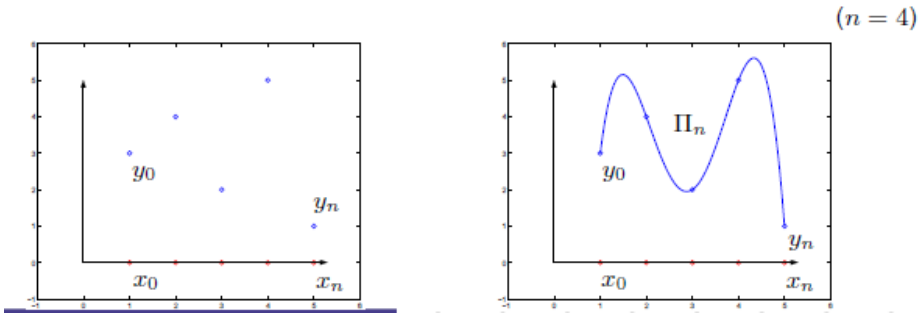


FIGURE 6.1 – polynôme d'interpolation

Lorsque les fi sont les valeurs d'une certaine fonction f aux points x_i, on parle de p_n comme de l'interpolant de f et on la note $\Pi_n f$.

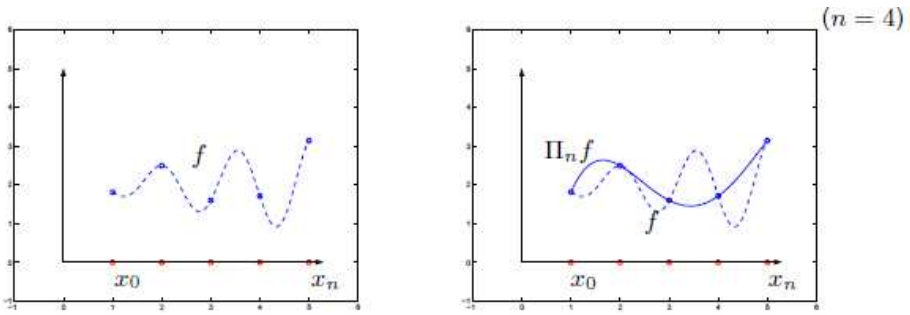


FIGURE 6.2 – interpolant

En principe il suffit de résoudre le système linéaire pour calculer les a_i, puis de calculer en chaque nouveau point x

$$p_n(x) = (a_0, a_1, \ldots, a_n) * \begin{pmatrix} 1 \\ x \\ \vdots \\ x^n \end{pmatrix}$$

Mais le système est très mal conditionné. Il vaut mieux programmer directement (6.2).

```
function [yy] = lagint(x, y, xx)
% LAGINT uses the points (x_i, y_i) for the Lagrange Form of the
% interpolating polynomial and interpolates the values
% yy_i = p_n(xx_i)
n = length(x); nn = length(xx);
for i = 1:nn,
    yy(i) = 0;
    for k =1:n
        yy(i) = yy(i)+y(k)*prod((xx(i) - x([1:k-1,k+1:n])))...
        /prod((x(k) - x([1:k-1,k+1:n])));
    end;
end;
```

6.1.1 Formulation barycentrique

Utiliser la formulation (6.2) mène à $\mathcal{O}(n^2)$ opérations pour chaque x. Nous définissons les coefficients

$$\lambda_i = \frac{1}{\prod_{j \neq i}(x_i - x_j).}$$

et nous réécrivons

$$p_n(x) = \sum_{i=0}^{n} \lambda_i \left(\prod_{j \neq i} (x - x_j) f_i \right) = \prod_j (x - x_j) \left(\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i} f_i \right)$$

Puisque la formule est exacte pour les polynômes de degré 0, on peut écrire pour $f \equiv 1$:

$$1 = \prod_j (x - x_j) \left(\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i} \right)$$

et donc

$$\prod_j (x - x_j) = \frac{1}{\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i}}$$

ce qui nous donne la formule barycentrique

$$p_n(x) = \frac{\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i} f_i}{\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i}}$$

$$p_n(x) = \frac{\displaystyle\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i} f_i}{\displaystyle\sum_{i=0}^{n} \frac{\lambda_i}{x - x_i}}$$

Pour l'utiliser nous calculons d'abord les λ_i en $\mathcal{O}(n^2)$ opérations,

```
function [lambda] = coeffbary(x)
% COEFFBARY computes the coefficients for the barycentric
% representation of the interpolating polynomial through
% the points (x_i, y_i)
n = length(x); x=x(:);
for k = 1:n,
lambda(k) = 1 / prod(x(k) - x([1:k-1,k+1:n]));
end;
```

Puis pour chaque x nous calculons les poids $\mu_i = \dfrac{\lambda_i}{x - x_i}$ et $p_n(x) = \dfrac{\sum_{i=0}^{n} \mu_i f_i}{\sum_{i=0}^{n} \mu_i}$ en seulement $\mathcal{O}(n)$ opérations.

```
function [yy] = intbary(x, y,lambda, xx)
%% INTBARY evaluates the interpolating polynomial
%% through (x_i, y_i) for the values xx: yy = P_n(xx)
x=x(:); y=y(:); xx = xx(:);
nn = length(xx);
for i = 1:nn,
z = (xx(i)-x) + 1e-30; % prevents a division by zero
mue=lambda'./z;
yy(i)=mue'*y/sum(mue);
end;
```

6.1.2 Formule de Newton

On se donne les $n+1$ points $x_0, \cdots, x_n$. Pour tout k plus petit que n, on note p_k le polynôme d'interpolation de f aux points $x_0, \cdots, x_k$. On a

$$p_k - p_{k-1} = C(x - x_0) \cdots (x - x_{k-1})$$

Définition 6.1 *Pour $k+1$ points $y_0, \cdots, y_k$, on note $f[y_0, \cdots, y_k]$ le coefficient de degré k du polynôme d'interpolation de f aux points $y_0, \cdots, y_k$.*

Lemme 6.1

$$p_k - p_{k-1} = f[x_0, \cdots, x_k](x - x_0) \cdots (x - x_{k-1})$$

Démonstration

■

Théorème 6.2 (Formule de Newton)

$$p_n(x) = f(x_0) + \sum_{k=0}^{n} f[x_0, \cdots, x_k](x - x_0) \cdots (x - x_{k-1}) \tag{6.3}$$

Démonstration Il suffit de sommer la formule de récurrence précédente. ■

Lemme 6.2 (Formule des différences divisées)

$$\forall k \geq 1, f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0} \tag{6.4}$$

Démonstration

Soit $q_{k-1} \in \mathbf{P}_{k-1}$ le polynôme d'interpolation de f aux points $x_1, \ldots, x_k$. Posons

$$\tilde{p}_k = \frac{(x - x_0)q_{k-1} - (x - x_k)p_{k-1}}{x_k - x_0}$$

Alors $\tilde{p}_k = p_k$. En effet

et il ne reste plus qu'à égaler les coefficients directeurs dans la formule de $\tilde{p}_k$. ■

Table de calcul

$$\begin{array}{llllll}
x_0 & f(x_0) = f[x_0] & & & & \\
x_1 & f(x_1) = f[x_1] & f[x_0, x_1] & & & \\
x_2 & f(x_2) = f[x_2] & f[x_1, x_2] & f[x_0, x_1, x_2] & & \\
\vdots & \vdots & \vdots & \vdots & \ddots & \\
x_n & f(x_n) = f[x_n] & f[x_{n-1}, x_n] & f[x_{n-2}, x_{n-1}, x_n] & \cdots & f[x_0, \ldots, x_n]
\end{array}$$

FIGURE 6.3 – table des différences divisées

Voici l'algorithme matlab

```
function [d,D] = coeffnewton(x, y)
% COEFFNEWTON computes the divided differences needed for
% constructing the interpolating polynomial through (x_i,y_i)
n = length(x)-1; % degree of interpolating polynomial
The Interpolation Polynomial 335
% divided differences
for i=1:n+1
D(i,1) = y(i);
for j = 1:i-1
D(i,j+1) = (D(i,j)-D(i-1,j))/(x(i)-x(i-j));
end
end
d = diag(D);
```

Une fois les d_i calculés, pour les utiliser nous couplons avec l'algorithme de Hörner, en réécrivant le polynôme p_n sous la forme

$$p_n(x) = d_0 + (x - x_0)(d_1 + (x - x_1)(d_2 + \cdots + (x - x_{n-2})(d_{n-1} + (x - x_{n-1})d_n)))$$

```
function y = intnewton(x,d,z)
% INTNEWTON evaluates the Newton interpolating polynomial
% at the new points z: y = P_n(z) using the Horner form
% and the diagonal d of the divided difference scheme.
n = length(x)-1;
y = d(n+1)
for i= n:-1:1
y = y.*(z-x(i))+d(i);
end;
```

En Matlab, on utilise la fonction *polyfit* pour l'interpolation polynomiale. Cette fonction utilise une interpolation au sens des moindres carrés discrets (voir partie 3).

6.1.3 estimation d'erreur

Théorème 6.3 *si $f \in \mathcal{C}^{n+1}([a,b])$, $\forall x \in [a,b]$, $\exists \zeta_x$ appartenant au plus petit intervalle ouvert contenant $x, x_0, \cdots, x_n$, tel que*

$$f(x) - p_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\zeta_x) \Pi_{n+1}(x) \tag{6.5}$$

où $\Pi_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$.

Démonstration On remarque d'abord que l'égalité est vraie si x est l'un des x_i. On suppose ensuite que x est fixé, non égal à l'un des x_i. On applique le théorème de Rolle $n+1$ fois à la fonction définie pour x fixé par

$$F(t) = f(t) - p_n(t) - C\Pi_n(t)$$

où C est défini par $F(x) = 0$. ■

On déduit de ce théorème d'abord que les différences divisées sont des approximations des dérivées : il existe un ξ dans l'intervalle $(inf(x_i), sup(x_i))$ tel que

$$f[x_0, \cdots, x_n] = \frac{f^{(n)}(\zeta)}{(n)!}$$

En effet écrivons d'après le lemme 6.1

$$p_n(x_n) - p_{n-1}(x_n) = f[x_0, \cdots, x_n](x_n - x_0) \cdots (x_n - x_{n-1})$$

et d'après l'estimation d'erreur aux points $x_0, \ldots, x_{n-1}$,

$$f(x_n) - p_{n-1}(x_n) = \frac{1}{(n)!} f^{(n)}(\zeta)(x_n - x_0) \cdots (x_n - x_{n-1}).$$

Egalons ces deux expressions pour conclure.

Nous en déduisons aussi une estimation d'erreur grossière. On note pour une fonction φ, $\|\varphi\|_\infty = \sup_{x\in[a,b]} |\varphi(x)|$, et on a

$$\|f - p_n\|_\infty \leq \frac{1}{(n+1)!} \|F^{n+1}\|_\infty \|\Pi_{n+1}\|_\infty \tag{6.6}$$

Pour une fonction f donnée, on minimise l'erreur en choisissant bien les points d'interpolation :

Théorème 6.4 *Sur un intervalle* $[a, b]$, $\|\Pi_{n+1}\|_\infty$ *est minimale pour le choix des points*

$$x_i^T = \frac{a+b}{2} + \frac{b-a}{2} y_i^{n+1}, \quad 0 \leq i \leq n$$

Les $y_i^{n+1} = \cos(\frac{2i+1}{2(n+1)}\pi)$ *sont les zéros du polynôme de Chebyshev* T_{n+1}.

Polynômes de Chebyshev

Pour tout k on définit sur $[-1,1]$ la fonction

$$T_k(y) = \cos(k\mathrm{Arc}\cos y). \tag{6.7}$$

T_k est en fait un polynôme de degré k. Pour le voir, on établit la formule de récurrence

$$T_{k+1}(y) = 2y\,T_k - T_{k-1}, \quad T_0 = 1, \quad T_1 = y, \tag{6.8}$$

ce qui permet de les définir sur tout $\mathbb{R}$. Le coefficient dominant de T_k est $2^{k-1}y^k$, ses zéros sont les $y_i^k = \cos\,(\frac{2i+1}{2k}\pi)$ pour $0 \le i \le k-1$, et les extrema valent $(-1)^i$, atteints aux points $\tilde{y}_i^k = \cos\,(\frac{i}{k}\pi)$ pour $0 \le i \le k$.

Lemme 6.3 *Pour tout $p \in \mathbf{P}_k$ unitaire (i.e. $p = y^k + \cdots$), on a*

$$\|p\|_\infty \ge \|\frac{T_k}{2^{k-1}}\|_\infty = \frac{1}{2^{k-1}}$$

Démonstration Pour n'importe quel choix des points d'interpolation $x_0, \cdots, x_n$, faisons un changement de variable

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}y_i,\ 0 \le i \le n, \quad x = \frac{a+b}{2} + \frac{b-a}{2}y$$

Lorsque x varie dans l'intervalle $[a,b]$, y varie dans l'intervalle $[-1,1]$ et

$$\Pi_{n+1}(x) = \left(\frac{b-a}{2}\right)^{n+1} \prod_{i=0}^{n}(y - y_i)$$

Minimiser l'erreur est donc minimiser la norme infinie d'un polynôme unitaire sur $(-1,1]$, et

$$\inf_{\{x_i\}} \|\Pi_{n+1}\|_\infty = \|\prod(x - x_i^T)\|_\infty = 2\left(\frac{b-a}{4}\right)^{n+1}$$

Admis, cf Demailly, *analyse numérique et équations différentielles.* ■

Remarque 6.1 *Pour une division en points équidistants, i.e. $x_j = a + \frac{b-a}{n}j$, on montre que*

$$\|\Pi_{n+1}\|_\infty \sim (b-a)^{n+1}\frac{e^{-n}}{\sqrt{n}\ln n}$$

Pour $(a,b) = (-1,1)$, la figure suivante montre le logarithme des erreurs en fonction de n dans les deux cas

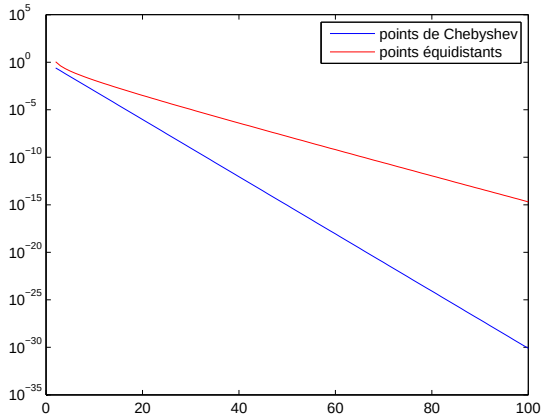


FIGURE 6.4 – Comparaison de $\|\Pi_{n+1}\|_\infty$ pour deux ensembles de points

6.1.4 Convergence de p_n vers f

Considérons la fonction $f(x) = \frac{x+1}{5}\sin(x)$ sur [0,6] (ex de Quarteroni).

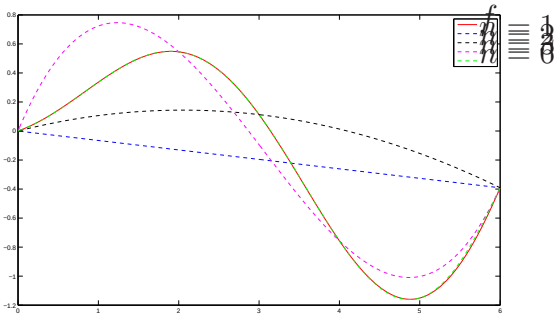


FIGURE 6.5 – interpolant

On constate dans ce cas que la suite $\Pi_n(f)$ converge vers f. Ce n'est pas vrai en général. Le contre-exemple classique est celui de la fonction de Runge $f(x) = \dfrac{1}{1+15x^2}$:

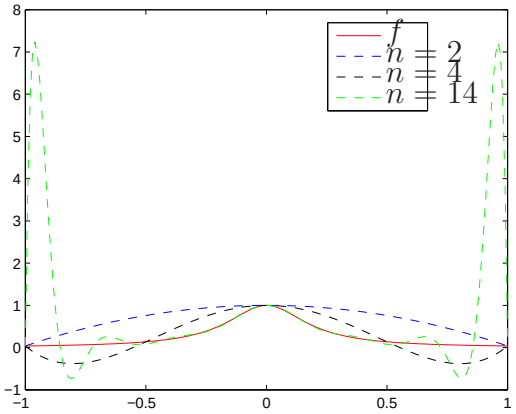


FIGURE 6.6 – interpolant

Bien que la fonction soit tout à fait régulière, on voit que l'erreur en 1 tend vers l'infini.

6.2 Interpolation d'Hermite

f est toujours une fonction suffisamment régulière sur le segment $[a, b]$. On se donne $k+1$ points $x_0, \cdots, x_k$ dans $[a, b]$.

Théorème 6.5 *Posons $n = 2k+1$. Il existe un et un seul polynôme $p_n \in \mathbf{P}_n$ tel que*

$$\forall j,\ 0 \leq j \leq k, \quad p_n(x_j) = f(x_j) \text{ et } p'_n(x_j) = f'(x_j).$$

Théorème 6.6 *si $f \in \mathcal{C}^{n+1}([a, b])$, $\forall x \in [a, b]$, $\exists \zeta_x$ appartenant au plus petit intervalle ouvert contenant $x, x_0, \cdots, x_k$, tel que*

$$f(x) - p_n(x) = \frac{1}{(n+1)!} F^{n+1}(\zeta_x) \Pi_{n+1}(x) \tag{6.9}$$

où $\Pi_{n+1}(x) = \prod_{i=0}^{k} (x - x_i)^2$.

p_n dépend de $2k+2$ coefficients, nous allons l'exprimer sous la forme

$$p_n(x) = \sum_{i=0}^{k} f(x_i) q_i(x) + \sum_{i=0}^{k} f'(x_i) r_i(x) \tag{6.10}$$

où les polynômes q_i et r_i sont définis par

$$\left| \begin{array}{l} q_i(x_j) = \delta_{ij}, \\ q_i'(x_j) = 0, \end{array} \right. \quad \left| \begin{array}{l} r_i(x_j) = 0, \\ r_i'(x_j) = \delta_{ij}. \end{array} \right. \tag{6.11}$$

On peut les déterminer en fonction des polynômes d'interpolation de Lagrange ℓ_i :

$$q_i(x) = (1 + 2(x_i - x)\ell_i'(x_i))\ell_i^2(x), \quad r_i(x) = (x - x_i)\ell_i^2(x).$$

6.3 Interpolation par morceaux

Soient $a \equiv a_0 < a_1 < \cdots < a_N \equiv b$ des points qui divisent l'intervalle $I = [a, b]$ en sous-intervalles $I_j = [a_j, a_{j+1}]$ de longueur $H = \frac{b-a}{N}$, soit $a_j = a + jH$.

6.3.1 Interpolation affine

Sur chaque intervalle I_j, on interpole f par un polynôme de degré inférieur ou égal à 1. On obtient un polynôme par morceaux, noté $\Pi_1^H f$. Il s'écrit

$$\Pi_1^H f(x) = f(a_j) + f[a_j, a_{j+1}](x - a_j), \quad x \in I_j$$

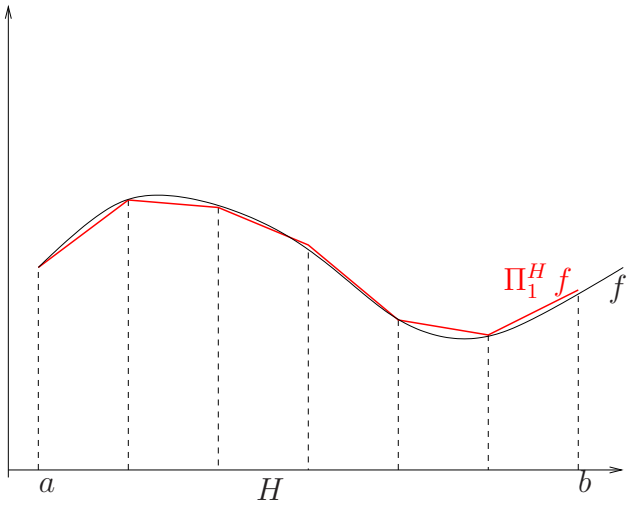


FIGURE 6.7 – interpolation affine par morceaux

Théorème 6.7 *Si $f \in \mathcal{C}^2(I)$, alors*

$$\sup_{x \in I} |f(x) - \Pi_1^H f(x)| \leq \frac{H^2}{8} \sup_{x \in I} |f"(x)|. \tag{6.12}$$

Démonstration Il suffit d'appliquer l'estimation d'erreur (6.6). ■

Remarque 6.2 *Si $f \in \mathcal{C}^{n+1}(I)$, on peut faire de même une interpolation par des polynômes de degré inférieur ou égal à n dans chaque sous-domaine et on obtient l'estimation d'erreur*

$$\sup_{x\in I} |f(x) - \Pi_n^H f(x)| \leq \frac{H^{n+1}}{4(n+1)} \sup_{x\in I} |f^{(n+1)}(x)|. \tag{6.13}$$

6.3.2 Interpolation par fonctions splines

L'inconvénient de la démarche précédente est que l'approximation de f manque de régularité. Ici nous nous donnons $y_i = f(a_i)$ et aussi des valeurs y_i' que nous choisirons ensuite. Dans chaque sous-intervalle, nous interpolons la fonction f par un polynôme $p_i \in \mathbf{P}_3$ tel que

$$p_i(a_i) = y_i,\ p_i(a_{i+1}) = y_{i+1}, \quad p_i'(a_i) = y_i',\ p_i'(a_{i+1}) = y_{i+1}',$$

Faisons le changement de variable $y = (x - a_i)/H$, et posons $p(x) = P(y)$. On doit donc avoir

$$P_i(0) = y_i, \quad P_i(1) = y_{i+1}; P_i'(0) = Hy_i', \quad P_i'(1) = Hy_{i+1}';$$

Nous utilisons les formules données pour les polynômes d'Hermite.

$$P_i = y_i q_0 + y_{i+1} q_1 + y_i' r_0 + y_{i+1}' r_1$$

Les polynôme de Lagrange aux points 0 et 1 sont $\ell_0 = 1 - y$, $\ell_1 = y$, et les polynômes q_i et r_i sont donnés par

$$q_0(y) = (2y-1)(1-y)^2,\ q_0(y) = (2y-1)y^2, \quad r_0(y) = y(1-y)^2,\ r_i(y) = (1-y)y^2.$$

Comment maintenant calculer la valeur de p_i en un point x ?

1. Déterminer l'intervalle $[a_i, a_{i+1}]$ où se trouve x.
2. Calculer la variable locale $y = (x - a_i)/H$.
3. Évaluer $P_i(y)$, de préférence par l'algorithme de Hörner.

Pour déterminer l'intervalle où se trouve x, on utiliser un algorithme de recherche binaire si les intervalles ne sont pas de même taille. Sinon bien sûr on prend la partie entière de x/H.

Les y_i' doivent approcher les dérivées $f'(a_i)$ qui ne sont pas données en général. On peut alors approcher par exemple $f'(a_i)$ par des différences divisées $y_i' = f[a_{i-1}, a_{i+1}]$ pour $1 \leq i \leq N-1$. Aux deux extrémités on peut prendre des dérivées décentrées $y_0' = f[a_0, a_1]$ et $y_N' = f[a_{N-1}, a_N]$.

Peut-on déterminer les y_i' de façon à être encore plus régulier ? Par exemple que les dérivées secondes soient aussi continues ? La réponse est oui, ce sont les vrais splines cubiques historiques.

Chapitre 7

Approximation par des polynômes

7.1 Théorèmes généraux

Soit E un espace vectoriel normé (e.v.n.) sur $\mathbb{R}$ ou $\mathbb{C}$, muni d'une norme $\|\cdot\|$.

Théorème 7.1 *Si M est un sous-espace vectoriel* ***de dimension finie*** *de E, alors pour tout g dans E, il existe au moins un y dans M tel que $\|g-y\| = \inf_{x\in M} \|g-x\|$.*

Démonstration On se fixe x_0 dans M. On définit $K = \{x \in M, \|g-x\| \leq \|g-x_0\|\}$. Alors $\inf_M = \inf_K$. La fonction $x \mapsto \|g-x\|$ est une fonction continue sur K compact, elle admet une borne inférieure d'après le théorème de Weirstrass. ■

Théorème 7.2 (Théorème de projection dans un Hilbert) *Soit E est un espace de Hilbert muni du produit scalaire $(\cdot,\cdot)$, $\|\cdot\|$ est la norme associée. Soit M un sous-espace vectoriel de dimension finie de E, alors pour tout g dans E, il existe* ***un unique*** *y dans M tel que $\|g-y\| = \inf_{x\in M} \|g-x\|$. On le note $P_M g$. C'est la projection de g sur M, caractérisée par*

$$\forall z \in M, \quad (g - P_M g, z) = 0$$

Démonstration Vu en L2. ■

On va appliquer ces théorèmes à $M = \mathbf{P}_n$. D'abord un résultat d'approximation

Théorème 7.3 (Théorème de Weierstrass) *Si $(a, b]$ est compact, toute fonction continue sur $(a, b]$ peut être approchée uniformément par des polynômes, ou encore l'espace des polynômes est dense dans $\mathcal{C}^0((a, b])$ pour la norme de L^∞.*

Démonstration La démonstration est un peu longue, elle s'appuie sur les polynômes de Bernestein

$$B_n(f) = \sum_{k=0}^{n} \binom{n}{k} f(\frac{k}{n}) x^k (1-x)^{n-k}$$

■

L'équivalent pour les fonctions périodiques est très utile :

Théorème 7.4 *Soit f une fonction continue de période 2π. Alors il existe des coefficients réels $a_0, \cdots, a_n, \cdots$ et $b_1, \cdots, b_n, \cdots$ tels que*

$$S_n(t) = a_0 + \sum_{k=1}^{n} (a_k \cos\, kt + b_k \sin\, kt)$$

converge uniformément vers f sur $\mathbb{R}$.

7.2 Polynômes orthogonaux, moindres carrés

L'espace $L^2(a, b)$ des fonctions de carré intégrable sur (a, b) est un espace de Hilbert pour le produit scalaire $(f, g) = \int_a^b f(x)g(x)\, dx$. Il contient les polynômes.

Définition 7.1 *On dit qu'une suite de polynômes $p_0, \cdots, p_n, \cdots$ forme une suite de polynômes orthogonaux si*
- *$d^\circ p_n = n$ pour tout n,*
- *$(p_i, p_j) = 0$ pour $i \neq j$.*

Il existe une unique suite, à une constante multiplicative près. Exemple : sur $[-1, 1]$, les polynômes de Legendre sont définis par $L_n(1) = 1$. Ils sont définis également par la formule de récurrence

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - n\, L_{n-1}(x), \quad L_0 = 1,\ L_1 = x.$$

Ils sont aussi solution de l'équation différentielle

$$(1-x^2)y" - 2xy' + n(n+1)y = 0.$$

Leur norme est égale à $2/(2n+1)$.

Soit maintenant f une fonction de $L^2(a,b)$. Par le théorème de projection, il existe un unique P_n dans $\mathbf{P}_n$, projection de f sur $\mathbf{P}_n$. Décomposons le sur la base des polynômes orthogonaux p_j :

$$P_n = \sum_{j=0}^{n} \alpha_j^n p_j.$$

Par la caractérisation de la projection, on doit avoir pour $0 \leq j \leq n$:

$$(f, p_j) = (Q_n, p_j) = \alpha_j^n \|p_j\|^2$$

et donc α_j^n ne dépend pas de n et

$$\alpha_j^n = \alpha_j = \frac{(f, p_j)}{\|p_j\|^2}$$

Théorème 7.5 *Soit f une fonction de $L^2(a,b)$.*

1. *Pour tout n positif, il existe un unique polynôme $P \in \mathbf{P}_n$ tel que $\|f - P_n\| = \inf_{P \in \mathbf{P}_n} \|f - P\|$. Il est donné par*

$$P_n = \sum_{k=0}^{n} \frac{(f, p_j)}{\|p_j\|^2} p_j$$

2. *Si de plus $[a,b]$ est compact et f est continue, alors P_n tend vers f dans L^2 et*

$$\|f\|^2 = \sum_{k=0}^{\infty} \frac{(f, p_j)^2}{\|p_j\|^2}$$

7.3 Moindres carrés discrets

On se place maintenant dans $E = \mathbb{R}^N$. On se donne N points x_i, N mesures f_i, et on cherche $p \in \mathbf{P}_{n-1}$ qui "approche" les f_i aux points x_i. Il est clair que si $N >> n$, il n'existe en général pas de polynôme qui passe par tous les points. On va alors chercher à passer "le plus près possible", c'est-à-dire à minimiser la distance entre les $p(x_i)$ et les f_i : on cherche donc p_{n-1} qui minimise $\sum_{i=1}^{N} |p_{n-1}(x_i) - f_i|^2$. On cherche p_n sous la forme

$$p_{n-1}(x) = \sum_{k=0}^{n-1} a_k x^k,$$

c'est-à-dire qu'on cherche les a_k, et et on minimise $\sum_{i=1}^{N} |\sum_{k=0}^{n-1} a_k x_i^k - f_i|^2$. Notons A la matrice des $a_{ik} = x_i^{k-1}$, $1 \leq i \leq N$, $1 \leq k \leq n$, $y = (a_0, \cdots, a_{n-1})$, $b = (f_1, \cdots, f_N)$.

$$Trouver\ y \in \mathbb{R}^n,\ \|Ay - b\| = \inf_{z \in \mathbb{R}^n} \|Az - b\| \tag{7.1}$$

Théorème 7.6 *Soit A une matrice $N \times n$. Le problème de minimisation (7.1) admet une solution, caractérisée par $A^T Ay = A^T b$. La solution est unique si et seulement si A est injective (i.e. $\mathrm{rg\,A = n}$).*

Ce sont les équations normales. Pour résoudre le problème de moindres carrés, on n'a donc qu'à résoudre un système linéaire de taille n. Mais ce problème est mal conditionné. Un exemple :

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

alors

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

Définition 7.2 *Les valeurs singulières de A sont les racines carrées positives des n valeurs propres de $A^T A$.*

Théorème 7.7 *Soit A une matrice $N \times n$ avec $N \geq n$. Alors il existe 2 matrices orthogonales U $(N \times N)$ et V $(n \times n)$, et un matrice Σ de taille $N \times n$*

$$\Sigma = \left(\begin{array}{ccc} \sigma_1 & & \\ & \sigma_2 & 0 \\ 0 & \ddots & \\ & & \sigma_n \\ \hline & 0 & \end{array}\right)$$

telles que $A = U\Sigma V^T$.

Démonstration

$$V^T(A^T A)V = \mathrm{diag}(\sigma_i^2)$$

Soit c_j le j-ème vecteur de AV. On a

$$c_i^T c_j = \sigma_i^2 \delta_{ij}$$

On ordonne les $\sigma_i : \sigma_1, \cdots, \sigma_r$ non nulles. Donc $c_i \equiv 0$ pour $i > r$. On pose

$$u_j = \frac{c_j}{\sigma_j}$$

pour $j \leq r$. Ils forment un système orthonormé, qu'on complète en une base orthonormée de $\mathbb{R}^N$. Soit U la matrice des u_j. ■

En corollaire, le rang de A est égal au nombre de valeurs singulières de A.

On a $A = \sum \sigma_i u_i v_i^T$, $A^T A = \sum \sigma_i^2 v_i v_i^T$. On en déduit que $u_1, \cdots, u_r$ forment une base de ImA, $v_{r+1}, \cdots, v_n$ une base de $\ker A$, $v_1, \cdots, v_r$ une base de $\mathrm{ImA^T} = (\ker \mathrm{A})^\perp$.

On appelle maintenant pseudo-inverse de Σ la matrice

$$\Sigma^\dagger = \left(\begin{array}{ccccc|c} 1/\sigma_1 & & 0 & & & \\ & 1/\sigma_2 & & & & \\ & & \ddots & & 0 & 0 \\ & & & 1/\sigma_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{array}\right)$$

On définit alors la pseudo-inverse de A par

$$A^\dagger = V \Sigma^\dagger U^T$$

On a maintenant $A^\dagger = \sum 1/\sigma_i v_i u_i^T$. On en déduit que $AA^\dagger = \sum u_i u_i^T$ est la matrice de la projection orthogonale sur ImA et $A^\dagger A = \sum v_i v_i^T$ la matrice de la projection orthogonale sur $\mathrm{ImA^T}$. Revenons à notre système de moindres carrés. Les équations normales ont maintenant une interprétation agréable. Si b n'appartient pas à ImA, nous le projetons sur ImA en $AA^\dagger b$ et nous résolvons alors $Ax = AA^\dagger b$, ou encore $A(x - A^\dagger b) = 0$, ou $x - A^\dagger b \in \ker A$, ou $x - A^\dagger b = (I - A^\dagger A)w$ pour un quelconque w.

Théorème 7.8 *La solution générale du problème de moindres carrés discrets s'écrit*

$$y = A^\dagger b + (I - A^\dagger A)w$$

Si A est de rang n, il y a une solution unique. Si $\mathrm{rgA} < \mathrm{n}$*, l'ensemble des solutions est un espace vectoriel de dimension $n - r$.*

Dans le deuxième cas, on choisit dans l'ensemble des solutions celle de norme minimale, i.e. on cherche $w \in \ker A$ tel que

$$\|A^\dagger b + w\| = \inf_{z \in \ker A} \|A^\dagger b + z\|$$

ce qui revient à projeter $-A^\dagger b$ sur $\ker A$.

7.4 Régression linéaire

On a mesuré, sur N individus, $n+1$ variables $Y, X_1, \cdots, X_n$. Appelons $\bar{x}_{ij}$ la mesure de la variable X_j sur l'individu i, et $\bar{y}_i$ celle de la variable Y. On cherche à reconstituer la loi de Y à partir de celles des X_i, supposées linéairement indépendantes, par une formule linéaire

$$Y = b_0 + b_1 X_1 + \cdots + b_n X_n$$

Soit y le vecteur des $y_i = b_0 + b_1 \bar{x}_{i1} + \cdots + b_n \bar{x}_{in}$, et $\bar{y}$ le vecteur de composantes $\bar{y}_i$. On cherche $b = (b_0, \cdots, b_n)$ qui minimise la norme de $y - \bar{y}$. EN notant $\bar{X}$ la matrice

$$\bar{X} = \begin{pmatrix} 1 & \bar{x}_{11} & \cdots & \bar{x}_{1in} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \bar{x}_{i1} & \cdots & \bar{x}_{in} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & \bar{x}_{N1} & \cdots & \bar{x}_{Nn} \end{pmatrix}$$

on est ramenés à minimiser $\|\bar{X}b - \bar{y}\|$.

7.5 Résolution des équations normales

7.5.1 Méthode de Cholewski

Supposons la matrice A injective. La matrice $B = A^T A$ est alors symétrique définie positive. On écrit sa décomposition de Cholewski : il existe une unique matrice S triangulaire supérieure à coefficients diagonaux strictement positifs, telle que $B = S^T S$. On résout alors successivement les deux problèmes triangulaires.

Malheureusement, les équations normales ne sont pas bien conditionnées, et la décomposition n'est valable que si A est injective. On va donc faire différemment.

7.5.2 Décomposition QR

Toujours en supposant la matrice A injective, écrivons successivement

$$\begin{gathered} S^T S = A^T A \\ (S^T)^{-1} A^T A S^{-1} = I \\ (AS^{-1})^T (AS^{-1}) = I \end{gathered}$$

ce qui montre que la matrice $Q_1 = AS^{-1}$ est orthogonale. On peut aussi écrire

$$A = Q_1 S$$

Augmentons Q_1 (de taille $N \times n$) par une matrice Q_2 en une matrice carrée de taille N : $Q = (Q_1|Q_2)$, alors nous pouvons écrire l'égalité précédente comme

$$A = (Q_1|Q_2) \begin{pmatrix} S \\ \hline 0 \end{pmatrix} = QR$$

En fait cette décomposition peut être obtenue par d'autre moyen (voir plus bas), et ne nécessite pas que A soit injective. Puisque la matrice Q est orthogonale on a pour tout z, d'après le théorème de Pythagore,

$$\|Az - b\|^2 = \|Q^T(Az - b)\|^2 = \|Rz - Q^T b\|^2 = \|Sz - (Q^T b)_1\|^2 + \|(Q^T b)_2\|^2$$

Si R est inversible, c'est-à-dire si le rang r de A est égal à n, l'équation $Sz = (Q^T b)_1$ a une seule solution y, et le minimum est atteint pour y :

$$\inf_z \|Az - b\| = \|AS^{-1}(Q^T b)_1\| = \|(Q^T b)_1\|.$$

Si $r < n$, notons que $\ker A = \ker S$: il y a une infinité de solution, comme mentionné dans le théorème 7.8. Pour en trouver une, nous effectuons une factorisation QR de S^T, sous la forme

$$S^T = PV$$

où P^T est une matrice orthogonale $n \times n$ et V de taille $n \times r$ triangulaire supérieure de rang r

$$V = \begin{pmatrix} v_1 & \times & \times & \times \\ & v_2 & \times & \times \\ 0 & & \ddots & \times \\ & & & v_r \\ \hline & & 0 & \end{pmatrix} = \begin{pmatrix} \bar{V}^T \\ \hline 0_{n-r,r} \end{pmatrix}$$

La matrice $\bar{V}$ est donc triangulaire inférieure de rang r. D'où

$$S = V^T P = (\bar{V}|0_{r,n-r})\, P$$

Décomposons $\tilde{z} = Pz$ sous la forme

$$\tilde{z} = \begin{pmatrix} \tilde{z}_1 \\ \hline \tilde{z}_2 \end{pmatrix}, \quad \tilde{z}_1 \in \mathbb{R}^r,\ \tilde{z}_2 \in \mathbb{R}^{n-r}.$$

Donc

$$Sz = \begin{pmatrix} \bar{V}\tilde{z}_1 \\ \hline 0_{r,n-r} \end{pmatrix}$$

et

$$\|Sz - (Q^T b)_1\|^2 = \|\bar{V}\tilde{z}_1 - (Q^T b)_1\|^2,$$

d'où

$$\|Az - b\|^2 = \|\bar{V}\tilde{z}_1 - (Q^T b)_1\|^2 + \|(Q^T b)_2\|^2.$$

La matrice $\bar{V}$ est inversible, donc $\|Az-b\|$ est minimal pour $\bar{V}\tilde{y}_1-(Q^T b)_1 = 0$ et le minimum est de nouveau $\|(Q^T b)_2\|$. Choisissons

$$\tilde{y} = \begin{pmatrix} \tilde{y}_1 \\ \hline 0 \end{pmatrix}, \quad y = P^T\tilde{y}$$

Alors y est solution du problème de minimisation. y est de norme minimale : toutes les autres solutions s'écrivent sous la forme

$$z = P^T\tilde{z}, \quad \tilde{z} = \begin{pmatrix} \tilde{y}_1 \\ \hline \tilde{z}_2 \end{pmatrix}$$

si bien que (puisque la matrice P^T est orthogonale),

$$\begin{array}{rcl} \|z\|^2 & = & \|\tilde{z}\|^2 \\ & = & \|\tilde{y}_1\|^2 + \|\tilde{z}_2\|^2 \\ & = & \|\tilde{y}\|^2 + \|\tilde{z}_2\|^2 \\ & = & \|y\|^2 + \|\tilde{z}_2\|^2 \\ & \geq & \|y\|^2 \end{array}$$

7.5.3 Décomposition QR par matrices de Householder

Définition 7.3 *On appelle matrice de Householder associée au vecteur $u \in \mathbb{R}^p$ de norme $\sqrt{2}$ la matrice $p \times p$ donnée par*

$$H_u = I - u\,u^T$$

Propriétés 7.1 *Pour tout u dans $\mathbb{R}^p$ de norme $\sqrt{2}$, la matrice H_u est symétrique, orthogonale. C'est la matrice de la réflexion sur l'hyperplan orthogonal à u.*

Pour effectuer la décomposition QR de la matrice A, on va procéder comme dans la méthode de Gauss : on va multiplier successivement la matrice A à gauche par des matrices élémentaires de façon à mettre successivement des zéros sous la diagonale de A.

Lemme 7.1 *Soit x un vecteur non colinéaire à e_1 (premier vecteur de base). Alors il existe un nombre σ réel, et une matrice H_u telle que $H_u x = \sigma e_1$.*

σ est donné par $|\sigma| = \|x\|$, et son signe est l'opposé de celui de x_1. Le vecteur u est alors

$$u = \sqrt{2}\frac{x - \sigma e_1}{\|x - \sigma e_1\|}$$

Ecrivons la matrice A à l'aide de ses vecteurs colonne :

$$A = (a^1 \dots a^n), \quad H_u A = (H_u a^1 \dots H_u a^n),$$

Supposons que a^1 n'est pas colinéaire à e_1 (sinon on ne fait rien). Utilisons le lemme pour trouver $(\sigma_1, u_1) \in \mathbb{R} \times \mathbb{R}^N$ tel que $H_{u_1} a^1 = \sigma_1 a^1$, et définissons $A_1 = H_{u_1} A$. La première colonne de la matrice A_1 est le vecteur

$$\begin{pmatrix} \sigma_1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

et la matrice A_1 se décompose par blocs :

$$A_1 = \left(\begin{array}{c|ccc} \sigma_1 & \times & \dots & \times \\ \hline 0 & & & \\ \vdots & & \bar{A}_1 & \\ 0 & & & \end{array}\right)$$

Nous renouvelons la construction sur $\bar{A}_1$: construisons $(\sigma_2, \bar{u}_2) \in \mathbb{R} \times \mathbb{R}^{N-1}$ tel que

$$H_{\bar{u}_2} \bar{A}_1 = \left(\begin{array}{c|ccc} \sigma_2 & \times & \dots & \times \\ \hline 0 & & & \\ \vdots & & \bar{A}_2 & \\ 0 & & & \end{array}\right)$$

Attention la matrice $\bar{A}_2$ est de taille $(N-2) \times (N-2)$. Notons que si nous définissons le vecteur

$$u_2 = \begin{pmatrix} 0 \\ \hline \bar{u}_2 \end{pmatrix}$$

alors on a

$$\left(\begin{array}{c|ccc} 1 & \times & \dots & \times \\ \hline 0 & & & \\ \vdots & & H_{\bar{u}_2} & \\ 0 & & & \end{array}\right) = H_{u_2}$$

et

$$A_2 = H_{u_2} A_1 = \left(\begin{array}{c|ccc} \sigma_1 & \times & \dots & \times \\ \hline 0 & & & \\ \vdots & & H_{\bar{u}_2} \bar{A}_1 & \\ 0 & & & \end{array}\right) = \left(\begin{array}{c|cc|c} \sigma_1 & \times & \dots & \times \\ \hline 0 & \sigma_2 & \times \dots \times & \\ \hline \vdots & 0 & & \\ \vdots & \vdots & \bar{A}_2 & \\ 0 & 0 & & \end{array}\right)$$

On itère jusqu'à A_{n-1} qui est triangulaire supérieure, avec la construction d'une famille $u_1, \dots, u_{n-1}$ et

$$A_{n-1} = H_{u_{n-1}} \dots H_{u_1} A$$

et

$$A = (H_{u_{n-1}} \dots H_{u_1})^{-1} A_{n-1} = (H_{u_1} \dots H_{u_{n-1}}) A_{n-1}$$

puisque les matrices H_u sont orthogonales et symétriques. La matrice $P = H_{u_1} \dots H_{u_{n-1}}$ est orthogonale et nous avons fini la construction.

7.5.4 Lien avec l'orthogonalisation de Gram-Schmidt

Notons a^j les vecteurs colonne de A et q^j les vecteurs colonne de Q. Alors

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

La décomposition QR correspond donc à l'orthogonalisation des colonnes de A.

Chapitre 8

Formules de quadrature

Les formules de quadrature sont des formules approchées de calcul d'intégrales de Riemann du type

$$I := \int_a^b f(x)\,dx.$$

La formule la plus connue est la formule des trapèzes. elle consiste à introduire des points équidistants a_i dans l'intervalle, $a =: a_0 < a_1 < \cdots a_N < a_{N+1} := b$, avec $a_{i+1} - a_i = h$, et à remplacer l'intégrale (l'aire de la portion de plan située entre la courbe et l'axe des x) par la somme suivante

$$I_N := \frac{h}{2} f(a) + h(f(a_1) + \cdots + f(a_i) + \cdots + f(a_N)) + \frac{h}{2} f(b)$$

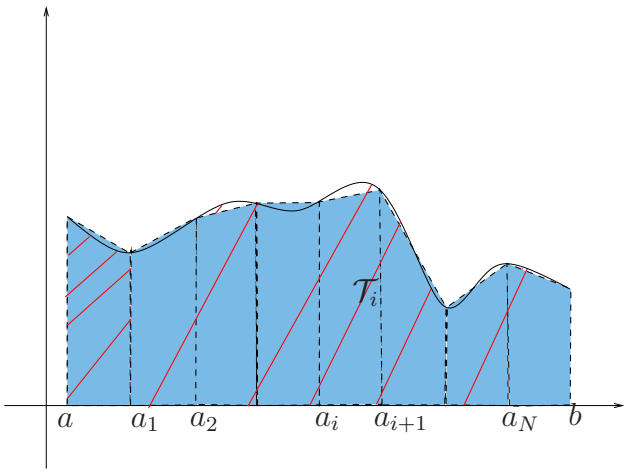


FIGURE 8.1 – calcul approché par la formule des trapèzes

Réécrivons la formule comme

$$\begin{aligned} S_N := \quad & \frac{h}{2}\left(f(a_0)+f(a_1)\right) \\ + \quad & \frac{h}{2}\left(f(a_1)+f(a_2)\right) \\ + \quad & \cdots \\ + \quad & \frac{h}{2}\left(f(a_i)+f(a_{i+1})\right) \\ + \quad & \cdots \\ + \quad & \frac{h}{2}\left(f(a_N)+f(a_{N+1})\right) \end{aligned}$$

La quantité $\frac{h}{2}\left(f(a_i)+f(a_{i+1})\right)$ représente l'aire $\mathcal{A}_i$ du trapèze $\mathcal{T}_i$. On a alors

$$S_N = \mathcal{T}_0 + \mathcal{T}_1 + \cdots \mathcal{T}_i + \cdots \mathcal{T}_N. \tag{8.1}$$

Maintenant pourquoi garder un partage équidistant ? On peut avoir avantage à généraliser la formule (8.1), avec des points distribués différemment, $a_{i+1} - a_i = h_i$, et $\mathcal{A}_i = \frac{h_i}{2}\left(f(a_i)+f(a_{i+1})\right)$.

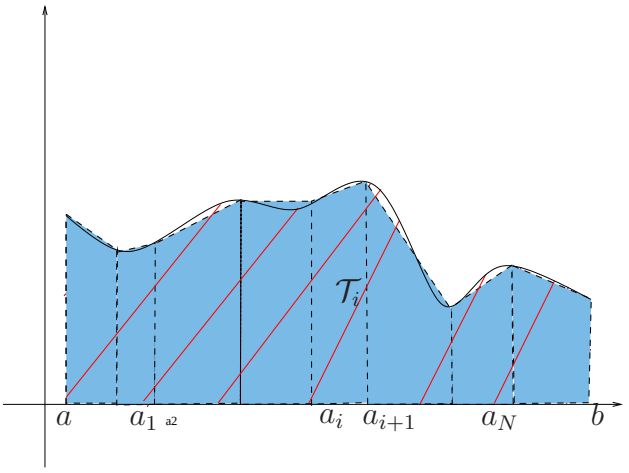


FIGURE 8.2 – calcul approché par la formule des trapèzes, choix des points non équidistants

On connait également les sommes de Riemann :

$$\sum_{j=0}^{N}(a_{j+1}-a_j)f(b_j), \quad b_j \in [a_j, a_{j+1}]$$

L'intégrale de Riemann d'une fonction réglée est définie comme la limite de telles sommes.

Questions

1. Peut-on évaluer l'erreur en fonction de h ?
2. Peut-on en trouver d'autres ?
3. Peut-on les comparer ?

8.1 Formules de quadrature élémentaires

Ce sont les formules qui permettent de calculer dans un sous-intervalle. Reprenons la formule

$$I_j = \int_{a_j}^{a_{j+1}} f(x)\,dx \sim \frac{h_j}{2}\left(f(a_j) + f(a_{j+1})\right)$$

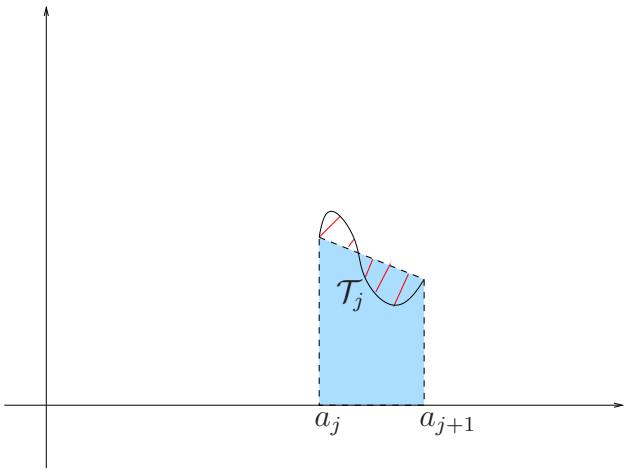


FIGURE 8.3 – calcul élémentaire par la formule des trapèzes

On voit bien que si f est affine sur l'intervalle, les deux quantités coincident. D'où une autre façon d'obtenir la formule élémentaire. On remplace f par son polynôme d'interpolation de degré ≤ 1 sur (a_j, a_{j+1}).

$$f(x) = p_1(x) + \frac{1}{2}f"(\zeta_x)(x - a_j)(x - a_{j+1}), \quad \zeta_x \in]a_j, a_{j+1}[$$

et on écrit

$$I_j = \int_{a_j}^{a_{j+1}} p_1(x)\,dx + \frac{1}{2}\int_{a_j}^{a_{j+1}} f"(\zeta_x)(x - a_j)(x - a_{j+1})\,dx$$

On applique la formule de la moyenne au dernier terme, et on obtient

$$I_j = \int_{a_j}^{a_{j+1}} p_1(x)\,dx + \frac{1}{2}f"(\zeta_j)\int_{a_j}^{a_{j+1}} (x-a_j)(x-a_{j+1})\,dx = \int_{a_j}^{a_{j+1}} p_1(x)\,dx - \frac{h_j^3}{12}f"(\zeta_j)$$

D'autre part $p_1 = \dfrac{1}{h_j}(f(a_{j+1})(x-a_j) - f(a_j)(x-a_{j+1}))$, et

$$\int_{a_j}^{a_{j+1}} p_1(x)\,dx = \frac{h_j}{2}(f(a_j)+f(a_{j+1}))$$

Méthode des trapèzes $\quad I_j = \dfrac{h_j}{2}(f(a_j)+f(a_{j+1})) - \dfrac{h_j^3}{12}f"(\zeta_j)$

Si nous interpolons dans $\mathbf{P}_0$, nous obtenons les 3 formules, suivant que nous interpolons à gauche, à droite ou au point milieu

Méthode	Formule	Erreur
formule des rectangles à gauche	$I_j \sim h_j f(a_j)$	$\dfrac{h_j^2}{2}f'(\zeta_j)$
formule des rectangles à droite	$I_j \sim h_j f(a_{j+1})$	$\dfrac{h_j^2}{2}f'(\zeta_j)$
formule du point milieu	$I_j \sim h_j f(\dfrac{a_j+a_{j+1}}{2})$	$\dfrac{h_j^3}{24}f"(\zeta_j)$

La méthode de Simpson, utilise l'interpolation dans $\mathbf{P}_2$ aux points a_j, a_{j+1}, et $\frac{a_j+a_{j+1}}{2}$. On démontre que l'on a

Méthode de Simpson $\quad I_j = h_j\left(\dfrac{1}{6}f(a_j) + \dfrac{2}{3}f(\dfrac{a_j+a_{j+1}}{2}) + \dfrac{1}{6}f(a_{j+1})\right) - \dfrac{h_j^5}{2880}f^{(4)}(\zeta_j)$

On note sur cette formule qu'elle est en fait exacte pour des polynômes de degré inférieur ou égal à 3.

On appelle formules de Newton-Cotes toutes les formules qu'on obtient de cette manière. Pour systématiser on fait le changement de variable dans I_j :

$$\begin{aligned} [-1,1] &\to [a_j, a_{j+1}] \\ y &\mapsto x = \frac{a_j, a_{j+1}}{2} + \frac{h_j}{2}y \end{aligned}$$

et donc

$$I_j = \frac{h_j}{2}\int_{-1}^{1} f(\frac{a_j, a_{j+1}}{2} + \frac{h_j}{2}y)\,dy.$$

On notera $\varphi_j(y) = f(\frac{a_j,a_{j+1}}{2} + \frac{h_j}{2}y)$.

On se donne des points $\tau_i = -1 + 2i/n$. Pour les formules de Newton-Cotes fermées, i varie de 0 à n. Pour les formules ouvertes i varie de 1 à $n-1$. Commençons par les formules fermées. On écrit pour tout f dans $[-1, 1]$,

$$\int_{-1}^{1} \varphi(y)\, dy = \sum_{i=0}^{n} \omega_i \varphi(\tau_i) + E(\varphi).$$

Avec $n + 1$ coefficients à déterminer, on peut réclamer que la formule soit exacte dans $\mathbf{P}_n$.

Théorème 8.1 – *Il existe une et une seule formule de quadrature exacte dans $\mathbf{P}_n$. Les poids sont donnés par*

$$\omega_i = \int_{-1}^{1} \ell_i(y)\, dy$$

– *Si n est pair, la formule est exacte dans $\mathbf{P}_{n+1}$.*

–

$$E(\varphi) = \begin{cases} \dfrac{\varphi^{(n+2)}(\xi)}{(n+2)!} \displaystyle\int_{-1}^{1} y\Pi_{n+1}(y)\, dy & \text{si } n \text{ est pair,} \\ \dfrac{\varphi^{(n+1)}(\xi)}{(n+1)!} \displaystyle\int_{-1}^{1} \Pi_{n+1}(y)\, dy & \text{si } n \text{ est impair,} \end{cases}$$

Démonstration Prenons pour φ le polynôme d'interpolation de Lagrange au point τ_i. On en déduit les coefficients par la formule précédente, ce qui donne aussi l'unicité. Le deuxième point relève de considérations de parité. Pour l'erreur on écrit pour n impair la formule d'erreur pour le polynôme d'interpolation de φ aux points τ_i :

$$\varphi(y) = p_n(y) + \frac{\varphi^{(n+1)}(\zeta_y)}{(n+1)!}\Pi_{n+1}(y).$$

Il faut alors intégrer sur $[-1, 1]$, utilisant le fait que $\int_{-1}^{1} p_n(y)\, dy = \sum_{i=0}^{n} \omega_i \varphi(\tau_i)$. L'intégration précise est alors un peu difficile, voir dans les livres. ■

Le cas $n = 1$ correspond à la formule des trapèzes, le cas $n = 2$ à la formule de Simpson.

Les formules de Newton-Cotes ouvertes ne sont utilisées que dans le cas du point-milieu.

On définit l'ordre r des formules de Newton-Cotes comme le plus grand entier tel que la formule est exacte pour $f \in \mathbf{P}_{r-1}$.

Théorème 8.2 *Si n est pair, $r = n+1$, Si n est impair, $r = n$.*

Méthode	nombre de points	ordre
rectangles	1	1
point-milieu	1	2
trapèzes	2	2
Simpson	3	4

8.2 Méthode composite

On recolle maintenant les intégrales élémentaires. Pour cela les points $\xi_{j,i}$ se déduisent tous des points $\tau_i = -1 + 2i/n$ au moyen de la transformation affine et

$$I = \sum_{j=0}^{N} \frac{h_j}{2} \left(\sum_{i=0}^{n} \omega_i f(\xi_{j,i}) + E(\varphi_j) \right)$$

avec $\sum_{i=0}^{n} \omega_i = 2$. L'erreur globale $E(f)$ est donc

$$E(f) = \sum_{j=0}^{N} \frac{h_j}{2} \begin{cases} \left(\frac{h_j}{2}\right)^{n+2} \dfrac{f^{(n+2)}(\xi_j)}{(n+2)!} \displaystyle\int_{-1}^{1} y \Pi_{n+1}(y)\, dy & \text{si } n \text{ est pair,} \\ \left(\frac{h_j}{2}\right)^{n+1} \dfrac{f^{(n+1)}(\xi_j)}{(n+1)!} \displaystyle\int_{-1}^{1} \Pi_{n+1}(y)\, dy & \text{si } n \text{ est impair,} \end{cases}$$

Que se passe t-il pour la méthode des trapèzes ? L'erreur s'écrit

$$E(f) = -\sum_{j=0}^{N} \frac{h_j^3}{12} f"(\zeta_j), \quad \zeta_j \in]a_j, a_{j+1}[$$

Si $h_j \equiv h$, on peut montrer qu'il existe $\eta \in]a, b[$ tel que

$$\boxed{\int_a^b f(x)\, dx = \frac{h}{2} f(a) + h(f(a_1) + \cdots + f(a_i) + \cdots + f(a_N)) + \frac{h}{2} f(b) - \frac{(b-a)^2 h^2}{12} f"(\eta)}$$

Théorème 8.3 *On suppose que la formule de quadrature élémentaire est d'ordre r et que $f \in \mathcal{C}^{r+1}([a,b])$. Soit $h = \sup h_j$. Alors*

$$\left| \int_a^b f(x)\, dx - \sum_{j=0}^{N} \frac{h_j}{2} \sum_{i=0}^{n} \omega_i f(\xi_{j,i}) \right| \leq C h^{r+1} \sup_{[a,b]} |f^{(r+1)}(x)|$$

Mise en œuvre :

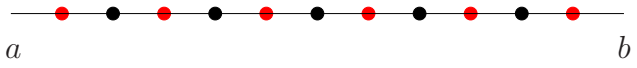


FIGURE 8.4 – N œuds de la formule de quadrature

Supposons que pour le pas h_N les nœuds intérieurs soient les ronds noirs du dessin, $a + i * h_N$. Pour le pas $h_{N+1} = h_N/2$, on ajoute les points rouges.

$$S_{N+1} = S_N/2 + h_{N+1}(f(a+h_{N+1}) + \cdots + f(a+3h_{N+1}) + \cdots + f(a+(2N-1)h_{N+1})$$

```
 function T = trapez(f,a, b, tol);
% TRAPEZ(f,a, b, tol) tries to integrate int_a^b f(x) dx
% to a relative tolerance tol using the composite
% trapezoidal rule.
%
h = b - a; s = (f(a) + f(b)) / 2;
tnew = h * s; zh = 1; told = 2*tnew;
while abs (told - tnew) > tol * abs (tnew),
told = tnew; zh = 2 * zh;
h = h / 2;
s = s + sum(f(a + [1:2:zh]*h));
tnew = h * s;
end;
T = tnew;
```

Qu'en est-il pour Simpson ? Les extrémités sont affectées du poids 1/6, les points noirs du poids 2/6 et les points rouge du poids 4/6. Mise en œuvre :

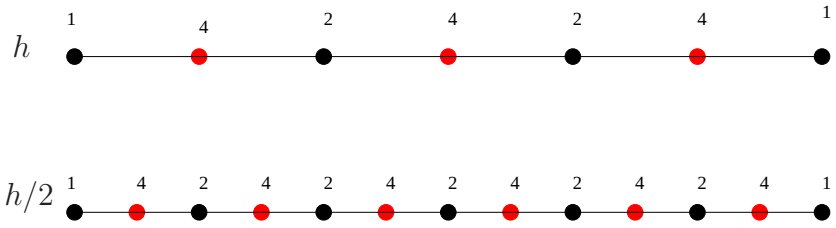


FIGURE 8.5 – N œuds de la formule de quadrature avec les poids pour Simpson

On écrit

$$S(h) = \frac{h}{6}(s_1 + 2s_2 + 4s_4)$$

Notons s_j^{new} les nouvelles valeurs, pour $h_1 = h/2$.

$$s_1^{new} = s_1, \quad s_2^{new} = s_2 + s_4, \quad s_4^{new} = f(a + h_1/2) + f(a + 3h_1/2) + \cdots$$

On écrit alors

```
function S = simpson (f,a,b,tol);
% SIMPSON (f,a, b,tol) tries to integrate int_a^b f(x) dx
% to a relative tolerance tol using the composite
% Simpson rule.
%
h = (b-a)/2; s1 = f(a) + f(b); s2 = 0;
s4 = f(a + h); snew = h*(s1 + 4 * s4)/6
zh = 2; sold=2*snew;
while abs(sold-snew)>tol*abs(snew),
sold = snew; zh = 2 * zh; h = h / 2;
s2 = s2 + s4;
s4 = sum(f(a +[1:4:zh]*h));
snew = h*(s1 + 2*s2 + 4*s4)/6;
end
S = snew;
```

Résultats pour le calcul de

$$\int_0^1 \frac{x\,e^x}{(x+1)^2}\,dx = \frac{e}{e-2} = 0.3591409142...$$

i	h_i	$T(h_i)$	$\frac{T(h_i)-I}{T(h_{i-1})-I}$	$S(h_i)$	$\frac{S(h_i)-I}{S(h_{i-1})-I}$
0	1	0.339785228		0.357516745	
1	$\frac{1}{2}$	0.353083866	0.31293	0.358992305	0.09149
2	$\frac{1}{4}$	0.357515195	0.26840	0.359130237	0.07184
3	$\frac{1}{8}$	0.358726477	0.25492	0.359140219	0.06511
4	$\frac{1}{16}$	0.359036783	0.25125	0.359140870	0.06317
5	$\frac{1}{32}$	0.359114848	0.25031	0.359140911	0.06267
6	$\frac{1}{64}$	0.359134395	0.25007	0.359140914	0.06254

Erreur divisée par 4

Erreur divisée par 16

FIGURE 8.6 – Comparaison des méthodes des trapèzes et de Simpson

8.3 Méthode de Gauss

Jusqu'ici nous nous sommes fixés les points et nous avons cherché les poids pour obtenir un ordre. On peut chercher à optimiser aussi les points pour maximiser encore l'ordre de la formule de quadrature. On se place sur $[-1, 1]$ et on cherche les points τ_j et les poids ω_j pour minimiser

$$\int_{-1}^1 f(x)\,dx - \sum_{j=0}^{n} \omega_j f(\tau_j)$$

Théorème 8.4 *Il existe un choix et un seul des points τ_j et des poids ω_j de sorte que la méthode soit d'ordre $r = 2n + 1$. Les points τ_j sont les zéros du polynôme de Legendre L_{n+1}. Les poids sont donnés par*

$$\omega_j = \int_{-1}^{1} \ell_j(x)dx,$$

l'erreur est donnée par

$$E(f) = C\frac{f^{(2n+2)}(\xi)}{(2n+2)!}$$

Exemple, $n = 1$. Si on connait la formule de récurrence

$$(n+1)L_{n+1}(x) = (2n+1)xL_n(x) - n\,L_{n-1}(x), \quad L_0 = 1,\ L_1 = x.$$

on a $L_0 = 1$, $L_1 = x$, et $L_2 = 3x^2/2 - 1/2$. Les nœuds sont donc $\pm 1/\sqrt{3}$. les poids sont égaux à 1.

$$\boxed{\int_{-1}^{1} f(x)\,dx \sim f(\frac{1}{\sqrt{3}}) + f(-\frac{1}{\sqrt{3}})}$$

Cas général, calcul des nœuds : on réécrit la formule de récurrence sous la forme

$$xL_n(x) = \frac{n+1}{2n+1}L_{n+1}(x) + \frac{n}{2n+1}L_{n-1}(x).$$

et on constate que x est valeur propre d'une matrice tridiagonale :

Calcul des coefficients : Ecrivons $L_{n+1}(x) = a_{n+1}\prod_{j=0}^{n}(\tau_j)$. On a alors $L'_{n+1}(\tau_i) = a_{n+1}\prod_{\substack{j=0\\ j\neq i}}^{n}(\tau_j)$, et on a

$$\ell_i(x) = \frac{L_{n+1}(x)}{L'_{n+1}(\tau_i)(x-\tau_i)}$$

et

$$\omega_i = \frac{1}{L'_{n+1}(\tau_i)} \int_{-1}^{1} \frac{L_{n+1}(x)}{(x-\tau_i)} dx$$

Théorème 8.5 *Definissons* $\Phi_0 = 0$, *et*

$$\Phi_i(t) = \int_{-1}^{1} \frac{L_i(x) - L_i(t)}{(x-t)} dx$$

Alors les Φ_i *satisfont la même relation de récurrence que les* L_i *et on a*

$$\omega_i = \frac{\Phi_i(\tau_i)}{L'_{n+1}(\tau_i)}$$

Chapitre 9

Calcul de vecteurs propres et valeurs propres

9.1 Généralités, outils matriciels

9.1.1 Matrices de Householder

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

$$H(v) = I - 2\frac{vv^*}{v^*v} \tag{9.1}$$

$H(v)$ 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 9.1 *Pour tout x dans $\mathbb{C}^n$, on a $(x - H(v)x)^*(x + H(v)x) = 0$.*

Lemme 9.2 *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*

$$v = \lambda(x - \omega y) \text{ et } \bar{\omega} y^* x = \omega x^* y \tag{9.2}$$

On en déduit :

Proposition 9.1 *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*

$$H(v)x = \omega y \tag{9.3}$$

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.

9.1.2 Quotients de Rayleigh

Définition 9.1 *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 9.1 *On a*

$$\lambda_n = \inf_{x \neq 0} r_A(x), \quad \lambda_1 = \sup_{x \neq 0} r_A(x)$$
$$\lambda_k = \sup_{dim\, V = k} \inf_{x \in V - 0} r_A(x), \lambda_k = \inf_{dim\, W = n-k+1} \sup_{x \in W - 0} r_A(x), \ 1 \leq k \leq n$$

9.1.3 Conditionnement d'un problème de valeurs propres

Théorème 9.2 *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 $1 \leq k \leq n$,*

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

9.2 Décompositions

9.2.1 Décomposition QR

Théorème 9.3 *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

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

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 9.1, il existe une matrice de Householder $H(v^{(1)})$ et un nombre complexe ω_1 de module 1 tels que

$$H(v^{(1)})\, a^1 = \omega_1 \|a^1\| e_1 \tag{9.4}$$

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

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

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

$$A^{(k+1)} = H^{(k)} A^{(k)}, k = 1, \cdots, n. \tag{9.5}$$

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 ${}^t Q = H^{(n-1)} \cdots H^{(1)}$, Q est une matrice orthogonale et $A = {}^t Q^{-1} A^{(n)} = Q A^{(n)}$. On a ainsi construit les matrices Q et R.

Remarque 9.1 *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.

9.2.2 Tridiagonalisation d'une matrice symétrique

Théorème 9.4 *Soit $A \in \mathcal{M}_n(\mathbb{R})$ symétrique. Alors il existe une matrice P orthogonale telle que ${}^t PAP$ 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.

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

9.3.1 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 :

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

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 9.5 *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 $A^{(1)} = {}^tR_{p_1,q_1}(\theta_1) A R_{p_1,q_1}(\theta_1)$ 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 9.6 *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, $A^{(k)} = {}^tR_{p_k,q_k}(\theta_k) \cdots {}^tR_{p_1,q_1}(\theta_1) A R_{p_1,q_1} \cdots R_{p_k,q_k}(\theta_k) = {}^tO^{(k)} A O^{(k)}$, 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.

9.3.2 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 9.3 *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 9.7 *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.

9.4 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 9.8 *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}.

Sommaire