Université Abdelmalek Essaadi
Ecole Nationale des Sciences Appliquées
Al Hoceima, Maroc

Méthodes d'Analyse Numérique pour l'Année Préparatoire II

–Cours de Mathématiques Appliquées–
Analyse Numérique, Analyse Numérique matricielle

Mohamed ADDAM
Professeur de Mathématiques

École Nationale des Sciences Appliquées d'Al Hoceima
–ENSAH–
Année Universitaire 2020/2021
addam.mohamed@gmail.com
m.addam@uae.ac.ma

04 Avril 2021

Table des matières

Chapitre 1

Analyse numérique matricielle

1.1 Spectre et rayon spectral d'une matrice, Matrice positive

Soit $A = (a_{i,j})_{1\leq i,j\leq n}$ une matrice carrée de taille $n \times n$.

1. La **trace** de A est $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{i,i}$.
2. Les valeurs propres de A sont les n racines réelles ou complexes $(\lambda_i)_{1\leq i\leq n}$ du polynôme caractéristique P de A. Le **spectre** de A, noté $\mathrm{Sp}(A)$ est l'ensemble de tous les valeurs propres de A :
$$\mathrm{Sp}(A) = \{\lambda_i : \ 1 \leq i \leq n\}$$
3. La matrice A est **diagonale** si $a_{i,j} = 0$ pour $i \neq j$, on la note
$$A = \mathrm{diag}(a_{ii}) = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{nn}).$$

On rappelle les propriétés suivantes :

1. $\mathrm{tr}(A) = \sum_{i=1}^{n} \lambda_i$, dét$(A) = \prod_{i=1}^{n} \lambda_i$.
2. $\mathrm{tr}(AB) = \mathrm{tr}(BA)$, $\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B)$.
3. det$(AB) =$ det$(BA) =$ det(A)det(B).

Définition 1.1.1 *On appelle le rayon spectral de la matrice A, noté $\varrho(A)$, le nombre réel positif*
$$\varrho(A) = \max\{|\lambda_i| : \ 1 \leq i \leq n\}$$

Définition 1.1.2 *Une matrice A est*

1. **Symétrique** *si A est réelle et $A = A^T$;*
2. **hermitienne** *si $A = A^*$;*
3. **Orthogonale** *si A est réelle et $AA^T = A^TA = I$;*
4. **Unitaire** *si $AA^* = A^*A = I$;*
5. **Normale** *si $AA^* = A^*A$.*

une matrice A est dite **singulière** *si elle n'est pas inversible.*

Propriété 1.1.1 *Si A et B sont deux matrices inversibles, alors $(AB)^{-1} = B^{-1}A^{-1}$, $(A^T)^{-1} = (A^{-1})^T$, $(A^*)^{-1} = (A^{-1})^*$.*

1.1.1 Matrice positive et matrice définie positive

Définition 1.1.3 *Soit A une matrice*

1. *La matrice A est* **définie positive** *si*

$$(Ax, x) > 0, \qquad \forall x \in E - \{0_E\}$$

   et $(Ax, x) = 00, \quad \Leftrightarrow \quad x = 0_E$.

2. *La matrice A est* **positive** *où* **semi-définie positive** *si*

$$(Ax, x) \geq 0, \qquad \forall x \in E - \{0\}$$

Théorème 1.1.1 *Une matrice hermitienne A est définie positive (resp. positive), si et seulement si toutes ses valeurs propres sont > 0 (resp. ≥ 0).*

Démonstration. soit A une matrice hermitienne et $x \neq 0$ un vecteur dans E.

$$(Ax, x) = \lambda(x, x) = \lambda \|x\|_E$$

□

1.1.2 Valeurs singulières d'une matrice

Définition 1.1.4 *Soit A une matrice carrée. On appelle* **valeurs singulières** *d'une matrice carrée A, les racines carrées positives des valeurs propres de la matrice hermitienne A^*A (ou A^TA si la matrice A est réelle).*

Remarque 1.1.1 *Les valeurs propres de la matrice hermitienne A^*A sont toujours ≥ 0 puisque*

$$A^*Ax = \lambda x, \quad x \neq 0 \quad \Rightarrow \quad (A^*Ax, x) = \lambda \|x\|_E,$$

les valeurs singulières sont toutes > 0 si et seulement si la matrice A est **inversible**

1.1.3 Décomposition en valeurs singulières

Définition 1.1.5 *Soit A une matrice carrée. On appelle* **valeurs singulières** *d'une matrice carrée A, les racines carrées positives des valeurs propres de la matrice hermitienne A^*A (ou A^TA si la matrice A est réelle).*

Toute matrice peut être réduite sous forme diagonale en la multipliant à droite et à gauche par des matrices unitaires bien choisies. Plus précisément on a le résultat suivant :

Propriété 1.1.2 *Soit $A \in \mathbb{C}^{m\times n}$. Il existe deux matrices unitaires $U \in \mathbb{C}^{m\times m}$ et $V \in\in \mathbb{C}^{n\times n}$ telles que*

$$U^*AV = diag(\sigma_1, \sigma_2, \ldots, \sigma_p) \in \mathbb{C}^{m\times n}, \quad p = \min(m, n) \tag{0.1}$$

et $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p \geq 0$.
La relation (0.1) est appelée **décomposition en valeurs singulières** *(SVD) de A et les scalaire (σ_i) sont appelés* **valeurs singulières** *de A.*
♣ *Si A est une matrice réelle, U et V sont aussi des matrices réelles et on peut remplacer U^* par U^T.*

Les valeurs singulières sont caractérisées par

$$\sigma_i = \sqrt{\lambda_i}, \quad \text{où} \quad \lambda_i \in \text{Sp}(A^*A), \quad i = 1, \ldots, p. \tag{0.2}$$

– Si $A \in \mathbb{C}^{n\times n}$ est une matrice hermitienne de valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$, alors d'après (0.2) les valeurs singulières de A coïncident avec les modules des valeurs propres de A. En effet, puisque $AA^* = A^2$, on a $\sigma_i = \sqrt{\lambda_i^2} = |\lambda_i|$ pour $i = 1, \ldots, n$.

– Si

$$\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r \geq \sigma_{r+1} = \ldots = \sigma_p = 0,$$

alors le rang de A est r (rang$(A) = r$), le noyau de A est le sous-espace vectoriel engendré par les vecteurs colonnes de V (Ker$(A) = \overline{\{v_{r+1}, \ldots, v_n\}}$), et l'image de A est le sous-espace vectoriel engendré par les vecteurs colonnes de U (Im$(A) = \overline{\{u_1, \ldots, u_r\}}$).

1.1.4 Pseudo-inverse de Moore-Penrose

Définition 1.1.6 *Supposons $A \in \mathbb{C}^{m\times n}$ soit de rang r et qu'elle admette une décomposition en valeurs singulières du type $U^*AV = \Sigma$. La matrice $A^\dagger = V\Sigma^\dagger U^*$ est appelée matrice* **pseudo-inverse** *de* **Moore-Penrose**, *où*

$$\Sigma^\dagger = \textit{diag}\left(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0\right) \in \mathbb{C}^{m\times n}, \quad p = \min(m, n) \tag{0.3}$$

la matrice $A^\dagger$ est aussi appelée **matrice inverse généralisée** *de A, on a*

$$A^\dagger = \begin{cases} (A^TA)^{-1}A^T, & \textit{si} \quad \textit{rang}(A) = n < m, \\ A^{-1}, & \textit{si} \quad \textit{rang}(A) = n = m. \end{cases}$$

Exemple 1.1.1 *On peut se demander si l'inverse de Moore-Penrose peut être utilisée dans des cas pratique. On cherche à estimer une valeur en x_0 à partir de deux valeurs situées au même point x_1. On considère le système simple*

$$\begin{pmatrix} a & a \\ a & a \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} b \\ b \end{pmatrix}$$

où a et b sont des données réelles.

La matrice A étant singulière, on va recourir à l'inverse généralisée de Moore-Penrose. Soit

$$AA^T = A^TA = \begin{pmatrix} 2a^2 & 2a^2 \\ 2a^2 & 2a^2 \end{pmatrix} = C$$

On a

$$\begin{aligned} \textit{dét}(C) = 0 \quad &\Rightarrow \quad \lambda_2 = 0; \\ \textit{tr}(C) = 4a^2 \quad &\Rightarrow \quad \lambda_1 = 4a^2, \end{aligned}$$

et une matrice de vecteurs propres normés

$$Q = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

La solution du système au sens de l'inverse généralisée est

$$w = A^\dagger b = Q\Sigma^\dagger Q^T b = \begin{pmatrix} \frac{b}{\sqrt{2}c} \\ \frac{b}{\sqrt{2}c} \end{pmatrix} = \begin{pmatrix} \frac{b}{2a} \\ \frac{b}{2a} \end{pmatrix}$$

1.2 Normes matricielles

soit $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$.

Définition 1.2.1 *Une* **norme matricielle** *est une application* $\|\,.\,\| : \mathbb{K}^{(m\times n)} \longrightarrow [0, +\infty[$ *telle que :*

i) $\|A\| \geq 0, \quad \forall A \in \mathbb{K}^{(m\times n)}$ *et* $\|A\| = 0$ *si et seulement si* $A = 0$ *;*

ii) $\|\alpha A\| = |\alpha|\|A\|, \quad \forall A \in \mathbb{K}^{(m\times n)}, \forall \alpha \in \mathbb{K}$ *(Propriété d'homogénéité) ;*

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

Définition 1.2.2 *On dit que la norme matricielle* $\|\,.\,\|$ *est* **compatible** *ou* **consistante** *avec la norme vectorielle* $\|.\|$ *si*

$$\|Ax\| \leq \|A\|\|x\|, \quad \forall A \in \mathbb{K}^{(m\times n)} \quad \forall x \in \mathbb{K}^n.$$

Plus généralement, on dit que trois normes, toutes notées $\|\,.\,\|$ et respectivement définies sur $\mathbb{K}^m$, $\mathbb{K}^n$, et $\mathbb{K}^{(m\times n)}$, sont **consistantes** si $\forall x \in \mathbb{K}^n$, $\forall y \in \mathbb{K}^m$ et $A \in \mathbb{K}^{(m\times n)}$ tels que $Ax = y$, on a $\|y\| \leq \|A\|\|x\|$. ■
Pour qu'une norme matricielle soit intéressante dans la pratique, on demande généralement qu'elle possède la propriété suivante :

Définition 1.2.3 *On dit qu'une norme matricielle* $\|\,.\,\|$ *est* **sous-multiplicative** *si* $\forall A \in \mathbb{K}^{(n\times m)}$, $\forall B \in \mathbb{K}^{(m\times q)}$

$$\|AB\| \leq \|A\|\|B\|. \tag{0.4}$$

■

Exemple 1.2.1 *Soit* $\|\,.\,\|_\blacktriangle$ *l'application définie par*

$$\|A\|_\blacktriangle = \max_{i,j} |a_{ij}|.$$

On considère deux matrices A *et* B *données par*

$$A = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} = B.$$

On peut vérifier facilement que $\|\,.\,\|_\blacktriangle$ *est une norme et qu'elle ne satisfait pas l'inégalité (0.4) puisque*

$$2 = \|AB\|_\blacktriangle > \|A\|_\blacktriangle\|B\|_\blacktriangle = 1.$$

D'où la norme $\|\,.\,\|_\blacktriangle$ *n'est pas une norme sous-multiplicative.*

Norme de Frobenius

La norme

$$\|A\|_F = \left(\sum_{j=1}^{m} \sum_{i=1}^{n} |a_{ij}|^2 \right)^{1/2} = \sqrt{\text{tr}(A^* A)}$$

est une norme matricielle appelée **norme de Frobenius** (ou norme euclidienne dans $\mathbb{C}^{(n\times n)}$ et elle est compatible avec la norme vectorielle euclidienne $\| \, . \, \|_2$. En effet,

$$\|Ax\|_2^2 = \sum_{i=1}^{m} \left| \sum_{j=1}^{n} a_{ij} x_j \right|^2 \leq \sum_{i=1}^{m} \left(\sum_{j=1}^{n} |a_{ij}|^2 \sum_{j=1}^{n} |x_j|^2 \right) = \|A\|_F^2 \|x\|_2^2.$$

On peut remarquer que pour cette norme, on a $\|I_n\|_F = \sqrt{n}$.

Théorème 1.2.1 *Soit $\|.\|$ une norme vectorielle. La fonction*

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{0.5}$$

est une norme matricielle. On l'appelle **norme matricielle subordonnée** *ou associée à la norme vectorielle $\|.\|$. On l'appelle aussi parfois norme matricielle naturelle, ou encore norme matricielle induite par la norme vectorielle $\|.\|$.*

Démonstration. Commençons par remarquer que (0.5) est équivalente à

$$\|A\| = \sup_{\|x\|=1} \|Ax\|. \tag{0.6}$$

Pour tout $x \neq 0$, on peut définir un vecteur unitaire $u = \frac{x}{\|x\|}$, de sorte que (0.5) s'écrive

$$\|A\| = \sup_{\|u\|=1} \|Au\| = \|Aw\|, \quad \|w\| = 1.$$

cela étant, vérifions que (0.5) est effectivement une norme, en utilisant les conditions d'une norme matricielle de la définition (1.2.1). □

Exemples de normes remarquables

D'autres exemples de normes remarquables, il s'agit de normes matricielles subordonnées fournies par les p-normes :

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} \tag{0.7}$$

La 1-norme et la norme infinie se calculent facilement :

$$\|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{m} |a_{ij}|, \tag{0.8}$$

$$\|A\|_\infty = \max_{1 \leq i \leq m} \sum_{j=1}^{n} |a_{ij}|, \tag{0.9}$$

On a les propriétés suivantes :

1. $\|A\|_1 = \|A^T\|_\infty$
2. Si A est matrice symétrique réelle, alors $\|A\|_1 = \|A\|_\infty$. ■

La 2-norme ou **norme spectrale** mérite une discussion particulière vu son intérêt pour des applications pratiques.

Théorème 1.2.2 *Soit σ_1 la plus grande valeur singulière de A. Alors*

$$\|A\|_2 = \sqrt{\varrho(A^*A)} = \sqrt{\varrho(AA^*)} = \sigma_1.$$

En particulier, si A est hermitienne (ou symmétrique réelle), alors

$$\|A\|_2 = \varrho(A),$$

tandis que si A est unitaire alors $\|A\|_2 = \varrho(A) = 1$.

Démonstration. Puisque A^*A est hermitienne, alors il existe une matrice unitaire U telle que

$$U^*A^*AU = \text{diag}(\mu_1, \mu_2, \ldots, \mu_n)$$

où μ_i sont les valeurs propres positive de A^*A. Soit $y = U^*x$, alors

$$\begin{aligned} \|A\|_2 &= \sup_{x\neq 0} \sqrt{\frac{(A^*Ax, x)}{(x, x)}} = \sup_{x\neq 0} \sqrt{\frac{(U^*A^*AUy, y)}{(y, y)}} \\ &= \sup_{x\neq 0} \sqrt{\sum_{i=1}^n \mu_i |y_i|^2 / \sum_{i=1}^n |y_i|^2} = \sqrt{\max_{1\leq i\leq n} |\mu_i|} = \sigma_1. \end{aligned}$$

Si A est hermitienne, les mêmes considérations s'appliquent directement à A. En fin si A est unitaire

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^*Ax) = \|x\|_2^2$$

et donc $\|A\|_2 = 1$. Enfin, si A est unitaire □

Exercice 1.2.1

1. *Soit B une matrice carrée. Montrer que les conditions suivantes sont équivalentes :*
   - *(a)* $\lim_{k\to+\infty} B^k = 0$ *;*
   - *(b)* $\lim_{k\to+\infty} B^k x = 0$ *pour tout vecteur x ;*
   - *(c)* $\varrho(B) < 1$ *;*
   - *(d)* $\|B\| < 1$ *pour au moins une norme matricielle subordonnée $\|\,.\,\|$.*
2. *Soit B une matrice carrée, et $\|\,.\,\|$ une norme matricielle quelconque. Montrer que*

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

1.3 Conditionnement

Les sources des erreurs numériques sont :

1. Erreur d'arrondi dues à la représentation des réels sur la machine,
2. Erreur sur les données (données qui proviennent d'autres calculs)
3. Erreur de troncature (faites dans les méthodes itératives : on remplace la valeur exacte par une valeur approchée)

1.3.1 Représentation des entiers et des réels sur ordinateur

Représentation des entiers

Soit $a \in \mathbb{N}$,

$$a = \sum_{i=0}^{n-1} a_i 2^i, \quad \text{avec} \quad a_i \in \{0, 1\}$$

la représentation binaire (représentation en base 2) de a est

$$a = a_{n-1}a_{n-2}\ldots a_2 a_1 a_0.$$

Exemple 1.3.1 *1.* $a = 13 = 2^3 + 2^2 + 2^0 = 1 \times 2^3 + 1 \times 2^2 + 0 \times 2^1 + 1 \times 2^0$
a est représenté sur la machine par (1101)

$$13 = (1101)$$

2. $226 = 2^7 + 2^6 + 2^5 + 0 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 2^1 + 0 \times 2^0$
donc $226 = (11100010)$

Le chiffre 0 où 1 représente un **bit**.
bit=*binary disit=chiffre binaire.*

Exemple 1.3.2 *Sur l'ordinateur, les entiers sont représentés par un ensemble de bits.*

1. $a = 13$ *est représenté par 4 bits.*
2. $a = 226$ *est représenté par 8 bits*

Exemple 1.3.3 *Avec une représentation sur 16 bits.*

$$a = a_{15}a_{14}\ldots a_2 a_1 a_0, \quad a_i \in \{0; 1\}$$

le binaire a_{15} désigne le signe $\begin{cases} 1 & \textit{entier négatif}, \\ 0 & \textit{entier positif} \end{cases}$

Exemple 1.3.4 $a = 226$ *est représente sur 16 bits par*

$$a = (0000000011100010)$$

$a = -226$ *est représenté sur 16 bits par*

$$a = (1000000011100010)$$

le plus grand entier qu'on peut représenter sur la machine avec 16 bits est :

$$g = 2^{14} = 2^{13} + \ldots + 2^2 + 2^1 + 1 = 2^{15} - 1 = 32768 - 1 = 32767$$

Remarque 1.3.1 *La somme de deux entiers sur la machine peut donner un nombre négatif*

Exemple 1.3.5 *Mathématiquement* $32767 + 1 = 32768$.
Sur la machine :

$$\begin{array}{rcl} 32767 & = & 0111111111111111 \\ & + & \\ 1 & = & 0000000000000001 \\ \hline 32768 & = & 1000000000000000 \end{array}$$

ce nombre est négatif car le bit de a_{15} est 1.

Représentation des réels

Représentation en vergule flottantes : une opération élémentaire sur des réels est appelée **flop** où **floaling point**.
soit $x \in \mathbb{R}$, x est représenté par

$$x = m \times b^p$$

où m est la montisse, b est la base et p est l'exposant, avec la condition

$$\frac{1}{b} \leq |m| < 1$$

donc $m = 0.d_1 d_2 \dots d_t$ avec $0 \leq d_i < b$ et $d_1 \neq 0$.
t désigne la précision (le nombre des chiffres après la vergule)

$$m = \sum_{i=1}^{t} d_i b^{-i}.$$

Exemple 1.3.6 *En base de* 10 *c'est-à-dire* $b = 10$.
Le réel $x = 455.321$ *est représenté sur la machine avec* $\widetilde{x}$ *où*

$$\widetilde{x} = 0.455321 \underbrace{e3}_{10^3}$$

ici e *signifie l'***exposant**, *et* $t = 6$ *est la précision de 6 chiffres.*

Exercice 1.3.1 *Supposons que* $\ell \leq p \leq u$. *Calculer le plus grand et le plus petit réel.*

Erreurs de représentation

La représentation des réels sur la machine n'est pas exacte.

Exemple 1.3.7 *supposons que* $b = 10$ *et* $t = 3$

1. $y = \frac{1}{3} = 0.333333 \dots 3$ *est représenté par* $\widetilde{y} = 0.333$
2. $y = \frac{3254}{100} = 32.54$ *est représenté par* $\widetilde{y} = 0.325e2$

on ne représente que les valeurs approchées.

Erreurs d'arrondies sur les opérations élémentaires $(+, -, *, /)$

les opérations flottants ne sont pas associatives

$$(\widetilde{x} + \widetilde{y}) + \widetilde{z} \neq \widetilde{x} + (\widetilde{y} + \widetilde{z}).$$

Exemple 1.3.8 $b = 10$, $t = 3$, $x = 10^{-3}$, $y = 1$ *et* $z = -1$.
Mathématiquement on a $x + y + z = 10^{-3}$.
Su la machine on a $\widetilde{x + y} = 1.001 \quad \rightarrow \quad 0.1001e1 = 0.1e1$ *car la précision* $t = 3$.

1. $(\widetilde{x + y}) + \widetilde{z} = 0$
2. $\widetilde{x} + (\widetilde{y} + \widetilde{z}) = 10^{-3} = 0.1e - 2.$

1.3.2 Effet de la représentation des réels et les erreurs d'arrondi sur la résolution de $Ax = b$

Exemple 1.3.9 *soit le système suivant :* $\left\{ \begin{array}{rcl} 3x - 7.0001y & = & 0.9998, \\ 3x - 7y & = & 1 \end{array} \right.$ *admet la solution unique*

$$\left\{ \begin{array}{rcl} x & = & \frac{1}{3}, \\ y & = & \frac{1-0.9998}{0.0001} \end{array} \right.$$

$$A = \begin{pmatrix} 3 & -7.0001 \\ 3 & -7 \end{pmatrix}, \quad b = \begin{pmatrix} 0.9998 \\ 1 \end{pmatrix}$$

supposons qu'on travaille sur une machine avec $t = 4$*, alors* 7.0001 *sera représentée par* $0.70001e1 \rightarrow 0.7e1$.
sur la machine, on a à résoudre le système suivant qui est singulier

$$\left\{ \begin{array}{rcl} 0.3e1x - 0.7e1y & = & 0.9998, \\ 0.3e1x - 0.7e1y & = & 1 \end{array} \right.$$

ce système n'a pas de solution.

$$\widetilde{A} = \begin{pmatrix} 3 & -7 \\ 3 & -7 \end{pmatrix}, \quad \widetilde{b} = \begin{pmatrix} 0.9998 \\ 1 \end{pmatrix}$$

ceci montre qu'une perturbation sur la matrice A *induit une matrice* $\widetilde{A}$ *où le système n'a pas de solution.*
Perturbation sur b **:** *Une perturbation sur* b *peut conduire à des résultats qui ne sont pas justes :*

$$A = \begin{pmatrix} 3 & -7.0001 \\ 3 & -7 \end{pmatrix}, \quad \widetilde{b} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

le système suivant $\left\{ \begin{array}{rcl} 3x - 7.0001y & = & 1, \\ 3x - 7y & = & 1 \end{array} \right.$ *admet la solution unique* $\left\{ \begin{array}{rcl} x & = & \frac{1}{3}, \\ y & = & 0 \end{array} \right.$ *mais elle n'est pas une solution juste.*
Les mauvais résultats obtenus sont dûs au fait que la matrice A *est* **mal conditionnée**.

Définition 1.3.1 *une matrice est* **mal conditionnée** *si une petite perturbation (modification des données) conduit à des résultats différents.*

1^{er} cas : Soit à résoudre le système $Ax = b$. Soit $\triangle b$ la perturbation sur b, on résout donc le système

$$\left\{ \begin{array}{l} A(x + \triangle x) = b + \triangle b, \\ \triangle x : \quad \text{perturbation sur } x \end{array} \right.$$

$x^* = x + \triangle x$ est la solution obtenue après modification de b.

$$\begin{array}{rcl} A(x + \triangle x) = b + \triangle b & & \\ Ax + A\triangle x = b + \triangle b & \Rightarrow & A\triangle x = \triangle b \\ & \Rightarrow & \triangle x = A^{-1}\triangle b \end{array}$$

soit $\| \cdot \|$ une norme matricielle induite :

$$\|\triangle x\| \leq \|A^{-1}\| \|\triangle b\|$$

d'autre part, on a

$$\|b\| = \|Ax\| \leq \|A\|\|x\| \quad \Rightarrow \quad \frac{1}{\|x\|} \leq \frac{\|A\|}{\|b\|}$$

on déduit que

$$\underbrace{\frac{\|\triangle x\|}{\|x\|}}_{\text{erreur relative sur } x} \leq \underbrace{\|A\|\|A^{-1}\|}_{\kappa(A)=\text{cond}(A)} \underbrace{\frac{\|\triangle b\|}{\|b\|}}_{\text{erreur relative sur } b}$$

La quantité $\kappa(A) = \text{cond}(A) = \|A\|\|A^{-1}\|$ s'appelle le conditionnement de $A \geq 1$. **Si $\kappa(A)$ est plus proche de 1 alors, une petite perturbation sur b ($\frac{\|\triangle b\|}{\|b\|}$ très petit) entraine des petites perturbations sur x ($\frac{\|\triangle x\|}{\|x\|}$ très petit)**

2^{eme} cas : Supposons qu'on a une perturbation sur A

$$\begin{cases} (A + \triangle A)(x + \triangle x) = b, \\ \triangle A : \quad \text{perturbation sur } A \end{cases}$$

$$\begin{aligned} Ax + A\triangle x + \triangle A(x + \triangle x) &= b \\ Ax + A\triangle x + \triangle A(x + \triangle x) = b \quad &\Rightarrow \quad \triangle A(x + \triangle x) = -A\triangle x \\ &\Rightarrow \quad \triangle x = -A^{-1}\triangle A(x + \triangle x) \end{aligned}$$

soit $\|\cdot\|$ une norme matricielle induite :

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

on déduit que

$$\underbrace{\frac{\|\triangle x\|}{\|x + \triangle x\|}}_{\text{erreur relative sur } x} \leq \underbrace{\|A\|\|A^{-1}\|}_{\kappa(A)=\text{cond}(A)} \underbrace{\frac{\|\triangle A\|}{\|A\|}}_{\text{erreur relative sur } A}$$

Si $\kappa(A)$ est plus proche de 1, alors une petite perturbation sur A entraine des petites perturbations sur le résultat x.

Définition 1.3.2 *Soit A une matrice.*

1. *La matrice A est* **bien conditionnée** *si $\kappa(A)$ est plus proche de 1.*
2. *La matrice A est* **mal conditionnée** *si $\kappa(A)$ est très grand.*

1.3.3 Propriétés du conditionnement

Soit A une matrice matrice carrée inversible.

1. $\kappa(A) \geq 1$, $\kappa(A) = \kappa(A^{-1})$ et $\kappa(\alpha A) = \kappa(A)$ pour tout $\alpha \in \mathbb{K}$.
2. $\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\sigma_1}{\sigma_n}$ où σ_1 est la plus grande valeur singulière de A et σ_n est la plus petite valeur singulière de A.
3. Le matrices orthogonales sont bien conditionnées, $\kappa_2(A) = 1$.

Exemple 1.3.10

$$A = \begin{pmatrix} 3 & -7.0001 \\ 3 & -7 \end{pmatrix}, \quad A^{-1} = \frac{1}{3.10^{-4}} \begin{pmatrix} -7 & 7.0001 \\ -3 & 3 \end{pmatrix}$$

On a $\kappa_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty = \frac{10.0001 \times 14.0001}{3} \times 10^4$ qui est très grand.

Chapitre 2

Résolution de systèmes : Méthodes directes et méthodes itératives

Soit A une matrice carrée d'ordre n, inversibles à coefficients dans $\mathbb{R}$ et soit b un vecteur à n composantes. l'objectif de ce chapitre est de résoudre le système linéaire $Ax = b$ à n équations.

2.1 Méthodes directes

On appelle **méthode directe** de résolution du système linéaire $Ax = b$, toute méthode permettant d'obtenir la solution en un nombre fini d'opérations arithmitiques élémentaires sur des nombres réels (additions, soustractions, multiplications, divisions) et éventuellement l'extraction des racines carrées.

2.1.1 résolution du système triangulaire

Supposons que A est une matrice triangulaire suppérieure

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & \dots & a_{1,n} \\ 0 & a_{2,2} & a_{2,3} & & a_{2,n} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & 0 & \ddots & \ddots & a_{n-1,n} \\ 0 & \dots & \dots & 0 & a_{n,n} \end{pmatrix}$$

La résolution du système $Ax = b$ s'effectue par (back substitution où backward substitution).
Elle consiste à calculer x_i en fonction de $x_n, x_{n-1}, \dots, x_{i+1}$.
A étant une matrice inversible, alors $det(A) = \prod_{i=1}^{n} a_{i,i} \neq 0$ donc $a_{i,i} \neq 0$ pour tout $1 \leq i \leq n$.

- Calculons d'abord $x_n = \dfrac{b_n}{a_{n,n}}$
- On reporte la valeur dans l'équation précédente pour calculer

$$x_{n-1} = \frac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}} = \frac{b_{n-1}a_{n,n} - a_{n-1,n}b_n}{a_{n,n}a_{n-1,n-1}}$$

– Ainsi de suite, plus généralement, on obtient

$$x_i = \frac{b_i - \sum_{j=i+1}^{n} a_{i,j}x_j}{a_{i,i}}, \quad \text{pour} \quad i = n, n-1, \ldots, 1$$

le coût de la résolution est mesuré en nombre d'opérations élémentaires appelé aussi **coût de calcul**. le calcul de x_i ($1 \leq i \leq n$) nécessite $2(n-i)+1$ opérations. On déduit que le coût de calcul des x_i est

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

opérations.

2.1.2 Principe des méthodes directes

Les résolutions directes consistent à transformer le système $Ax = b$ en le système $Rx = c$ où R est une matrice triangulaire supérieure qu'on sait résoudre facilement.
On multiplie A par des matrices bien choisies, soit M le produit des ces matrices, on transforme alors le système $Ax = b$ en un nouveau système

$$MAx = Mb = c,$$

On détermine M de telle sorte que MA soit triangulaire où diagonale.

2.1.3 Élimination de Gauss

Pour résoudre le système $Ax = b$, le principe de la méthode de Gauss consiste à :

1. **Phase délimination** : prémultiplier A et le second membre b par des matrices bien choisies pour transformer le système $Ax = b$ en un système triangulaire supérieur ($Rx = c$) donc facile à résoudre (ohase de triangularisation)
2. **Remontée par back substitution** : Résoudre par la méthode de la remontée back substitution du système $Rx = c$ obtenu.

Description de la méthode :

$$Ax = b \quad \Leftrightarrow \quad (S^{(1)}) : \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n &= b_1, \quad (\mathbf{1}) \\ a_{2,1}x_1 + a_{2,2}x_2 + \ldots + a_{2,n}x_n &= b_2, \quad (\mathbf{2}) \\ \vdots &= \vdots \quad (\mathbf{i}) \\ a_{n,1}x_1 + a_{n,2}x_2 + \ldots + a_{n,n}x_n &= b_n, \quad (\mathbf{n}) \end{cases}$$

avec $A = (a_{i,j})_{\substack{1 \leq i \leq n \\ 1 \leq j \leq n}}$, $x = (x_1, x_2, \ldots, x_n)^T$ et $b = (b_1, b_2, \ldots, b_n)^T$.

– 1^{er}**-étape** : élimination de x_1 des équations $(\mathbf{2})$ jusqu'à $(\mathbf{n})$. On suppose que $a_{1,1} \neq 0$ (sinon on permute l'équation $(\mathbf{1})$ avec une équation $(\mathbf{i})$ du système tel que $a_{i,1} \neq 0$).
Puisque a est inversible, alors il existe i tel que $a_{i,1} \neq 0$ et pour éliminer x_1 de l'équation $(\mathbf{i})$ ($2 \leq i \leq n$), on effectue la combinaison linéaire suivante entre l'équation $(\mathbf{1})$ et l'équation $(\mathbf{i})$: on remplace l'équation $(\mathbf{i})$ par l'équation

$$\boxed{\text{équation}(\mathbf{i}) - \frac{\text{équation}(\mathbf{1})}{a_{11}} \ . \ a_{i,1},}$$

Donc l'équation $(\mathbf{i})$ devient sous la forme suivante

$$\begin{aligned}\left(a_{i,1} - \frac{a_{1,1}}{a_{1,1}}a_{i,1}\right) x_1 \quad &+ \quad \left(a_{i,2} - \frac{a_{1,2}}{a_{1,1}}a_{i,1}\right) x_2 + \ldots + \left(a_{i,j} - \frac{a_{1,j}}{a_{1,1}}a_{i,1}\right) x_j \\ &+ \quad \left(a_{i,n} - \frac{a_{1,n}}{a_{1,1}}a_{i,1}\right) x_n = b_i - \frac{b_1}{a_{1,1}}a_{i,1}\end{aligned}$$

Posons

$$\ell_{i,1} = \frac{a_{i,1}}{a_{1,1}}, \quad a_{i,j}^{(2)} = a_{i,j} - \ell_{i,1}a_{1,j} \quad \text{et} \quad b_i^{(2)} = b_i - \ell_{i,1}b_1$$

alors le système devient :

$$(S^{(2)}) : \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + a_{1,n}x_n &= b_1, \\ 0 \ \ x_1 + a_{2,2}^{(2)}x_2 + \ldots + a_{2,n}^{(2)}x_n &= b_2^{(2)}, \\ \vdots &= \vdots \\ 0 \ \ x_1 + a_{n,2}^{(2)}x_2 + \ldots + a_{n,n}^{(2)}x_n &= b_n^{(2)}, \end{cases}$$

Formulation matricielle : Posons

$$L_1 = \begin{pmatrix} 1 & 0 & \ldots & \ldots & 0 \\ -\ell_{2,1} & 1 & 0 & \ldots & 0 \\ \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ -\ell_{n,1} & 0 & \ldots & 0 & 1 \end{pmatrix} \quad \text{avec} \quad \ell_{i,1} = \frac{a_{i,1}}{a_{1,1}}, \quad 2 \leq i \leq n$$

$$A^{(1)} = A, \quad b^{(1)} = b$$
$$A^{(2)} = L_1 \, A,$$

le système $(S^{(2)})$ équivalent à

$$A^{(2)}x = L_1 \, A^{(1)}x = L_1 b^{(1)} = b^{(2)}$$

D'où on obtient le système

$$Ax = b \quad \Leftrightarrow \quad A^{(2)}x = b^{(2)}$$

où

$$A^{(2)} = \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & \ldots & a_{1,n} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & & a_{2,n}^{(2)} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & a_{n-1,n}^{(2)} \\ 0 & a_{n,2}^{(2)} & \ldots & \ldots & a_{n,n}^{(2)} \end{pmatrix} \quad \text{et} \quad b^{(2)} = \begin{pmatrix} b_1 \\ b_2^{(2)} \\ \vdots \\ b_{n-1}^{(2)} \\ b_n^{(2)} \end{pmatrix}$$

– 2^{eme}**-étape** : élimination de x_2 des équations $(\mathbf{3})$ jusqu'à $(\mathbf{n})$. On suppose que $a_{2,2}^{(2)} \neq 0$ (sinon on permute l'équation $(\mathbf{2})$ avec une équation $(\mathbf{i})$ du système tel que $a_{i,2}^{(2)} \neq 0$).
Pour éliminer x_2 de l'équation $(\mathbf{i})$ ($3 \leq i \leq n$) du système $A^{(2)}x = b^{(2)}$, on effectue la combinaison linéaire suivante entre l'équation $(\mathbf{2})$ et l'équation $(\mathbf{i})$: on remplace l'équation $(\mathbf{i})$ par l'équation

$$\boxed{\text{équation}(\mathbf{i}) - \frac{\text{équation}(\mathbf{2})}{a_{2,2}^{(2)}} \cdot a_{i,2}^{(2)}}$$

Posons

$$\ell_{i,2} = \frac{a_{i,2}^{(2)}}{a_{2,2}^{(2)}}, \quad a_{i,j}^{(3)} = a_{i,j}^{(2)} - \ell_{i,2} a_{2,j}^{(2)} \quad \text{et} \quad b_i^{(3)} = b_i^{(2)} - \ell_{i,2} b_2^{(2)}$$

alors le système devient :

$$(S^{(3)}) : \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + \ldots + a_{1,n}x_n &= b_1, \\ 0\ x_1 + a_{2,2}^{(2)}x_2 + \ldots + \ldots + a_{2,n}^{(2)}x_n &= b_2^{(2)}, \\ 0\ x_1 + 0\ x_2 + a_{3,3}^{(3)}x_3 + \ldots + a_{3,n}^{(3)}x_n &= b_3^{(3)}, \\ \vdots &= \vdots \\ 0\ x_1 + 0\ x_2 + a_{n,3}^{(3)}x_3 + \ldots + a_{n,n}^{(3)}x_n &= b_n^{(3)}, \end{cases}$$

Formulation matricielle : Posons

$$L_2 = \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & 0 \\ 0 & 1 & 0 & \ldots & \ldots & 0 \\ 0 & -\ell_{3,2} & 1 & 0 & \ldots & 0 \\ 0 & -\ell_{4,2} & 0 & \ddots & \ddots & \vdots \\ 0 & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & -\ell_{n,2} & 0 & \ldots & 0 & 1 \end{pmatrix} \quad \text{avec} \quad \ell_{i,2} = \frac{a_{i,2}^{(2)}}{a_{2,2}^{(2)}}, \quad 3 \leq i \leq n$$

$$A^{(1)} = A, \quad b^{(1)} = b$$

$$A^{(2)} = L_1\, A, \quad \text{et} \quad A^{(3)} = L_2 A^{(2)} = L_2 L_1\, A,$$

le système $(S^{(3)})$ équivalent à

$$A^{(3)}x = L_2 L_1\, A^{(1)}x = L_2 L_1 b^{(1)} = b^{(3)}$$

D'où on obtient le système

$$Ax = b \quad \Leftrightarrow \quad A^{(3)}x = b^{(3)}$$

où

$$A^{(3)} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \ldots & a_{1,n} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \ldots & a_{2,n}^{(2)} \\ \vdots & 0 & a_{3,3}^{(3)} & \ldots & a_{3,n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & a_{n,3}^{(3)} & \ldots & a_{n,n}^{(3)} \end{pmatrix} \quad \text{et} \quad b^{(3)} = \begin{pmatrix} b_1 \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_{n-1}^{(3)} \\ b_n^{(3)} \end{pmatrix}$$

– k^{eme}**-étape** : élimination de x_k des équations $(\mathbf{k+1})$ jusqu'à $(\mathbf{n})$. On suppose que $a_{k,k}^{(k)} \neq 0$ (sinon on permute l'équation $(\mathbf{k})$ avec une équation $(\mathbf{i})$ ($i > k$) du système tel que $a_{i,k}^{(k)} \neq 0$).
Pour éliminer x_k de l'équation $(\mathbf{i})$ ($k \leq i \leq n$) du système $A^{(k)}x = b^{(k)}$, on effectue la combinaison linéaire suivante entre l'équation $(\mathbf{k})$ et l'équation $(\mathbf{i})$: on remplace l'équation $(\mathbf{i})$ par l'équation

$$\boxed{\text{équation}(\mathbf{i}) - \frac{\text{équation}(\mathbf{k})}{a_{k,k}^{(k)}} \cdot a_{i,k}^{(k)}}$$

Posons

$$\ell_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \quad a_{i,j}^{(k+1)} = a_{i,j}^{(k)} - \ell_{i,k} a_{k,j}^{(k)} \quad \text{et} \quad b_i^{(k+1)} = b_i^{(k)} - \ell_{i,k} b_k^{(k)}$$

alors le système devient :

$$(S^{(k+1)}) : \begin{cases} a_{1,1}x_1 + a_{1,2}x_2 + \ldots + \ldots + \ldots + a_{1,n}x_n &= b_1, \\ 0\ x_1 + a_{2,2}^{(2)}x_2 + \ldots + \ldots + \ldots + a_{2,n}^{(2)}x_n &= b_2^{(2)}, \\ 0\ x_1 + 0\ x_2 + a_{3,3}^{(3)}x_3 + \ldots + \ldots + a_{3,n}^{(3)}x_n &= b_3^{(3)}, \\ \ddots & \vdots \\ 0\ x_1 + 0\ x_2 + \ldots + a_{k,k}^{(k)}x_k + \ldots + a_{k,n}^{(k)}x_n &= b_k^{(k)}, \\ \vdots & \vdots \\ 0\ x_1 + 0\ x_2 + \ldots + a_{n,k}^{(k)}x_k + \ldots + a_{n,n}^{(k)}x_n &= b_n^{(k)}, \end{cases}$$

Formulation matricielle : Posons

$$L_k = \begin{pmatrix} 1 & 0 & \ldots & \ldots & \ldots & \ldots & \ldots & 0 \\ 0 & 1 & 0 & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & -\ell_{k+1,k} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & -\ell_{n,k} & 0 & \ldots & 0 & 1 \end{pmatrix} \quad \text{avec} \quad \ell_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}, \quad k+1 \leq i \leq n$$

$$A^{(1)} = A, \quad b^{(1)} = b$$

$$A^{(2)} = L_1 A, \quad \ldots \quad , A^{(k+1)} = L_k A^{(k)} = L_k \ldots L_2 L_1 A,$$

le système $(S^{(k+1)})$ équivalent à

$$A^{(k+1)}x = L_k \ldots L_2 L_1 A^{(1)}x = L_k \ldots L_2 L_1 b^{(1)} = b^{(k+1)}$$

D'où on obtient le système

$$Ax = b \quad \Leftrightarrow \quad A^{(k+1)}x = b^{(k+1)}$$

Au bout de $(n-1)$-étape, on aboutit au système suivant

$$A^{(1)} = A, \quad b^{(1)} = b$$

$$A^{(2)} = L_1 A, \quad \ldots \quad , A^{(n)} = L_{n-1} A^{(n-1)} = L_{n-1} \ldots L_2 L_1 A,$$

le système $(S^{(n)})$ équivalent à

$$A^{(n)}x = L_{n-1} \ldots L_2 L_1 A^{(1)}x = L_{n-1} \ldots L_2 L_1 b^{(1)} = b^{(n)}$$

D'où on obtient le système

$$Ax = b \quad \Leftrightarrow \quad A^{(n)}x = b^{(n)}$$

avec $A^{(n)}$ est une matrice triangulaire supérieure.
Pour résoudre le système $Ax = b$, on résout le système équivalent $Rx = c$ avec

$$R = A^{(n)} = L_{n-1}L_{n-2}\dots L_2L_1\,A = M\,A \quad \text{où} \quad M = L_{n-1}L_{n-2}\dots L_2L_1$$

et

$$c = b^{(n)} = L_{n-1}L_{n-2}\dots L_2L_1\,b = M\,b \quad \text{où} \quad M = L_{n-1}L_{n-2}\dots L_2L_1.$$

Dans la pratique : On ne calcule pas le produit $L_{n-1}L_{n-2}\dots L_2L_1$ mais plutot $R = L_{n-1}L_{n-2}\dots L_2L_1$. Le calcul de R se fait par étape (où bien selon plusieurs algorithmes).
On ne stocke que la matrice A (n^2 éléments de type réel) à la fin de la triangularisation on n'aura plus besoin de la matrice A, mais, plutot on travaillera sur la matrice R. Donc A sera stochée dans la partie mémoire resrvée pour le stockage de A.
L'algorithme (étapes de calcul) est donné dans un language naturel. Pour pouvoir le mettre en œuvre sur un ordinateur il faut le traduire en un language de programmation tel que : Matlab, C, C++, Fortron, Java, ...

1. **Algorithme 1** :
   Pour chaque étape k **:** $k = 1, \dots, n-1$

   $$\textbf{on calcule} \quad \begin{cases} L_k\,A^{(k)} & : \quad A^{(k+1)} \leftarrow L_k\,A^{(k)}, \\ L_k\,b^{(k)} & : \quad b^{(k+1)} \leftarrow L_k\,b^{(k)}, \end{cases}$$

   fin pour.
   Pour pouvoir traduire l'algorithme 1 en language de programmation, on doit l'affiner, c'est-à-dire l'écrire à l'aide des opérations élémentaires, soit l'algorithme 2 obtenu après avoir affiné l'algorithme 1.
2. **Algorithme 2** : On considère les deux phrases mathématiques suivantes :

   $$(*) \quad a_{i,j}^{(k+1)} \leftarrow a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.a_{k,j}^{(k)}$$

   $$(**) \quad b_i^{(k+1)} \leftarrow b_i^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.b_k^{(k)}$$

```
Pour k=1...n-1, faire
      Pour chaque ligne i (k<i<= n), faire
            pour chaque indice j de la ligne i, faire
               écrire ici la phrase mathématique (*)
            Fin pour j
               écrire ici la phrase mathématique (**)
      Fin pour i
Fin pour k
```

   en réalité, les instructions

   $$a_{i,j}^{(k+1)} \leftarrow a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.a_{k,j}^{(k)} \quad \text{et} \quad b_i^{(k+1)} \leftarrow b_i^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.b_k^{(k)}$$

   seront remplacées par les instructions (affectation)

   $$a_{i,j} \leftarrow a_{i,j} - \frac{a_{i,k}}{a_{k,k}}.a_{k,j} \quad \text{et} \quad b_i \leftarrow b_i - \frac{a_{i,k}}{a_{k,k}}.b_k$$

Le signe $\leftarrow$ où bien $:=$ veut dire que $x = a_{i,j}^{(k+1)}$ sera remplacé par $y = a_{i,j}^{(k)} - \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}.a_{k,j}^{(k)}$ Autrement dit : dans la zone mémoire qui représente x on met la valeur de y.

```
Pour k=1...n-1, faire
      Pour chaque ligne i (k<i<= n), faire
            pour chaque indice j de la ligne i, faire
               a[i,j]:=a[i,j]-a[i,k]/a[k,k]*a[k,j]
            Fin pour j
               b[i]:=b[i]-a[i,k]/a[k,k]*b[k]
      Fin pour i
Fin pour k
```

3. **Algorithme 3** :

```
For k=1..n-1,
      For i=k+1..n,
            For j=k+1..n,
               a[i,j]:=a[i,j]-a[i,k]/a[k,k]*a[k,j]
            End j
               b[i]:=b[i]-a[i,k]/a[k,k]*b[k]
      End i
End k
```

2.1.4 Factorisation LU d'une matrice

Supposons qu'à chaque étape k de la méthode de Gauss $a_{k,k}^{(k)} \neq 0$. Dans ce cas, on a montré qu'il existe une matrice M telle que :

$$MA = R$$

où R est une matrice triangulaire supérieur et $M = L_{n-1}L_{n-2}\dots L_2L_1$ avec

$$L_k = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & & & & & \vdots \\ \vdots & \ddots & 1 & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & -\ell_{k+1,k} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & -\ell_{n,k} & 0 & \dots & 0 & 1 \end{pmatrix}$$

Propriété 2.1.1 *(**Propriété de la matrice** M)*

1. *M est inversible puisque* $det(M) = \prod_{k=1}^{n-1} det(L_k) = 1$ *car* $det(L_k) = 1 \neq 0$.

2. M^{-1} *est une matrice triangulaire inférieure et*

$$M^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ \ell_{2,1} & \ddots & 0 & & & & & \vdots \\ \ell_{3,1} & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & \ddots & 1 & \ddots & & & \vdots \\ \vdots & & & \ell_{k+1,k} & 1 & \ddots & & \vdots \\ \vdots & & & \vdots & \ell_{k+2,k+1} & \ddots & \ddots & \vdots \\ \vdots & & & \vdots & \vdots & \ddots & \ddots & 0 \\ \ell_{n,1} & \dots & \dots & \ell_{n,k} & \ell_{n,k+1} & \dots & \ell_{n,n-1} & 1 \end{pmatrix}$$

En effet, $M^{-1} = L_1^{-1} L_2^{-1} \dots L_{n-2}^{-1} L_{n-1}^{-1}$ *avec*

$$L_k^{-1} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \ddots & 0 & & & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & & & \vdots \\ \vdots & & 0 & 1 & \ddots & & & \vdots \\ \vdots & & \vdots & \ell_{k+1,k} & 1 & \ddots & & \vdots \\ \vdots & & \vdots & \vdots & 0 & \ddots & \ddots & \vdots \\ \vdots & & \vdots & \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \ell_{n,k} & 0 & \dots & 0 & 1 \end{pmatrix}$$

de la relation $M\ A = R$ *on déduit que* $A = M^{-1}\ R$.
Posons $L = M^{-1}$ **Lower : triangulaire inférieure**
et $R = U$ **Upper : triangulaire supérieure**.

Finalement on a le théorème suivant :

Théorème 2.1.1 *Si à chaque étape* k *de la méthode de Gauss on a le pivot* $a_{k,k}^{(k)} \neq 0$*, alors il existe une matrice triangulaire inférieur* L *et une matrice triangulaire supérieure* U *telles que* $A = L\,U$ *qui s'appelle la factorisation* $L\,U$ *de la matrice* A *avec* $\ell_{i,i} = 1$.

D'une manière générale, on a le théorème suivant :

Théorème 2.1.2 *Si* A *est inversible, alors il existe une matrice de permutation* P *telle que* $P\,A = L\,U$ *où* L *est une matrice triangulaire inférieur* L *et* U *une matrice triangulaire supérieure* U *avec* L *est une matrice à diagonale unitée* $\ell_{i,i} = 1$.

Théorème 2.1.3 *Soit* A *une matrice inversible.*
A *possède la factorisation* $L\,U$*, avec* L *est une matrice triangulaire inférieur* L *et* U *une matrice triangulaire supérieure* U *avec* L *est une matrice à diagonale unitée* $\ell_{i,i} = 1$*, si et seulement si toutes les matrices principales de* A *sont inversible.*

Théorème 2.1.4 *Si une* A *est inversible et possède la factorisation* $A = L\,U$*, où* L *est une matrice triangulaire inférieure et* U *une matrice triangulaire supérieure avec* L *est une matrice à diagonale unitée* $\ell_{i,i} = 1$*, alors cette factorisation est unique.*

Interprétation : La méthode de Gauss utilisée pour résoudre un système régulier $Ax = b$ (A inversible) consiste à factoriser la matrice $P\,A$ (P est une matrice de permutation bien choisie) en un produit $L\,U$ avec L est une matrice triangulaire inférieure et U une matrice triangulaire supérieure avec L est une matrice à diagonale unitée $\ell_{i,i} = 1$, en suite résoudre les systèmes suivants :

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

2.1.5 Factorisation de Cholesky où bien factoristion BB^T

Définition 2.1.1 *Soit A une matrice inversible*
Une factorisation régulière de Cholesky de A est une factorisation $A = BB^T$ où B est une matrice triangulaire inférieure régulière.
**Si les coefficients diagonaux de L sont positifs, on dit que l'on a une factorisation positive de Cholesky.*

Lemme 2.1.1 *Si A est une matrice symétrique définie positive, alors A possède la factorisation $L\,U$.*

Démonstration. Montrons que toutes les matrices principales d'ordre $1 \leq k \leq n$ sont inversible

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

décomposition en bloc A_{11} est une matrice principale d'ordre k.
Montrons que $y^T A_{11} y > 0$ pour tout $y \neq 0$ avec $y = (y_1, \ldots, y_k)^T$. Posons $x = (x_1, \ldots, x_n)^T$ avec $\begin{cases} x_i = y_i, \quad 1 \leq i \leq k \\ x_i = 0, \quad i > k \end{cases}$
Alors $x^T Ax = y^T A_{11} y$ et comme A est définie positive, on déduit que $y^T A_{11} y > 0$ ce qui montre que A_{11} est définie positive et A_{11} est symétrique puisque A l'est.
D'où on a la factorisation $L\,U$ de A. □

Théorème 2.1.5 *Soit A une matrice régulière (inversible).*
A possède une factorisation régulière de Cholesky $A = B\,B^T$ si et seulement si A est symétrique définie positive.
**Dans ce cas, elle possède une factorisation positive unique.*

Démonstration. $\Rightarrow\rfloor$ $A = B\,B^T$. Soit $x \in \mathbb{R}^n$, $x \neq 0$:

$$x^T Ax = x^T B\,B^T x = (B^T x)^T B^T x = (B^T x, B^T x) = \|B^T x\| \geq 0$$

comme $x \neq 0$ et B est régulière alors $L^T x \neq 0$, par conséquent on a

$$(B^T x)^T B^T x = (B^T x, B^T x) = \|B^T x\| > 0$$

donc A est symétrique définie positive.
$\Leftarrow\rfloor$ A est symétrique définie positive alors les valeurs propres $\mu_{ii} > 0$ pour $1 \leq i_l eqn$.
Soit $D = \text{diag}(\mu_{11}, \ldots, \mu_{nn})$ et $R = D^{-1}U$ triangulaire supérieure à diagonale unité, On déduit que

$$A = L\,D\,R$$

est une décomposition unique.
Comme A est symétrique, alors on a

$$L\,D\,R = A = A^T = R^T\,D^T\,L^T = R^T\,D\,L^T$$

puisque la factorisation $L\,U$ est unique, on déduit que $L = R^T$, par conséquent on a

$$A = L\,D\,L^T$$

Posons $\Lambda = \text{diag}(\sqrt{\mu_{11}}, \ldots, \sqrt{\mu_{nn}})$, alors

$$A = L\,\Lambda\,\Lambda\,L^T = BB^T \quad {}_t extrmavec \quad B = L\,\Lambda \quad \text{et} \quad b_{ii} > 0$$

et la factorisation $B\,B^T$ de A est unique. □

méthode pratique pour calculer B

Le calcul de la factorisation de Cholesky $A = BB^T$ peut se faire par identification. Comme A est symétrique, on refait l'identification de coefficients de la partie triangulaire inférieure de A, on a $A = (a_{ij})_{1\leq i,j\leq n}$ et $B = (b_{ij})_{1\leq i,j\leq n}$:

$$a_{ij} = \sum_{k=1}^{n} b_{ik}b_{jk}, \quad \text{pour} \quad i \geq j \quad \text{avec} \quad (b_{ij} = 0 \quad \text{pour} \quad i < j)$$

Calcul de B par colonne :

– Pour $j = 1$:

$$\begin{aligned}
a_{11} &= b_{11}^2 \quad \Rightarrow \quad b_{11} = \sqrt{a_{11}} \\
a_{21} &= b_{21}b_{11} \quad \Rightarrow \quad b_{21} = \frac{a_{21}}{\sqrt{a_{11}}} \\
&\vdots \\
a_{i1} &= b_{i1}b_{11} \quad \Rightarrow \quad b_{i1} = \frac{a_{i1}}{\sqrt{a_{11}}} \\
&\vdots \\
a_{n1} &= b_{n1}b_{11} \quad \Rightarrow \quad b_{n1} = \frac{a_{n1}}{\sqrt{a_{11}}}
\end{aligned}$$

– Pour $j = 2$

$$\begin{aligned}
a_{22} &= b_{21}b_{21} + b_{22}b_{22} \quad \Rightarrow \quad b_{22} = \sqrt{a_{22} - b_{21}^2} \\
a_{32} &= b_{31}b_{21} + b_{32}b_{22} \quad \Rightarrow \quad b_{32} = \frac{a_{32} - b_{31}b_{21}}{b_{22}}
\end{aligned}$$

d'une manière générale, on trouve la relation

$$a_{i2} = b_{i1}b_{21} + b_{i2}b_{22} \quad \Rightarrow \quad b_{i2} = \frac{a_{i2} - b_{i1}b_{21}}{b_{22}}$$

– Pour $j > 1$, la j^{eme} colonne est calculée à partir des colonnes $1, \ldots, (j-1)$ de la façon suivante :

$$\begin{aligned} b_{jj} &= \sqrt{a_{jj} - \sum_{k=1}^{j-1} b_{jk}^2} \\ b_{ij} &= \frac{a_{ij} - \sum_{k=1}^{j-1} b_{jk} b_{ik}}{b_{jj}} \quad \text{pour} \quad i \geq j+1. \end{aligned}$$

– Le calcul de b_{jj} nécessite $2(j-1)$ opérations élémentaires et une racine carrée.
– Le calcul de b_{ij} nécessite $2(j-1)$ opérations élémentaires $(+,-)$ et une division.
Donc le calcul de la colonne j nécessite :

$$2(n-j+1)(j-1) + n - j \;\; (+,-,/) \quad \text{et une racine carrée}$$

Le coût total de la méthode est

$$\frac{n^3}{3} + \frac{n^2}{2} - \frac{5n}{6} \;\; (+,-,/) \quad \text{et n racines carrées.}$$

2.2 Méthodes itératives

2.2.1 Généralités

Soit $A \in \mathbb{K}^{(n\times n)}$ et $b \in \mathbb{K}^n$. On veut résoudre le système linéaire $Ax = b$ avec A est une matrice inversible.
Résoudre le système linéaire $Ax = b$ par une méthode itérative consiste à construire une suite de vecteurs $(x^k)_{k\in\mathbb{N}}$ où $x^k \in \mathbb{K}^n$ tel que

$$\lim_{k\to+\infty} x^k = x = A^{-1}b.$$

Les méthodes itératives classiques sont de la forme :

$$\begin{cases} x^O, & \text{valeur initiale :donnée} \\ & x^{k+1} = B\,x^k + c \end{cases}$$

où $B \in \mathbb{K}^{(n\times n)}$ et $c \in \mathbb{K}^n$ sont donnés en fonction de A et b.

Définition 2.2.1 *Une méthode itérative est dite convergente si pour toute valeur initiale x^O, on a*

$$\lim_{k\to+\infty} x^k = x = A^{-1}b.$$

Pratiquement : on s'arrête lorsque $\|x^k - x\|$ est suffisamment petit.

Définition 2.2.2 *Une méthode itérative est dite consistante si la limite de la suite x^k lorsqu'elle existe, est solution du système linéaire $Ax = b$.*

Définition 2.2.3 *Une méthode itérative classique de la forme $x^{k+1} = B\,x^k + c$ est consistante si et seulement si*

$$\begin{cases} c = (I-B)A^{-1}b, \\ I - B \text{ inversible}. \end{cases}$$

Remarque 2.2.1 *Une méthode itérative consistante n'est pas toujours convergente.*

Exemple 2.2.1 *Soit à résoudre le système $Ax = b$ avec $a = \frac{1}{2}I$ par la méthode itérative $x^{k+1} = B\,x^k + c$ avec*

$$\left\{ \begin{array}{c} c = -b, \\ B = I + A\,. \end{array} \right.$$

- *cette méthode est consistante car $B = I + A \quad \Rightarrow \quad I - B = -A$ qui est inversible et $c = (I - B)A^{-1}b = -b$.*
- *cette méthode n'est pas convergente car :*

$$x^{k+1} - x = B(x^k - x) = \frac{3}{2}(x^k - x)$$

Par itération du schéma on obtient

$$x^{k+1} - x = \left(\frac{3}{2}\right)^{k+1} (x^O - x) \quad \textit{qui diverge}$$

Théorème 2.2.1 *On considère une méthode itérative consistante $x^{k+1} = B\,x^k + c$. Le système linéaire $Ax = b$ admet une solution unique si les assertions suivantes sont équivalentes :*

1. $\varrho(B) < 1$,
2. $\lim_{k\to+\infty} B^k x = 0$ *pour tout vecteur* x,
3. *La méthode itérative est convergente.*

Démonstration. $\underline{1) \Rightarrow 2)}$: $\varrho(B) < 1 \Rightarrow$ il existe au moins une norme matricielle telle que $\|B\| < 1$, car

$$\varrho(B) = \inf\{\|B\| \,/\, \|\cdot\| \text{ est une norme matricielle induite}\}$$

$\Rightarrow \quad \lim_{k\to+\infty} \|B^k\| = 0$ car $\|B^k\| \leq \|B\|^k$,
$\Rightarrow \quad \lim_{k\to+\infty} B^k = 0$.
$\underline{2) \Rightarrow 3)}$:

$$\begin{array}{rcl} e^k = x^k - x & = & B\,x^{k-1} - x + c \\ & = & B\,x^{k-1} - x + (I - B)A^{-1}b \\ & = & B\,x^{k-1} - x + (I - B)x \\ & = & B(x^{k-1} - x) \\ & = & Be^{k-1} \\ & \Rightarrow & e^k = B^k e^0 \quad (e^0 \text{donnée}) \\ & \Rightarrow & \lim\limits_{k\to+\infty} e^k = \lim\limits_{k\to+infty} (B^k e^0) = 0 \quad (\text{car } \lim\limits_{k\to+\infty} B^k = 0 \end{array}$$

d'où $\lim_{k\to+\infty} x^k = x$,
en suite la méthode itérative est convergente.
$\underline{3) \Rightarrow 1)}$: Pour tout x^0 donnée, on a $\lim_{k\to+\infty} x^k - x = 0$,
d'où pour tout x^0, $\lim_{k\to+\infty} e^k = 0$ avec $e^k = x^k - x$.
Soit $\lambda \in \text{Sp}(B)$ associée au vecteur propre v.
alors $Bv = \lambda v \quad \Rightarrow \quad B^k v = \lambda^k v$
pour $x^0 = x - v$, on obtient

$$B^k(x - x^0) = \lambda^k(x - x^0) \quad \Rightarrow \quad B^k e^0 = \lambda^k e^0$$

or $B^k e^0 = e^k$, on en déduit que $\lim_{k\to+\infty} B^k = 0$ car $\lim_{k\to+\infty} e^k = 0$,
comme $B^k e^0 = \lambda^k e^0$, on déduit que $\lim_{k\to+\infty} \lambda^k = 0$
ce qui prouve que $|\lambda| < 1, \quad \forall \lambda \in \text{Sp}(B)$, d'où $\varrho(B) < 1$. □

2.2.2 Comparaison des méthodes itératives

On considère la méthode itérative convergente définie par

$$\begin{cases} x^O, & \text{valeur initiale :donnée} \\ & x^{k+1} = B\,x^k + c \end{cases}$$

posons $e^k = x^k - x$ l'erreur à la k^{eme} itérative.
On s'arrête lorsque $\|e^k\|$ est suffisamment petit avec $e^0 = x^0 - x$ est l'erreur initiale.
On obtient donc $e^k = be^{k-1}$ et par conséquence $e^k = Be^{k-1}$.

$$\text{méthode convergente} \quad \Leftrightarrow \quad \lim_{k\to+_infty} B^k = 0 \quad \Leftrightarrow \quad \lim_{k\to+_infty} e^k = 0.$$

Soit $\|\cdot\|$ une norme matricielle : $\|e^k\| \leq \|B^k\|\|e^0\|$
et donc

$$\frac{\|e^k\|}{\|e^0\|} \leq \|B^k\| = \sup_{y\neq 0} \frac{\|B^k y\|}{\|y\|}.$$

Il est important de pouvoir estimer le nombre d'itérations (vitesse de convergence) nécessaires à l'obtention d'une approximation acceptable de la solution. **Problème 1** *(T)*

rouver k tel que $\frac{\|e^k\|}{\|e^0\|} \leq \varepsilon$: erreur permise. Il suffit que : $\|B^k\| \leq \varepsilon$, c'est-à-dire :

$$\begin{aligned} \ln(\|B^k\|) &\leq \ln(\varepsilon) \\ k\ln(\|B^k\|^{1/k}) &\leq \ln(\varepsilon) \\ k &\geq \frac{\ln(\varepsilon)}{\ln(\|B^k\|^{1/k})}, \quad \text{car} \quad \|B\| < 1 \end{aligned}$$

Définition 2.2.4 *Soit la méthode itérative convergente suivante*

$$\begin{cases} x^O, & \textit{valeur initiale :donnée} \\ & x^{k+1} = B\,x^k + c \end{cases}$$

1. *On appelle* **taux moyen de convergence** *pour k^{eme} itérative, d'une méthode itérative convergente, le nombre*
$$\mathcal{R}_k(B) = -\ln(\|B^k\|^{1/k}).$$
2. *On appelle* **vitesse de convergence**, *le nombre*
$$v(B) = \lim_{k\to+\infty} \mathcal{R}_k(B).$$

Théorème 2.2.2 *Soit B une matrice carrée. Pour toute norme matricielle, on a :*

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

et par suite

$$v(B) = -\ln(\varrho(B)).$$

Démonstration.

– Montrons que $\forall \varepsilon > 0$, $\exists k_0$ tel que : $\forall k \geq k_0$, on a $|\|B^k\|^{1/k} - \varrho(B)| < \varepsilon$?
Soit $\lambda \in \text{Sp}(B)$ associée à un vecteur propre $x \neq 0$
On a $Bx = \lambda x \quad \Rightarrow \quad B^k x = \lambda^k x$,
on déduit que

$$\|\lambda^k x\| \leq \|B^k\|\|x\| \quad \Rightarrow \quad |\lambda^k|\|x\| \leq \|B^k\|\|x\|$$

d'où

$$|\lambda| \leq \|B^k\|^{1/k}, \quad \forall \lambda \in \text{Sp}(B) \quad \Rightarrow \quad \varrho(B) \leq \|B^k\|^{1/k}$$

– Soit $\varepsilon > 0$, posons $B_\varepsilon = \frac{1}{\varrho(B)+\varepsilon} B$, on a donc

$$\varrho(B_\varepsilon) \leq \frac{\varrho(B)}{\varrho(B)+\varepsilon} < 1 \quad \Leftrightarrow \quad \lim_{k \to +\infty} B_\varepsilon^k = 0$$

par conséquent : $\lim_{k\to+\infty} \|B_\varepsilon^k\| = 0$.
Ceci veut dire que :

$$\forall \varepsilon' > 0, \ \exists k_0, \quad \forall k \geq k_0, \quad \text{on a :} \quad \|B_\varepsilon^k\| < \varepsilon'.$$

Pour $\varepsilon' = 1$, pour $\varepsilon > 0$, on a $\exists k_0, \quad \forall k \geq k_0, \quad$ on a : $\quad \|B_\varepsilon^k\| < 1$

$$\|B_\varepsilon^k\| = \frac{\|B^k\|}{(\varrho(B)+\varepsilon)^k} \quad \Rightarrow \quad \|B_\varepsilon^k\|^{1/k} < \varrho(B) + \varepsilon$$

d'où $\forall \varepsilon > 0, \ \exists k_0, \quad \forall k \geq k_0, \quad$ on a

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

Soit M_1 : $x^{k+1} = B_1\, x^k + c_1$,
et M_2 : $x^{k+1} = B_2\, x^k + c_2$,
deux méthodes itératives convergente pour résoudre $Ax = b$.
Si $\varrho(B_1) < \varrho(B_2)$, alors on a $v(B_1) > v(B_2)$ c'est-à-dire qu'on dit que la méthode M_1 est plus rapide que la méthode M_2.

□

2.2.3 Principales méthodes itératives classiques

Elles sont basées sur la décomposition de A sous la forme $A = M - N$ (A inversible), avec $M \in \mathbb{K}^{(n\times n)}$ facille à inverser.

$$\begin{aligned} Ax = b \quad &\Leftrightarrow \quad Mx = Nx + b \\ &\Leftrightarrow \quad x = M^{-1}Nx + M^{-1}b \\ &\Leftrightarrow \quad x = Bx + c \quad \text{avec} \quad B = M^{-1}N \ \text{ et } c = M^{-1}b \end{aligned}$$

On peut associer la méthode itérative

$$\begin{cases} x^O, & \text{valeur initiale :donnée} \\ & x^{k+1} = B\, x^k + c \end{cases}$$

la méthode itérative ainsi construite est consistante car $I - B = I - M^{-1}N = M^{-1}A$ est inversible, et $c = M^{-1}b = M^{-1}AA^{-1}b = (I - b)A^{-1}b$.
Puisque cette méthode itérative est consistante alors pour qu'elle soit convergente il suffit que

$$\varrho(B) < 1$$

Dans ce cas

$$\lim_{k\to+\infty} x^k = x$$

avec x est la solution du système $Ax = b$.

$$\varrho(B) < 1 \quad \Leftrightarrow \quad \varrho(M^{-1}N) < 1.$$

Méthode de Jacobi

Elle consiste à décomposer A sous la forme

$$A = D - E - F = M - N$$

avec $\begin{cases} M = D \\ N = E + F \end{cases}$

Où

$$D = \begin{pmatrix} a_{11} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & a_{nn} \end{pmatrix}, \quad -F = \begin{pmatrix} 0 & a_{12} & \dots & a_{1n} \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{(n-1)n} \\ 0 & \dots & 0 & 0 \end{pmatrix}$$

$$-E = \begin{pmatrix} 0 & 0 & \dots & 0 \\ a_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ a_{n1} & \dots & a_{n(n-1)} & 0 \end{pmatrix}$$

On suppose que $a_{ii} \neq 0, \forall 1 \leq i \leq n$, dans ce cas m est inversible.
la méthode itérative s'écrit

$$x^{k+1} = B\,x^k + c = D^{-1}(E + F)x^k + D^{-1}b$$

La matrice $B = J = D^{-1}(E + F)$ est appelée la **matrice de Jacobi** associée à A.

Proposition 2.2.1 *La méthode de Jacobi converge si et seulement si* $\varrho(J) < 1$.

Comme $A = D - E - F$ alors $E + F = D - A$.
La méthode itérative peut aussi s'écrire

$$x^{k+1} = D^{-1}(D - A)x^k + D^{-1}b = (I - D^{-1}A)x^k + D^{-1}b$$

à partir de $x \in \mathbb{K}$, on construit la suite (x^k) : les composantes de (x^k) sont données par :

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

qu'on peut aussi écrire sous la forme :

$$x_i^{k+1} - x_i^k = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j^k \right) = \frac{r_i^k}{a_{ii}}$$

où r_i^k est la i^{eme} composante du vecteur r^k défini par : $r^k = b - Ax^k$ appelé **vecteur résidu** à la k^{eme} itération.

Méthode de Gauss-Seidel

Elle consiste à prendre $M = D - E$ (M est triangulaire inférieure) et $N = F$:

$$A = M - N = D - E - F$$

M est inversible si $a_{ii} \neq 0$, $\forall 1 \leq i \leq n$.
La méthode itérative s'écrit

$$x^{k+1} = B\,x^k + c = (D-E)^{-1}Fx^k + (D-E)^{-1}b$$

avec $B = M^{-1}N = (D-E)^{-1}F$ et $c = M^{-1}b = (D-E)^{-1}b$.
d'où on obtient :

$$(D-E)x^{k+1} = Fx^k + b.$$

Supposons les composantes $x_1^{k+1}, \ldots, x_{i-1}^{k+1}$ sont calculées, alors la i^{eme} composante x_i^{k+1} est obtenue par

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

On pose $\mathcal{L}_1 = (D-E)^{-1}F$, cette matrice s'appelle la matrice de Gauss-Seidel.

Proposition 2.2.2 *La méthode de Gauss-Seidel converge si et seulement si* $\varrho(\mathcal{L}_1) < 1$.

2.2.4 Etude de la convergence

Définition 2.2.5 *Une matrice* $A = (a_{ij})_{1\leq i,j\leq n}$ *est à* **diagonale strictement dominante** *si on a*

$$|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|.$$

Théorème 2.2.3 *Si A est à diagonale strictement dominante, alors la méthode de Jacobi converge.*

Démonstration. A est à diagonale strictement dominante alors A est inversible. Soit $J = D^{-1}(E+F) = I - D^{-1}A$ la matrice de Jacobi :

$$b_{ij} = \begin{cases} -\dfrac{a_{ij}}{a_{ii}}, & \text{si} \quad i \neq j \\ \\ 0, & \text{si} \quad i = j \end{cases}$$

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \frac{1}{|a_{ii}|}\sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| < 1$$

On déduit que

$$\max_{1\leq i\leq n}\left(\sum_{j=1}^{n} |b_{ij}|\right) = \|J\|_\infty < 1$$

d'où $\varrho(J) < 1$. D'où la méthode de Jacobi est convergente. □

Théorème 2.2.4 *Soit A une matrice hermitienne (resp. symétrique) définie positive. Considérons la décomposition $A = M - N$ avec M inversible.*
Si la matrice $M^* + N$ (resp. la matrice symétrique $M^T + N$) est définie positive, alors $\varrho(M^{-1}N) < 1$.
De plus le schéma itératif classique $x^{k+1} = B\,x^k + c$ avec$B = M^{-1}N$ et $c = M^{-1}b$ est convergente.

Démonstration.

1. $M^* + N = M^* + M - A$
2. $(M^* = N)^* = (M^* + M - A)^* = M + M^* - A^* = M + M^* - A$, donc $M^* + N$ est hermitienne. Pour montrer que $\varrho(M^{-1}N) < 1$, on construit une norme vectorielle, notée $\| \cdot \|_A$ telle que la norme matricielle induite vérifie $\|M^{-1}N\| < 1$. Comme A est définie positive, on pose

$$\|x\|_A = (x^*Ax)^{1/2}$$

$\| \cdot \|_A$ est bien une norme vectorielle.
Considérons la norme matricielle induite

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

montrons que $\forall x \in \mathbb{K}^n$ tel que $\|x\|_A = 1$ on a $\|x - M^{-1}Ax\|_A < 1$?
Soit $x \in \mathbb{K}^n$ tel que $\|x\|_A = 1$, posons $M^{-1}Ax = y \quad \Leftrightarrow \quad Ax = My$.

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

Pour $x \neq 0$, on a $y \neq 0$ (car A et M sont inversible) donc

$$y^*(M^* + N)y > 0, \quad \text{car } M^* + N \text{ est définie positive}$$

on déduit que $\|x - M^{-1}Ax\|_A^2 < 1, \quad \forall x \in \mathbb{K}, \quad \|x\|_A = 1$

$$\Rightarrow \quad \|M^{-1}N\| < 1 \quad \Rightarrow \quad \varrho(M^{-1}N) < 1$$

□

Corollaire 2.2.1 *Soit A une matrice hermitienne. Si $2D - A$ est définie positive, alors la méthode de Jacobo converge si et seulement si A est définie positive.*

Cas des matrices tridiagonales : Soit A une matrice tridiagonale.

$$A = \begin{pmatrix} a_1 & b_1 & 0 & \dots & 0 \\ c_1 & a_2 & b_2 & \ddots & \vdots \\ 0 & c_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \dots & 0 & c_2 & a_n \end{pmatrix}$$

avec E et F définies par

$$E = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ c_1 & \ddots & 0 & \ddots & \vdots \\ 0 & c_2 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & c_n & 0 \end{pmatrix} \quad \text{et} \quad F = \begin{pmatrix} 0 & b_1 & 0 & \dots & 0 \\ 0 & \ddots & b_2 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & b_n \\ 0 & \dots & 0 & 0 & 0 \end{pmatrix}$$

Lemme 2.2.1 *Soit $\mu \in \mathbb{C}^*$, alors on a det$(A_\mu) =$ det(A), avec*

$$A_\mu = D - \mu E - \frac{1}{\mu} F.$$

Démonstration. On considère la matrice

$$Q_\mu = \begin{pmatrix} \mu & 0 & 0 & \dots & 0 \\ 0 & \mu^2 & 0 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 0 & \mu^n \end{pmatrix}, \quad \mu \neq 0$$

on vérifie que $A_\mu = Q_\mu A Q_\mu^{-1}$. Donc on a : det$(A_\mu) =$ det(A). □

Théorème 2.2.5 *Soit A une matrice tridiagonale dont les éléments de la diagonale sont non nul, alors*

$$\varrho(\mathcal{L}_1) = (\varrho(J))^2.$$

La méthode de Gauss-Seidel et la méthode de Jacobi convergent simultaniment.

Démonstration. Soit $J = D^{-1}(E+F)$ et $\mathcal{L}_1 = (D-E)^{-1}F$. On calcule le polynôme caractéristique de J et de $\mathcal{L}_1$:

$$\begin{aligned} P_J(\lambda) &= \text{det}(D^{-1}(E+F) - \lambda I_n) \\ &= \text{det}(D^{-1}(E+F-\lambda D)) \\ &= \text{det}(D^{-1})\text{det}(E+F-\lambda D) \\ &= (-1)^n \text{det}(D^{-1})\text{det}(\lambda D - E - F) \\ &= (-1)^n \text{det}(D^{-1})\text{det}(\lambda D + E + F) \quad \text{(d'après le lemme : } \quad \mu = -1) \\ &= (-1)^n \text{det}(\lambda I_n + D^{-1}(E+F)) \\ &= (-1)^n \text{det}(-(-\lambda) I_n + D^{-1}(E+F)) \\ &= (-1)^n P_J(-\lambda) \end{aligned}$$

donc λ est une valeur propre de J si et selement si $(-\lambda)$ est une valeur propre de J.

$$\begin{aligned} P_{\mathcal{L}_1}(\lambda) &= \text{det}((D-E)^{-1}F - \lambda I_n) \\ &= \text{det}((D-E)^{-1})\text{det}(F - \lambda D + \lambda E) \\ &= (-1)^n \text{det}((D-E)^{-1})\text{det}(\lambda D - \lambda E - F) \\ &= (-1)^n \text{det}((D-E)^{-1})\text{det}\left(\lambda D - \frac{\lambda}{\mu} E - \mu F\right), \quad \forall \mu \in \mathbb{C}^* \end{aligned}$$

on choisit $\mu = \lambda^{1/2} (\lambda \neq 0)$, $\lambda^{1/2}$ est un nombre complexe vérifiant $(\lambda^{1/2})^2 = \lambda$.
Pour $\lambda \neq 0$, on a

$$\begin{aligned} P_{\mathcal{L}_1}(\lambda) &= (-1)^n \det((D-E)^{-1}) \det\left(\lambda D - \lambda^{1/2} E - \lambda^{1/2} F\right) \\ &= (-1)^n \lambda^{n/2} \det((D-E)^{-1}) \det\left(\lambda^{1/2} D - (E+F)\right) \\ &= \lambda^{n/2} \det((D-E)^{-1}) \det(D) \det\left(D^{-1}(E+F) - \lambda^{1/2} I_n\right) \\ &= \lambda^{n/2} \det\left(J - \lambda^{1/2} I_n\right) \\ &= \lambda^{n/2} P_J\left(\lambda^{1/2}\right) \end{aligned}$$

car "$\det((D-E)^{-1}) \det(D) = 1$" puisque "$\det((D-E)^{-1}) = \det(D^{-1})$".
ceci montre que λ est une valeur propre de $\mathcal{L}_1$, avec $\lambda \neq 0$ alors

$$\{\lambda^{1/2}, -\lambda^{1/2}\} \subset \mathrm{Sp}(J).$$

Réciproquement : Si $\beta \neq 0$ est une valeur propre de J, alors β^2 est une valeur propre de $\mathcal{L}_1$. On en déduit que $\varrho(\mathcal{L}_1) = (\varrho(J))^2$. □

Corollaire 2.2.2

$$\varrho(J) < 1 \quad \Leftrightarrow \quad \varrho(\mathcal{L}_1) < 1.$$

alors dans ce cas les méthodes de Jacobi et de Gauss-Seidel convergent où divergent simultanément. Lorsque les deux méthodes convergent, alors la méthode de Gauss-Seidel converge plus vite que la méthode de Jacobi :

$$2 \ln(\varrho(J)) = \varrho(\mathcal{L}_1).$$

Chapitre 3

Interpolation et approximation polynômiale

3.1 Introduction

Ce chapitre traite de l'approximation d'une fonction dont on ne connaît les valeurs qu'en certains points.

Plus précisément, étant donné $n+1$ couples (x_i, y_i), le problème consiste à trouver une fonction $\Phi = \Phi(x)$ telle que $\Phi(x_i) = y_i$ pour $i = 0, \ldots, m$, où les y_i sont des valeurs données.

Définition 3.1.1 *On dit alors que* Φ **interpole** $\{y_i\}$ *aux nœuds* $\{x_i\}$.
*On parle d'***interpolation polynomiale** *quand* Φ *est un polynôme, d'***approximation trigonométrique** *quand* Φ *est un polynôme trigonométrique et d'***interpolation polynomiale par morceaux** *(ou d'***interpolation par fonctions splines** *ou d'***interpolation par fonctions à base radiales***) si* Φ *est polynomiale par morceaux.*

Les quantités y_i peuvent, par exemple, représenter les valeurs aux nœuds x_i d'une fonction f connue analytiquement ou des données expérimentales. Dans le premier cas, l'approximation a pour but de remplacer f par une fonction plus simple en vue d'un calcul numérique d'intégrale ou de dérivée. Dans l'autre cas, le but est d'avoir une représentation synthétique de données expérimentales dont le nombre peut être très élevé. Nous étudions dans ce chapitre l'interpolation polynomiale, polynomiale par morceaux et les splines paramétriques. Nous aborderons aussi les interpolations trigonométriques et interpolations basées sur les polynômes orthogonaux.

Vision sur une fonction

Soient $[a,\ b]$ un intervalle de $\mathbb{R}$, $\mathcal{S} = (x_i)_{0 \le i \le n}$ une subdivision de $[a,\ b]$ et $f : [a,\ b] \to \mathbb{R}$ une fonction connue aux $(n+1)$ points $x_i (i = 0, \ldots, n)$ de la subdivision $\mathcal{S}$, c'est-à- dire qu'on connaît les valeurs

$$y_i = f(x_i), \quad \text{pour} \quad i = 0, \ldots, n.$$

Définition 3.1.2 *On dit qu'un polynôme* $\mathcal{P}$ *de degrée inférieur ou égal à* n *(i.e., deg*$(\mathcal{P}) \le n$*) interpole* f *(ou encore interpole les valeurs* $y_0, \ldots, y_n$ *aux* $(n+1)$ *points* $x_0, \ldots, x_n$ *s'il vérifie les conditions d'interpolation suivantes :*

$$\mathcal{P}(x_i) = f(x_i), \quad (\textit{où encore} \ \ y_i = \mathcal{P}(x_i) \quad i = 0, \ldots, n$$

3.2 Interpolation polynomiale

Considérons $n+1$ couples (x_i, y_i). Le problème d'interpolation consiste de trouver un polynôme $\mathcal{P}_m \in \mathbb{P}_m$ ou $\mathbb{P}_m$ est l'espace des polynômes de degré m. Le polynôme $\mathcal{P}_m$ est appelé **polynôme d'interpolation** ou **polynôme interpolant**, tel que

$$\mathcal{P}_m(x_i) = a_0 + a_1 x_i + \ldots + a_m x_i^m = y_i, \quad i = 0, \ldots, n.$$

Les points x_i sont appelés **nœuds d'interpolation**.

3.2.1 Interpolation polynomiale de Lagrange

Soit $[a, b]$ un intervalle borné de $\mathbb{R}$. Soit $a = x_1 < x_2 < \ldots < x_n < x_{n+1} = b$ une subdivision de l'intervalle $[a, b]$. On se donne $n+1$ réels y_i.
Pour tout $i = 1, \ldots, n+1$, on appelle **polynôme de Lagrange** [1] d'indice i, le polynôme ℓ_i défini par

$$\ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}.$$

Pour tout $0 \leq i \leq n$, on a $\ell_i \in \mathbb{P}_n$ et que le système $\{\ell_0, \ell_1, \ldots, \ell_n\}$ forme une base de l'espace $\mathbb{P}_n$. On l'appelle base de Lagrange de $\mathbb{P}_n$.

Théorème 3.2.1 *Etant donné $n+1$ points distincts $x_0, \ldots, x_n$ et $n+1$ valeurs correspondantes $y_0, \ldots, y_n$, alors il existe un unique polynôme $\mathcal{P}_n \in \mathbb{P}_n$ tel que $\mathcal{P}_n(x_i) = y_i$ pour $i = 0, \ldots, n$.*

Démonstration. Pour prouver l'existence, on va construire explicitement $\mathcal{P}_n$. Posons

$$\ell_i \in \mathbb{P}_n : \quad \ell_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}, \quad i = 0, \ldots, n$$

$\{\ell_0, \ell_1, \ldots, \ell_n\}$ est une base de l'espace $\mathbb{P}_n$, alors on a

$$\mathcal{P}_n(x) = \sum_{i=0}^{n} c_i \ell_i(x),$$

d'où

$$\mathcal{P}_n(x_i) = \sum_{j=0}^{n} c_j \ell_j(x_i) = y_i, \quad i = 0, \ldots, n$$

Il est facile de voir que

$$\ell_j(x_i) = \delta_{ij} = \begin{cases} 1, & \text{si} \quad i = j, \\ 0, & \text{si} \quad i \neq j. \end{cases}$$

on en déduit immédiatement que $c_i = y_i$ pour tout $i = 0, \ldots, n$.

Par conséquent, le polynôme d'interpolation existe et s'écrit sous la forme suivante

$$\mathcal{P}_n(x) = \sum_{i=0}^{n} y_i \ell_i(x). \tag{1.1}$$

Pour montrer l'unicité, supposons qu'il existe un autre polynôme Ψ_m de degré $m \leq n$ tel que $\Psi_m(x_i) = y_i$ pour $i = 0, \ldots, n$. La différence $\mathcal{P}_n - \Psi_m$ s'annule alors en $n+1$ points distincts x_i, elle est donc nulle. Ainsi, $\mathcal{P}_n = \Psi_m$. □

1. J.L.Lagrange est un mathématicien Franco-Italien($1736 - 1813$)

Corollaire 3.2.1 *On a*

$$\mathcal{P}_n(x) = \sum_{i=0}^{n} \frac{\omega_{n+1}(x)}{(x - x_i)\omega'_{n+1}(x_i)} y_i$$

où ω_{n+1} *est le* **polynôme nodal** *de degré* $n+1$ *défini par*

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

la relation (1.1) est appelée **formule d'interpolation de Lagrange**, et les polynômes $\ell_i(x)$ sont les polynômes caractéristiques (de Lagrange).

Exemple 3.2.1 *On considère l'intervalle* $[-1, 1]$ *et une subdivision*

$$x_0 = -1 < x_1 < \ldots < x_n = 1.$$

On choisit $n = 4$*, on a* $x_0 = -1$*,* $x_1 = -0.5$*,* $x_2 = 0$*,* $x_3 = 0.5$ *et* $x_4 = 1$

$$\ell_0(x) = \prod_{\substack{j=0 \\ j\neq 1}}^{4} \frac{x - x_j}{x_0 - x_j} = \frac{x - x_1}{x_0 - x_1}\frac{x - x_2}{x_0 - x_2}\frac{x - x_3}{x_0 - x_3}\frac{x - x_4}{x_0 - x_4},$$

$$\ell_1(x) = \prod_{\substack{j=0 \\ j\neq 2}}^{4} \frac{x - x_j}{x_1 - x_j} = \frac{x - x_0}{x_1 - x_0}\frac{x - x_2}{x_1 - x_2}\frac{x - x_3}{x_1 - x_3}\frac{x - x_4}{x_1 - x_4},$$

$$\ell_2(x) = \prod_{\substack{j=0 \\ j\neq 3}}^{n} \frac{x - x_j}{x_2 - x_j} = \frac{x - x_0}{x_2 - x_0}\frac{x - x_1}{x_2 - x_1}\frac{x - x_3}{x_2 - x_3}\frac{x - x_4}{x_2 - x_4},$$

$$\ell_3(x) = \prod_{\substack{j=0 \\ j\neq 4}}^{4} \frac{x - x_j}{x_3 - x_j} = \frac{x - x_0}{x_3 - x_0}\frac{x - x_1}{x_3 - x_1}\frac{x - x_2}{x_3 - x_2}\frac{x - x_4}{x_3 - x_4},$$

$$\ell_4(x) = \prod_{\substack{j=0 \\ j\neq 4}}^{4} \frac{x - x_j}{x_4 - x_j} = \frac{x - x_0}{x_4 - x_0}\frac{x - x_1}{x_4 - x_1}\frac{x - x_2}{x_4 - x_2}\frac{x - x_3}{x_4 - x_3}.$$

3.3 Détermination du polynôme d'interpolation

3.3.1 Cas où $n = 2$

Il s'agit de déterminer une polynôme d'interpolation $\mathcal{P}$ de degré $n = 2$, c'est-à-dire $\mathcal{P} \in \mathbb{P}_2$. On écrit

$$\mathcal{P}(x) = a_0 + a_1 x + a_2 x^2.$$

L'interpolation se fait en trois nœuds x_0, x_1 et x_2, alors on écrit le système suivant :

$$(\mathcal{E}) : \begin{cases} \mathcal{P}(x_0) = y_0, \\ \mathcal{P}(x_1) = y_1, \\ \mathcal{P}(x_2) = y_2. \end{cases}$$

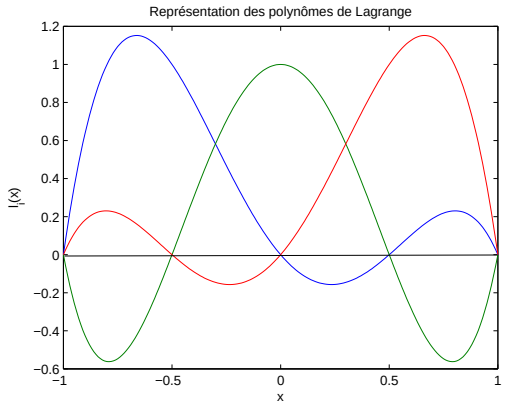

FIGURE 3.1 – Exemples de polynômes de Lagrange ℓ_2, ℓ_3 et ℓ_4.

y_0, y_1 et y_2 sont des données réelles et x_0, x_1 et x_2 sont des nœuds fixés.
Ainsi,le système $(\mathcal{E})$ est équivalent au système suivant dons les inconnus sont a_0, a_1 et a_2 tles que

$$(\mathcal{E}) : \begin{cases} a_0 + a_1 x_0 + a_2 x_0^2 = y_0, \\ a_0 + a_1 x_1 + a_2 x_1^2 = y_1, \\ a_0 + a_1 x_2 + a_2 x_2^2 = y_2. \end{cases}$$

Ce système analytique est équivalent à un système matricielle qu'on peut écrire sous la forme suivante :

$$\underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}}_{b}.$$

Ce système admet une unique solution si dét$(A) \neq 0$ (c'est-à-dire le discriminant $\Delta \neq 0$). On obtient un système matricielle $Ax = b$ à résoudre dont l'inconnu est le vecteur x, on a

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}^{-1} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}.$$

Le discriminant $\Delta =$ dét(A) qu'on calcule selon

$$\text{dét}(A) = \begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} = (x_2 - x_1)(x_2 - x_0)(x_1 - x_0).$$

Une fois les coefficients a_0, a_1 et a_2 sont calculés, on a :

$$\begin{aligned} \mathcal{P}(x) &= a_0 + a_1 x + a_2 x^2 \\ &= (1 \;\; x \;\; x^2) \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \\ &= (1 \;\; x \;\; x^2) \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}^{-1} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} \\ &= (\xi_0 \;\; \xi_1 \;\; \xi_2) \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} \end{aligned}$$

avec $(\xi_0 \quad \xi_1 \quad \xi_2) = (1 \quad x \quad x^2) \begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}^{-1}$. D'après le technique de transposition, on obtient le système matricielle suivant :

$$\begin{pmatrix} \xi_0 \\ \xi_1 \\ \xi_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}$$

En résolvant ce système pour ontenir

$$\xi_0 = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$$

$$\xi_1 = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

$$\xi_2 = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

On pose $\ell_0(x) = \xi_0$, $\ell_1(x) = \xi_1$ et $\ell_2(x) = \xi_2$ qui sont respectivement les polynômes de Lagrange associés à x_0, x_1 et x_2. Dans ce cas, on a :

$$\mathcal{P}(x) = a_0 + a_1 x + a_2 x^2 = y_0\ell_0(x) + y_1\ell_1(x) + y_2\ell_2(x).$$

Exemple 3.3.1 *Pour des valeurs données, de la fonction $x \mapsto f(x)$, aux points $x = x_0, x_2$ et x_2, on exprime le polynôme d'interpolation par*

$$\mathcal{P}_2(x) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2)$$

Pour $f(1) = 0$, $f(-1) = -3$ et $f(2) = 4$ avec $x_0 = 1$, $x_1 = -1$ et $x_2 = 2$ on obtient

$$\mathcal{P}_2(x) = \frac{(x+1)(x-2)}{(1+1)(1-2)} \times 0 + \frac{(x-1)(x-2)}{(-1-1)(-1-2)} \times (-3) + \frac{(x-1)(x+1)}{(2-1)(2+1)} \times 4$$

Après simplification, on a

$$\mathcal{P}_2(x) = \frac{1}{6}\left(5x^2 + 9x - 14\right).$$

On a bien $\mathcal{P}_2(1) = 0$, $\mathcal{P}_2(-1) = -3$ et $\mathcal{P}_2(2) = 4$.

3.3.2 Cas général

On cherche $\mathcal{P} \in \mathbb{P}_n$ vérifiant le problème d'interpolation $\mathcal{P}(x_i) = y_i$ pour $i = 0, \ldots, n$, c'est à dire, on cherche $a_0, \ldots, a_n$ tels que $\mathcal{P}(x) = a_0 + a_1 x + \ldots + a_n x^n$ et vérifiant les $(n+1)$ équations ci-dessous :

$$(\mathcal{E}) : \begin{cases} a_0 + a_1 x_0 + \ldots + a_n x_0^n &= y_0, \\ a_0 + a_1 x_1 + \ldots + a_n x_1^n &= y_1, \\ \vdots &= \vdots \\ a_0 + a_1 x_n + \ldots + a_n x_n^n &= y_n. \end{cases}$$

La forme matricielle de ce système est :

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

et on peut facilement montrer que

$$\Delta = \text{dét}(A) = \begin{vmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \dots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{vmatrix} = \prod_{0 \le i < j \le n} (x_i - x_j).$$

Exemple 3.3.2 *Cherchons $\mathcal{P}$ le polynôme interpolant la fonction $f(x) = \sqrt{x^2+1}$ aux points 0, 1, 4 et 9. En notant $x_0 = 0$, $x_1 = 1$, $x_2 = 4$ et $x_3 = 9$ et en utilisant la base de Lagrange, on peut affirmer que :*

$$\mathcal{P}(x) = f(x_0)\ell_0(x) + f(x_1)\ell_1(x) + f(x_2)\ell_2(x) + f(x_3)\ell_3(x).$$

On a $f(x_0) = 1$, $f(x_1) = \sqrt{2}$, $f(x_2) = \sqrt{17}$, $f(x_3) = \sqrt{82}$. On a bien

$$\begin{aligned} \ell_0(x) &= \frac{x-x_1}{x_0-x_1}\frac{x-x_2}{x_0-x_2}\frac{x-x_3}{x_0-x_3} = -\frac{1}{36}(x-1)(x-4)(x-9), \\ \ell_1(x) &= \frac{x-x_0}{x_1-x_0}\frac{x-x_2}{x_1-x_2}\frac{x-x_3}{x_1-x_3} = \frac{1}{24}x(x-4)(x-9), \\ \ell_2(x) &= \frac{x-x_0}{x_2-x_0}\frac{x-x_1}{x_2-x_1}\frac{x-x_3}{x_2-x_3} = -\frac{1}{60}x(x-1)(x-9), \\ \ell_3(x) &= \frac{x-x_0}{x_3-x_0}\frac{x-x_1}{x_3-x_1}\frac{x-x_2}{x_3-x_2} = \frac{1}{360}x(x-1)(x-4). \end{aligned}$$

Donc on obtient

$$\mathcal{P}(x) = \ell_0(x) + \sqrt{2}\ell_1(x) + \sqrt{17}\ell_2(x) + \sqrt{82}\ell_3(x)$$

3.4 Interpolation par les différences divisées

Dans cette section, nous discuterons une autre méthode d'interpolation polynômiale dans laquelle nous utiliserons les différences divisées. Cette méthode est due à Isaac Newton.
On note par $x_0, x_1, \dots, x_n$ les $(n+1)$ points distincts et on désigne par $\mathcal{P}_n$ le polynôme d'interpolation de la fonction f, qui est donné par l'expression suivante :

$$\mathcal{P}_n(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)(x-x_1) + \dots + a_n(x-x_0)(x-x_1)\dots(x-x_{n-1}).$$

Par la substitution $x = x_0, x_1, \dots, x_n$ dans l'expression explicite de $\mathcal{P}_n$, on obtient le système suivant :

$$\begin{aligned} f(x_0) &= a_0, \\ f(x_1) &= a_0 + a_1(x_1 - x_0), \\ f(x_2) &= a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1), \\ &\vdots \\ f(x_n) &= a_0 + a_1(x_n - x_0) + a_2(x_n - x_0)(x_n - x_1) + \dots \\ &\quad + a_n(x_n - x_0)(x_n - x_1)\dots(x_n - x_{n-1}). \end{aligned}$$

Ce système d'équations nous permet de déterminer les coefficients $a_0, a_1, \ldots, a_n$ de façon unique. La première équation détermine $a_0 = f(x_0)$, la deuxième détermine $a_1 = \frac{f(x_1)-a_0}{x_1-x_0}$ si $x_1 - x_0 \neq 0$, si $(x_2 - x_0)(x_2 - x_1) \neq 0$ alors $a_2 = \frac{f(x_2)-[a_0+a_1(x_2-x_0)]}{(x_2-x_0)(x_2-x_1)}$. Finallement, si on connait $a_0, a_1, \ldots, a_{n-1}$ alors la dernière équation du système nous donne a_n puisque $(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1}) \neq 0$.
Le polynôme $\mathcal{P}_n$ s'écrit donc de façon unique et la formule reste permanente. Mais, si on ajoute un autre point x_{n+1} à la suite $x_0, x_1, \ldots, x_n$ alors on construit un polynôme d'interpolation $\mathcal{P}_{n+1}$ lié à la nouvelle suite $x_0, x_1, \ldots, x_n, x_{n+1}$ de la forme suivante :

$$\begin{aligned}\mathcal{P}_{n+1}(x) &= b_0 + b_1(x - x_0) + b_2(x - x_0)(x - x_1) + \ldots + b_n(x - x_0)\ldots(x - x_{n-1}) \\ &\quad + b_{n+1}(x - x_0)\ldots(x - x_{n-1})(x - x_n).\end{aligned}$$

En utilisant le même principe qu'avant, on trouve $b_0 = a_0$, $b_1 = a_1, \ldots, b_n = a_n$ car par substitution des points $x = x_0, x_1, \ldots, x_n, x_{n+1}$ dans l'expression $\mathcal{P}_{n+1}$, on a :

$$\begin{aligned}f(x_0) &= b_0, \\ f(x_1) &= b_0 + b_1(x_1 - x_0), \\ f(x_2) &= b_0 + b_1(x_2 - x_0) + b_2(x_2 - x_0)(x_2 - x_1), \\ &\vdots \\ f(x_n) &= b_0 + b_1(x_n - x_0) + b_2(x_n - x_0)(x_n - x_1) + \ldots \\ &\quad + b_n(x_n - x_0)(x_n - x_1)\ldots(x_n - x_{n-1}), \\ f(x_{n+1}) &= b_0 + b_1(x_n - x_0) + b_2(x_n - x_0)(x_n - x_1) + \ldots \\ &\quad + b_{n+1}(x_{n+1} - x_0)(x_{n+1} - x_1)\ldots(x_{n+1} - x_n).\end{aligned}$$

Par la résolution de ce système d'équations, on aboutit à ce que

$$b_i = a_i, \quad \forall 1 \leq i \leq n$$

En général, on a l'expression

$$a_j = f[x_0, x_1, \ldots, x_j]$$

avec la relation suivante :

$$f[x_0, x_1, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{\prod\limits_{\substack{j=0 \\ j\neq i}}^{n}(x_i - x_j)}.$$

Cette formule consiste à écrire les valeurs $f(x_i)$ sous la forme d'une combinaison linéaire

Exemple 3.4.1 *Pour $n = 2$, on a*

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

Est-ce qu'on peut écrire l'expression $f[x_0, x_1, x_2, x_3]$ sous une forme plus ssimple ? Trouver α et β tels que

$$f[x_0, x_1, x_2, x_3] = \alpha f[x_0, x_1, x_2] + \beta f[x_1, x_2, x_3].$$

Par comparaison des deux termes, on obtient

$$\alpha = \frac{1}{x_0 - x_3} \quad \textit{et} \quad \beta = \frac{1}{x_3 - x_0}$$

$$f[x_0, x_1, x_2, x_3] = \frac{f[x_1, x_2, x_3] - f[x_0, x_1, x_2]}{x_3 - x_0}.$$

La formule générale des différences divisées est donnée par l'expression récurrente :

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}$$

Nous avons écris $f[x_i]$ au lieu de $f(x_i)$. D'où, par exemple, on calcule $f[x_2, x_3]$ par

x_0	$f[x_0]$			
		$f[x_0, x_1]$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		
x_3	$f[x_3]$			

TABLE 3.1 – Schéma pour le calcul des différences divisées

$$f[x_2, x_3] = \frac{f[x_3] - f[x_2]}{x_3 - x_2}$$

et pour $f[x_1, x_2, x_3]$ par

$$f[x_1, x_2, x_3] = \frac{f[x_2, x_3] - f[x_1, x_2]}{x_3 - x_1}.$$

D'après l'expression générale du polynôme d'interpolation de Newton, on obtient

$$\begin{aligned} \mathcal{P}_n(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots \\ &\quad + f[x_0, \ldots, x_n](x - x_0)(x - x_1)\ldots(x - x_{n-1}). \end{aligned}$$

Exemple 3.4.2 *En utilisant le principe des différences divisées pour obtenir le polynôme d'interpolation tel que*

$$f(1) = 0, \quad f(-1) = -3, \quad f(2) = 4$$

Comme l'ulistre le tableau, on remarque

x	$f(x)$	Différences divisées		
1	0			
		$\frac{-3-0}{-1-1} = \frac{3}{2}$		
				$\frac{\frac{7}{3} - \frac{3}{2}}{2-1} = \frac{5}{6}$
-1	-3			
		$\frac{4-(-3)}{2-(-1)} = \frac{7}{3}$		
2	4			

TABLE 3.2 – Schéma pour le calcul des différences divisées

$$f[x_0, x_1] = \frac{3}{2}, \quad f[x_1, x_2] = \frac{7}{3}, \quad f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = \frac{5}{6}.$$

Ainsi, on obtient

$$\begin{aligned} \mathcal{P}_2(x) &= f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &= 0 + \frac{3}{2}(x - 1) + \frac{5}{6}(x - 1)(x + 1) \\ &= \frac{1}{6}\left(5x^2 + 9x - 14\right). \end{aligned}$$

3.5 Erreur d'interpolation

Dans cette section, nous évaluons l'erreur d'interpolation faite quand on remplace une fonction f donnée par un polynôme $\mathcal{P}_n$ qui l'interpole aux nœuds $x_0, x_1, \ldots, x_n$.

Théorème 3.5.1 *Soient $x_0, x_1, \ldots, x_n$, $(n+1)$ nœuds distincts et soit x un point appartenant au domaine de définition de f. On suppose que f esst de classe $\mathcal{C}^{n+1}$ sur le plus petit intervalle I_x contenant $x_0, x_1, \ldots, x_n$ et x. L'erreur d'interpolation au point x est donnée par*

$$\mathcal{E}_n(x) = f(x) - \mathcal{P}_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x - x_i),$$

où $\xi \in I_x$.

Démonstration. Comme $f(x_i) = \mathcal{P}_n(x_i)$ pour $i = 0, 1, \ldots, n$, alors il existe une fonction réelle ϕ telle que

$$f(x) - \mathcal{P}_n(x) = \phi(x)\prod_{i=0}^{n}(x - x_i),$$

Pour x fixé et $x \neq x_i$ ($i = 0, 1, \ldots, n$), posons $g(t) = f(t) - \mathcal{P}_n(t) - c\prod_{i=0}^{n}(x - x_i)$ où c est un paramètre réel choisi tel que $g(x) = 0$. Dans ce cas, la condition $g(x) = 0$ entraîne

$$c = \frac{f(x) - \mathcal{P}_n(x)}{\prod_{i=0}^{n}(x - x_i)}.$$

Maintenant, remarquons que la fonction g est de classe $\mathcal{C}^{n+1}$ sur $[a, b]$ et de plus

$$\begin{cases} g(x) = 0, \\ g(x_i) = 0, \quad \text{pour} \quad i = 0, \ldots, n, \end{cases}$$

c'est à dire que g admet $n+2$ racines distinctes deux à deux et est de de classe $\mathcal{C}^{n+1}$ sur $[a, b]$. Cela entraîne que g'' admet $n+1$ racines distinctes deux à deux et est de de classe $\mathcal{C}^{n}$ sur $[a, b]$.
Et en réitérant, on a g'' admet n racines distinctes deux à deux et est de de classe $\mathcal{C}^{n-1}$ sur $[a, b]$ et finalement, on arrive à vérifier que $g^{(n+1)}$ admet une racine sur $]a, b[$. Ainsi,

$$\exists \xi_x \in]a, b[, \quad \text{tel que} \quad g^{(n+1)}(\xi_x) = 0.$$

Or, $g^{(n+1)}(t) = f^{(n+1)}(t) - \mathcal{P}_n^{(n+1)}(t) - c(n+1)!$, et puisque $\mathcal{P}_n^{(n+1)}(t) = 0$ (car $\deg(\mathcal{P}_n) \leq n$) et $g^{(n+1)}(\xi_x) = 0$, alors on vérifie que

$$c = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x - x_i),$$

c'est-à- dire que

$$f(x) - \mathcal{P}_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x - x_i).$$

□

Corollaire 3.5.1 *Dans les conditions du théorème ci-dessus, on*

$$\sup_{x\in[a,b]} |f(x) - \mathcal{P}_n(x)| \leq \frac{M_{n+1}(b-a)^{n+1}}{(n+1)!},$$

où $M_{n+1} = \sup_{x\in[a,b]} |f^{(n+1)}(x)|$.

Démonstration. Il suffit d'utiliser la technique de majoration. □

Remarque 3.5.1 *La majoration de l'erreur d'interpolation donnée ci-dessus peut laisser croire que si le nombre d'abscisses d'in- terpolation* n *est grand, alors le polynôme d'interpolation* $\mathcal{P}_n$ *de* f *aux points d'interpolation* $x_0, x_1, \ldots, x_n$ *tend vers* f*. En fait, on n'a pas nécessairement* $\lim_{n\to+\infty} f(x) - \mathcal{P}_n(x) = 0$ *pour tout* $x \in [a, b]$ *car cette limite dépend aussi de la façon dont la quantité* M_n *se comporte lorsque la valeur de* n *devient large. L'exemple ci-dessous permet d'illustrer cette remarque.*

Exemple 3.5.1 *(Phénomène de Runge)*
Supposons qu'on approche la fonction suivante

$$f(x) = \frac{1}{1+9x^2}, \quad a \leq x \leq b.$$

Chapitre 4

Intégration et dérivation numérique

4.1 Introduction et outils de base

Nous présentonsdans ce chapitre les méthodes les plus couramment utilisées pour l'intégration numérique. Bien que nous nous limitons essentiellement aux intégrales sur des intervalles bornés. Eventuellement, nous oborderons des extensions aux intervalles non bornés.

Théorème 4.1.1 *(Première formule de la moyenne)*
Soient u et v deux fonctions continues sur $[a,b]$ telles que u est de signe constant dans $[a,b]$. Alors

$$\exists \eta \in]a,b[\quad \textit{tel que} \quad \int_a^b u(x)v(x)dx = v(\eta)\int_a^b u(x)dx.$$

4.2 Formule de quadrature

Soit f une fonction réelle intégrable sur un intervalle $[a,b]$. Lee calcul explicite de l'intégrale définie $I(f) = \int_a^b f(x)dx$ peut être difficile, voire impossible.

Définition 4.2.1 *On appelle* **formule de quadrature** *ou* **formule d'intégration numérique** *toute formule permettant de calculer une approximation de $I(f)$.*

Une possibilité consiste à remplacer f par une approximation f_n, où n est un entier positif, et calculer $I(f_n)$ au lieu de $I(f)$. En posant $I_n(f) = I(f_n)$, on a

$$I_n(f) = \int_a^b f_n(x)dx, \quad n \geq 1. \tag{1.1}$$

La dépendance par rapport aux extrémités a et b sera toujours sous-entendue.
On écrira donc $I_n(f)$ au lieu de $I_n(f;a,b)$.

Si $f \in \mathcal{C}^0([a,b])$, l'erreur de quatrature $\mathcal{E}_n(f) = I(f) - I_n(f)$ satisfait

$$|\mathcal{E}_n(f)| \leq \int_a^b |f(x) - f_n(x)|dx \leq (b-a)\sup_{x\in[a,b]} |f(x) - f_n(x)| = (b-a)\|f - f_n\|_\infty.$$

Donc, si pour un certain n, $\|f - f_n\|_\infty < \varepsilon$, alors $|\mathcal{E}_n(f)| \leq \varepsilon(b-a)$.

L'approximation f_n doit être facilement intégrable, ce qui est le cas si, par exemple, $f_n \in \mathbb{P}_n$. Une méthode naturelle consiste à prendre $f_n = \Pi_n f$, le polynôme d'interpolation de Lagrange de f sur un ensemble de $n+1$ nœuds distincts $\{x_i, \ \ i = 0, \ldots, n\}$. Ainsi, on définit de (1.1) que

$$I_n(f) = \sum_{i=0}^{n} f(x_i) \int_a^b \ell_i(x)dx, \quad n \geq 1. \tag{1.2}$$

où ℓ_i est le polynôme caractéristique de Lagrange de degré n associé au nœuds x_i. On note que (1.2) est un cas particulier de la formule de quadrature suivante :

$$I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i), \quad n \geq 1. \tag{1.3}$$

où les coefficients α_i de la combinaison linéaire sont donnés par $\alpha_i = \int_a^b \ell_i(x)dx$.

Le système $\{(f(x_1), \alpha_1), \ldots, (f(x_n), \alpha_n)\}$ est un système pondéré de somme $I_n(f)$. On dit aussi que la formule (1.3) est une somme pondérée des valeurs de f aux points x_i, pour $i = 0, \ldots, n$. On dit que ces points sont les **nœuds** de la formule de quadrature et que les nombres $\alpha_i \in \mathbb{R}$ sont ses **coefficients** ou encore ses **poids**. Les poids est les nœuds dépendent en général de n.

Définition 4.2.2 *On appelle le* **degré d'exactitude** *d'une formule de quadrature, le plus grand entier $r \geq 1$ pour lequel*

$$I_n(f) = I(f), \quad \forall f_n \in \mathbb{P}_r$$

4.3 Quadratures interpolatoires

4.3.1 Formule du rectangle ou du point milieu

Cette formule est obtenue en remplaçant f par une constante égale à la valeur de f au milieu de $[a, b]$, ce qui donne

$$I_0(f) = (b-a) f\left(\frac{a+b}{2}\right). \tag{1.4}$$

Le poids est donc $\alpha_0 = b - a$ et le nœud $x_0 = (a+b)/2$.
Si $f \in \mathcal{C}^2([a, b])$, l'erreur de quadrature est

$$\mathcal{E}_0(f) = \frac{h^3}{3} f''(\xi), \quad h = \frac{b-a}{2}, \tag{1.5}$$

où $\xi \in]a, b[$. En effet, le développement de Taylor au second ordre de f en $c = (a+b)/2$ s'écrit

$$f(x) = f(c) + f'(c)(x-c) + f''(\eta(x))(x-c)^2/2$$

en intégrant sur $[a, b]$ et on obtient

$$\int_a^b f(x)dx = (b-a)f(c) + \frac{h^3}{3} f''(\xi).$$

Exemple 4.3.1 *On considère la fonction $f : x \mapsto f(x) = \sin(x)$, on veut intégrer cette fonction sur l'intervalle $[1, 1.2]$.*
La formule du point milieu implique

$$\int_1^{1.2} \sin(x)dx = (1.2-1)\sin((1+1.2)/2) \approx 0.2 \times 0.8912 = 0.1782.$$

Par contre la valeur théorique est

$$\int_1^{1.2} \sin(x)dx = -\cos(1.2) + \cos(1) = 0.1779.$$

L'erreur relative est

$$\frac{|0.1779 - 0.1782|}{0.1779}.100\% = 0.17\%$$

Théorème 4.3.1 *(Formule de la moyenne discrète)*
Soit $u \in \mathcal{C}^0([a,b])$, soient x_j les $(s+1)$ points de $[a,b]$ et δ_j les $(s+1)$ constantes, toutes de même signe. Alors, il existe $\eta \in]a,b[$ tel que

$$\sum_{j=0}^{s} \delta_j u(x_j) = u(\eta) \sum_{j=0}^{s} \delta_j.$$

Démonstration. Soit $u_m = \min_{x\in[a,b]} u(x) = u(x_m)$ et $u_M = \max_{x\in[a,b]} u(x) = u(x_M)$, où x_m et x_M sont deux points de $[a,b]$. Alors

$$u_m \sum_{j=0}^{s} \delta_j \leq \sum_{j=0}^{s} \delta_j u(x_j) \leq u_M \sum_{j=0}^{s} \delta_j. \tag{1.6}$$

On pose $\sigma_s = \sum_{j=0}^{s} \delta_j u(x_j)$ et on considère la fonction continue $U(x) = u(x) \sum_{j=0}^{s} \delta_j$. D'après (1.6), on a $U(x_m) \leq \sigma_s \leq U(x_M)$. Le théorème de la moyenne implique l'existence d'un point $\eta \in]a,b[$ tel que $U(\eta) = \sigma_s$. d'où le résultat □

***Formule composite du rectangle** : Supposons maintenant qu'on approche l'intégrale $I(f)$ en remplaçant f par son interpolation polynomiale composite de degré 0 sur $[a,b]$, construite sur m sous-intervalles de largeur $h = (b-a)/m$, avec $m \geq 1$. En introduisant les nœuds de quadrature $x_k = a + (2k+1)h/2$, pour $k = 0, \ldots, m-1$, on obtient la formule composite du point milieu :

$$I_m = h \sum_{k=0}^{m-1} f(x_k), \quad m \geq 1. \tag{1.7}$$

Si $f \in \mathcal{C}^2([a,b])$, l'erreur de quadrature $\mathcal{E}_m(f) = I(f) - I_m(f)$ est donnée par

$$\mathcal{E}_m(f) = \frac{b-a}{24} h^2 f''(\xi), \tag{1.8}$$

où $\xi \in]a,b[$. On déduit de (1.8) que (1.7) a un degré d'exactitude égal à 1 ; on peut montrer (1.8) en utilisant (1.5) et la linéarité de l'intégration.
En effet, pour $k = 0, \ldots, m-1$ et $\xi_k \in]a+kh, a+(k+1)h[$,

$$\mathcal{E}_m = \sum_{k=0}^{m-1} f''(\xi_k)(h/2)^3/3 = \sum_{k=0}^{m-1} f''(\xi_k) \frac{h^2}{24} \frac{b-a}{m},$$

d'après la formule de la moyenne discrète, pour $u = f''$ et $\delta_j = 1$ pour $j = 0, \ldots, m-1$, on obtient

$$\mathcal{E}_m = \frac{b-a}{24} h^2 f''(\xi).$$

D'où

$$\int_a^b f(x)dx = h \sum_{k=0}^{m-1} f(a + (2k+1)h/2) + \frac{b-a}{24} h^2 f''(\xi), \quad \text{où } \xi \in]a, b[.$$

4.3.2 Formule du trapèze

Cette formule est obtenue en remplaçant f par $\Pi_1 f$, son polynôme d'interpolation de Lagrange de degrée 1 aux noeuds $x_0 = a$ et $x_1 = b$. Les noeuds de la formule de quadrature sont alors $x_0 = a$, $x_1 = b$ et ses poids $\alpha_0 = \alpha_1 = (b-a)/2$:

$$I_1(f) = \int_a^b f(x)dx = \frac{b-a}{2}[f(a) + f(b)].$$

En effet, soit $A(a, f(a))$ et $B(b, f(b))$ les deux points dont les abscisses respectifs a et b. Les points $M(x, y)$ du segment $[AB]$ ont une abscisse $x \in [a, b]$ et une ordonnée

$$y(x) = f(a) + \frac{f(b) - f(a)}{b-a}(x - a)$$

L'approximation par la méthode des trapèze implique

$$\int_a^b f(x)dx \approx \int_a^b y(x)dx$$

Analytiquement, on obtient

$$\int_a^b y(x)dx = \int_a^b \left(f(a) + \frac{f(b) - f(a)}{b-a}(x-a) \right) dx = \frac{b-a}{2} \left(f(a) + f(b) \right).$$

Si $f \in \mathcal{C}^2([a, b])$, l'erreur de quadrature est donnée par

$$\mathcal{E}_1(f) = -\frac{h^3}{12} f''(\xi), \quad h = b - a,$$

où ξ est un point de l'intervalle d'intégration. C'est-à-dire que

$$I(f) = I_1(f) + \mathcal{E}_1(f).$$

En effet, l'expression de l'erreur d'interpolation implique

$$\mathcal{E}_1(f) = \int_a^b (f(x) - \Pi_1 f(x))dx = -\frac{1}{2} \int_a^b f''(\xi_x)(x-a)(b-x)dx.$$

Comme $\omega_2(x) = (x-a)(x-b) < 0$ sur $]a, b[$, alors d'après le théorème de la moyenne on a

$$\mathcal{E}_1(f) = \frac{1}{2} f''(\xi) \int_a^b \omega_2(x)dx = f''(\xi) \frac{(b-a)^3}{12},$$

pour un $\xi \in]a, b[$, d'où le résultat.$\diamondsuit$

Exemple 4.3.2 *On considère la fonction $f : x \mapsto f(x) = e^x$, on veut intégrer cette fonction sur l'intervalle $[1, 2]$. La valeur théorique de cette intégrale est*

$$\int_1^2 e^x dx = e^2 - e^1 = e(e-1) \approx 4.6708.$$

Maintenant, on calcule la valeur approximative de l'intégrale où bien l'intégration numérique en utilisant la méthode du trapèze. Soit $A(1, e)$ et $B(2, e^2)$ les deux points du plan ayant pour abscisses respectifs 1 et 2, les points du segmant $[AB]$ ont une abscisse $x \in [1, 2]$ et une ordonnées $y(x) = e[1 + (e-1)(x-1)]$. On a bien $y(1) = e$ et $y(2) = e^2$.

$$\int_1^2 y(x)dx = \int_1^2 e[1 + (e-1)(x-1)]dx = (2-1)/2(e^1 + e^2) \approx 5.0537$$

D'où l'erreur relative est

$$\frac{|4.6708 - 5.0537|}{4.6708}.100\% = 7.4\%$$

Exemple 4.3.3 *On considère la fonction $f : x \mapsto f(x) = \sin(x)$, on veut intégrer cette fonction sur l'intervalle $[1, 1.2]$.*
La formule des trapèze implique

$$\int_1^{1.2} \sin(x)dx = (1.2-1)/2(\sin(1) + \sin(1.2)) \approx 0.1 \times 1.7735 = 0.1774.$$

Par contre la valeur théorique est

$$\int_1^{1.2} \sin(x)dx = -\cos(1.2) + \cos(1) = 0.1779.$$

L'erreur relative est

$$\frac{|0.1779 - 0.1774|}{0.1779}.100\% = 0.28\%$$

On peut observer que par la relation de Chasles pour les intégrales, on a $[a, b] = [a, c] \cup [c, b]$ où $c = (a+b)/2$, et on a

$$\int_a^b f(x)dx = \int_a^c f(x)dx + \int_c^b f(x)dx.$$

D'après la formule de trapèze pour deux points, on obtient

$$\int_a^b f(x)dx \approx \left[\frac{c-a}{2}(f(a) + f(b))\right] + \left[\frac{b-c}{2}(f(b) + f(c))\right]$$

or, $c - a = b - c = (b-a)/2$, alors

$$\int_a^b f(x)dx \approx \frac{b-a}{4}\left[f(a) + 2f(c) + f(b)\right].$$

***Formule composite d'intégration par la méthode de trapèze** : Soit $[a, b]$ un intervalle de $\mathbb{R}$ et $n \geq 1$ un entier. Soit $h = \frac{b-a}{n}$ et on définit $x_i = a + ih$ pour $i = 0, \ldots, n$. On considère la subdivision suivante

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$$

Alors on a

$$\int_a^b f(x)dx = \int_{x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{n-1}}^{x_n} f(x)dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx$$

D'après la formule des trapèzes pour deux points x_i et x_{i+1} on a

$$\int_{x_i}^{x_{i+1}} f(x)dx = \frac{x_{i+1} - x_i}{2} [f(x_{i+1}) + f(x_i)] = \frac{h}{2} [f(x_{i+1}) + f(x_i)]$$

Alors

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx = \frac{h}{2} \sum_{i=0}^{n-1} [f(x_{i+1}) + f(x_i)]$$

finalement,

$$\int_a^b f(x)dx \approx \frac{h}{2} \left[f(x_0) + 2 \sum_{i=1}^{n-1} f(x_i) + f(x_n) \right]$$

où bien

$$\int_a^b f(x)dx = \frac{h}{2} \left[f(a) + 2 \sum_{i=1}^{n-1} f(x_i) + f(b) \right] - \frac{b-a}{12} h^2 f''(\xi),$$

où $\xi \in]a, b[$. Le degré d'exactitude est à nouveau égal à 1.

4.3.3 Formule de Cavalieri-Simpson

Soit $[a, b]$ un intervalle de $\mathbb{R}$. On définit $c = (b + a)/2$ et soit $p_1(x)$ un polynôme de degré ≤ 2 qui interpole f aux points a, c et b. Alors,

$$p_2(x) = A + Bx + Cx^2,$$

où A, B et C sont des constantes à déterminer tels que

$$p_2(a) = f(a), \;\; p_2(c) = f(c) \;\; p_2(b) = f(b)$$

Après avoir déterminé A, B et C, on obtient

$$\int_a^b p_2(x)dx = \frac{b-a}{6} [f(a) + 4f(c) + f(b)]$$

et le rôle de Simpson est

$$\int_a^b f(x)dx \approx \int_a^b p_2(x)dx = \frac{b-a}{6} [f(a) + 4f(c) + f(b)] .$$

On peut montrer que, si $f \in \mathcal{C}^4([a, b])$, l'erreur de quadrature est

$$\mathcal{E}_2(f) = -\frac{h^5}{90} f^{(4)}(\xi) = -\frac{(b-a)^5}{2880} f^{(4)}(\xi), \quad h = \frac{b-a}{2},$$

où $\xi \in]a, b[$. On en déduit que le degré d'exactitude est 3.

$$\int_a^b f(x)dx = \frac{b-a}{6} [f(a) + 4f(c) + f(b)] - \frac{(b-a)^5}{2880} f^{(4)}(\xi), \quad \xi \in]a, b[.$$

***Formule composite d'intégration par la méthode de Simpson** :
Pour $n \geq 0$, soit $h = (b-a)/n$ le pas d'une subdivision $\mathcal{S}$ de l'intervalle $[a, b]$. On définit $x_i = a + ih$ pour $i = 0, \ldots, n$ où $x_0 = a$ et $x_n = b$.

$$\begin{aligned} \int_a^b f(x)dx &= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx \\ &\approx \frac{h}{6} \sum_{i=1}^{n} \Big(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i) \Big) \end{aligned}$$

où $x_{i-\frac{1}{2}} = \frac{1}{2}(x_i + x_{i-1})$.
Si $f \in \mathcal{C}^4([a, b])$ alors l'erreur de quadrature associée à cette intégration numérique est donnée par

$$\mathcal{E}(f) = -\frac{b-a}{180 \times 2^4} h^4 f^{(4)}(\xi)$$

où $\xi \in]a, b[$. Et, dans ce cas on écrit

$$\int_a^b f(x)dx = \frac{h}{6} \sum_{i=1}^{n} \Big(f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i) \Big) - \frac{b-a}{2880} h^4 f^{(4)}(\xi).$$

***Autre formule composite de Simpson** :
Si on introduit les nœuds de quadrature $x_k = a + \frac{k}{2}h$ pour $k = 0, \ldots, 2n$, avec $h = (b-a)/n$ le pas d'une subdivision $\mathcal{S}$ de l'intervalle $[a, b]$. on a alors

$$\begin{aligned} \int_a^b f(x)dx &= \sum_{i=1}^{n} \int_{x_{i-1}}^{x_i} f(x)dx \\ &\approx \frac{h}{6} \left(f(a) + 2 \sum_{i=1}^{n-1} f(x_{2i}) + 4 \sum_{i=0}^{n-1} f(x_{2i+1}) + f(b) \right) \end{aligned}$$

où $x_0 = a$ et $x_{2n} = b$.

Chapitre 5

Méthode des moindres carrés et optimisation quadratique

Nous nous intéresserons dans cette partie au problème de maximisation ou minimisation des fonctions de plusieurs variables. Commençons par les fonctions de deux variables. Nous rappelons ici un théorème fondamental, qui va servir dans ce chapitre

Théorème 5.0.2 *Une matrice symétrique A est définie positive si et seulement si $det(A_k) > 0$ pour toutes les sous-matrices principales A_k de A.*

5.1 Maxima et minima de fonctions de deux variables

5.1.1 Gradient d'une application et Matrice hessienne d'une F.P.V

Soit $f : \mathbb{R}^n \longrightarrow \mathbb{R}^p$ une application définie par

$$f(X) = (f_1(X), f_2(X), \ldots, f_p(X))$$

où $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Supposons que l'application est différentiable sur $\mathbb{R}^n$.

- Pour $p = 1$. On appelle **Gradient** de f au point X_0, noté $\nabla f(X_0) = \overrightarrow{\text{grad}} f(X_0)$, est le vecteur pointant dans la direction de croissance maximale de la fonction f où bien il est le vecteur définie par

$$\nabla f(X_0) = \begin{pmatrix} \frac{\partial f}{\partial x_1}(X_0) \\ \\ \frac{\partial f}{\partial x_2}(X_0) \\ \vdots \\ \frac{\partial f}{\partial x_n}(X_0) \end{pmatrix}$$

- Pour $p > 1$. On appelle **Gradient** de f au point X_0, noté $\nabla f(X_0)$, la matrice jacobienne de taille $(n \times p)$ définie par

$$J_f(X_0) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(X_0) & \frac{\partial f_1}{\partial x_2}(X_0) & \ldots & \frac{\partial f_1}{\partial x_n}(X_0) \\ \\ \frac{\partial f_2}{\partial x_1}(X_0) & \frac{\partial f_2}{\partial x_2}(X_0) & \ldots & \frac{\partial f_2}{\partial x_n}(X_0) \\ \vdots & \vdots & \ddots & \vdots \\ \\ \frac{\partial f_p}{\partial x_1}(X_0) & \frac{\partial f_p}{\partial x_2}(X_0) & \ldots & \frac{\partial f_p}{\partial x_n}(X_0) \end{pmatrix}$$

Définition 5.1.1 *On appelle* **matrice hessienne** *de f au point X_0, noté $\mathcal{H}_f(X_0)$, la matrice de taille $(n \times n)$ définie par*

$$\mathcal{H}_f(X_0) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(X_0) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(X_0) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(X_0) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(X_0) & \frac{\partial^2 f}{\partial x_2^2}(X_0) & \dots & \frac{\partial f_2}{\partial x_2 \partial x_n}(X_0) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(X_0) & \frac{\partial^2 f}{\partial x_n \partial x_2}(X_0) & \dots & \frac{\partial^2 f}{\partial x_n^2}(X_0) \end{pmatrix}$$

5.1.2 Approximations linéaire et quadratique : Formule de Taylor

On considère $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ une fonction de plusieurs variables définie par

$$f(X) = f(x_1, \dots, x_n)$$

Supposons que f est de classe $\mathcal{C}^2$ au voisinage d'un point X_0. Soit $h \in \mathbb{R}^n$.
La formule de Taylor de f en le point x_0 (ou la formule d'approximation) est donnée par l'expression suivante :

$$f(X_0 + h) = f(X_0) + (\nabla f(X_0), h) + \frac{1}{2}(\mathcal{H}_f(X_0)h, h) + o(\|h\|^2).$$

Que se passe-t-il si $\nabla f(X_0) = 0_{\mathbb{R}^n}$? si $(\mathcal{H}_f(X_0)h, h) < 0$? si $(\mathcal{H}_f(X_0)h, h) > 0$? si $(\mathcal{H}_f(X_0)h, h) = 0$?

5.1.3 Points critiques d'une application

Considérons une fonction de deux variables définie par $f(x, y)$. Nous supposerons que f est de classe $\mathcal{C}^3$ (i.e. f et ses dérivées partielles d'ordre≤ 3 sont continues).

Définition 5.1.2 *On dit que (x_0, y_0) est un point critique de f si $f_x(x_0, y_0) = 0$ et $f_y(x_0, y_0) = 0$ où f_x et f_y dénotent les dérivées partielles de f par rapport à x et y respectivement.*

Soit $X_0 = (x_0, y_0)$ un point critique de f. En utilisant le formule de Taylor on obtient :

$$f(x_0 + h, y_0 + k) = f(X_0) + hf_x(X_0) + kf_y(X_0) + \frac{1}{2}\left[h^2 f_{xx}(X_0) + 2hkf_{xy}(X_0) + k^2 f_{yy}(X_0)\right] + R$$

où $R = \frac{1}{6}\left[h^3 f_{xxx}(X) + 3h^2 k f_{xxy}(X) + 3hk^2 f_{xyy}(X) + k^3 f_{yyy}(X)\right]$ avec $X = (x_0 + \theta h, y_0 + \theta k)$ $(0 < \theta < 1)$.
Or $f_x(x_0, y_0) = 0$ et $f_y(x_0, y_0) = 0$ puisque X_0 est un point critique. Il en résulte que

$$f(x_0 + h, y_0 + k) = f(X_0) + \frac{1}{2}\left[ah^2 + 2bhk + ck^2\right] + R$$

où on a posé $a = f_{xx}(X_0)$, $b = f_{xy}(X_0)$ et $c = f_{yy}(X_0)$. Si h et k sont suffisamment petits, on montre alors que

$$|R| \leq \frac{1}{2}|ah^2 + 2bhk + ck^2|$$

Par conséquent, $f(x_0 + h, y_0 + k) - f(x_0, y_0)$ a le même signe que $ah^2 + 2bhk + ck^2$ pour h et k petits, ce qui entraîne que (x_0, y_0) est minimum local si

$$ah^2 + 2bhk + ck^2 > 0, \quad \text{pour tout}(h, k),$$

est un maximum local si

$$ah^2 + 2bhk + ck^2 < 0, \quad \text{pour tout}(h,k),$$

ou est un point selle si $ah^2 + 2bhk + ck^2$ prend des valeurs positives et des valeurs négatives.
En d'autres mots, on a minimum local si la forme quadratique

$$q(h,k) = ah^2 + 2bhk + ck^2$$

est définie positive,un maximum local si elle est définie négative ou un point selle si elle est indéfinie. La forme quadratique s'écrit

$$q(h,k) = (h,k)\begin{pmatrix} f_{xx}(X_0) & f_{xy}(X_0) \\ f_{yx}(X_0) & f_{yy}(X_0) \end{pmatrix}\begin{pmatrix} h \\ k \end{pmatrix}$$

La matrice précédente

$$H(X_0) = \begin{pmatrix} f_{xx}(X_0) & f_{xy}(X_0) \\ f_{yx}(X_0) & f_{yy}(X_0) \end{pmatrix}$$

définissant la forme quadratique est appelée la matrice hessienne de f en X_0. Soient λ_1 et λ_2 les valeurs propres de $H(X_0)$. On déduit du théorème 5.0.2 :

1. (x_0, y_0) est un minimum local si $\lambda_1 > 0$ et $\lambda_2 > 0$.
2. (x_0, y_0) est un maximum local si $\lambda_1 < 0$ et $\lambda_2 < 0$.
3. (x_0, y_0) est un point selle si $\lambda_1\lambda_2 < 0$.

Comme on l'a vu au théorème 5.0.2, la matrice hessienne $H(X_0)$ est définie positive si ses déterminants principaux $f_{xx}(X_0)$ et $\Delta = det(H(X_0)) = f_{xx}(X_0)f_{yy}(X_0) - (f_{xy}(X_0)))^2$ sont > 0.
La matrice hessienne est définie négative si $-H(X_0)$ est définie positive, ce qui sera le cas si ses déterminants principaux $-f_{xx}(X_0)$ et $det(-H(X_0)) = det(H(X_0)) = \Delta$ sont > 0, c'est-à-dire $f_{xx}(X_0) < 0$ et $\Delta > 0$. Il résulte de ce qui précède que $H(X_0)$ est indéfinie si $\Delta < 0$. Ainsi, on obtient :

1. Si $\Delta > 0$ et $f_{xx}(X_0) > 0$, alors $X_0 = (x_0, y_0)$ est un minimum local.
2. Si $\Delta > 0$ et $f_{xx}(X_0) < 0$, alors $X_0 = (x_0, y_0)$ est un maximum local.
3. Si $\Delta < 0$, alors $X_0 = (x_0, y_0)$ est un point selle.

Remarque 5.1.1 *Si $\Delta = 0$, on peut rien conclure quant à la nature du point critique.*

Exemple 5.1.1 *Considérons la fonction donnée par*

$$f(x,y) = \frac{1}{3}x^3 + xy^2 - 4xy + 1.$$

Déterminons d'abord les points critiques.
Ce sont les solutions du système suivant :

$$\begin{cases} f_x(x,y) = \frac{\partial F}{\partial x}(x,y) = 0 \\ f_y(x,y) = \frac{\partial F}{\partial y}(x,y) = 0 \end{cases}$$

On a

$$\begin{cases} f_x(x,y) = \frac{\partial F}{\partial x}(x,y) &= x^2 + y^2 - 4y \\ f_y(x,y) = \frac{\partial F}{\partial y}(x,y) &= 2xy - 4x \end{cases}$$

On obtient les quatre points critiques : $X_1 = (0,0)$, $X_2 = (0,4)$, $X_3 = (2,2)$ et $X_4 = (-2,2)$. La matrice hessienne de f est

$$H(x,y) = \frac{\partial^2 F}{\partial x \partial y}(x,y) = \begin{pmatrix} 2x & 2y-4 \\ 2y-4 & 2x \end{pmatrix}.$$

En évaluant la matrice hessienne aux points $X_1, \ldots, X_4$ on a

$$H(0,0) = \begin{pmatrix} 0 & -4 \\ -4 & 0 \end{pmatrix}, \quad H(0,4) = \begin{pmatrix} 0 & 4 \\ 4 & 0 \end{pmatrix},$$

$$H(2,2) = \begin{pmatrix} -4 & 0 \\ 0 & -4 \end{pmatrix}, \quad H(-2,2) = \begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}.$$

Les valeurs propres de ces matrices sont :

1. *Pour $H(0,0)$ sont $\lambda_1 = 4$ et $\lambda_2 = -4$.*
2. *Pour $H(0,4)$ sont $\lambda_1 = 4$ et $\lambda_2 = -4$.*
3. *Pour $H(2,2)$ sont $\lambda_1 = 4$ et $\lambda_2 = 4$.*
4. *Pour $H(-2,2)$ sont $\lambda_1 = -4$ et $\lambda_2 = -4$.*

Par conséquent, X_1 et X_2 sont des points selles, X_3 est minimum local et X_4 est un maximum local.

5.1.4 Maxima et minima des fonctions de n variables

Considérons une fonctions de n variables définie par $F(x) = F(x_1, \ldots, x_n)$. Un point $a = (a_1, \ldots, a_n)$ est un point critique si $F_{x_i}(a) = 0$ pour $i = 1, \ldots, n$, où $F_{x_i} = \frac{\partial F}{\partial x_i}$ désigne la dérivée partielle par rapport à x_i.
C'est-à-dire que Les points critiques de l'application $F : (x_1, \ldots, x_n) \longmapsto F(x_1, \ldots, x_n)$ sont les solutions de l'équation vectorielle suivante :

$$\nabla F(x_1, \ldots, x_n) = 0$$

ce qui équivalent au système d'équations suivant

$$\begin{cases} \frac{\partial F}{\partial x_1}(x_1, \ldots, x_n) &= 0 \\ \\ \frac{\partial F}{\partial x_2}(x_1, \ldots, x_n) &= 0 \\ \\ \qquad\qquad\qquad\quad \vdots & \\ \\ \frac{\partial F}{\partial x_n}(x_1, \ldots, x_n) &= 0 \end{cases}$$

Utilisant, comme pour les fonctions de deux variables, la formule de Taylor on montre qu'un point critique X

1. est un minimum local si la matrice hessienne $H(X)$ est définie positive,
2. est un maximum local si $H(X)$ est définie négative
3. est un point selle si $H(X)$ est indéfinie.

Ici, la matrice hessienne est la matrice symétrique $n \times n$ définie par

$$H(X) = (F_{x_i x_j}(X)) = \left(\frac{\partial^2 F}{\partial x_i \partial x_j}(x_1, \ldots, x_n) \right),$$

où $i = 1, \ldots, n$ et $j = 1, \ldots, n$.

Exemple 5.1.2 *On considère la fonction définie par*

$$F(x,y,z) = x^2 + xz - 3\cos(y) + z^2.$$

Déterminons les points critiques. Les dérivées partielles sont

$$\begin{cases} \frac{\partial F}{\partial x}(x,y,z) &= 2x+z \\ \frac{\partial F}{\partial y}(x,y,z) &= 3\sin(y) \\ \frac{\partial F}{\partial z}(x,y,z) &= x+2z \end{cases}$$

Ainsi, les points critiques sont les solutions de

$$\begin{cases} 2x+z &= 0 \\ 3\sin(y) &= 0 \\ x+2z &= 0 \end{cases}$$

Il en résulte que les points critiques sont les points $X_k = (0, k\pi, 0)$, $k \in \mathbb{Z}$. La matrice hessienne au point X_k est

$$H(X_k) = \begin{pmatrix} 2 & 0 & 1 \\ 0 & 3(-1)^k & 0 \\ 1 & 0 & 2 \end{pmatrix}.$$

- *Si k est pair, les valeurs propres de $H(X_k)$ sont 3, 3 et 1. Ainsi, $H(X_k)$ est définie positive si k est pair.D'où il s'ensuit que X_k est un minimum local si k pair.*
- *si k est impair, les valeurs propres sont 3, -3 et 1. Ainsi, $H(X_k)$ est indéfinie si k est impair. D'où X_k est un point selle si k est impair.*

5.2 Fonctions quadratiques

5.2.1 Forme linéaires et bilinéaires

Définition 5.2.1 *Une forme linéaire ℓ sur un espace vectoriel réel E est une application linéaire de E dans $\mathbb{R}$.*
C'est-à-dire pour tout $x, y \in E$ et $\alpha, \beta \in \mathbb{R}$ on a

$$f(\alpha x + \beta y) = \alpha f(x) + \beta f(y).$$

Si E est dimension finie n, $(e_1, \ldots, e_n)$ sa base canonique et $L = (\ell(e_1), \ldots, \ell(e_n))$ le vecteur ligne des images de ℓ dans la base $(e_1, \ldots, e_n)$. Soit $x = x_1e_1 + \ldots + x_ne_n$ un vecteur dans E et $X = (x_1, \ldots, x_n)$, alors

$$\ell(x) = (L, X)$$

En dimension finie, toute forme linéaire est représentée par un produit scalaire.

Définition 5.2.2 *Une forme bilinéaire a sur un espace vectoriel réel E est une application de $E \times E$ dans $\mathbb{R}$, linéaire par rapport à chacune de ses deux arguments.*
Soit a la forme bilinéaire, on a donc :

$$a(u, \alpha v_1 + \beta v_2) = \alpha a(u, v_1) + \beta a(u, v_2) \quad (1.1)$$
$$a(\alpha u_1 + \beta u_2, v) = \alpha a(u_1, v) + \beta a(u_2, v) \quad (1.2)$$

Représentation matricielle

Toute forme bilinéaire sur un espace E de dimension finie n se représente, dans une basee $(e_1, \ldots, e_n)$ par une matrice carrée d'ordre n. Les coefficients A_{ij} de la matrice A représentant l'application a sont donnés par

$$A_{ij} = a(e_i, e_j)$$

Soit $u = u_1e_1 + \ldots + u_ne_n$ et $v = v_1e_1 + \ldots + v_ne_n$ on a

$$a(u, v) = (AU, V) = (U, A^TV)$$

où $U = (u_1, \ldots, u_n)^T$ et $V = (v_1, \ldots, v_n)^T$.
Si $(\cdot, \cdot)$ représente le produit scaalaire usuel de $\mathbb{R}^n$ et A^T est la matrice transposée de A définie par

$$A^T_{ij} = A_{ji}, \quad \forall i, j = 1, \ldots, n$$

Définition 5.2.3 *1. Une forme bilinéaire a sur un espace vectoriel réel E est symétrique si :*

$$\forall u, v \in E, \quad a(u, v) = a(v, u).$$

2. Une forme bilinéaire a sur un espace vectoriel réel E est définie positive si :

$$\forall u \in E, \quad a(u, u) \geq 0 \quad \textit{et} \quad a(u, u) = 0 \Leftrightarrow u = 0$$

Les formes bilinéaires symétriques sont représentées par des matrices symétriques $A_{ij} = A_{ji}$.
Les formes bilinéaires symétriques définies positives sont représentées par des matrices symétriques définies positives, qui vérifient donc

$$(AU, U) \geq 0, \quad \forall U \in \mathbb{R}^n \quad \textit{et} \quad (AU, U) = 0 \;\Rightarrow\; U = 0$$

Théorème 5.2.1 ***(Résultat important)***

1. *Les matrices symétriques réelles ont des valeurs propres réelles, sont diagonalisables et admettent une base de vecteurs propres orthonormés.*
2. *Les matrices symétriques réelles définies positives ont des valeurs propres strictement positives, et donc sont inversibles.*

5.2.2 Équivalence entre la résolution d'un système linéaire et la minimisation quadratique

On considère $a : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ une forme bilinéaire et A la matrice associée à a relativement à la base $(e_1, \ldots, e_n)$:

$$A_{ij} = a(e_i, e_j), \quad i, j = 1, \ldots, n$$

On définit la fonctionnelle $J : \mathbb{R}^n \to \mathbb{R}$ donnée pour tout $v \in \mathbb{R}^n$ par

$$J(v) = \frac{1}{2}(Av, v) - (b, v)$$

où $b \in \mathbb{R}^n$ est un vecteur donné.

Théorème 5.2.2 *Si A est une matrice symétrique définie positive, il y a équivalence entre les trois problèmes suivants :*

$$(1) \quad \left\{ \begin{array}{rcl} \textit{trouver } X & \in & \mathbb{R}^n \quad \textit{tel que} \\ AX & = & b \end{array} \right.$$

$$(2) \quad \left\{ \begin{array}{rcl} \textit{trouver } X & \in & \mathbb{R}^n \quad \textit{tel que} \\ (AX, Y) & = & (b, Y), \quad \forall Y \in \mathbb{R}^n \end{array} \right.$$

$$(3) \quad \left\{ \begin{array}{rcl} \textit{trouver } X & \in & \mathbb{R}^n \quad \textit{tel que} \\ J(X) & = & \frac{1}{2}(AX, X) - (b, X) \quad \textit{soit minimale}. \end{array} \right.$$

Démonstration. $(1) \Rightarrow (2)$ est évident.
$(2) \Rightarrow (1)$ en prenant pour Y les vecteurs de base e_i de $\mathbb{R}^n$.
$(2) \Rightarrow (3)$ On calcule $J(X + \lambda Y)$ pour tout λ réel et tout $Y \in \mathbb{R}^n$, on obtient

$$J(X + \lambda Y) = J(X) + \lambda \left((AX, Y) - (b, Y) \right) + \frac{1}{2}\lambda^2 (AY, Y)$$

en utilisant la symétrie de la matrice A.
On en déduit, si $(AX, Y) - (b, Y) = 0$ que $J(X + \lambda Y) = J(X) + \frac{1}{2}\lambda^2 (AY, Y)$
d'où en utilisant le fait que A est définie positive :

$$J(X + \lambda Y) > J(X)$$

si λ et Y sont non nuls. Donc on a montré que si X vérifie (2), X minimise J.
$(3) \Rightarrow (2)$, car X minimise J, on a

$$\lambda \left((AX, Y) - (b, Y) \right) + \frac{1}{2}\lambda^2 (AY, Y) \geq 0, \quad \forall \lambda, \ \forall Y$$

le trinôme en λ ci-dessus doit être toujours positif. Ceci entraîne que son discriminant soit toujours négatif ou nul. Or ce discriminant est

$$\triangle = \left((AX, Y) - (b, Y) \right)^2$$

ceci implique (2). □

5.3 Application aux moindres carrés

Considérons le problème générale d'un système linéaire sur-déterminé, c'est-à-dire dans lequel il y a plus d'équations que d'inconnues. C'est en particulier le cas dans le calcul de la droite des moindre carré ou plus généralement de polynômes d'approximation au sens des moindres carrés. On ne peut pas obtenir exactement l'égalité

$$AX = b$$

où A est une matrice rectangulaire de n lignes et m colonnes avec $n > m$. On définit

$$J(X) = \|AX - b\|^2 = (AX - b, AX - b).$$

On essaie alors de minimiser l'écart entre les vecteurs AX et B de

$$\min_{X \in \mathbb{R}^m} J(X)$$

en minimisant la norme euclidienne de leur différence, ou ce qui revient au même le carré de cette norme. On utilise les propriétés classiques du produit scalaire $(AU, V) = (U, A^T V)$ pour obtenir :

$$J(X) = (A^T AX, X) - 2(A^T b, X) + (b, b)$$

la matrice $A^T A$ est une matrice carrée $(m \times m)$ symétrique définie positivedès lors que la matrice rectangulaire A est de rang m. Le théorème (5.2.2) nous donne l'équivalence de ce problème de moindre carrés avec la résolution du système linééaire

$$A^T AX = A^T b$$

On retrouve ainsi le système carré de m équations à m inconnues, dit "système des équations normales".

5.3.1 Approximation par la droite des moindre carrés

Par exemple, dans le cas de la droite des moindre carrés, il s'agit de trouver la fonction affinie

$$y = \alpha + \beta x$$

qui représente "au mieux" une collection de n valeurs y_i associées aux n abscisses x_i. Au sens des moindres carrés , ceci revient à minimiser la somme

$$\sum_{i=1}^{n} (y_i - (\alpha + \beta x_i))^2$$

donc à minimiser la norme euclidienne de la différence

$$J(a) = \|Aa - b\|^2$$

où l'on a noté b le vecteur des n valeurs y_i, A la matrice rectangulaire à n lignes et 2 colonnes

$$A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{n-1} \\ 1 & x_n \end{pmatrix}$$

et $a = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$. En appliquant les résultats précédents, on obtient la solution en résolvant le système (2×2) suivant

$$A^T Aa = A^T b$$

5.3.2 Interprétation géométrique : projection sur un sous-espace

Reprenons le problème de l'approximation au sens des moindre carrés

$$J(x) = \|Ax - b\|^2$$

On peut interpréter ce problème comme celui de la recherche du vecteur de la forme Ax le plus proche au sens de la norme euclidienne d'un vecteur b donné dans $\mathbb{R}^n$.
Les vecteurs de la forme Ax sont une combinaison linéaire des m vecteurs colonnes de la matrice A. Ces

vecteurs colonnes sont indépendants car A est supposée de rang m.
Ils engendrent donc un sous-espace vectoriel F de $\mathbb{R}^n$ de dimension m. Et le problème s'interprète comme la recherche du vecteur du sous-espace F le plus proche du vecteur b.
On obtient donc x en écrivant que Ax est la projection orthogonale de b dans F donc

$$(Ax - b, V) = 0, \quad \forall V \in F$$

ceci car $\mathbb{R}^n = F \oplus F^\perp$ et que $Ax - b \in F^\perp$.
ce qui est équivalent, puisque les colonnes de A engendrent F, à

$$(Ax - b, A_j) = 0, \quad \forall j = 1, \ldots, n$$

où A_j est le j^{ime} vecteur colonne de A. On retrouve ainsi le résultat

$$A^T Ax = A^T b.$$