Analyse Numérique

Institut de Mathématiques de Bourgogne (IMB)
Université de Bourgogne Franche-Comté

Organisation

- CM: 24 heures, 12 séances. TD: 18 heures, 9 séances. TD: 8 heures, 4 séances.
- Page web :
  `http://jaramillo.perso.math.cnrs.fr/Courses/AnalyseNumerique/AnalyseNumerique.html`

Intervenants

- J.L. Jaramillo (resp. Bureau : A323), CM (`Jose-Luis.Jaramillo@u-bourgogne.fr`).
- X. Dupuis (Bureau: A329), TD et TP (`(Xavier.Dupuis@u-bourgogne.fr`).

Table des matières

I Problèmes linéaires 5

1 Introduction: problèmes linéaires 7
- 1.1 Le problème et les méthodes . . . 7
- 1.2 Diagonalisation de matrices réelles symétriques . . . 7
- 1.3 Exemples de matrices symétriques . . . 10

2 Méthodes directes 15
- 2.1 Décomposition *LU* (méthode de Gauss/ d'échelonnement) . . . 15
  - 2.1.1 Algorithme LU, sans permutation . . . 18
  - 2.1.2 Algorithme LU, avec permutation . . . 20
- 2.2 Décomposition de Choleski . . . 27

3 Normes et conditionnement d'une matrice 35
- 3.1 Norme matricielle, norme matricielle induite et rayon spectral . . . 35
  - 3.1.1 Rappel sur les normes . . . 35
  - 3.1.2 Normes matricielles . . . 36
- 3.2 Relation entre norme matricielle et rayon spectral . . . 39
  - 3.2.1 Convergence et rayon spectral . . . 39
- 3.3 Conditionnement d'une matrice . . . 41
  - 3.3.1 Conditionnement et ses propriétés élémentaires . . . 41
  - 3.3.2 Majoration des erreurs . . . 42

4 Méthodes itératives 45

5 Méthodes de descente 47

II Interpolation 49

6 Interpolation polynomiale 51

III Annexes 53

A Espaces linéaires 55
- A.1 Rappel sur les espaces linéaires . . . 55
- A.2 Application linéaire . . . 56
- A.3 Produit scalaire . . . 56

B Matrices **59**
B.1 Définitions . . . 59
B.1.1 Opérations sur les matrices . . . 60
B.1.2 Rang et noyau d'une matrice . . . 61
B.1.3 Inverse d'une matrice . . . 61
B.2 Applications linéaires et matrice d'une application linéaire . . . 62
B.2.1 Matrices semblables . . . 62
B.3 Matrices et géométrie . . . 63

C Factorisation de matrices **65**

D Diagonalisation **67**
D.1 Notion de diagonalisation . . . 67
D.2 Diagonalisation de matrices normales . . . 68

Part I

Problèmes linéaires

Chapitre 1

Introduction: problèmes linéaires

1.1 Le problème et les méthodes

Le problème fondamental que nous allons aborder est la résolution de l'équation

$$A\boldsymbol{x} = \boldsymbol{b} \ , \tag{1.1}$$

avec A une matrice et $\boldsymbol{b}$ un vecteur donnés dans des espaces linéaires de dimension fini. Un problème lié qui va aussi nous occuper est la résolution du problème spectral de la matrice A

$$A\boldsymbol{x} = \lambda\boldsymbol{x} \ . \tag{1.2}$$

Les appendices A-C offrent un rappel des éléments de base sur les espaces linéaires et sujets liés.

La résolution "efficace" de l'équation (1.1) pour des dimensions élevées pose un problème difficile dont la résolution demande le concours de différentes techniques mathématiques. L'esprit ici est de souligner l'enchevêtrement de l'Algèbre Linéaire, la Géométrie et l'Analyse Mathématique dans l'Analyse Numérique. Le problème abordé dans la section suivante illustre ce point.

1.2 Diagonalisation de matrices réelles symétriques

Les matrices carrées réelles et symétriques sont diagonalisables, avec des valeurs et vecteurs propres réels. D'un côté, ceci est un résultat fondamental dans le contexte du problème abordé dans ce cours. D'un autre côté, la preuve de ce résultat nous permet d'illustrer concrètement la puissance de la combinaison d'Analyse Mathématique et de l'Algèbre Linéaire, un leitmotiv dans ce cours.

Lemme 1.1. (Une matrice symétrique réelle est diagonalisable sur $\mathbb{R}$). *Soit V un espace vectoriel sur $\mathbb{R}$ de dimension finie: $\dim(V) = n, n \in \mathbb{N}^*$, muni d'un produit scalaire $\langle\ ,\ \rangle; V \times V \to \mathbb{R}$, avec norme associée $||\boldsymbol{x}|| = \langle \boldsymbol{x}, \boldsymbol{x}\rangle^{\frac{1}{2}}, \forall \boldsymbol{x} \in V$. Soit T une application linéaire $T : V \to V$, symétrique*

$$\langle T(\boldsymbol{x}), \boldsymbol{y}\rangle = \langle T(\boldsymbol{y}), \boldsymbol{x}\rangle, \tag{1.3}$$

Alors, il existe une base orthonormée $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ de V (c.à.d $\langle \boldsymbol{e}_i, \boldsymbol{e}_j\rangle = \delta_{ij}$) et $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$ tels que $T\boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i, \forall i \in \{1, \ldots, n\}$.

Preuve : Par récurrence sur $n = \dim(V)$.

i) Nous supposons d'abord $\dim(V) = 1$.

Soit $\boldsymbol{v} \in V$, $\boldsymbol{v} \neq 0$, alors $V = \mathrm{Lin}_{\mathbb{R}}(\boldsymbol{v}) = \mathrm{Lin}_{\mathbb{R}}(\boldsymbol{e}_1)$ avec $e_1 = \frac{\boldsymbol{v}}{||\boldsymbol{v}||}$. En particulier $\langle \boldsymbol{e}_1, \boldsymbol{e}_1 \rangle = 1$. Soit $T : V \to V$ linéaire (symétrique), on a $T\boldsymbol{e}_1 \in V = \mathrm{Lin}_{\mathbb{R}}(\boldsymbol{e}_1) \Longrightarrow \exists \lambda_1 \in \mathbb{R}$ tel que $T\boldsymbol{e}_1 = \lambda_1 \boldsymbol{e}_1$.

Notez que le raisonnement peut être étendu au cas complexe[1].

ii) On suppose que le résultat est vrai pour $\dim(V) < n$ (hypothèse de récurrence). On le montre pour $\dim(V) = n$.

Soit V un espace linéaire avec produit scalaire, avec $\dim(V) = n < \infty$ et $T : V \to V$ linéaire symétrique. On définit :

$$\begin{aligned} \varphi : V &\to \mathbb{R} \\ x &\mapsto \langle T\boldsymbol{x}, \boldsymbol{x} \rangle \ . \end{aligned}$$

Cette application est continue. En particulier, sa restriction à la sphère unité $S_1 = \{\boldsymbol{x} \in V \ ; \ ||\boldsymbol{x}|| = 1\}$, qui est compacte (fermée et bornée, parce $\dim(V) = n < \infty$), est aussi continue. L'image de φ est alors compacte dans $\mathbb{R}$ (intervalle fermé) alors φ atteint son maximum. C'est-à-dire $\exists \boldsymbol{v} \in S_1$ tel que

$$\varphi(\boldsymbol{x}) \leq \varphi(\boldsymbol{v}) = \langle T\boldsymbol{v}, \boldsymbol{v} \rangle := \lambda, \forall \boldsymbol{x} \in S_1. \tag{1.4}$$

On va montrer que $T\boldsymbol{v} = \lambda \boldsymbol{v}$.

Soit $\boldsymbol{y} \in V \setminus \{0\}$ et soit $t \in]0, \frac{1}{||\boldsymbol{y}||}[$, alors $\boldsymbol{v} + t\boldsymbol{y} \neq 0$[2] On en déduit :

$$\frac{\boldsymbol{v} + t\boldsymbol{y}}{||\boldsymbol{v} + t\boldsymbol{y}||} \in S_1 \tag{1.5}$$

et donc :

$$\underbrace{\varphi(\boldsymbol{v})}_{\lambda} \geq \langle T \frac{\boldsymbol{v} + t\boldsymbol{y}}{||\boldsymbol{v} + t\boldsymbol{y}||}, \frac{\boldsymbol{v} + t\boldsymbol{y}}{||\boldsymbol{v} + t\boldsymbol{y}||} \rangle \ \Leftrightarrow \ \lambda ||\boldsymbol{v} + t\boldsymbol{y}||^2 \geq \langle T(\boldsymbol{v} + t\boldsymbol{y}), \boldsymbol{v} + t\boldsymbol{y} \rangle \ . \tag{1.6}$$

En développant à gauche :

$$\lambda ||\boldsymbol{v} + t\boldsymbol{y}||^2 = \lambda \langle (\boldsymbol{v} + t\boldsymbol{y}), \boldsymbol{v} + t\boldsymbol{y} \rangle = \lambda \left(\underbrace{||v||^2}_{=1} + 2t\langle \boldsymbol{v}, \boldsymbol{y} \rangle) + t^2 ||\boldsymbol{u}||^2 \right) \ , \tag{1.7}$$

et à droite :

$$\langle T(\boldsymbol{v} + t\boldsymbol{y}), \boldsymbol{v} + t\boldsymbol{y} \rangle = \underbrace{\langle T\boldsymbol{v}, \boldsymbol{v} \rangle}_{\lambda} + t\langle T\boldsymbol{v}, \boldsymbol{y} \rangle + t \underbrace{\langle T\boldsymbol{y}, \boldsymbol{v} \rangle}_{\langle \boldsymbol{y}, T\boldsymbol{v} \rangle = \langle T\boldsymbol{v}, \boldsymbol{y} \rangle} + t^2 \langle T\boldsymbol{y}, \boldsymbol{y} \rangle \ , \tag{1.8}$$

et on conclut (en utilisant $t > 0$ pour pouvoir simplifier par t) :

$$\lambda \left(2\langle \boldsymbol{v}, \boldsymbol{y} \rangle + t||\boldsymbol{y}||^2 \right) \geq 2\langle T\boldsymbol{v}, \boldsymbol{y} \rangle + t\langle T\boldsymbol{y}, \boldsymbol{y} \rangle \ . \tag{1.9}$$

[1] Il faut utiliser une application linéaire T symétrique (hermitien) par rapport au produit scalaire hermitien (voir Appendix A.3).

[2] Avec $|||\boldsymbol{x}|| - ||\boldsymbol{y}||| \leq ||x - y||$, on a $||\boldsymbol{v} + t\boldsymbol{y}|| \geq |||\boldsymbol{v}|| - ||t\boldsymbol{y}||| = |1 - \underbrace{t||y||}_{\neq 0}| \neq 0$.

Si on prend la limite lorsque $t \to 0^+$, on a :

$$2\lambda\langle \boldsymbol{v}, \boldsymbol{y}\rangle \geq 2\langle T\boldsymbol{v}, \boldsymbol{y}\rangle \;\Rightarrow\; 0 \geq \langle T\boldsymbol{v} - \lambda\boldsymbol{v}, \boldsymbol{y}\rangle \;,\; \forall \boldsymbol{y} \in V \setminus \{0\} \;. \tag{1.10}$$

Si on prend maintenant $\boldsymbol{z} = -\boldsymbol{y}$, on dérive aussi :

$$\begin{aligned} 0 \geq \langle T\boldsymbol{v} - \lambda\boldsymbol{v}, \boldsymbol{z}\rangle = \langle T\boldsymbol{v} - \lambda\boldsymbol{v}, -\boldsymbol{y}\rangle = -\langle T\boldsymbol{v} - \lambda\boldsymbol{v}, \boldsymbol{y}\rangle \\ \Rightarrow \langle T\boldsymbol{v} - \lambda\boldsymbol{v}, \boldsymbol{y}\rangle \geq 0 \;,\; \forall \boldsymbol{y} \in V \setminus \{0\} \end{aligned} \tag{1.11}$$

On a alors :

$$\langle T\boldsymbol{v} - \lambda\boldsymbol{v}, \boldsymbol{y}\rangle = 0 \;,\; \forall \boldsymbol{y} \in V \setminus \{0\} \Rightarrow T\boldsymbol{v} - \lambda\boldsymbol{v} = 0 \Leftrightarrow T\boldsymbol{v} = \lambda\boldsymbol{v} \tag{1.12}$$

On pose $\boldsymbol{e}_n = \boldsymbol{v}$ et $\lambda_n = \lambda$.

Maintenant, on utilise l'hypothèse de récurrence. Soit $W = \{\boldsymbol{x} \in V; \langle \boldsymbol{x}, \boldsymbol{e}_n\rangle = 0\}$. Alors $V \neq W$ et $V = W \oplus \mathrm{Lin}_{\mathbb{R}}(\boldsymbol{e}_n)$. On peut décomposer $\boldsymbol{x} \in V$

$$\boldsymbol{x} = \underbrace{\boldsymbol{x} - \langle \boldsymbol{x}, \boldsymbol{e}_n\rangle \boldsymbol{e}_n}_{\in W} + \langle \boldsymbol{x}, \boldsymbol{e}_n\rangle \boldsymbol{e}_n \;. \tag{1.13}$$

Si on note $T|_W$ la restriction de T à W, l'application $T|_W : W \to W$ est linéaire et symétrique (exercice). On peut alors faire usage de l'hypothèse de récurrence sur $T|_W$:

$$\exists \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{n-1}\} \text{ et } (\lambda_1, \ldots, \lambda_{n-1}) \in \mathbb{R}^{n-1} \;, \tag{1.14}$$

tels que :

$$T|_W \boldsymbol{e}_i = T|_W \boldsymbol{e}_i = \lambda_i \boldsymbol{e}_i \;,\; \forall i \in \{1, \ldots, n-1\} \;,\; \langle \boldsymbol{e}_i, \boldsymbol{e}_j\rangle = 0 \;. \tag{1.15}$$

On choisit finalement :

$$\{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{n-1}, \boldsymbol{e}_n\} \text{ et } (\lambda_1, \ldots, \lambda_{n-1}, \lambda_n) \;, \tag{1.16}$$

qui montrent le théorème.

□

Corollaire 1.1. *Une matrice $A \in M_n(\mathbb{R})$ symétrique est diagonalisable sur $\mathbb{R}$.*

Preuve :
On considère $V = \mathbb{R}^n$ avec le produit scalaire canonique

$$\langle \boldsymbol{x}, \boldsymbol{y}\rangle = \boldsymbol{x}^t \cdot \boldsymbol{y} = \sum_{i=1}^{n} x^i y^i \;,$$

où $\boldsymbol{x} = x^i \boldsymbol{e}_i$ avec $B = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ la base canonique. Si $A \in M_n(\mathbb{R})$ est une matrice symétrique, l'application

$$\begin{array}{rcl} T : \mathbb{R}^n & \to & \mathbb{R}^n \\ \boldsymbol{x} & \mapsto & A \cdot \boldsymbol{x} \quad , \quad \text{où } \boldsymbol{x} = x^i \boldsymbol{e}_i, B = \{\boldsymbol{e}_{i \in \mathbb{N}}\} \text{ base canonique} \end{array} \tag{1.17}$$

est linéaire et symétrique. Notez que A est la matrice de T dans la base canonique

$$A = M(T, B, B) \tag{1.18}$$

Pour la symétrie,

$$\begin{aligned} \langle \boldsymbol{x}, T\boldsymbol{y} \rangle &= \boldsymbol{x}^t \cdot A \cdot \boldsymbol{y} = \left(\boldsymbol{x}^t \cdot A \cdot \boldsymbol{y}\right)^t = (A \cdot \boldsymbol{y})^t \cdot \left(\boldsymbol{x}^t\right)^t = \boldsymbol{y}^t \cdot A \cdot \boldsymbol{x} = \left(\boldsymbol{y}^t \cdot A \cdot \boldsymbol{x}\right)^t \\ &= (A \cdot \boldsymbol{x})^t \cdot \boldsymbol{y} = \langle T\boldsymbol{x}, \boldsymbol{y} \rangle \ . \end{aligned} \tag{1.19}$$

Par le lemme précédent, $\exists B' = \{\boldsymbol{e}'_1, \ldots, \boldsymbol{e}'_n\}$ et $\{\lambda_1, \ldots, \lambda_n\}$ tels que

$$A\boldsymbol{e}'_i = T\boldsymbol{e}'_i = \lambda_i \boldsymbol{e}'_i \ , \text{ et } \langle \boldsymbol{e}'_i, \boldsymbol{e}'_j \rangle = \delta_{ij} \tag{1.20}$$

Par la définition D.2 on conclut que A est diagonalisable

□

Remarque 1.1. La matrice A est semblable à une matrice diagonale. En effet, on a

$$A = M(T, B) = M(T, B, B) \text{ et } D = M(T, B') = M(T, B', B') \ , \tag{1.21}$$

et on peut écrire

$$\begin{aligned} A &= M(T, B, B) = M(\mathrm{I}, B', B) M(T, B', B') M(\mathrm{I}, B, B') \\ &= M(B' \to B) M(T, B', B') M(B \to B') = PDP^{-1} \end{aligned} \tag{1.22}$$

où P est la matrice de changement de base $P = M(B' \to B)$. On a

$$A = PDP^{-1} \quad , \quad D = P^{-1}AP \tag{1.23}$$

Étant donné que P est la matrice de changement de base entre des bases orthonormées, la matrice P est orthogonale, c.à.d $P^{-1} = P^t$.

1.3 Exemples de matrices symétriques

Exemple 1.1 (Opérateur Laplacien: discrétisation)**.** Le premier exemple de matrice symétrique est placé dans le contexte des équations aux dérivées partielles (EDP), en particulier dans une classe de problèmes qui amènent à une équation

$$A\boldsymbol{x} = \boldsymbol{b} \ , \tag{1.24}$$

avec A symétrique et le vecteur $\boldsymbol{b}$ donnés.

a) **EDP et Laplacien**:

L'équation de Poisson est

$$-\Delta u = f \ , \text{ où } \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2} \ . \tag{1.25}$$

L'opérateur Δ est le Laplacien et cette équation est un exemple d'EDP de type dit elliptique. Elle joue un rôle fondamental dans plusieurs contextes mathématiques et physiques. Voici quelques exemples :

– Détermination du potentiel gravitationnel de Newton

$$-\Delta\phi_{\text{grav}} = -4\pi G\rho_{\text{masse}} \ . \tag{1.26}$$

– Potentiel électrostatique

$$-\Delta\phi_{\text{lec}} = \frac{\rho_{\text{charge}}}{\epsilon} \ . \tag{1.27}$$

– états stationnaires de l'équation de la chaleur, avec une densité de flux de chaleur q

$$\rho c_p \frac{\partial T}{\partial t} = k\Delta T + q \tag{1.28}$$

Celle-ci est une EDP de type *parabolique*.

– états stationnaires de l'équation d'ondes avec forçage g

$$-\frac{\partial^2 u}{\partial t^2} + c^2\Delta u = g \ . \tag{1.29}$$

L'équation d'ondes est le prototype d'EDP de type hyperbolique.

b) Le Laplacien: un opérateur symétrique, (en fait, auto-adjoint [3]). Nous considérons l'espace linéaire $L^2(D \subset \mathbb{R}^n)$ de fonctions de carré intégrable, c.à.d $f \in L^2(D)$ ssi $\int_D f^2 d^n x < \infty$. $L^2(D \subset \mathbb{R}^n)$ est un espace de Hilbert avec produit scalaire

$$\langle f, g\rangle = \int_D f g d^n x \ . \tag{1.30}$$

Remarque 1.2. espace de Hilbert complexe. Nous pouvons aussi considérer des fonctions $\varphi : D \subset \mathbb{R}^n \to \mathbb{C}$, avec $\int_D |\varphi|^2 d^n x < \infty$, qui est aussi un espace de Hilbert avec produit hermitien (sesquilinéaire)

$$\langle \varphi, \phi\rangle = \int_D \bar{\varphi}\phi d^n x \tag{1.31}$$

On va étudier une discrétisation de cet opérateur, notamment ce qu'on appelle des différences finies. Pour simplifier, nous allons étudier le cas en dimension 1, avec des condiitions aux bords dites "Dirichlet honogèenes". L'équation (1.25) devient

$$\begin{cases} -\dfrac{d^2u}{dx^2} = f(x) \\ u(0) = u(1) = 0 \end{cases} \tag{1.32}$$

avec $f \in C([0,1], \mathbb{R})$. On peut montrer que $u \in C^2([0, \mathbb{R}])$.

Étant donné $N \in \mathbb{N}$, on définit le pas de la discrétisation $h = \frac{1}{N+1}$, et on introduit les points :

$$x_i = ih = \frac{i}{N+1} \quad , \quad \forall i \in \{0, 1, \ldots, N+1\} \tag{1.33}$$

[3] La notion d'opérateur auto-adjoint dans un espace linéaire de dimension infinie implique des aspects subtils sur le domaine de l'opérateur. Dans ce cours, nous ne discutons pas de ces questions.

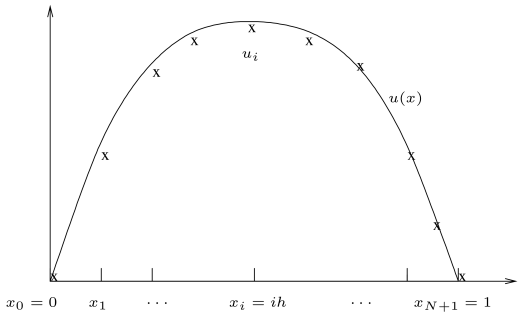

Figure 1.1: Schéma de la discrétrisation par différences finies.

On évalue l'équation (1.32) dans les points $x_i \in \{1, \ldots, N\}$. Notons que $u(x_0) = u(x_{N+1}) = 0$. Ainsi :

$$-u''(x_i) = f(x_i) \quad , \quad \forall i \in \{0, 1, \ldots, N+1\} \tag{1.34}$$

où $f(x_i)$ sont les *valeurs exactes* de la solution $u(x)$ evalué sur x_i. On va suposser que $u \in C^4([0,1], \mathbb{R})$. On peut alors écrire le développement de Taylor-Lagrange :

$$u(x_{i+1}) = u(x_i + h) = u(x_i) + hu'(x_i) + \frac{h^2}{2}u''(x_i) + \frac{h^3}{6}u^{(3)} + \frac{h^4}{24}u^{(4)}(\xi_i) \ , \ \xi_i \in]x_i, x_{i+1}[\ . \tag{1.35}$$

De la même manière :

$$u(x_{i-1}) = u(x_i - h) = u(x_i) - hu'(x_i) + \frac{h^2}{2}u''(x_i) - \frac{h^3}{6}u^{(3)} + \frac{h^4}{24}u^{(4)}(\eta_i) \ , \ \eta_i \in]x_{i-1}, x_i[\ . \tag{1.36}$$

Si on ajoute les deux expressions, on obtient :

$$u(x_{i+1}) + u(x_{i-1}) = 2u(x_i) + h^2u''(x_i) + \frac{h^4}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)) \ . \tag{1.37}$$

Alors, on peut écrire :

$$-u''(x_i) - \frac{-u(x_{i+1} + 2u(x_i) - u(x_{i-1})}{h^2} = \frac{h^2}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)) \tag{1.38}$$

Si on définit maintenant *l'erreur de consistance* R_i par :

$$R_i = -u''(x_i) - \frac{-u(x_{i+1} + 2u(x_i) - u(x_{i-1})}{h^2} \ , \tag{1.39}$$

on a

$$R_i = \frac{h^2}{24}(u^{(4)}(\xi_i) + u^{(4)}(\eta_i)) \ , \tag{1.40}$$

de telle manière que :

$$\begin{aligned} |R_i| &= \frac{h^2}{24}|u^{(4)}(\xi_i) + u^{(4)}(\eta_i)| \leq \frac{h^2}{24}\left(\underbrace{|u^{(4)}(\xi_i)|}_{\leq \sup_{x\in]0,1[}|u^{(4)}(x)|} + \underbrace{|u^{(4)}(\eta_i)|}_{\leq \sup_{x\in]0,1[}|u^{(4)}(x)|} \right) \\ &= \frac{h^2}{12}\sup_{x\in]0,1[}|u^{(4)}(x)| = \frac{h^2}{12}\underbrace{||u^{(4)}||_\infty}_{<\infty} \ . \end{aligned} \tag{1.41}$$

Cela nous dit que l'erreur de consistance tend vers 0 comme h^2 (on dit que le schéma est consistant à l'ordre 2).

L'idée maintenant est de se rapprocher du problème originel par un problème discret, en faisant

$$\frac{d^2u}{dx^2} \sim \frac{-u(x_{i+1}) + 2u(x_i) - u(x_{i-1})}{h^2} \ . \tag{1.42}$$

Néanmoins, cette expression utilise les valeurs exactes $u(x_i)$, mais ces valeurs sont précisément les inconnues qu'on cherche à trouver. À ce point, nous introduisons des valeurs u_i qui vont représenter des *approximations* des vraies valeurs de $u(x_i)$. Si on a aussi $b_i = f(x_i)$ (valeurs exactes du membre à droite!), nous pouvons définir le problème discret

$$\begin{cases} -\dfrac{-u_{i+1} + 2u_i - u_{i-1}}{h^2} = b_i \quad , \quad i \in \{1, \dots, N\} \\ u_0 = u_{N+1} = 0 \end{cases} \tag{1.43}$$

Si on définit la matrice Δ_N, par ses éléments matriciels :

$$(\Delta_N)_{ij} = \begin{cases} \frac{2}{h^2} & , \ i = j \ , i, j \in \{1, \dots, N\} \\ -\frac{1}{h^2} & , \ |i-j| = 1 \ , i, j \in \{1, \dots, N\} \\ 0 & , \ |i-j| > 1 \ , i, j \in \{1, \dots, N\} \end{cases} \tag{1.44}$$

on peut écrire le problème (1.43) comme

$$\Delta_N \boldsymbol{u} = \boldsymbol{b} \tag{1.45}$$

Pour avoir une idée, les premières matrices Δ_N sont de la forme :

$$\Delta_2 = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \quad , \quad \Delta_3 = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix} \quad , \quad \Delta_4 = \begin{pmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{pmatrix},$$

On verra que cette matrice est inversible (en fait elle symétrique définie positive), ce qui nous permet de trouver la solution approchée comme :

$$\boldsymbol{u} = (\Delta_N)^{-1} \boldsymbol{b} \ . \tag{1.46}$$

Cette équation différentielle nous amène à notre premier problème matriciel de type $A\boldsymbol{x} = \boldsymbol{b}$, et on peut déjà voir que si N devient très grand, le problème d'inversion de Δ_N devient non-trivial. Ce problème d'inversion des matrices de grande taille, et en particulier des matrices symétriques, sera un des sujets fondamentaux de ce cours.

Un autre aspect important, déjà présent dans ce problème? concerne la convergence de la méthode. Notamment, l'attente est que, lorsque $h \to 0$, la solution approchée u_i tend vers $u(x_i)$. C'est-à-dire que, si on définit l'erreur comme :

$$e_i = u(x_i) - u_i \ , \ i \in \{1, \dots, N\} \ , \tag{1.47}$$

on peut voir en appliquant les résultats précédents que :

$$\Delta_n \boldsymbol{e} = \boldsymbol{R} \ , \tag{1.48}$$

où R_i est donné par (1.40). Alors,

$$\boldsymbol{e} = (\Delta_N)^{-1}\boldsymbol{R}\ . \tag{1.49}$$

Le fait que $|R_i|$ tende vers 0 quand $h \to 0$ est bon, mais ce n'est pas suffisant pour garantir que $\boldsymbol{e} \to 0$. Pour ça, il faut montrer que Δ_N, qui dépend de h, est telle que l'on puisse donner un sens à la *norme* $||(\Delta_N)^{-1}||$ et elle qu'elle soir, notamment bornée indépendamment de h. Ce type de question liant la convergence d'un schéma avec la norme d'une matrice approprié sera aussi l'objet de notre cours.

Chapitre 2

Méthodes directes

Définition 2.1 (méthode directe)**.** On appelle méthode directe de résolution de

$$A\boldsymbol{x} = \boldsymbol{b}\ ,\ A \in M_n(\mathbb{R})\ , \boldsymbol{x}, \boldsymbol{b} \in \mathbb{R}^n \tag{2.1}$$

une méthode qui donne exactement $\boldsymbol{x}$ (A et $\boldsymbol{b}$ étant connus), solution de (3.34) après un nombre fini d'opérations élémentaires.

2.1 Décomposition LU (méthode de Gauss/ d'échelonnement)

Le principe de la méthode est de transformer un système linéaire, par des combinaisons linéaires et des permutations de lignes, dans un système triangulaire équivalent. On va rappeler la procédure de cette méthode, étudiée dans les cours précédents avec un exemple.

Exemple 2.1. On considère le problème (3.34) avec :

$$A = \begin{pmatrix} 4 & -2 & 6 \\ 2 & -4 & 9 \\ -4 & 5 & -9 \end{pmatrix} \quad , \quad \boldsymbol{b} = \begin{pmatrix} 12 \\ -24 \\ 12 \end{pmatrix} \ . \tag{2.2}$$

On construit la matrice augmentée:

$$\tilde{A} = \begin{pmatrix} A & b \end{pmatrix} = \left(\begin{array}{ccc|c} 4 & -2 & 6 & 12 \\ 2 & -4 & 9 & -24 \\ -4 & 5 & -9 & 12 \end{array}\right) \ , \tag{2.3}$$

Et on procède avec la méthode de pivot déjà étudiée pour transformer A en une matrice triangulaire supérieure. Simplement, dans ce process, on va garder systématiquement l'information des lignes qu'on retranche à chaque pas de l'itération. Concrètement, vu que $\ell_1(1) = 4 \neq 0$, nous pouvons l'utiliser comme *pivot*. Si on note $\tilde{A}_0{=}\tilde{A}$

$$\tilde{A}_0 \overset{\ell_2 \to \ell_2 - \frac{2}{4}\ell_1}{\longrightarrow} \tilde{A}_1 = \left(\begin{array}{ccc|c} 4 & -2 & 6 & 12 \\ 0 & -3 & 6 & -30 \\ -4 & 5 & -9 & 12 \end{array}\right) \ . \tag{2.4}$$

Notons que cette transformation de $\tilde{A}$ et $\boldsymbol{b}$, qui change la deuxième ligne en lui retranchant la première ligne multipliée par un facteur $\frac{1}{2}$ est le résultat de multiplier $\tilde{A}_0$ par la gauche par la

matrice :

$$E_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} - \frac{1}{2} \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{2.5}$$

On procède de la même manière pour échelonner la matrice :

$$\tilde{A}_1 \underset{E_2}{\overset{\ell_3 \to \ell_3 - \frac{-4}{4}\ell_1}{\longrightarrow}} \tilde{A}_2 = \left(\begin{array}{ccc|c} 4 & -2 & 6 & 12 \\ 0 & -3 & 6 & -30 \\ 0 & 3 & -3 & 24 \end{array}\right) \underset{E_3}{\overset{\ell_3 \to \ell_3 - \frac{-3}{-3}\ell_1}{\longrightarrow}} \tilde{A}_3 = \left(\begin{array}{ccc|c} 4 & -2 & 6 & 12 \\ 0 & -3 & 6 & -30 \\ 0 & 0 & 3 & -6 \end{array}\right) , \tag{2.6}$$

avec

$$E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} , \; E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix} \tag{2.7}$$

Comme résultat de la procédure, autant que tous les pivots qu'on trouve soient différents de zéro, on construit un système linéaire équivalent $\tilde{A}_3 \boldsymbol{x} = \tilde{\boldsymbol{b}}$ échelonné qu'on peut résoudre facilement par remontée. Le résultat est

$$\boldsymbol{x} = \begin{pmatrix} 9 \\ 6 \\ -2 \end{pmatrix} \tag{2.8}$$

Si on a besoin de résoudre ce système pour différents vecteurs $\boldsymbol{b}$, dans chaque cas on doit répéter le calcul du vecteur $\tilde{\boldsymbol{b}}$ (dans le process d'échelonnement de A). Si le nombre d'équations à résoudre est élevé, cette répétition est clairement à éviter en cherchant une méthode plus efficace. Si on regarde notre exemple, on note que c'est ce qu'on a fait, on peut l'écrire :

$$A\boldsymbol{x} = \boldsymbol{b} \; \longrightarrow \underbrace{E_3 E_2 E_1 A}_{U} \boldsymbol{x} = E_3 E_2 E_1 \boldsymbol{b} \; , \tag{2.9}$$

où U est une matrice triangulaire supérieure (c.à.d. $u_{ij} = 0$, si $i > j$). Le matrices E_i sont inversibles et on peut alors multiplier l'expression à gauche

$$A\boldsymbol{x} = (E_3 E_2 E_1)^{-1} U\boldsymbol{x} = \boldsymbol{b} = \underbrace{E_1^{-1} E_2^{-1} E_3^{-1}}_{L} U\boldsymbol{x} = LU\boldsymbol{x} = \boldsymbol{b} \tag{2.10}$$

avec

$$(E_1)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} , \; E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix} , \; E_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \tag{2.11}$$

Ce qui amène à :

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \tag{2.12}$$

L est une matrice triangulaire inférieure ($l_{ij} = 0, i < j$) avec des 1 dans la diagonale principale et les éléments non nuls sont les coefficients qu'on a utilisé pour retrancher chaque ligne dont

le pivot est non nul dans le process d'échelonnement [1]. En particulier nous avons décomposé la matrice A comme le produit d'une matrice triangulaire inférieure L et d'une matrice triangulaire supérieure U

$$A = LU \ . \tag{2.15}$$

En particulier, dans notre exemple

$$A = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ -1 & -1 & 1 \end{pmatrix} \begin{pmatrix} 4 & -2 & 6 \\ 0 & -3 & 6 \\ 0 & 0 & 3 \end{pmatrix} \tag{2.16}$$

Notez que la décomposition de A comme produit d'une matrice triangulaire inférieure et d'une matrice triangulaire supérieure n'est pas unique. Par exemple, on peut multiplier la première par 2 et la deuxième par $\frac{1}{2}$

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & 2 & 0 \\ -2 & -2 & 2 \end{pmatrix} \begin{pmatrix} 2 & -1 & 3 \\ 0 & -\frac{3}{2} & 3 \\ 0 & 0 & \frac{3}{2} \end{pmatrix} \ , \tag{2.17}$$

et on peut aussi trouver d'autres paires de matrices

$$A = \begin{pmatrix} 2 & 0 & 0 \\ 1 & -3 & 0 \\ -2 & 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 & 3 \\ 0 & 1 & -2 \\ 0 & 0 & 3 \end{pmatrix} \ . \tag{2.18}$$

Ce qui est particulier de la décomposition (2.16) c'est que les éléments de la diagonale principale de la matrice L sont seulement des 1. Comme va le voir plus tard, une telle décomposition est unique et on l'appelle décomposition LU de la matrice A.

Si on utilise cette décomposition, on peut résoudre le système original

$$\begin{aligned} A\boldsymbol{x} &= \boldsymbol{b} \\ LU\boldsymbol{x} &= \boldsymbol{b} \end{aligned} \tag{2.19}$$

en deux pas, en introduisant un vecteur $\boldsymbol{y}$ intermédiaire

$$\begin{aligned} L\boldsymbol{y} &= \boldsymbol{b} \\ U\boldsymbol{x} &= \boldsymbol{y} \end{aligned} \tag{2.20}$$

[1] En général, la structure de matrice sera toujours

$$\begin{aligned} (E_1) &= \begin{pmatrix} 1 & 0 & 0 \\ -a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ , \quad E_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -b & 0 & 1 \end{pmatrix} \ , \quad E_3 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -c & 1 \end{pmatrix} \\ (E_1)^{-1} &= \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \ , \quad E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ b & 0 & 1 \end{pmatrix} \ , \quad E_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c & 1 \end{pmatrix} \end{aligned} \tag{2.13}$$

Ce qui amène à :

$$L = E_1^{-1} E_2^{-1} E_3^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ b & c & 1 \end{pmatrix} \tag{2.14}$$

Étant donné $\boldsymbol{b}$, la résolution pour $\boldsymbol{y}$ est directe car L est échelonnée. En particulier, vu que L est triangulaire inférieure, on va résoudre par descente. Une fois $\boldsymbol{y}$ calculé, on résout pour $\boldsymbol{x}$ par remontée. La procédure peut être répétée en commençant toujours avec $\boldsymbol{b}$.

2.1.1 Algorithme LU, sans permutation

Nous présentons maintenant les pas pour résoudre le système par factorisation, dans le cas que tous les pivots qu'on trouve soient différents de zéro.

1. *Factorisation.*

   On passe en $n-1$ pas de la matrice $A^{(1)}$ à une matrice $A^{(n)} = U$ triangulaire supérieure. Dans le process, on construit la matrice L.

   i) Pas $k = 1$: On initialise la matrice $A^{(1)} = A$ et on définit la première ligne de U. Si $a_{11} \neq 0$ (pivot non-nul), on fait

   $$\begin{aligned} a_{ij}^{(1)} &= a_{ij} \ , \ i,j \in \{1,\ldots,n\} \\ u_{1j} &= a_{1j}^{(1)} \ , \ j \in \{1,\ldots,n\} \end{aligned} \tag{2.21}$$

   ii) Pas $k+1 < n$: On construit la colonne k de L, la matrice $A^{(k+1)}$ et la ligne $k+1$ de U. Si $a_{k,k} \neq 0$ (pivot non-nul)

   a) Pour L:

   $$l_{i,k} = 0 \ , \ i < k \quad , \quad l_{kk} = 1 \quad , \quad l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \ , \ i > k \ . \tag{2.22}$$

   b) Pour $A^{(k+1)}$:

   $$\begin{aligned} a_{ij}^{(k+1)} &= a_{ij}^{(k)} \ , \ i \in \{1,\ldots,k\}, j \in \{1,\ldots,n\} \\ a_{ij}^{(k+1)} &= a_{ij}^{(k)} - l_{ik}\, a_{kj}^{(k)} \ , \ i \in \{k+1,\ldots,n\}, j \in \{1,\ldots,n\} \ . \end{aligned} \tag{2.23}$$

   c) Pour U:

   $$u_{k+1,j} = a_{k+1,j}^{(k+1)} \ , \ j \in \{1,\ldots,n\} \tag{2.24}$$

   iii) Pas $k = n$: On agit comme dans le cas $k+1 < n$ et on finit la dernière colonne de L

   $$l_{i,n} = 0 \ , \ i < n \quad , \quad l_{nn} = 1 \ . \tag{2.25}$$

   On constate que $U = A^{(n)}$.

2. *Descente*: On calcule $\boldsymbol{y}$ (où $L\boldsymbol{y} = \boldsymbol{b}$)

   $$y_i = b_i - \sum_{k=1}^{i-1} l_{ik}\, y_k \ , \ i \in \{1,\ldots,n\} \tag{2.26}$$

3. *Remontée*: On calcule $\boldsymbol{x}$ (où $U\boldsymbol{x} = \boldsymbol{y}$)

   $$x_i = \frac{1}{u_{ii}} \left(y_i - \sum_{k=i+1}^{n} u_{ik}\, x_k \right) \ , \ i \in \{n,\ldots,1\} \tag{2.27}$$

Remarque 2.1 (Factorisation, version matricielle). Le pas 1 de la procédure précédente peut s'écrire d'une façon plus compacte :

i) Pas $k = 1$:

$$A^{(1)} = A \; . \tag{2.28}$$

ii) Pas $k+1 \leq n$: On introduit la matrice E_k avec des zéros partout sauf dans la diagonale principale (diag$[1 \ldots 1]$) et en dessous de la diagonale principale dans la colonne k:

$$E_k = \begin{pmatrix} 1 & \ldots & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & & 1 & 0 & & \vdots \\ \vdots & & \frac{-a^{(k)}_{k+1,k}}{a^{(k)}_{kk}} & 1 & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & \ldots & \frac{-a^{(k)}_{n,k}}{a^{(k)}_{kk}} & \ldots & 0 & 1 \end{pmatrix} \quad , \quad E_k^{-1} = \begin{pmatrix} 1 & \ldots & 0 & \ldots & \ldots & 0 \\ 0 & \ddots & 0 & & & \vdots \\ \vdots & & 1 & 0 & & \vdots \\ \vdots & & \frac{a^{(k)}_{k+1,k}}{a^{(k)}_{kk}} & 1 & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ 0 & \ldots & \frac{a^{(k)}_{n,k}}{a^{(k)}_{kk}} & \ldots & 0 & 1 \end{pmatrix} \tag{2.29}$$

On a alors

$$A^{(k+1)} = E_k A^{(k)} \tag{2.30}$$

iii) Une fois le process fini, on construit :

$$\begin{aligned} U &= A^{(n)} \\ L &= E_1^{-1} \cdot \ldots \cdot E_n^{-1} = \begin{pmatrix} 1 & \ldots & 0 & \ldots & \ldots & 0 \\ \frac{a^{(1)}_{2,1}}{a^{(1)}_{11}} & \ddots & 0 & & & \vdots \\ \vdots & & 1 & 0 & & \vdots \\ \vdots & & \frac{a^{(k)}_{k+1,k}}{a^{(k)}_{kk}} & 1 & & \vdots \\ \vdots & & \vdots & & \ddots & 0 \\ \frac{a^{(1)}_{n,1}}{a^{(1)}_{11}} & \ldots & \frac{a^{(k)}_{n,k}}{a^{(k)}_{kk}} & \ldots & \frac{a^{(n-1)}_{n,n-1}}{a^{(n-1)}_{n-1,n-1}} & 1 \end{pmatrix} \end{aligned} \tag{2.31}$$

Notez:

a) Comme on peut le voir dans la notation adoptée dans le pas ii), on a regroupé dans la matrice E_k tous les coefficients utilisés pour obtenir des zéros sous la diagonale principale à la colonne k. En général, une matrice E_k est une matrice triangulaire inférieure, avec des 1 dans la diagonale principale et des éléments non nuls seulement dans la colonne k en dessous de la diagonale.

b) La matrice L a une forme très simple: elle est triangulaire inférieure, avec pour diagonale diag$(1, \ldots, 1)$ et l'élément l_{ij} est le coefficient par lequel on a multiplié la ligne de pivot $a^{(j)}_{jj}$ avant de la retrancher de la ligne i.

2.1.2 Algorithme LU, avec permutation

Dans la discussion précédente, on a supposé qu'à chaque pas de l'itération, on trouve un pivot non nul. Ceci n'est pas le cas en général. Pour illustrer la manière de procéder quand on trouve un pivot nul, on commence avec un exemple.

Exemple 2.2 (Une première décomposition LU avec permutation)**.** On considère l'exemple de la matrice :

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 2 & 1 & 1 \end{pmatrix} . \tag{2.32}$$

Le premier pas de la méthode LU :

$$A^{(1)} = A \xrightarrow[E_1=\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}]{\substack{\ell_2 \to \ell_2 - \ell_1 \\ \ell_3 \to \ell_3 - 2\ell_1}} A^{(2)} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & -2 \\ 0 & -1 & -1 \end{pmatrix} \tag{2.33}$$

produit un zéro pour le deuxième pivot. Pour pouvoir continuer, on interchange simplement la deuxième et la troisième ligne. En effet, cette opération ne change le système linéaire associé. Concrètement :

$$A^{(2)} \xrightarrow[P_2=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}]{\ell_2 \leftrightarrow \ell_3} \tilde{A}^{(3)} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -1 & -1 \\ 0 & 0 & -2 \end{pmatrix} \tag{2.34}$$

Dans ce cas, l'itération a déjà abouti à une matrice triangulaire supérieure. On peut écrire les pas suivants :

$$P_2 E_1 A = U \tag{2.35}$$

Notons que $P_2 E_1$ n'est pas triangulaire inférieure. En effet,

$$P_2 E_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix} \neq L \tag{2.36}$$

Notez, néanmoins, qu'on peut écrire :

$$P_2 E_1 = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 0 & 1 \\ -1 & 1 & 0 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}}_{E_1'} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} = E_1' P_2 \tag{2.37}$$

où E_1' est obtenue à partir de E_1 en interchangeant, dans la colonne 1, deux éléments (dans ce cas, les deuxième et troisième). Ainsi, on peut réécrire (2.35)

$$\begin{aligned} E_1' P_2 A &= U \\ P_2 A &= \underbrace{(E_1')^{-1}}_{L} U = LU . \end{aligned} \tag{2.38}$$

Autrement dit,

$$\underbrace{\begin{pmatrix}1&0&0\\0&0&1\\0&1&0\end{pmatrix}}_{P_2}\underbrace{\begin{pmatrix}1&1&1\\1&1&-1\\2&1&1\end{pmatrix}}_{A}=\underbrace{\begin{pmatrix}1&0&0\\2&1&0\\1&0&1\end{pmatrix}}_{L}\underbrace{\begin{pmatrix}1&1&1\\0&-1&-1\\0&0&-2\end{pmatrix}}_{U} \tag{2.39}$$

On a pas réussi[2] à écrire $A = LU$. Par contre, on a trouvé un réarrangement des lignes de A, donné par P_2A, qui admet en fait la décomposition $P_2A = LU$. Dans le théorème 2.1, on va montrer qu'on peut toujours trouver une telle décomposition.

La matrice P_2 dans l'exemple précédent est un cas de matrice de permutation P, qui nous a servi à réordonner les lignes dans la procédure de Gauss pour trouver un pivot non nul.

Erreur d'arrondi est choix de pivot.
Même si nous ne sommes pas obligés de réarranger les lignes pour obtenir un pivot non nul, nous pouvons nous intéresser à un tel changement d'ordre. En général, étant donnée une matrice A, il n'y a pas une seule matrice de permutation P telle qu'on puisse écrire $PA = LU$. Différents P sont possibles et chaque P amène génériquement a un couple (L, U) différent.

Un choix privilégié du point de vue numérique consiste à arranger les lignes de telle manière qu'on choisisse comme pivot, à chaque pas de l'itération de la méthode de Gauss, le coefficient avec la plus grande valeur absolue. La raison est la limitation des erreurs d'arrondi: si on peut garder seulement un nombre fini de chiffres significatifs, comme c'est le cas avec un ordinateur, c'est mieux de travailler avec des nombres avec des petites valeurs absolues plutôt qu'avec des nombres des grandes valeurs absolues. Vu que, dans la méthode de Gauss, il faut diviser à chaque pas le pivot, ça nous amène au choix signalé.

On reprend l'exemple 2.2 pour illustrer ce point.

Exemple 2.3 (Non-unicité de P dans la décomposition $PA = LU$)**.** Si on considère à nouveau la matrice (2.32), en regardant la colonne 1, on voit que, dans le premier pas de l'itération, on peux interchanger les lignes 1 et 3 de manière à ce que le pivot choisi soit celui avec la valeur absolue la plus grande. Explicitement, cela donne :

$$\begin{aligned}
A^{(1)} \;=\; A=\begin{pmatrix}1&1&1\\1&1&-1\\2&1&1\end{pmatrix} &\xrightarrow[P_1=\begin{pmatrix}0&0&1\\0&1&0\\1&0&0\end{pmatrix}]{\ell_1\leftrightarrow\ell_3} \tilde{A}^{(1)}=\begin{pmatrix}2&1&1\\1&1&-1\\1&1&1\end{pmatrix}\\
&\xrightarrow[E_1=\begin{pmatrix}1&0&0\\-\frac{1}{2}&1&0\\-\frac{1}{2}&0&1\end{pmatrix}]{\substack{\ell_2\to\ell_2-\frac{1}{2}\ell_1\\ \ell_3\to\ell_3-\frac{1}{2}\ell_1}} A^{(2)}=\begin{pmatrix}2&1&1\\0&\frac{1}{2}&-\frac{3}{2}\\0&\frac{1}{2}&\frac{1}{2}\end{pmatrix}\\
&\xrightarrow[E_2=\begin{pmatrix}1&0&0\\0&1&0\\0&-1&1\end{pmatrix}]{\ell_3\to\ell_3-\ell_2} A^{(3)}=\begin{pmatrix}2&1&1\\0&\frac{1}{2}&-\frac{3}{2}\\0&0&2\end{pmatrix}
\end{aligned} \tag{2.40}$$

[2] En fait, on peut montrer que pour cette matrice, ce n'est possible à faire (voir caractérisation dans le théorème 2.1).

Finalement, on a obtenu la matrice triangulaire supérieure $U = A^{(3)}$ en faisant

$$E_2 E_1 P_1 A \quad = \quad U \ , \tag{2.41}$$

qu'on peut réexprimer comme suit :

$$P_1 A = \underbrace{E_1^{-1} E_2^{-1}}_{L} U \tag{2.42}$$

$$\underbrace{\begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}}_{P_1} \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & -1 \\ 2 & 1 & 1 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 1 & 0 \\ \frac{1}{2} & 1 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 2 & 1 & 1 \\ 0 & \frac{1}{2} & -\frac{3}{2} \\ 0 & 0 & 2 \end{pmatrix}}_{U} \tag{2.43}$$

Ici, P_1 est à nouveau une matrice de permutations mais elle est différente de celle de l'exemple (2.2).

Définition 2.2 (*matrice de permutation).* Une matrice $P \in M_n(\mathbb{N})$ est une matrice dont les éléments sont des 0 ou des 1 et telle qu'il y a un seul 1 par ligne et par colonne. Autrement dit, étant donnée une permutation $\{\sigma(1), \ldots, \sigma(n)\}$ des éléments $\{1, \ldots, n\}$, les éléments de la matrice de permutation associée à σ P^σ s'écrivent :

$$P^\sigma_{ij} = \delta_{\sigma(i),j} = \begin{cases} 1, & \sigma(i) = j \\ 0, & \sigma(i) \neq j \end{cases} \tag{2.44}$$

Remarque 2.2. Dans notre contexte, nous commentons quelques points:

i) Si on voit P^σ comme la matrice d'une application linéaire dans une base $B = (e_1, ..., e_n)$, la matrice P^σ réordonne simplement les éléments de la base pour donner une nouvelle base ordonnée $B' = (e_{\sigma(1)}, \ldots, e_{\sigma(n)})$.

ii) étant donnée une matrice $A \in M_n(\mathbb{R})$,en multipliant par P à gauche $(A \to PA)$, on réordonne les lignes de A, tandis qu'en multipliant à droite $(A \to AP)$, on réordonne les colonnes.

iii) *Transpositions.* Pour notre problème, il suffit de considérer des transpositions, c'est-à-dire des permutations qui interchangent simplement deux éléments. Dans ce cas, l'expression matricielle est particulièrement simple :

$$P^{(i,j)} = \begin{pmatrix} 1 & & & & & & & \\ & \ddots & & & & & & \\ & & 0 & \cdots & \cdots & \cdots & 1 & \\ & & \vdots & 1 & & & \vdots & \\ & & \vdots & & \ddots & & \vdots & \\ & & \vdots & & & 1 & \vdots & \\ & & 1 & \cdots & \cdots & \cdots & 0 & \\ & & & & & & & \ddots \\ & & & & & & & & 1 \end{pmatrix} \tag{2.45}$$

Avec cette notation, la matrice P_2 de l'exemple 2.2 est écrite $P_2 = P^{(23)}$

Nous pouvons maintenant présenter le procédé de factorisation dans la décomposition LU dans le cas général d'un possible arrangement des lignes à chaque pas de l'itération.

Factorisation dans l'algorithme LU avec permutations (version matricielle).
Nous suivons les trois pas dans 2.1:

i) Pas $k = 1$:

$$A^{(1)} = A \ . \tag{2.46}$$

ii) Pas $k + 1 \leq n$: D'abord, on arrange les lignes de la matrice en multipliant à gauche par une permutation P_k (notamment une transposition) qui change seulement l'ordre des lignes $\ell(i)$ avec $i \geq k$. Après, on multiplie (à gauche) par la matrice E_k de (2.29). On a alors

$$A^{(k+1)} = E_k P_k A^{(k)} \tag{2.47}$$

iii) Une fois le procédé fini, on construit une matrice triangulaire supérieure

$$(E_{n-1}P_{n-1}) \dots (E_2P_2)(E_1P_1)A = A^{(n)} = U \ . \tag{2.48}$$

La différence fondamentale avec le cas sans permutation est que les pas précédents n'amènent pas directement, en général, à une matrice triangulaire inférieure comme $L = (E_1)^{-1} \dots (E_{n-1})^{-1}$. En effet, en général, les matrices E_k et P_{k+l} (avec $l \geq 1$) ne commutent pas, ce qui empêche de factoriser la matrice $E_{n-1} \dots E_1$ à gauche dans (2.48). On peut illustrer ceci avec un exemple. Si on considère le cas $n = 4$, on peut écrire (2.48) comme :

$$(E_3P_3)(E_2P_2)(E_1P_1)A = U \tag{2.49}$$

Comme on l'a dit, les matrices E_i et P_j ne commutent pas en général, alors on ne peut pas écrire $(E_3E_2E_1)(P_3P_2P_1)A = U$. En effet, si on considère par exemple le troisième pas de l'algorithme, on peut écrire :

$$E_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \end{pmatrix} \quad , \quad P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} \ , \tag{2.50}$$

pour certains a et b. Si on calcule P_3E_2 et E_2P_3, on obtient :

$$P_3E_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 0 & 1 \\ 0 & a & 1 & 0 \end{pmatrix} \quad , \quad E_2P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 0 & 1 \\ 0 & b & 1 & 0 \end{pmatrix} \tag{2.51}$$

et on constate que $P_3E_2 - E_2P_3 \neq 0$. Néanmoins on peut observer que si dans P_3E_2, on interchange la troisième et la quatrième colonne (ce qu'on fait en multipliant *à droite* par P_3), on récupère une matrice "type E_k". Plus spécifiquement, on a :

$$P_3E_2 = P_3E_2 \underbrace{(P_3{P_3}^{-1})}_{I_4} = E_2'P_3 \quad , \tag{2.52}$$

où (on a aussi utilisé ${P_3}^{-1} = P_3$)

$$E_2' = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 1 & 0 \\ 0 & a & 0 & 1 \end{pmatrix} \tag{2.53}$$

est, en effet, une matrice type E_k dans (2.29), obtenue de E_2 précisément en interchangeant les coefficients 3 et 4 (ceux interchangés par P_3) dans la colonne 2.

Cette propriété de commutation n'est pas valable pour n'importe quelle permutation et matrice "type E_k", mais il n'est pas difficile à se convaincre qu'elle est en fait valide pour les matrices spécifiques qu'on est en train considérer: notamment, les permutations P_{k+l} ($l \geq 1$) correspondant à l'arrangement de lignes après la ligne dont le pivot correspond à E_k. Dans ce cas, on a

$$P_{k+l} E_k = E_k' P_{k+l} \quad , \quad l \geq 1 \ , \tag{2.54}$$

où E_k' est obtenue à partir de E_k en interchangeant les éléments dans la colonne k sous la diagonale principale, d'accord avec la permutation P_{k+l} (cf. E_2 et E_2' en (2.50) et (2.53), respectivement). Pour l'expression générale, pour le cas des transpositions, on a :

$$P_{k+l}^{(i_o,j_o)} E_k = E_k' P_{k+l}^{(i_o,j_o)} \ , \tag{2.55}$$

avec

$$\begin{aligned} (E_k')_{i_o,k} &= (E_k)_{j_o,k} \\ (E_k')_{j_o,k} &= (E_k)_{i_o,k} \\ (E_k')_{i,j} &= (E_k)_{i,j} \quad \text{pour le reste, c.à.d.} \quad j \neq k \text{ ou } (i \neq i_o, j_o) \ . \end{aligned} \tag{2.56}$$

Nous sommes maintenant en condition pour énoncer et démontrer le théorème de décomposition LU avec permutations.

Théorème 2.1 (Décomposition LU)**.** *Soit $A \in M_n(\mathbb{R})$ une matrice inversible. Alors il existe une permutation P (non unique) telle que, pour ce P, il existe un et un seul couple de matrices (L, U) où L est triangulaire inférieure de termes diagonaux égaux à 1 et U triangulaire supérieure, avec*

$$PA = LU \ . \tag{2.57}$$

Preuve :

- *Existence.*

  L'existence découle directement de l'expression (2.48) comme résultat de la factorisation dans l'algorithme LU avec permutations. On peut alors écrire :

  $$E_{n-1} P_{n-1} \ldots E_2 P_2 \; E_1 P_1 \; A = U. \tag{2.58}$$

  Si on utilise maintenant (2.54) pour passer les permutations P_i vers la droite, on a :

  $$\begin{aligned} & E_{n-1} E_{n-2}' P_{n-1} \ldots E_2' P_3 E_1' P_2 P_1 \; A = U \\ & \ldots \\ & (E_{n-1} E_{n-2}' \ldots E_2^{n-3)} E_1^{n-2)}) \; (P_{n-1} P_{n-2} \ldots P_3 P_2 P_1) A = U \ . \end{aligned} \tag{2.59}$$

Si on définit maintenant L et P comme :

$$\begin{aligned} L &= (E_{n-1}E'_{n-2}\dots E_2^{n-3)}E_1^{n-2)})^{-1} = (E_1^{n-2)})^{-1}(E_2^{n-3)})^{-1}(E'_{n-2})^{-1}(E_{n-1})^{-1} \\ P &= P_{n-1}P_{n-2}\dots P_3P_2P_1 \ , \end{aligned} \tag{2.60}$$

on peut écrire

$$PA = LU \ . \tag{2.61}$$

Notez que la preuve d'existence n'utilise pas le fait que A soit inversible.

- *Unicité.*

  On suppose que pour P donné, il existe deux couples (L_1, U_1) et (L_2, U_2), tels que:

$$PA = L_1U_1 \quad , \quad PA = L_2U_2 \ , \tag{2.62}$$

alors

$$L_1U_1 = L_2U_2 \tag{2.63}$$

Si A est inversible, vu que L_1 et L_2 sont inversibles, on conclut que U_1 et U_2 sont inversibles. Ainsi, on peut multiplier (2.63) à gauche par L_2^{-1} et à droite par U_1^{-1}. On obtient

$$L_2^{-1}L_1 = U_2U_1^{-1} \tag{2.64}$$

Maintenant, on utilise[3] le fait que $L_2^{-1}L_1$ est une matrice triangulaire inférieure avec diag$(1, \dots, 1)$ pour diagonale principale et $U_2U_1^{-1}$ est une matrice triangulaire supérieure. Alors, nécessairement, on a :

$$L_2^{-1}L_1 = I_n = U_2U_1^{-1} \ , \tag{2.65}$$

et on conclut

$$L_1 = L_2 \quad , \quad U_1 = U_2 \tag{2.66}$$

□

Remarque 2.3. Faisons quelques remarques générales sur le résultat de la décomposition $PA = LU$:

i) Dans le résultat de la décomposition $PA = LU$, on affirme l'existence de P, mais pas son unicité.

ii) étant donnée P, si A est inversible, L et U sont uniques. Mais si A n'est pas inversible, la décomposition $PA = LU$ peut toujours se faire, mais elle n'est pas unique.

[3] Exercice! (Matrices triangulaires). Montrer que pour L_1, L_2 matrices triangulaires inférieures (respectivement U_1, U_2 matrices triangulaires supérieures), les matrices $L_1^{-1}, L_2^{-1}, L_1 \cdot L_2$ sont triangulaires inférieures (respectivement, $U_1^{-1}, U_2^{-1}, U_1 \cdot U_2$ sont triangulaires supérieures).

Exemple 2.4 (Non-unicité de $PA = LU$, pour A non-inversible). Si on considère

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{2.67}$$

on peut vérifier $\forall \lambda \in \mathbb{R}$

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 0 & 1-\lambda \end{pmatrix} \tag{2.68}$$

Cette non-unicité est consistent avec la non-inversibilité de A

i) *Pivot partiel versus pivot total.* Dans l'algorithme qui amène à la construction des matrices L et U (partie d'existence), on interchange seulement les lignes, et pas les colonnes, pour trouver le pivot avec la valeur absolue la plus grande. On parle de *pivot partiel.* Pour le pivot total, on chercherait le plus grand pivot en changeant aussi les colonnes.

ii) L'algorithme du pivot partiel dans le problème $A\boldsymbol{x} = \boldsymbol{b}$, ne réarrange pas les composantes de $\boldsymbol{x}$ vu qu'on ne change pas le colonnes de A. Le vecteur $\boldsymbol{b}$, par contre, est en général réarrangé.

Caractérisation $A = LU$ sans permutation

Avant de considérer plus spécifiquement le cas de matrices symétriques définies positives, nous allons donner une caractérisation de la possibilité d'écrire $A = LU$ sans matrice de permutation P. D'abord, nous devons introduire la définition suivante.

Définition 2.3 (matrice principale, mineur principal)**.** Soit $A \in M_n(\mathbb{R})$ et $k \in \{1, \ldots, k\}$. On appelle:

i) Matrice principale d'ordre k de A, la matrice $A_k \in M_k(\mathbb{R})$ définie par $(A_k)_{ij} = A_{ij}, \forall i, j \in \{1, \ldots, k\}$.

ii) Mineur principal d'ordre k, le déterminant de la matrice principale d'ordre k.

Avec ces éléments, nous pouvons énoncer la caractérisation suivante.

Proposition 2.1 (Caractérisation de $A = LU$, sans permutation)**.** *Soit $A \in M_n(\mathbb{R})$. Les deux propriétés suivantes sont équivalentes:*

i) Il existe une unique couple (L, U), avec L matrice triangulaire inférieure dont les coefficients diagonaux sont égaux à 1 et U une matrice inversible triangulaire supérieure, tel que $A = LU$.

ii) Les mineurs principaux de A sont tous non nuls.

Avant de démontrer ce résultat, on va introduire le lemme technique suivant.

Lemme 2.1. *Soit $A \in M_n(\mathbb{R})$ et $k \in \{1, \ldots, n\}$. On suppose que pour chaque matrice principale A_k de A, il existe une matrice $L_k \in M_k(\mathbb{R})$ triangulaire inférieure avec des coefficients diagonaux tous égaux à 1 et une matrice triangulaire supérieure $U_k \in M_k(\mathbb{R})$ inversible, telles que $A_k = L_k U_k$. On a alors:*

a) On peut écrire A sous la forme :

$$A = \begin{pmatrix} L_k & 0_{k\times(n-k)} \\ C_k & I_{n-k} \end{pmatrix} \begin{pmatrix} U_k & B_k \\ 0_{(n-k)\times k} & D_k \end{pmatrix} \quad , \tag{2.69}$$

où $B_k \in M_{k,n-k}(\mathbb{R})$, $C_k \in M_{n-k;k}(\mathbb{R})$ et $D_k \in M_{n-k;n-k}(\mathbb{R})$.

b) En particulier, la matrice principale d'ordre $k+1$ s'écrit sous la forme

$$A_{k+1} = \begin{pmatrix} L_k & 0_{1\times k} \\ c_k & 1 \end{pmatrix} \begin{pmatrix} U_k & b_k \\ 0_{k\times 1} & d_k \end{pmatrix} \quad , \tag{2.70}$$

où $(b_k) \in M_{k,1}(\mathbb{R})$ est la première colonne de la matrice B_k , $(c_k) \in M_{,k}(\mathbb{R})$ est la première ligne de la matrice C_k, et d_k est le coefficient de la ligne 1 et colonne 1 de D_k.

Preuve: À compléter. □

Preuve (de la proposition 2.1): À compléter.

□

2.2 Décomposition de Choleski

Dans cette section, nous allons nous concentrer sur le cas des matrices symétriques définies positives.

Définition 2.4 (Matrice définie positive)**.** $A \in M_n(\mathbb{R})$ est symétrique définie positive si :

i) $A = A^t$ (symétrique).

ii) $Ax \cdot x > 0, \forall x \in \mathbb{R}^n \setminus \{0\}$ (définie positive).

Un exemple important de ce type de matrice est la discrétisation du Laplacien qu'on a étudié dans l'exemple 1.1.

Une caractérisation très utile des matrices symétriques définies positives est donnée en termes de mineurs principaux.

Proposition 2.2 (Caractérisation d'une matrice symétrique définie positive par ses mineurs principaux)**.** *$A \in M_n(\mathbb{R})$ est symétrique définie positive si et seulement si tous ses mineurs principaux sont strictement positifs.*

En particulier, comme conséquence de la proposition 2.1, une matrice A symétrique définie positive admet toujours une (unique) décomposition $A = LU$ sans permutation. Dans cette section nous allons discuter une autre factorisation spécialement adaptée au caractère symétrique et défini positif de ces matrices. Nous commençons avec un exemple qui part de la décomposition LU.

Exemple 2.5 (Premier contact avec la factorisation de Choleski)**.** Nous considérons la matrice

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 5 & 6 \\ 2 & 6 & 17 \end{pmatrix} \quad . \tag{2.71}$$

La matrice A est symétrique et, comme on peut le vérifier à partir e.g. du calcul de ses mineurs principaux, elle est aussi définie positive. Alors, elle admet une décomposition $A = LU$ unique. On va construire explicitement une telle décomposition.

$$\begin{pmatrix} 1 & 1 & 2 \\ 1 & 5 & 6 \\ 2 & 6 & 17 \end{pmatrix} \xrightarrow[\substack{\ell_3 \to \ell_3 - 2\ell_1}]{\ell_2 \to \ell_2 - \ell_1} \underset{(E_1)^{-1}=\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix}}{\begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & 4 \\ 0 & 4 & 13 \end{pmatrix}} \xrightarrow{\ell_3 \to \ell_3 - \ell_2} \underset{(E_2)^{-1}=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}}{\begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & 4 \\ 0 & 0 & 9 \end{pmatrix}} . \tag{2.72}$$

Alors, on peut écrire

$$\begin{pmatrix} 1 & 1 & 2 \\ 1 & 5 & 6 \\ 2 & 6 & 17 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & 4 \\ 0 & 0 & 9 \end{pmatrix}}_{U} \tag{2.73}$$

Notez qu'on peut écrire :

$$U = \begin{pmatrix} 1 & 1 & 2 \\ 0 & 4 & 4 \\ 0 & 0 & 9 \end{pmatrix} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{L^t} \tag{2.74}$$

où D est la diagonale principale de U. Alors,

$$A = LDL^t \tag{2.75}$$

Nous notons que tous les coefficients de D sont strictement positifs. Alors, nous pouvons introduire la matrice racine carrée

$$D^{\frac{1}{2}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \tag{2.76}$$

pour écrire :

$$\begin{aligned} A = LD^{\frac{1}{2}}D^{\frac{1}{2}}L^t &= \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix}}_{\tilde{L}} \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}}_{\tilde{L}^t} \\ &= \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 2 & 3 \end{pmatrix}}_{\tilde{L}} \underbrace{\begin{pmatrix} 1 & 1 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix}}_{\tilde{L}^t} \end{aligned} \tag{2.77}$$

C'est-à-dire, on a écrit

$$A = \tilde{L}\tilde{L}^t \quad , \quad \text{avec} \quad \tilde{L} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 2 & 2 & 3 \end{pmatrix} . \tag{2.78}$$

La matrice $\tilde{L}$ est triangulaire inférieure avec coefficients positifs dans la diagonale principale. L'expression (2.78) s'appelle la décomposition de Choleski de A.

Le pas clé, dans l'exemple précédent, est la possibilité de prendre (dans les réels) la racine carrée de la matrice D définie pour la diagonale principale de U. Dans l'exemple suivant, le cas de la décomposition de Cholesky dans $M_2(\mathbb{R})$, on voit que ceci est une conséquence du caractère symétrique défini positif des matrices considérées. Après l'exemple, ceci est énoncé en toute généralité.

Exemple 2.6 (Décomposition de Cholesky: cas $n = 2$)**.** On considère une matrice $A \in M_2(\mathbb{R})$ symétrique définie positive, on peut toujours écrire :

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} . \tag{2.79}$$

Si on prends $\boldsymbol{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, le caractère défini positif de A implique

$$A\boldsymbol{x} \cdot \boldsymbol{x} = \boldsymbol{x}^t A\boldsymbol{x} = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = a > 0 . \tag{2.80}$$

Autrement dit, le mineur principal $A_1 = a > 0$, comme il en résulte de la caractérisation 2.2. Ainsi, étant donné que $a > 0$, on peut procéder à la décomposition LU

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix} \overset{\substack{\ell_2 \to \ell_2 - \frac{b}{a}\ell_1 \\ \longrightarrow \\ L = \begin{pmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{pmatrix}}}{} \begin{pmatrix} a & b \\ 0 & c - \frac{b^2}{a} \end{pmatrix} . \tag{2.81}$$

Alors, on peut écrire :

$$A = \underbrace{\begin{pmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} a & b \\ 0 & c - \frac{b^2}{a} \end{pmatrix}}_{U} = \underbrace{\begin{pmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} a & 0 \\ 0 & c - \frac{b^2}{a} \end{pmatrix}}_{D} \underbrace{\begin{pmatrix} 1 & \frac{b}{a} \\ 0 & 1 \end{pmatrix}}_{L^t} \tag{2.82}$$

Si on utilise maintenant le reste de l'information sur A, notamment la positivité du mineur $A_2 = ac - b^2 > 0$, on conclut (avec $a > 0$) $c - \dfrac{b^2}{a} > 0$. On peut alors écrire :

$$A = \underbrace{\begin{pmatrix} 1 & 0 \\ \frac{b}{a} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} \sqrt{a} & 0 \\ 0 & \sqrt{c - \frac{b^2}{a}} \end{pmatrix}}_{D^{\frac{1}{2}}} \underbrace{\begin{pmatrix} \sqrt{a} & 0 \\ 0 & \sqrt{c - \frac{b^2}{a}} \end{pmatrix}}_{D^{\frac{1}{2}}} \underbrace{\begin{pmatrix} 1 & \frac{b}{a} \\ 0 & 1 \end{pmatrix}}_{L^t} \tag{2.83}$$

$$= \underbrace{\begin{pmatrix} \sqrt{a} & 0 \\ \frac{b}{\sqrt{a}} & \sqrt{c - \frac{b^2}{a}} \end{pmatrix}}_{\tilde{L}} \underbrace{\begin{pmatrix} \sqrt{a} & \frac{b}{\sqrt{a}} \\ 0 & \sqrt{c - \frac{b^2}{a}} \end{pmatrix}}_{\tilde{L}^t} \tag{2.84}$$

Comme on l'a annoncé ci-dessus, la proposition suivante nous garantit que $D^{\frac{1}{2}}$ peut toujours être construite comme matrice réelle, pour des matrices symétriques définies positives.

Proposition 2.3 (Caractérisation d'une matrice symétrique définie positive à partir de sa décomposition LU)**.** *Soit $A \in M_n(\mathbb{R})$ une matrice symétrique admettant une décomposition LU sans permutation. Alors A est symétrique définie positive si et seulement si tous les pivots (c.à.d. les coefficients diagonaux de la matrice U) sont strictement positifs.*

Preuve: À compléter. □

Pour continuer, nous allons énoncer et prouver le théorème de décomposition de Choleski, qui présente une partie d'existence et une autre d'unicité. En particulier, la partie d'existence suit l'esprit du schéma de l'exemple 2.6 pour le cas $n = 2$, dont le contenu est essentiellement celui de la proposition 2.3. La partie de l'unicité, qui utilise d'une manière basique le résultat d'existence, proportionne l'algorithme de Choleski proprement dit.

Théorème 2.2 (Décomposition de Choleski)**.** *Soit $A \in M_n(\mathbb{R})$, une matrice symétrique définie positive. Alors il existe une unique matrice $\tilde{L} \in M_n(\mathbb{R})$ telle que:*

i) $\tilde{L}$ *est triangulaire inférieure.*

ii) $\tilde{l}_{ii} > 0$, $\forall i \in \{1, \ldots, n\}$.

iii) $A = \tilde{L}\tilde{L}^t$.

Preuve:

- *Existence.* Par récurrence sur la dimension n.

  1) Cas $n = 1$.
  On a $A = (a_{11})$. A est automatiquement symétrique et elle est définie positive si et seulement si $a_{11} > 0$. On définit $\tilde{L} = (\tilde{l}_{11})$, avec $\tilde{l}_{11} = \sqrt{a_{11}}$ et on a

  $$A = \tilde{L}\tilde{L}^t \ . \tag{2.85}$$

  2) On suppose $A = \tilde{L}\tilde{L}^t$, avec $\tilde{L}$ triangulaire inférieure et $\tilde{l}_{ii} > 0, \forall A \in M_p(\mathbb{R}), 1 \leq p \leq n$. On va démontrer que, pour $A \in M_{n+1}(\mathbb{R})$ symétrique et définie positive, il existe une décomposition de Choleski . On considère $A \in M_{n+1}(\mathbb{R})$, que l'on peut écrire :

  $$A = \left(\begin{array}{c|c} B & \boldsymbol{a} \\ \hline \boldsymbol{a}^t & \alpha \end{array}\right) \tag{2.86}$$

  où $\boldsymbol{a} \in \mathbb{R}^n$ est un vecteur colonne, B est symétrique et $\alpha \in \mathbb{R}$. D'abord, on montre que B est définie positive à partir du fait que A est définie positive, i.e. $\boldsymbol{x}^t A\boldsymbol{x} > 0, \forall \boldsymbol{x} \in \mathbb{R}^{n+1}, x \neq 0$. Si on prend en particulier $\boldsymbol{x} = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix}$, avec $\boldsymbol{y} \in \mathbb{R}^n$, on a :

  $$0 < \boldsymbol{x}^t A\boldsymbol{x} = \begin{pmatrix} \boldsymbol{y} & 0 \end{pmatrix} \left(\begin{array}{c|c} B & \boldsymbol{a} \\ \hline \boldsymbol{a}^t & \alpha \end{array}\right) \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix} = \boldsymbol{y}^t B\boldsymbol{y} \quad , \quad \forall \boldsymbol{y} \in \mathbb{R}^n \ . \tag{2.87}$$

  Alors B est symétrique définie positive. Par hypothèse de récurrence $\exists M \in M_n(\mathbb{R})$ triangulaire inférieure telle que

  $$B = MM^t \quad , \quad m_{ii} > 0 \ , \ \forall i \in \{1, \ldots, n\} \ . \tag{2.88}$$

  On cherche une matrice $\tilde{L}$ avec l'hypothèse (Ansatz) de la forme :

  $$\tilde{L} = \left(\begin{array}{c|c} M & 0 \\ \hline \boldsymbol{b}^t & \lambda \end{array}\right) \tag{2.89}$$

où $\boldsymbol{b} \in \mathbb{R}^n$ et $\lambda \in \mathbb{R}$ (on veut que $\lambda > 0$). On veut imposer :

$$A = \tilde{L}\tilde{L}^t \ . \tag{2.90}$$

D'abord, nous évaluons :

$$\tilde{L}\tilde{L}^t = \left(\begin{array}{c|c} M & 0 \\ \hline \boldsymbol{b}^t & \lambda \end{array}\right)\left(\begin{array}{c|c} M^t & \boldsymbol{b} \\ \hline 0 & \lambda \end{array}\right) = \left(\begin{array}{c|c} MM^t & M\boldsymbol{b} \\ \hline \boldsymbol{b}^t M^t & \boldsymbol{b}^t\boldsymbol{b} + \lambda^2 \end{array}\right) = \left(\begin{array}{c|c} B & M\boldsymbol{b} \\ \hline (M\boldsymbol{b})^t & \boldsymbol{b}^t\boldsymbol{b} + \lambda^2 \end{array}\right) \tag{2.91}$$

où on a utilisé (2.88). Maintenant, si on impose (2.90) en utilisant (2.86), on obtient :

$$\begin{aligned} M\boldsymbol{b} &= \boldsymbol{a} \\ \boldsymbol{b}^t\boldsymbol{b} + \lambda^2 &= \alpha \ . \end{aligned} \tag{2.92}$$

Par hypothèse de récurrence, M est inversible ($m_{ii} > 0$), alors :

$$\boldsymbol{b} = M^{-1}\boldsymbol{a} \ , \tag{2.93}$$

et

$$\alpha = (M^{-1}\boldsymbol{a})^t(M^{-1}\boldsymbol{a}) + \lambda^2 = \ldots = \boldsymbol{a}^t(\underbrace{MM^t}_{B})^{-1}\boldsymbol{a} + \lambda^2 = \boldsymbol{a}^t B^{-1}\boldsymbol{a} + \lambda^2 \ , \tag{2.94}$$

et on peut écrire :

$$\lambda^2 = \alpha - \boldsymbol{a}^t B^{-1}\boldsymbol{a} \ . \tag{2.95}$$

On va montrer que $\alpha - \boldsymbol{a}^t B^{-1}\boldsymbol{a} > 0$, en utilisant la pièce d'information qu'on n'a pas encore utilisé, notamment que A est aussi définie positive pour un vecteur $\boldsymbol{x}$ avec la composante $x^{n+1} \neq 0$. Pour ça, on choisit:

$$\boldsymbol{x} = \begin{pmatrix} B^{-1}\boldsymbol{a} \\ -1 \end{pmatrix} \tag{2.96}$$

On a $\boldsymbol{x} \neq 0$, alors :

$$\begin{aligned} 0 < \boldsymbol{x}^t A\boldsymbol{x} &= \left((B^{-1}\boldsymbol{a})^t \quad -1\right)\left(\begin{array}{c|c} B & \boldsymbol{a} \\ \hline \boldsymbol{a}^t & \alpha \end{array}\right)\begin{pmatrix} B^{-1}\boldsymbol{a} \\ -1 \end{pmatrix} = \left((B^{-1}\boldsymbol{a})^t \quad -1\right)\begin{pmatrix} \overbrace{BB^{-1}}^{I_n}\boldsymbol{a} - \boldsymbol{a} \\ \boldsymbol{a}^t B^{-1}\boldsymbol{a} - \alpha \end{pmatrix} \\ &= \left((B^{-1}\boldsymbol{a})^t \quad -1\right)\begin{pmatrix} 0 \\ \boldsymbol{a}^t B^{-1}\boldsymbol{a} - \alpha \end{pmatrix} = \alpha - \boldsymbol{a}^t B^{-1}\boldsymbol{a} \end{aligned} \tag{2.97}$$

Alors, de (2.95) on peut écrire

$$\lambda = \sqrt{\alpha - \boldsymbol{a}^t B^{-1}\boldsymbol{a}} \ , \tag{2.98}$$

et, finalement

$$\tilde{L} = \left(\begin{array}{c|c} M & 0 \\ \hline (M^{-1}\boldsymbol{a})^t & \sqrt{\alpha - \boldsymbol{a}^t B^{-1}\boldsymbol{a}} \end{array}\right) \ , \tag{2.99}$$

est une matrice triangulaire inférieure, avec $\tilde{l}_{ii}, \forall i \in \{1, \ldots, n+1\}$ et telle que (2.90) est satisfaite.

- *Unicité.*

  Dans la deuxième de partie de la preuve, on utilise de manière fondamentale l'existence montrée dans la première partie. Cela nous garantit, en particulier, de prendre la racine carrée des expressions qu'on va trouver dans l'algorithme.

  Étant donnée A symétrique définie positive, dans la première partie on a montré l'existence d'une matrice $\tilde{L}$ telle que :

$$A = \tilde{L}\tilde{L}^t \;\; ; \;\; \tilde{l}_{ij} = 0 \; , \; j > i \;\; ; \;\; \tilde{l}_{ii} > 0 \; \forall i \; . \tag{2.100}$$

  Alors, si on écrit le produit de matrices en termes des coefficients :

$$a_{ij} = \sum_{k=1}^{n} \tilde{l}_{ik}\tilde{l}^t_{kj} = \sum_{k=1}^{n} \tilde{l}_{ik}\tilde{l}_{jk} \tag{2.101}$$

  On va employer cette expression pour déterminer d'une manière constructive l'expression des coefficients $\tilde{l}_{ij}$ en termes des données a_{ij}, en particulier *de forme unique.*

  On va procéder colonne par colonne.

  i) On commence avec la première colonne. Si on fait $j = 1$ dans (2.101) on a, pour la première ligne

$$a_{11} = \sum_{k=1}^{n} \tilde{l}_{1k}\tilde{l}_{1k} = \tilde{l}_{11}\tilde{l}_{11} + \sum_{k=2}^{n} \tilde{l}_{1k} \underbrace{\tilde{l}_{1k}}_{\substack{=0 \\ (k>1)}} = (\tilde{l}_{11})^2 \; . \tag{2.102}$$

     Alors

$$l_{11} = \sqrt{a_{11}} > 0 \; . \tag{2.103}$$

     On a la garantie qu'on peut prendre la racine carrée de a_{11} parce qu'on a montré dans la première partie de la preuve l'existence d'un $l_{11} > 0$: dans le cas de trouver $a_{11} \leq 0$, il y aurait donc une contradiction avec le résultat déjà prouvé.
     On continue avec le reste des lignes dans la première colonne :

$$\begin{aligned}
a_{21} &= \sum_{k=1}^{n} \tilde{l}_{2k}\tilde{l}_{1k} = \tilde{l}_{21}\tilde{l}_{11} + \sum_{k=2}^{n} \tilde{l}_{2k} \underbrace{\tilde{l}_{1k}}_{\substack{=0 \\ (k>1)}} = \tilde{l}_{21}\tilde{l}_{11} \\
\tilde{l}_{21} &= \frac{a_{21}}{\tilde{l}_{11}} = \frac{a_{21}}{\sqrt{a_{11}}} \\
&\vdots \\
a_{i1} &= \sum_{k=1}^{n} \tilde{l}_{ik}\tilde{l}_{1k} = \ldots = \tilde{l}_{i1}\tilde{l}_{11} \\
\tilde{l}_{i1} &= \frac{a_{i1}}{\sqrt{a_{11}}}
\end{aligned} \tag{2.104}$$

     Ceci fixe la première colonne de manière unique.

ii) Nous continuons avec le reste des colonnes, de façon récursive. Si on a déjà calculé les q premières colonnes, on calcule la colonne $q+1$. Pour ça, on commence dans la ligne $q \tilde{+} 1$, dans la diagonale principale, car dans le résultat d'existence, on a montré que la matrice est diagonale inférieure. Ainsi

$$\begin{aligned} a_{q+1,q+1} &= \sum_{k=1}^{n} \tilde{l}_{q+1,k}\tilde{l}_{q+1,k} = \sum_{k=1}^{q+1} (\tilde{l}_{q+1,k})^2 = (\tilde{l}_{q+1,q+1})^2 + \sum_{k=1}^{q} (\tilde{l}_{q+1,k})^2 \\ \tilde{l}_{q+1,q+1} &= \left(a_{q+1,q+1} - \sum_{k=1}^{q} (\tilde{l}_{q+1,k})^2 \right)^{\frac{1}{2}} > 0 \end{aligned} \tag{2.105}$$

On note que l'expression sous la racine carrée est positive: sinon on serait en contradiction avec le résultat d'existence de la matrice $\tilde{L}$. On continue avec le reste de la colonne. Pour $i \in \{q+2, \ldots, n\}$

$$\begin{aligned} a_{i,q+1} &= \sum_{k=1}^{n} \tilde{l}_{i,k}\tilde{l}_{q+1,k} = \sum_{k=1}^{q+1} \tilde{l}_{i,k}\tilde{l}_{q+1,k} = \tilde{l}_{i,q+1} \underbrace{\tilde{l}_{q+1,q+1}}_{>0} + \sum_{k=1}^{q} \tilde{l}_{i,k}\tilde{l}_{q+1,k} \\ \tilde{l}_{i,q+1} &= \frac{1}{\tilde{l}_{q+1,q+1}} \left(a_{i,q+1} - \sum_{k=1}^{q} \tilde{l}_{i,k}\tilde{l}_{q+1,k} \right) \end{aligned} \tag{2.106}$$

De cette manière, nous pouvons calculer toutes les colonnes de $\tilde{L}$ et celle-ci est complètement fixée de manière unique.

□

Remarque 2.4 (Décomposition de Choleski: quelques points). Nous soulignons les points suivants:

i) En général, toute matrice symétrique admet une décomposition unique $PA = LU$. Si A est symétrique définie positive, on peut choisir $P = I_n$, alors on a $A = LU = \tilde{L}\tilde{L}^t$, avec $\tilde{L} = L\sqrt{D}$ où $D = \text{diag(U)}$.

ii) Coût de la méthode Choleski: la méthode de Choleski (avec l'algorithme donné dans la partie d'unicité du théorème) est deux fois plus rapide que celui de la méthode de Gauss (décomposition LU).

iii) *Conservation du profil de A.* Étant donnée la matrice symetrique A on peut definir pour chaque ligne i l'entier j_i

$$j_i = \min \{ j \in \{1, \ldots\ n\} \} \quad \text{tels que } a_{ij} \neq 0 \ . \tag{2.107}$$

L'ensemble $\{j_i\}_{i \in \{1,\ldots,n\}}$ détermine le profil de la matrice. En particulier (pour A symétrique), seuls les éléments a_{ij} avec $j \in \{j_i, \ldots, i\}$ doivent être stockés. Une propriété importante de la décomposition de Choleski est qu'elle préserve le profil $\{j_i\}_{i \in \{1,\ldots,n\}}$ de la matrice A. Ceci peut être démontré à partir de la preuve d'unicité de la décomposition de Choleski en appliquant à chaque ligne i un raisonnement par l'absurde ou par récurrence sur les colonnes.

iv) Si on étudie le problème $A\boldsymbol{x} = \boldsymbol{b}$ avec A inversible mais non-symétrique, on peut utiliser la méthode de Choleski en notant que la matrice AA^t est symétrique définie positive:

$$A\boldsymbol{x} = \boldsymbol{b} \Leftrightarrow A^t (A\boldsymbol{x} = \boldsymbol{b}) \Leftrightarrow A^t A\boldsymbol{x} = A^t\boldsymbol{b} \tag{2.108}$$

La matrice A^tA, admet une décomposition de Choleski $A^tA = \tilde{L}\tilde{L}$.

Chapitre 3

Normes et conditionnement d'une matrice

Dans ce chapitre, nous introduisons des éléments de base autour les notions de norme matricielle et de conditionnement d'une matrice. Le but fondamental est la formulation de conditions suffisantes pour la convergence des suites définies par l'itération de l'action d'une matrice. Un objectif secondaire est la discussion de la majoration des erreurs d'arrondis dans la résolution du problème

$$A\boldsymbol{x} = \boldsymbol{b}\ , \tag{3.1}$$

en conséquence d'une erreur dans A ou dans $\boldsymbol{b}$.

3.1 Norme matricielle, norme matricielle induite et rayon spectral

3.1.1 Rappel sur les normes

Nous commençons par rappeler la notion de norme dans un espace linéaire (sur $\mathbb{C}$).

Définition 3.1 (Norme)**.** Étant donné un espace linéaire V sur $\mathbb{C}$, une norme est une application $||\cdot|| : V \to \mathbb{R}$ telle que

i) $||\alpha\boldsymbol{x}|| = |\alpha|||\boldsymbol{x}||\ \ ,\forall\alpha \in \mathbb{C},\ \boldsymbol{x} \in V.$

ii) $||\boldsymbol{x} + \boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||\ \ ,\forall \boldsymbol{x}, \boldsymbol{y} \in V$

iii) $||\boldsymbol{x}|| = 0 \Leftrightarrow \boldsymbol{x} = 0.$

Nous allons travailler avec des espaces linéaires réels de dimension fini.

Exemple 3.1 (Exemples de normes sur $\mathbb{R}^n$)**.** Nous donnons trois exemples des normes en $\mathbb{R}^n$. Étant donnée $\boldsymbol{x} = (x_1, \ldots, x_n)$ dans la base canonique, on définit :

i) La norme $||\boldsymbol{x}||_\infty$: $\quad ||\boldsymbol{x}||_\infty = \max\{|x_1|, \ldots, |x_n|\}.$

ii) La norme $||\boldsymbol{x}||_1$: $\quad ||\boldsymbol{x}||_1 = |x_1| + \ldots + |x_n| = \sum_{i=1}^{n} |x_i|.$

iii) La norme $||\boldsymbol{x}||_2$: $\quad ||\boldsymbol{x}||_2 = \left(|x_1|^2 + \ldots + |x_n|^2\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$.

iv) Les deux dernières sont cas particuliers de la norme $||\boldsymbol{x}||_p$, avec $p \in \mathbb{R}, p \geq 1$:

$$||\boldsymbol{x}||_p = (|x_1|^p + \ldots + |x_n|^p)^{\frac{1}{p}} = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}.$$

Définition 3.2 (Normes équivalentes)**.** Deux normes $||\cdot||_1$ et $||\cdot||_2$ sur V sont équivalentes s'il existe deux constantes réelles $C_1 > 0$ et $C_2 > 0$ telles que

$$C_1||\boldsymbol{x}||_1 \leq ||\boldsymbol{x}||_2 \leq C_2||\boldsymbol{x}||_2 \quad , \quad \forall \boldsymbol{x} \in V \ . \tag{3.2}$$

Remarque 3.1. Dans la suite, on va utiliser les deux propriétés suivantes :

- Deux normes équivalentes induisent la même topologie sur V.
- Si V est de dimension finie, toutes les normes sont équivalentes.

3.1.2 Normes matricielles

Définition 3.3 (Norme matricielle)**.** Étant donnée l'espace vectoriel $M_n(\mathbb{R})$:

i) On appelle *em norme matricielle* sur $M_n(\mathbb{R})$ une norme $||\cdot||$ sur $\mathbb{R}$ telle que

$$||AB|| \leq ||A||\ ||B|| \quad , \forall A, B \in M_n(\mathbb{R}) \ . \tag{3.3}$$

ii) Si $\mathbb{R}^n$ est muni d'une norme $||\cdot||$, on appelle *norme matricielle induite (ou subordonnée)* sur $M_n(\mathbb{R})$ par $||\cdot||$ à la norme sur $M_n(\mathbb{R})$ définie par

$$||A|| = \sup_{\boldsymbol{x} \in \mathbb{R}^n, ||\boldsymbol{x}||=1} \{||A\boldsymbol{x}||\} \ , \forall A \in M_n(\mathbb{R}) \ . \tag{3.4}$$

Proposition 3.1 (propriété de la norme induite)**.** . *Soit $M_n(\mathbb{R}^n)$ muni d'une norme induite $||\cdot||$. Alors, pour tout $A \in M_n(\mathbb{R})$ on a:*

i) $||A\boldsymbol{x}|| \leq ||A||\ ||\boldsymbol{x}|| \quad , \forall \boldsymbol{x} \in \mathbb{R}^n$.

ii) $||A|| = \max_{\boldsymbol{x} \in \mathbb{R}^n, ||\boldsymbol{x}||=1} \{||A\boldsymbol{x}||\}$.

iii) $||A|| = \max\left\{ \dfrac{||A\boldsymbol{x}||}{||\boldsymbol{x}||} \ ; \boldsymbol{x} \in R^n \setminus \{0\} \right\}$

iv) $||\cdot||$ *est une norme matricielle.*

Preuve:

i) Soit $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$. On prend $\boldsymbol{y} = \frac{\boldsymbol{x}}{||\boldsymbol{x}||}$, alors $||\boldsymbol{y}|| = 1$. De la définition de norme induite on a $A\boldsymbol{y} \leq ||A||(= \sup\{||A\boldsymbol{y}||; ||\boldsymbol{y}|| = 1\})$. Alors

$$||A|| \geq ||A\frac{\boldsymbol{x}}{||\boldsymbol{x}||}|| = \frac{1}{||\boldsymbol{x}||}||A\boldsymbol{x}|| \Leftrightarrow ||A||\ ||\boldsymbol{x}|| \geq ||A\boldsymbol{x}||, \forall \boldsymbol{x} \in \mathbb{R}^n \setminus 0 \ . \tag{3.5}$$

Si $\boldsymbol{x} = 0$, il se suit $A\boldsymbol{x} = 0$ et $||\boldsymbol{x}||$, alors $||A\boldsymbol{x}|| \leq ||A||\ ||\boldsymbol{x}||$ est vérifié.

ii) On définie $\varphi : \mathbb{R}^n \to \mathbb{R}$, avec $\varphi(\boldsymbol{x}) = ||A\boldsymbol{x}||$. L'application φ est continue sur $S^1 = \{\boldsymbol{x} \in \mathbb{R}^n, ||\boldsymbol{x}|| = 1\}$, que c'est un compact dans $\mathbb{R}^n$. Alors, l'image de S^1 par φ est un compacte, dons φ est bornée est atteint ses bornes. On conclut qu'il existe $\boldsymbol{x}_{\max}$, tel que $||A|| = ||A\boldsymbol{x}_{\max}||$.

iii) On prend $||\boldsymbol{x} \neq 0||$. Alors

$$\frac{||A\boldsymbol{x}||}{||\boldsymbol{x}||} = A\frac{||\boldsymbol{x}||}{||\boldsymbol{x}||}, \boldsymbol{x} \in S^1 \Rightarrow \max_{x\in\mathbb{R}^n\setminus\{0\}} \left\{ \frac{||A\boldsymbol{x}||}{||\boldsymbol{x}||} \right\} = \max_{x\in S^1}\{||A\boldsymbol{x}||\} = ||A||. \tag{3.6}$$

iv) Étant donnés A et $B \in M_n(\mathbb{R})$, on a $||AB|| = \max\{||AB\boldsymbol{x}|| : ||\boldsymbol{x}|| = 1, \boldsymbol{x} \in \mathbb{R}^n\}$. Or, en utilisant deux fois la propriété dans le point i):

$$||AB\boldsymbol{x}|| \leq ||A||\,||B\boldsymbol{x}|| \leq ||A||\,||B||\,||\boldsymbol{x}|| = ||A||\,||B|| \tag{3.7}$$

Alors

$$||AB|| = \max\{||AB\boldsymbol{x}|| : ||\boldsymbol{x}|| = 1, \boldsymbol{x} \in \mathbb{R}^n\} \leq ||A||\,||B|| \ . \tag{3.8}$$

On conclut que $||\cdot||$ est une norme matricielle

□

Définition 3.4 (Rayon spectral)**.** Soit $A \in M_n(\mathbb{R})$ inversible. On appelle *rayon spectral* $\rho(A)$ de A à:

$$\rho(A) = \max\{|\lambda|; \lambda \in \mathbb{C}, \lambda \text{ valeur propre de } A\} \tag{3.9}$$

Proposition 3.2 (Caractérisation de de normes induites; calculs effectifs))**.** *Soit* $A = (a_{ij}) \in M_n(\mathbb{R})$. *Montrer:*

i) On munit $\mathbb{R}^n$ *de la norme* $||\cdot||_\infty$ *et* $M_n(\mathbb{R})$ *de la norme induite correspondante; notée aussi* $||\cdot||_\infty$. *Alors*

$$||A||_\infty = \max_{i\in\{1,\ldots,n\}} \sum_{j=1}^{n} |a_{ij}|$$

ii) On munit $\mathbb{R}^n$ *de la norme* $||\cdot||_1$ *et* $M_n(\mathbb{R})$ *de la norme induite correspondante; notée aussi* $||\cdot||_1$. *Alors*

$$||A||_1 = \max_{j\in\{1,\ldots,n\}} \sum_{i=1}^{n} |a_{ij}|$$

iii) On munit $\mathbb{R}^n$ *de la norme* $||\cdot||_2$ *et* $M_n(\mathbb{R})$ *de la norme induite correspondante; notée aussi* $||\cdot||_2$. *Alors*

$$||A||_2 = \left(\rho(A^tA)\right)^{\frac{1}{2}}$$

En particulier; si A *est symétrique,* $||A||_2 = \rho(A)$.

Preuve: i) et ii) Exercice.
iii) Par définition

$$\begin{aligned} ||A||_2^2 &= \left(\max_{\boldsymbol{x}\in\mathbb{R}^n,||\boldsymbol{x}||=1}\{||A\boldsymbol{x}||_2\}\right)^2 = \left(\max_{\boldsymbol{x}\in\mathbb{R}^n,||\boldsymbol{x}||=1}\{(A\boldsymbol{x}\cdot A\boldsymbol{x})^{\frac{1}{2}}\}\right)^2 \\ &= \max_{\boldsymbol{x}\in\mathbb{R}^n,||\boldsymbol{x}||=1}\{A\boldsymbol{x}\cdot A\boldsymbol{x}\} = \max_{\boldsymbol{x}\in\mathbb{R}^n,||\boldsymbol{x}||=1}\{A^tA\boldsymbol{x}\cdot\boldsymbol{x}\} \end{aligned} \tag{3.10}$$

La matrice A^tA est symétrique de "non-negative". En effet

$$\boldsymbol{x}^tA^tA\boldsymbol{x} = \boldsymbol{x}\cdot A^tA\boldsymbol{x} = A\boldsymbol{x}\cdot A\boldsymbol{x} \geq 0\ . \tag{3.11}$$

(Noter que $A\boldsymbol{x}$=0 n'implique pas $\boldsymbol{x}$ en générale, au moins que A soit inversible). Alors, on peut diagonaliser A^tA sur une base orthonormée $B = \{\boldsymbol{e}_i, i \in \{1,\ldots,n\}\}$ avec valeurs propres non-négatifs :

$$\begin{aligned} &A^tA\boldsymbol{e}_i = \lambda_i\boldsymbol{e}_i \quad , \quad \boldsymbol{e}_i\cdot\boldsymbol{e}_j\delta_{ij} \\ &0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n\ . \end{aligned} \tag{3.12}$$

On considère $\boldsymbol{x} = \sum_{i=1}^{n} x^i\boldsymbol{e}_i$. Alors

$$\begin{aligned} A^tA\boldsymbol{x}\cdot\boldsymbol{x} &= \left(A^tA\sum_i x^i\boldsymbol{e}_i\right)\cdot\left(\sum_j x^j\boldsymbol{e}_j\right) = \left(\sum_i x^i\lambda_i\boldsymbol{e}_i\right)\cdot\left(\sum_j x^j\boldsymbol{e}_j\right) \\ &= \sum_{i,j}\lambda_i x^i x^j\underbrace{\boldsymbol{e}_i\cdot\boldsymbol{e}_j}_{\delta_{ij}} = \sum_i\lambda_i(x^i)^2 \leq \sum_i\lambda_n(x^i)^2 = \lambda_n\underbrace{\sum_i(x^i)^2}_{||\boldsymbol{x}||_2^2} \end{aligned} \tag{3.13}$$

Avec ceci, on conclut

$$\langle A^tA\frac{\boldsymbol{x}}{||\boldsymbol{x}||},\frac{\boldsymbol{x}}{||\boldsymbol{x}||}\rangle = \frac{A^tA\boldsymbol{x}\cdot\boldsymbol{x}}{||\boldsymbol{x}||^2} \leq \lambda_n = \max\{|\lambda|,\lambda \text{ valeur propre de } A^tA\} = \rho(A^tA)\ , \tag{3.14}$$

Avec, (3.10) on a

$$||A||_2^2 \leq \rho(A^tA)\ . \tag{3.15}$$

Pour l'égalité, on considère $\boldsymbol{x} = \boldsymbol{e}_n$, $||\boldsymbol{e}|| = 1$, et on a

$$||A\boldsymbol{e}_n||_2^2 = \langle A\boldsymbol{e}_n, A\boldsymbol{e}_n\rangle = \langle A^tA\boldsymbol{e}_n,\boldsymbol{e}_n\rangle = \langle\lambda_n\boldsymbol{e}_n,\boldsymbol{e}_n\rangle = \lambda_n = \rho(A^tA)\ . \tag{3.16}$$

Alors

$$||A||_2^2 = \rho(A^tA)\ , \tag{3.17}$$

qui montre le résultat. En particulier, si A est symétrique elle diagonalisable avec des valeurs propres λ_i^A, on a $A^tA = A^2$ et $\lambda_i = (\lambda_i^A)^2, \forall i \in \{1,\ldots,n\}$. En particulier, pour $i = n$ on a $\lambda_n = (|\lambda_n^A|)^2$, où $|\lambda_n^A| = \rho(A)$ et on conclut

$$||A||_2^2 = \rho(A^tA) = \lambda_n = (\rho(A))^2 \Rightarrow ||A||_2 = \rho(A)\ . \tag{3.18}$$

□

3.2 Relation entre norme matricielle et rayon spectral

Théorème 3.1 (Approximation spectral du rayon spectral par une norme induite)**.** *On considère* $\mathbb{R}^n$ *et* $M_n(\mathbb{R})$*:*

i) Soit $||\cdot||$ *une norme induite en* $M_n(\mathbb{R})$*. Alors*

$$\rho(A) \leq ||A|| \ , \quad \forall A \in M_n(\mathbb{R}) \ . \tag{3.19}$$

ii) Étant donnés $A \in M_n(\mathbb{R}^n)$ *et* $\varepsilon > 0$*, il exixte une norme sur* $\mathbb{R}^n$ *(qui dépend de* A *est de* ε*) telle que la correspondante norme induite sur* $M_n(\mathbb{R})$*, notée* $||\cdot||_{A,\varepsilon}$*, vérifie*

$$||A||_{A,\varepsilon} \leq \rho(A) + \varepsilon \ , \quad \forall A \in M_n(\mathbb{R}) \ . \tag{3.20}$$

Preuve: Pour i), étant donnée A on considère $\lambda \in \mathbb{C}$ valeur propre de A, c'est-à-dire, $\exists \boldsymbol{x} \neq 0$ tel que $A\boldsymbol{x} = \lambda \boldsymbol{x}$. Alors

$$||A\boldsymbol{x}|| = ||\lambda \boldsymbol{x}|| = |\lambda| ||\boldsymbol{x}|| \quad , \quad ||A\boldsymbol{x}|| \leq ||\boldsymbol{A}|| ||x|| \ , \tag{3.21}$$

donc $|\lambda| \leq ||A||$ et ceci $\forall \lambda$ valeur propre de A. Alors $\rho(A) \leq ||A||$.

On laisse ii) comme exercice.

□

Remarque 3.2. Après le théorème précédent, pour une matrice $A \in M_n(\mathbb{R})$ donnée, on peut considérer $\rho(A)$ comme le infimum sur l'ensemble $\{||A||;$ avec $||\cdot||$ norme induite sur $M_n(R)\}$.

3.2.1 Convergence et rayon spectral

Corollaire 3.1 (Convergence et rayon spectral)**.** *Soit* $A \in M_n(\mathbb{R})$*. A condition nécessaire et suffisante pour la convergence à zéro de la suite de matrices* $\{A_k\}_{k\in\mathbb{N}}$ *est:*

$$\rho(A) < 1 \Leftrightarrow A^k \to 0 \ . \tag{3.22}$$

Preuve: On montre d'abord la direction $\Rightarrow$ (condition suffisante). Si $\rho(A) < 1$, on peut toujours choisir $\varepsilon < 0$ tel que $\rho(A) < 1 - 2\varepsilon$. Par le théorème 3.1 d'approximation du rayon spectral, il existe une norme induite $||\cdot||_{A,\varepsilon}$ telle que:

$$||A||_{A,\varepsilon} \equiv \mu \leq \rho(A) + \varepsilon < 1 - 2\varepsilon + \varepsilon = 1 - \varepsilon < 1 \ . \tag{3.23}$$

Étant donné que la norme induite est une norme matricielle, on a

$$||A^k||_{A,\varepsilon} \leq ||A||_{A,\varepsilon} ||A^{k-1}||_{A,\varepsilon} \leq \ldots \leq ||A||^k_{A,\varepsilon} = \mu^k \to 0 (\text{quand} \ \ k \to \infty) \ . \tag{3.24}$$

Notons que dans l'espace $M_n(\mathbb{R})$, de dimension finie, toutes les normes sont équivalentes (génèrent la même topologie, alors la même convergence) alors, pour n'importe quelle norme $||A^k|| \to 0||$ $(k \to 0)$ $\Rightarrow A^k \to 0$.

Pour la condition nécessaire ($\Leftarrow$), supposons $A^k \to 0$ $(k \to 0)$. On considère λ une valeur propre de A et $\boldsymbol{x} \neq 0$ un vecteur propre associé. Alors

$$A^k \boldsymbol{x} = \lambda^k \boldsymbol{x} \ . \tag{3.25}$$

Si $A^k \to 0$, on a $A^k \boldsymbol{x} \to 0$ et alors $\lambda^k \boldsymbol{x}$. Vu que $\boldsymbol{x} \neq 0$, $\lambda^k \to 0$, ce qui est vrai si et seulement si $|\lambda| < 1$. Ceci est vrai pour n'importe quel valeur propre λ. Alors

$$\rho(A) = \max\{|\lambda|, \lambda \text{ valeur propre de } A\} < 1 \ . \tag{3.26}$$

□

Remarque 3.3 (Convergence de suites de vecteurs). Du corollaire précédent on déduit que la suite $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ définie par $\boldsymbol{x}^{(0)} = \boldsymbol{x}_o$ et $\boldsymbol{x}^{(k+1)} = A\boldsymbol{x}^{(k)}$, on a

$$\boldsymbol{x}^k \to 0, \forall \boldsymbol{x}^{(0)} = \boldsymbol{x}_o \Leftrightarrow \rho(A) < 1 \ . \tag{3.27}$$

Cette caractérisation sera utilisé dans l'étude des méthodes itératifs pour la résolution d'un problème linéaire.

Nous venons de voir que pour montrer la convergence d'une suite de vecteurs $\{\boldsymbol{x}^{(k)}\}_{k\in\mathbb{N}}$ donnée par $\boldsymbol{x}^{(k+1)} = A\boldsymbol{x}^{(k)}$, il faut contrôler la valeur du rayon spectrale. Le calcul de $\rho(A)$ peut être difficile mais, après le théorème 3.1, il suffit de trouver une norme induite $||\cdot||$ pour laquelle $||A|| < 1$. Nous allons donner maintenant une condition suffisante plus générale, en montrant qu'on peut relaxer la condition sur le caractère *induit* de la norme matricielle employée. Pour montrer ça nous donons d'abord la caractérisation suivante du rayon spectrale (valable aussi en dimension infinie).

Proposition 3.3 (Caractérisation du rayon spectrale)**.** *On munit $M_n(\mathbb{R})$ d'une norme* générale*, notée $||\cdot||$. Soit $A \in M_n(\mathbb{R})$. Alors*

$$\rho(A) = \lim_{k\to\infty} ||A^k||^{\frac{1}{k}} \ . \tag{3.28}$$

Preuve: Exercice.

Remarque 3.4. Noter que la norme dans la proposition est une norme générale, pas nécessairement matricielle.

Nous pouvons maintenant prouver le résultat annoncé.

Corollaire 3.2 (Comparaison rayon spectrale et norme matricielle)**.** . *On munit $M_n(\mathbb{R})$ d'une norme matricielle générale, notée $||\cdot||$. Soit $A \in M_n(\mathbb{R})$. Alors*

$$\rho(A) \leq ||A|| \tag{3.29}$$

Preuve: Si $||\cdot||$ est matricielle, on a $||A^k|| \leq ||A||^k$ et donc, en utilisant la caractérisation (3.28) du rayon spectral, on a

$$\rho(A) = \lim_{k\to\infty} ||A^k||^{\frac{1}{k}} \leq \lim_{k\to\infty} \left(||A||^k\right)^{\frac{1}{k}} = ||A|| \tag{3.30}$$

. □

Remarque 3.5. C'est fondamental que la norme $||\cdot||$ soit matricielle. Ça ne marche pas pour une norme générique non-matricielle (exercice !).

Nous finissons cette section en énonçant un résultat sur les matrices "perturbation" de l'identité.

Théorème 3.2 (Matrices de la forme $I_n + A$)**.** *Soit $||\cdot||$ une norme matricielle induite sur $M_n(\mathbb{R})$.*

i) Étant donnés $I_n \in M_n(\mathbb{R})$ la matrice identité et $A \in M_n(\mathbb{R})$ telle que $||A|| < 1$, on a que la matrice $I_n + A$ est inversible et

$$||(I_n + A)^{-1}|| \leq \frac{1}{1 - ||A||} \ . \tag{3.31}$$

ii) Si un matrice de la forme $I_n + A \in M_n(\mathbb{R})$ est singulière, alors $||A|| \geq 1$ pour tout norme matricielle.

Preuve:

i) Exercice.

ii) Si $I_n + A \in M_n(\mathbb{R})$ est singulière, alors $\exists \boldsymbol{x} \neq 0$ tel que

$$(I_n + A)\,\boldsymbol{x} = 0 \Leftrightarrow A\boldsymbol{x} = -\boldsymbol{x} \Rightarrow \lambda = -1 \text{ est un valeur propre.} \tag{3.32}$$

Alors $\rho(A) \geq 1$. Si on utilise (3.29), étant donnée que $||\cdot||$ est matricielle, on a

$$||A|| \geq \rho(A) \geq 1\ . \tag{3.33}$$

□

3.3 Conditionnement d'une matrice

Quand on étudie le problème

$$A\boldsymbol{x} = \boldsymbol{b}\ , \tag{3.34}$$

en particulier sa résolution numérique par un ordinateur, un problème crucial à considérer est posé par la question suivante : quel est le changement de $\boldsymbol{x}$ si on fait un petit erreur sur A ou $\boldsymbol{b}$ par rapport à ses vrais valeurs ?

Si un petit changement δA dans A ou un petit erreur $\delta\boldsymbol{b}$ entraînent un grand changement $\delta\boldsymbol{x}$ dans la valeur de $\delta\boldsymbol{x}$, alors le problème est très mal posé pour être résolu par un ordinateur, étant donné que la résolution dans A et $\boldsymbol{b}$ par l'ordinateur sera toujours finie.

D'une manière plus précise, on veut estimer l'erreur $\delta\boldsymbol{x}$ en $\boldsymbol{x}$, si on fait des erreurs δA en A et $\delta\boldsymbol{b}$ en $\boldsymbol{b}$, en résolvant l'équation

$$(A + \delta A)\,(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b}\ . \tag{3.35}$$

Si δA est suffisamment petite, la matrice $A + \delta A$ est inversible et le problème posé par (3.51) a du sens. D'une manière plus précise, si $\left|\left|A^{-1}\delta A\right|\right| < 1$, le théorème 3.2 garantie que la matrice $I_n + A^{-1}\delta A$ est inversible, donc $A + \delta A = A\left(I_n + A^{-1}\delta A\right)$ est aussi inversible. Dans ces conditions on peut estimer $\delta\boldsymbol{x}$ dans (3.51) en fonction de δA et $\delta\boldsymbol{b}$.

3.3.1 Conditionnement et ses propriétés élémentaires

Définition 3.5. Soit $\mathbb{R}^n$ muni d'une norme $||\cdot||$ et $M_n(\mathbb{R})$ muni de la norme matricielle induite. Étant donnée $A \in M_n(\mathbb{R})$ inversible, on appelle conditionnement de A par rapport à la norme $||\cdot||$ le nombre réel positif

$$\text{cond}(A) = ||A||\ ||A^{-1}||\ . \tag{3.36}$$

Proposition 3.4 (propriétés générales du conditionnement)**.** *. Soit $\mathbb{R}^n$ muni d'une norme $||\cdot||$ et $M_n(\mathbb{R})$ muni de la norme matricielle induite, et $A, B \in M_n(\mathbb{R})$ inversibles et $\alpha \in \mathbb{R}^*$. On a :*

i) $\text{cond}(A) \geq 1$.

ii) $\text{cond}(\alpha A) = \text{cond}(A)$.

iii) $\text{cond}(AB) \leq \text{cond}(A)\ \text{cond}(B)$.

Preuve:

i) On utilise que $||\cdot||$ est matricielle. Alors

$$||I_n|| = ||AA^{-1}|| \leq ||A||\ ||A^{-1}|| = \text{cond}(A)\ . \tag{3.37}$$

D'autre part

$$||I_n|| = \sup_{||\boldsymbol{x}||=1} \{||I_n \boldsymbol{x}||\} = 1 \tag{3.38}$$

Alors $1 \leq \text{cond}(A)$ (noter qu'on a utilisé que la norme est induite, et pas seulement matricielle).

ii) En utilisant la définition (3.36)

$$\text{cond}(\alpha A) = ||\alpha A||\ ||(\alpha A)^{-1}|| = |\alpha| ||A|| \cdot \frac{1}{|\alpha|} ||A^{-1}|| = ||A||\ ||A^{-1}|| = \text{cond}(A) \tag{3.39}$$

iii) Étant donné que A et B sont inversibles, alors $A\ B$ est aussi inversible et, utilisant le fait que $||\cdot||$ est matricielle, on a

$$\begin{aligned} \text{cond}(AB) &= ||AB||\ ||(AB)^{-1}|| = ||AB||\ ||B^{-1}A^{-1}|| \\ &\leq ||A||\ ||B||\ ||B^{-1}||\ ||A^{-1}|| = \text{cond}(A)\ \text{cond}(B)\ . \end{aligned} \tag{3.40}$$

□

3.3.2 Majoration des erreurs

Nous allons étudier le cas d'une perturbation seulement par δA et le cas d'une perturbation seulement par $\delta \boldsymbol{b}$. Le cas plus général d'une perturbation simultanée par δA et $\delta \boldsymbol{b}$ suit une stratégie pareille, mais avec une expression plus compliquée.

Proposition 3.5 (majoration de l'erreur relative par rapport à une erreur dans le deuxième membre)**.** *Soit $A \in M_n(\mathbb{R})$ inversible et $\boldsymbol{b} \in \mathbb{R}^n$, avec $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $||\cdot||$ et $M_n(\mathbb{R})$ de la norme matricielle induite associée. Étant donné $\delta \boldsymbol{b} \in \mathbb{R}^n$, si $\boldsymbol{x}$ est solution de*

$$A\boldsymbol{x} = \boldsymbol{b} \quad , \quad \boldsymbol{b} \neq 0\ , \tag{3.41}$$

et $\boldsymbol{x} + \delta \boldsymbol{x}$ est solution de

$$A\,(\boldsymbol{x} + \delta \boldsymbol{x}) = \boldsymbol{b} + \delta \boldsymbol{b}\ , \tag{3.42}$$

on a la majoration suivante de l'erreur relative en $\boldsymbol{x}$

$$\frac{||\delta \boldsymbol{x}||}{||\boldsymbol{x}||} \leq \text{cond}(A) \frac{||\delta \boldsymbol{b}||}{||\boldsymbol{b}||} \tag{3.43}$$

Preuve: On considère

$$\begin{aligned} A\boldsymbol{x} &= \boldsymbol{b} \\ A\left(\boldsymbol{x}+\delta\boldsymbol{x}\right) &= \boldsymbol{b}+\delta\boldsymbol{b}\ . \end{aligned} \tag{3.44}$$

Si on retranche la première à la deuxième, on a :

$$A\delta\boldsymbol{x} = \delta\boldsymbol{b} \Leftrightarrow \delta\boldsymbol{x} = A^{-1}\delta\boldsymbol{b}\ , \tag{3.45}$$

donc

$$||\delta\boldsymbol{x}|| \leq ||A^{-1}||||\delta\boldsymbol{b}||\ . \tag{3.46}$$

D'autre part, de $\boldsymbol{b} = A\boldsymbol{x}$, on a

$$||b|| = ||A\boldsymbol{x}|| \leq ||A||||\boldsymbol{x}||\ . \tag{3.47}$$

Si on utilise que $\boldsymbol{b} \neq 0$, on a $\boldsymbol{x} \neq 0$, donc on peut écrire

$$\frac{1}{||\boldsymbol{x}||} \leq \frac{||A||}{||\boldsymbol{b}||}\ . \tag{3.48}$$

Si on multiplie, membre à membre, cette expression par celle dans (3.46) on obtient

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}||} \leq \underbrace{||A||\ ||A^{-1}||}_{\mathrm{cond}(A)}\frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||} = \mathrm{cond}(A)\frac{||\delta\boldsymbol{b}||}{||\boldsymbol{b}||}\ . \tag{3.49}$$

□

Remarque 3.6. L'inégalité (3.52) est optimale, c'est à dire, il y a des choix que la rendre une égalité.

Proposition 3.6 (majoration de l'erreur relative par rapport à une erreur sur la matrice)**.** *Soit $A \in M_n(\mathbb{R})$ inversible et $\boldsymbol{b} \in \mathbb{R}^n$, avec $\boldsymbol{b} \neq 0$. On munit $\mathbb{R}^n$ d'une norme $||\cdot||$ et $M_n(\mathbb{R})$ de la norme matricielle induite associée. Étant donné $\delta A \in M_n(\mathbb{R})$, on suppose que $A+\delta A$ est inversible. Si $\boldsymbol{x}$ est solution de*

$$A\boldsymbol{x} = \boldsymbol{b}\quad ,\quad \boldsymbol{b} \neq 0\ , \tag{3.50}$$

et $\boldsymbol{x}+\delta\boldsymbol{x}$ est solution de

$$\left(A+\delta A\right)\left(\boldsymbol{x}+\delta\boldsymbol{x}\right) = \boldsymbol{b}\ , \tag{3.51}$$

alors

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x}+\delta\boldsymbol{x}||} \leq \mathrm{cond}(A)\frac{||\delta A||}{||A||}\ . \tag{3.52}$$

Preuve: On part de

$$(A + \delta A)(\boldsymbol{x} + \delta \boldsymbol{x}) = \boldsymbol{b} \ . \tag{3.53}$$

En le développant

$$A\boldsymbol{x} + A\delta\boldsymbol{x} + \delta A\boldsymbol{x} + \delta A\delta\boldsymbol{x} = \boldsymbol{b} \ . \tag{3.54}$$

Si on retranche $A\boldsymbol{x} = \boldsymbol{b}$ on trouve

$$A\delta\boldsymbol{x} + \delta A\boldsymbol{x} + \delta A\delta\boldsymbol{x} = 0 \ , \tag{3.55}$$

c'est à dire

$$\begin{aligned} A\delta\boldsymbol{x} &= -A\delta A(\boldsymbol{x} + \delta\boldsymbol{x}) \\ \delta\boldsymbol{x} &= A^{-1}\delta A(\boldsymbol{x} + \delta\boldsymbol{x}) \ . \end{aligned} \tag{3.56}$$

Alors,

$$||\delta\boldsymbol{x}|| \leq ||A^{-1}||\ ||\delta A||\ ||\boldsymbol{x} + \delta\boldsymbol{x}|| \ . \tag{3.57}$$

Si on utilise maintenant que $\boldsymbol{b} \neq 0$ et $A + \delta A$ est inversible, on conclut $\boldsymbol{x} + \delta\boldsymbol{x} \neq 0$ et on diviser par $||\boldsymbol{x} + \delta\boldsymbol{x}||$ pour trouver

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x} + \delta\boldsymbol{x}||} \leq ||A^{-1}||\ ||\delta A|| \ . \tag{3.58}$$

Si on multiplie maintenant le deuxième par $\dfrac{||A||}{||A||} = 1$

$$\frac{||\delta\boldsymbol{x}||}{||\boldsymbol{x} + \delta\boldsymbol{x}||} \leq \underbrace{||A^{-1}||\ ||A||}_{\mathrm{cond}(A)} \frac{||\delta A||}{||A||} \leq \mathrm{cond}(A)\frac{||\delta A||}{||A||} \ . \tag{3.59}$$

□

Chapitre 4

Méthodes itératives

Chapitre 5

Méthodes de descente

Part II

Interpolation

Chapitre 6

Interpolation polynomiale

Part III

Annexes

Appendix A

Espaces linéaires

A.1 Rappel sur les espaces linéaires

Définition A.1 (*espace linéaire ou vectoriel).* Un espace linéaire sur un corps $\mathbb{K}$ (ici $\mathbb{K} = \mathbb{R}$ ou $\mathbb{K} = \mathbb{C}$) est un ensemble non vide V, dont les éléments sont appelés vecteurs et où on a deux opérations définies l'*addition* et la *multiplication par un scalaire*, avec les propriétés:

1. L'addition est commutative et associative.
2. Il existe un élément $\mathbf{0}$ (*vecteur zéro* ou *nul*) tel que $\mathbf{v} + \mathbf{0} = \mathbf{v}$, $\forall \mathbf{v} \in V$.
3. $0 \cdot \mathbf{v} = 0, 1 \cdot \mathbf{v} = \mathbf{v}$, $\forall \mathbf{v} \in V$, où 0 et 1 sont le zéro et l'unité de $\mathbb{K}$.
4. Pour chaque $\mathbf{v} \in V$ il exite un unique opposé, $-\mathbf{v} \in V$, tel que $\mathbf{v} - \mathbf{v} = \mathbf{0}$.
5. Proprieté distributive:

$$\alpha(\mathbf{v} + \mathbf{w}) = \alpha\mathbf{v} + \alpha\mathbf{w}, \quad \forall \alpha \in \mathbb{K}, \forall \mathbf{v}, \mathbf{w} \in V \tag{A.1}$$

$$(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}, \quad \forall \alpha, \beta \in \mathbf{K}, \forall \mathbf{v} \in V \tag{A.2}$$

6. Proprieté associative:

$$(\alpha\beta)\mathbf{v} = \alpha(\beta\mathbf{v}) \, \forall \alpha, \beta \in \mathbf{K}, \forall \mathbf{v} \in V \tag{A.3}$$

Exemple A.1. On donne plusieurs exemples d'espace linéaire.

i) $V = \mathbb{R}^n$, $V = \mathbb{R}^n$, $M_{m\times n}(\mathbb{R}) \equiv \mathbb{R}^{m\times n}$.

ii) $V = \mathbb{P}_n$, ensemble de polynômes $p_n(x) = \sum_{i=0}^{n} a_i x^i$

iii) $V = C^p([a, b])$, ensemble des fonctions réelles (complexes) continues jusqu'à la derivée p-ième.

Définition A.2 (espace linéaire ou vectoriel)**.** Un sous-espace non vide W de V es un sous-espace linéaire de V si W est un espace linéaire sur $\mathbb{K}$.

Exemple A.2. $V = \mathbb{P}_n$ est un sous espace linéaire de $V = C^\infty(\mathbb{R})$.

Définition A.3 (système de générateurs)**.** Étant donné un ensemble $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\}$, le sous-espace

$$W = \text{Lin}\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\} = \langle \boldsymbol{v}_1, \ldots, \boldsymbol{v}_p \rangle = \{\boldsymbol{w} = \alpha_1 \boldsymbol{v}_1 + \cdots + \alpha_1 \boldsymbol{v}_p, \text{ avec } \alpha_i \in\} \tag{A.4}$$

est un espace linéaire et $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p\}$ est appelé *système de générateurs* de W. Si $W_1, \ldots, W_m$ sont des sous-espaces linéaires de V, et $W_i \cap W_j = \{\boldsymbol{0}\} \forall i \neq j$, l'ensemble

$$S = \{\boldsymbol{w} \in V, \text{ tel que } \boldsymbol{w} = \boldsymbol{w}_1 + \ldots \boldsymbol{w}_m, \text{ avec } \boldsymbol{w}_i \in W_i\} \tag{A.5}$$

est un espace linéaire appelé *somme directe* et on note $S = W_1 \oplus \cdots \oplus W_m$.

Définition A.4 (Indépendance linéaire)**.** Un système de vecteurs $\{\mathbf{v}_1, \ldots, \mathbf{v}_m\}$ es dit *linéairement indépendant* si

$$\alpha_1 \mathbf{v}_1 + \cdots + \alpha_1 \mathbf{v}_m = 0 \ , \ \alpha_i \in \mathbb{K} \tag{A.6}$$

implique $\alpha_1 = \cdots = \alpha_m = 0$. Dans le cas contraire, le système est dit *linéairement dépendant*.

Définition A.5 (base)**.** Une base de V est un système linéairement indépendant de générateurs de V.

Propriétés (base)**.** Les bases d'un espace linéaire satisfont:

i) Si $B = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ (avec $n < \infty$) est une base de V, l'expression de $\boldsymbol{v} \in V$ $\boldsymbol{v} = v_1 \boldsymbol{e}_1 + \ldots v_n \boldsymbol{e}_n$ est appelée la décomposition de $\boldsymbol{v}$ en B. Les *composants* v_i sont uniques.

ii) Dans le cas où $n < \infty$, chaque système de vecteurs linéairement indépendants a un maximum de n vecteurs. En particulier, toutes les bases de V on le même nombre des vecteurs $n < \infty$, appeleé dimension de V: $\dim(V) = n$.

iii) Si pour chaque n il y a un système de vecteurs linéairement indépendant, l'espace linéaire V est dit de dimension infinie. Ici, on va se focaliser sur des V de dimension finie.

A.2 Application linéaire

Définition A.6 (application linéaire)**.** Étant donnés deux espaces linéaires V et W, une application $\mathcal{A} : V \to W$ est dite linéaire si elle respecte la structure de espace linéaire, ceci signifie que

$$\mathcal{A}(\alpha \mathbf{v} + \beta \mathbf{w}) = \alpha \mathcal{A}(\mathbf{v}) + \beta \mathcal{A}(\mathbf{w}) \quad , \quad \mathbf{v}, \mathbf{w} \in V, \alpha, \beta \in \mathbb{K}. \tag{A.7}$$

A.3 Produit scalaire

Définition (produit scalaire). Un produit scalaire est une application bilinéaire $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$, symétrique et positivement définie. Cela veut dire

i) Application bilinéaire:

$$\begin{aligned} \langle \alpha \mathbf{v} + \beta \mathbf{w}, \mathbf{x} \rangle &= \alpha \langle \mathbf{v}, \mathbf{x} \rangle + \beta \langle \mathbf{w}, \mathbf{x} \rangle \\ \langle \mathbf{x}, \alpha \mathbf{v} \rangle + \beta \mathbf{w} \rangle &= \alpha \langle \mathbf{x}, \mathbf{v} \rangle + \beta \langle \mathbf{x}, \mathbf{w} \rangle \quad , \quad \mathbf{v}, \mathbf{w}, \mathbf{x} \in V, \alpha, \beta \in \mathbb{K} \end{aligned} \tag{A.8}$$

ii) Symétrique:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \langle \mathbf{w}, \mathbf{v} \rangle \quad , \quad \mathbf{v}, \mathbf{w} \in V \tag{A.9}$$

iii) Définie positive:

$$\langle \mathbf{v}, \mathbf{v} \rangle \geq 0 \ \text{ et } \ \langle \mathbf{v}, \mathbf{v} \rangle = 0 \text{ sii } \mathbf{v} = \mathbf{0}. \tag{A.10}$$

Définition (produit scalaire hermitien). Un produit scalaire est une application $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{C}$ qui satisfait:

i) Linéaire dans la deuxième variable et anti-linéaire dans la première variable:

$$\begin{aligned} \langle \alpha \mathbf{v} + \beta \mathbf{w}, \mathbf{x} \rangle &= \bar{\alpha} \langle \mathbf{v}, \mathbf{x} \rangle + \bar{\beta} \langle \mathbf{w}, \mathbf{x} \rangle \\ \langle \mathbf{x}, \alpha \mathbf{v} + \beta \mathbf{w} \rangle &= \alpha \langle \mathbf{x}, \mathbf{v} \rangle + \beta \langle \mathbf{x}, \mathbf{w} \rangle \quad , \quad \mathbf{v}, \mathbf{w}, \mathbf{x} \in V, \alpha, \beta \in \mathbb{K} \end{aligned} \tag{A.11}$$

ii) Symétrique hermitienne:

$$\langle \mathbf{v}, \mathbf{w} \rangle = \overline{\langle \mathbf{w}, \mathbf{v} \rangle} \quad , \quad \mathbf{v}, \mathbf{w} \in V \tag{A.12}$$

iii) Définie positive:

$$\langle \mathbf{v}, \mathbf{v} \rangle \geq 0 \ \text{ et } \ \langle \mathbf{v}, \mathbf{v} \rangle = 0 \text{ sii } \mathbf{v} = \mathbf{0}. \tag{A.13}$$

Appendix B

Matrices

B.1 Définitions

Définition B.1 (matrice)**.** Étant donnés m et n deux entiers positifs, on appelle matrice A à m lignes et n colonnes l'ensemble ordonné de $m \cdot n$ éléments de $\mathbb{K}$

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{m1} & \cdots & a_{mn} \end{pmatrix} \tag{B.1}$$

On note par $\mathrm{M}_{m,n}(\mathbb{K})$ l'ensemble de matrices de m lignes et n colonnes, avec des coefficients dans $\mathbb{K}$.

Remarques :

i) Notation: $A = (A_{ij})$ avec $i = 1, \ldots, m$,$j = 1, \ldots, n$.

ii) Matrice carrée: $m = n$. Diagonale principale: $\mathrm{diag}(a_{11}, \ldots, a_{nn})$.

iii) Une matrice avec une ligne est un *vecteur ligne* et une matrice avec une seule colonne est un *vecteur colonne*

$$\boldsymbol{v} = \begin{pmatrix} v^1 \\ \vdots \\ v^n \end{pmatrix} \quad , \quad \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) \quad , \quad \boldsymbol{\alpha}^t = \begin{pmatrix} \alpha_1 & \ldots & \alpha_n \end{pmatrix} \tag{B.2}$$

iv)

$$\text{Ligne-i} : \boldsymbol{\ell}_i = (A_{i1}, \ldots, A_{in}) \quad , \quad \text{Colonne-j} : \boldsymbol{c}_j = \begin{pmatrix} A_{1j} \\ \vdots \\ A_{mj} \end{pmatrix} \tag{B.3}$$

iv) $M_{m,n}(\mathbb{K}) = \{(A_{ij}),\ a_{ij} \in \mathbb{K};\ i = 1, \ldots, m, j = 1, \ldots, n\}$

B.1.1 Opérations sur les matrices

i) Somme. Pour $A \in M_{m\times n}(\mathbb{K})$, $B \in M_{m\times n}(\mathbb{K})$:

$$C = A + B \quad , \quad (c_{ij}) = (a_{ij}) + (b_{ij}) \tag{B.4}$$

ii) Produit par un scalaire. Pour $A \in M_{m\times n}(\mathbb{K})$ et $\lambda \in \mathbb{K}$:

$$C = \lambda A \quad , \quad (c_{ij}) = (\lambda a_{ij}) \tag{B.5}$$

iii) Produit de matrices. Pour $A \in M_{m\times n}(\mathbb{K})$, $B \in M_{n\times p}(\mathbb{K})$, on définit la matrice $A \in M_{m\times p}(\mathbb{K})$

$$C = A \cdot B \quad , \quad (c_{ij}) = (\sum_{k=1}^{n} a_{ik} b_{kj}) \tag{B.6}$$

Exercice B.1. Montrer que le produit de matrices est associatif est distributif par rapport à l'addition.

Remarque B.1. :

i) Pour des matrices carrées d'ordre n, la matrice $\mathrm{I}_n = (\delta_{ij})$ est la *matrice unité*, ce qui veut dire $A \cdot \mathrm{I}_n = \mathrm{I}_n \cdot A = A$.

ii) Matrice diagonale: $D = (d_{ii}\delta_{ij})$.

iii) $A^p = \underbrace{A \cdot \ldots \cdot A}_{p fois}$

iv) **Définition** (matrice transposée). Étant donnée une matrice $A \in M_{m\times n}(\mathbb{R})$, on appelle *matrice transposée* de A la matrice $A^t \in M_{n\times m}(\mathbb{R})$ telle que

$$({a^t}_{ij}) = (a_{ji}) \tag{B.7}$$

Proprietés:

$$(A^t)^t = A \tag{B.8}$$
$$(A \cdot B)^t = B^t \cdot A^t \tag{B.9}$$

Définition (matrice conjuguée transposée). Étant donnée une matrice $A \in M_{m\times n}(\mathbb{C})$, on appelle *matrice conjuguée transposée*de A (ou *adjointe*, voir ci-après) la matrice $A^\dagger \in M_{n\times m}(\mathbb{C})$ telle que

$$({a^\dagger}_{ij}) = (\bar{a}_{ji}) \tag{B.10}$$

où $\bar{z}$ le complexe conjugué de $z \in \mathbb{C}$ (l'adjointe de A est souvent notée A^*.) Proprietés:

$$(A^\dagger)^\dagger = A \tag{B.11}$$
$$(A \cdot B)^\dagger = B^\dagger \cdot A^\dagger \tag{B.12}$$

v) Un vecteur $\boldsymbol{v} \in \mathbb{K}^n$ sera consideré comme un vecteur colonne. On peut définir un produit scalaire naturel en $\mathbb{R}^n$[1]

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v} \cdot \boldsymbol{w} := \boldsymbol{v}^t \boldsymbol{w} = \sum_i^n v^i w^i \ . \tag{B.13}$$

Un vecteur ligne sera considéré comme un élément de l'espace dual $\alpha \in (\mathbb{K}^n)^*$, avec la relation de dualité

$$\boldsymbol{\alpha}(\boldsymbol{v}) = \boldsymbol{\alpha}\boldsymbol{v} = \boldsymbol{\alpha}^t \cdot \boldsymbol{v} \tag{B.14}$$

vi) L'expression (B.6) peut être écrite $c_{ij} = \boldsymbol{\ell}_i \boldsymbol{c}_j$.

vii) Les colonnes de AB sont des combinaison linéaires des colonnes de A.
Les lignes de AB sont de combinaisons linéaires de lignes de B.

$$\begin{aligned} \boldsymbol{\ell}_i(AB) &= \sum_{k=1}^n a_{ik} \boldsymbol{\ell}_k(B) \\ \boldsymbol{c}_j(AB) &= \sum_{k=1}^n b_{kj} \boldsymbol{c}_k(A) \end{aligned}$$

Rélévant pour la méthode d'élimination de Gauss.

Exemple B.1.

B.1.2 Rang et noyau d'une matrice

Définition B.2 (rang)**.** Étant donnée $A \in \mathrm{M}_{m,n}$, l'image de A (notée $\mathrm{Im}(A)$) est donnée par $\mathrm{Im}(A) = \{y \in \mathbb{K}^m; \boldsymbol{y} = A\boldsymbol{x}, \boldsymbol{x} \in \mathbb{K}^n\}$. Le rang de A est défini par $\mathrm{rang}(A) = \dim(\mathrm{Im}(A))$.

Définition B.3 (Noyau)**.** Étant donnée $A \in \mathrm{M}_{m,n}$, le noyau de A (notée par $\mathrm{Ker}(A)$) est donné par $\mathrm{Ker}(A) = \{\boldsymbol{x} \in \mathbb{K}^n; A\boldsymbol{x} = 0\}$.

Proposition B.1. *Étant donnée $A \in \mathrm{M}_{m,n}$, les relations suivantes sont satisfaites*

i) $\mathrm{rang}(A) = \mathrm{rang}(A^t)$.

ii) $\mathrm{rang}(A) + \mathrm{Ker}(A) = n$.

B.1.3 Inverse d'une matrice

Définition B.4 (matrice inverse)**.** Una matrice carrée d'ordre n est *inversible* (ou régulière ou non-singulière) s'il existe une matrice B telle que: $A \cdot B = B \cdot A = I_n$. On note $B = A^{-1}$.

Une matrice qui n'est pas inversible est dite *singulière*. Propriétés:

$$(A^{-1})^{-1} = A \tag{B.15}$$

$$(A \cdot B)^{-1} = B^{-1} \cdot A^{-1} \tag{B.16}$$

[1] Cas complexe: produit scalaire hermitien

B.2 Applications linéaires et matrice d'une application linéaire

Définition B.5 (matrice d'une application unitaire dans bases données)**.** Soit une application linéaire $\mathcal{A} : V \to W$, avec $\dim(V) = n$ et $\dim(W) = m$. Soient $B = (\mathbf{e}_1, \ldots, \mathbf{e}_n)$ et $B' = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$ des bases de V et W. On introduit la matrice de $\mathcal{A}$ dans les bases B et B', $A = (a_{ij})$

$$\mathcal{A}\mathbf{e}_j = \sum_i A_{ij}\mathbf{f}_i \tag{B.17}$$

Alors, si les composantes de $\mathbf{v} \in V$ dans B sont $\mathbf{v} = v^i\mathbf{e}_i$, les composantes de $\mathbf{w} = \mathcal{A}\mathbf{v}$ dans B' sont ($\mathbf{w} = \sum_i w^i\mathbf{f}_i$)

$$\mathcal{A}\mathbf{v} = \sum_i \left(\sum_j a_{ij}v^j \right) \mathbf{f}_i \tag{B.18}$$

$$w^i = \sum_j a_{ij}v^j \tag{B.19}$$

$$\begin{pmatrix} w_1 \\ \vdots \\ w_m \end{pmatrix} = A \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \tag{B.20}$$

La matrice A ainsi construite est nommée matrice de $\mathcal{A}$ dans les bases B et B' et on peut la noter $A = M(\mathcal{A}, B, B')$. En particulier, l'action de la matrice $M(\mathcal{A}, B, B')$ sur le vecteur dans $\mathbf{R}^n$ donnée par les composantes de $\mathbf{v}$ en B, renvoient le vecteur w dans $\mathbf{R}^m$ avec les composantes de $\mathbf{w} = \mathcal{A}\mathbf{v}$ dans B'

$$\begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \mapsto \begin{pmatrix} w_1 \\ \vdots \\ w_n \end{pmatrix} = A \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = M(\mathcal{A}, B, B') \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \tag{B.21}$$

En particulier, si $V = W$ et $\mathcal{A} = \mathrm{I}_n$, la matrice $M(\mathrm{I}_n, B, B')$ est la matrice de changement de base de B à B' dans le sens (passif) qu'elle prend les composantes en B' et donne les composantes en B'. Ainsi, si $\mathbf{v} = \sum_i v^i\mathbf{e}_i = \sum_i (v')^i\mathbf{e}'_i$ et si on note $M(\mathrm{I}_n, B, B') = M(B \to B')$, on a

$$\begin{pmatrix} v'_1 \\ \vdots \\ v'_n \end{pmatrix} = M(\mathrm{I}_n, B, B') \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = M(B \to B') \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} \tag{B.22}$$

où les colonnes de $M(\mathrm{I}_n, B, B') = M(B \to B')$ sont données par les vecteurs colonnes de e_i exprimées dans la base $B' = (\mathbf{f}_1, \ldots, \mathbf{f}_n)$

$$M(B \to B') = \begin{pmatrix} \mathbf{e}_1 & \cdots & \mathbf{e}_n \end{pmatrix} \tag{B.23}$$

B.2.1 Matrices semblables

Définition B.6 (matrices semblables)**.** Deux matrices carrées $A, B \in \mathrm{M}_n(\mathbb{K})$ sont semblables ssi $\exists P$ inversible, telle que

$$A = P \cdot B \cdot P^{-1} \tag{B.24}$$

Proposition B.2. *Deux matrices semblables correspondent à l'expression matricielle d'un même endomorphisme dans deux bases différentes.*

Remarque B.2. La notion de matrice semblable est un cas particulier dans le cas de matrices rectangulaires *équivalentes*, qui correspondent à l'expression matricielle d'une même application linéaire dans des bases différentes.

B.3 Matrices et géométrie

En utilisant le produit de matrices, l'application suivante définit un produit scalaire en $\mathbb{C}^n$.

$$\begin{aligned} \langle \cdot, \cdot \rangle : \mathbb{C}^n \times \mathbb{R}^n &\rightarrow \mathbb{R} && \text{(B.25)} \\ (\mathbf{v}, \mathbf{v}) &\mapsto \langle \mathbf{v}, \mathbf{v} \rangle = \mathbf{v}^t \cdot \mathbf{v} && \text{(B.26)} \end{aligned}$$

Définition (matrice symétrique). Une matrice A est dite symétrique si

$$A = A^t \quad \Leftrightarrow a_{ij} = a_{ji} \tag{B.27}$$

Définition (matrice hermitienne). Une matrice A est dite hermitienne si

$$A = A^* \quad \Leftrightarrow a_{ij} = \bar{a}_{ji} \tag{B.28}$$

Définition (matrice orthogonale). Une matrice U est dite orthogonale si

$$OO^t = \mathrm{I} \tag{B.29}$$

Définition (matrice unitaire). Une matrice U est dite unitaire si

$$UU^* = \mathrm{I} \tag{B.30}$$

Définition (matrice normale). Une matrice A est dite normale si

$$AA^* = A^*A \tag{B.31}$$

Appendix C

Factorisation de matrices

Théorème C.1 (Factorisation unitaire d'une matrice: Décomposition de Schur)**.** *Toute matrice carrée (réelle ou complexes) peut s'écrire*

$$A = UTU^* \ , \tag{C.1}$$

avec U une matrice unitaire $U^{-1} = U^*$ et T une matrice triangulaire supérieure.

Appendix D

Diagonalisation

D.1 Notion de diagonalisation

On commence par introduire la notion de valeur propre.

Définition D.1. (*valeur propres*). Soit $A \in \mathcal{M}_n(\mathbb{K})$. On appelle valeur propre de A tout $\lambda \in \mathbb{C}$ tel qu'il existe $\boldsymbol{x} \in \mathbb{C}^n$, $\boldsymbol{x} \neq 0$, qui satisfait $A\boldsymbol{x} = \lambda\boldsymbol{x}$. L'élément $\boldsymbol{x}$ est appelé vecteur propre ("à droite") de A associé à λ.

Ceci nous amène à la notion de diagonalisation.

Définition D.2. (*diagonalisation/réduction des endomorphismes*). Soit $A \in \mathcal{M}_n(\mathbb{K})$. On dit que A est diagonalisable dans $\mathbb{K}$ s'il existe une base $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ avec $v_i \in \mathbb{K}^n$ et des valeurs $\lambda_1, \ldots, \lambda_n \in \mathbb{K}$ telles que $A\lambda_i = \lambda_i v_i$.

Lemme D.1. *Une matrice diagonalisable est semblable à une matrice diagonale D donnée par*

$$D = \begin{pmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix} , \tag{D.1}$$

c.à.d., s'il existe une matrice inversible P, telle que $A = PDP^{-1}$.

Lemme D.2. *Les valeurs propres sont les racines du polynôme caractéristique $p_A(\lambda) = \det(A - \lambda\mathrm{I})$.*

Preuve (idée) : Le déterminant de $A - \lambda\mathrm{I}$ s'annule si son noyau n'est pas trivial. On montre que $x \in \mathrm{Ker}(A - \lambda\mathrm{I})$ ssi λ est une racine de $p_A(\lambda)$. □

Dans le cas $\mathbb{K} = \mathbb{C}$, le théorème fondamental de l'algèbre nous garantit l'existence de n solutions complexes pour le polynôme caracteristique de $A \in \mathcal{M}_n(\mathbb{C})$. La proposition suivante caractérise la possibilité de diagonaliser une matrice.

Lemme D.3. *Étant donné $A \in M_n(\mathbb{C})$, une condition nécessaire et suffisante pour pouvoir diagonaliser A est que la dimension géometrique de chaque valeur propre λ, c.à.d.* $\dim E_\lambda$*, avec $E_\lambda = A - \lambda\mathrm{I}$ le sous-espace propre de λ, coïncide avec sa dimension algébrique (multiplicité comme racine de $p_A(\lambda)$.*

Exemple D.1. Nous considérons les exemples suivants:

i) La matrice σ_x

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \in \mathcal{M}_2(\mathbb{R}) \ , \tag{D.2}$$

est réelle et symétrique. Ses valeurs propres sont $\lambda_1 = 1$ et $\lambda_2 = -1$. On peut vérifier que A est diagonalisable dans $\mathbb{R}$.

ii) La matrice A

$$A = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \in \mathcal{M}_2(\mathbb{R}) \ , \tag{D.3}$$

n'est pas diagonalisable en $\mathbb{R}$. En effet, ses valeurs propres sont $\lambda_1 = i$ et $\lambda_2 = -i$. Elle est diagonalisable dans $\mathbb{C}$.

iii) La matrice σ_y

$$\sigma_y = \begin{pmatrix} 0 & i \\ -i & 0 \end{pmatrix} \in \mathcal{M}_2(\mathbb{C}) \ , \tag{D.4}$$

est une matrice hermitienne. Ses valeurs propres sont $\lambda_1 = 1$ et $\lambda_2 = -1$, alors la matrice est diagonalisable dans C. Noter que λ_1 et λ_2 sont réels. Néanmoins, les vecteurs propres sont complexes.

iv) La matrice N

$$N = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \in \mathcal{M}_2(\mathbb{R}) \ , \tag{D.5}$$

n'est pas diagonalisable dans $\mathbb{C}$. En effet, sa seule valeur propre est $\lambda = 0$. Si elle était diagonalisable, elle serait la matrice nulle, et ce n'est pas le cas. De la même manière, la matrice J

$$N = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix} \in \mathcal{M}_2(\mathbb{C}) \ , \tag{D.6}$$

n'est pas diagonalisable dans $\mathbb{C}$. En effet, la seule valeur propre et $\lambda = 1$. Si J était diagonalisable, elle serait semblable à la matrice unité, ce qui amène à une contradiction.

v) On peut construire des exemples de manière systématique en utilisant la *forme normale de Jordan*, valable pour n'importe quelle matrice complexe carrée (preuve par récurrence).

D.2 Diagonalisation de matrices normales

Théorème D.1. *Une matrice A est normale si et seulement si il existe U une matrice unitaire telle que*

$$A = UDU^* \tag{D.7}$$

avec D la matrice diagonale formée des valeurs propres. C'est-à-dire, une matrice normale est diagonalisable et ses vecteurs propres sont orthonormés.

Dém: (voir Nantes)