

Chapitre 1

INTRODUCTION À L'ANALYSE NUMÉRIQUE

Chapitre 1

INTRODUCTION À L'ANALYSE NUMÉRIQUE

1.1 Introduction générale

La maîtrise des méthodes numériques est une nécessité pour la résolution des problèmes que l'ingénieur est amené à rencontrer au cours de sa vie professionnelle.

Les exemples énumérés ci-dessous à titre d'introduction illustrent cette nécessité.

Exemple 1

Soit Ω un ouvert de $\mathbb{R}^n$, ($n = 2$ ou 3) et soit l'équation aux dérivées partielles (EDP) suivante :

$$(1.1) \quad \begin{cases} \frac{\partial \theta}{\partial t}(t, x) - \Delta \theta(t, x) = f(t, x), & t > 0, x \in \Omega \\ \theta(t, x) = \theta_0(t, x), & t > 0, x \in \partial\Omega \quad (\text{conditions aux limites}) \\ \theta(0, x) = \theta_1(x), & x \in \Omega \quad (\text{conditions initiales}) \end{cases}$$

appelée *équation de la chaleur*, où :

$\partial\Omega$ est la frontière de Ω ;

f désigne une source de chaleur ;

$\theta(x, t)$ désigne la température en un point $x = (x_1, \dots, x_n) \in \Omega$ à l'instant t ;

θ_0 est une température donnée, imposée sur la frontière de Ω ;

$\theta_1(x)$, $x \in \Omega$, est la température à l'instant initial $t = 0$;

Δ est l'opérateur Laplacien ; $\Delta \theta = \sum_{i=1}^n \frac{\partial^2 \theta}{\partial x_i^2}$.

Mis à part le cas où la géométrie de Ω est très simple, f , θ_0 et θ_1 ont des expressions simples, on ne sait pas trouver la solution de l'équation aux dérivées partielles ci-dessus. ■

Exemple 2

Dans le cas stationnaire ($\partial\theta/\partial t = 0$), l'équation de la chaleur (1.1) s'écrit :

$$(1.2) \quad \begin{cases} -\Delta \theta(x) = f(x), & x \in \Omega \\ \theta(x) = \theta_0(x), & x \in \partial\Omega \end{cases}$$

Dans le cas général, on ne sait pas résoudre explicitement cette équation. ■

On doit trouver une approximation de la solution de l'équation (1.1) (ou (1.2)). Pour cela, on peut utiliser par exemple la méthode des *éléments finis* ou celle des *différences finies*. Toutes ces méthodes conduisent à des systèmes linéaires à résoudre (si les problèmes de départ sont linéaires). La méthode des différences finies sera présentée dans l'exemple 6 ci-dessous.

Exemple 3

On considère une plaque encastree dans le plan (x_1, x_2) , soumise à une distribution surfacique de forces f . Si on note u le déplacement transversal de la plaque, alors u est solution de l'équation :

$$\begin{cases} \Delta \Delta u = f & \text{dans } \Omega \\ u = \frac{\partial u}{\partial \nu} = 0 & \text{sur } \partial\Omega \end{cases}$$

où ν désigne la normale unitaire extérieure à Ω .

Sous certaines hypothèses, le problème de vibration de la plaque conduit à étudier l'équation

$$\Delta \Delta u = \lambda u$$

qu'on ne sait pas résoudre dans le cas général. La recherche de solutions approchées par la méthode des éléments finis ou celle des différences finies conduit à des problèmes de recherche de valeurs propres de matrices.

Exemple 4 : (Notion de conditionnement d'un système linéaire)

On considère le système linéaire suivant :

$$(1.3) \quad \begin{cases} 2x_1 + 6x_2 = 8 \\ 2x_1 + 6,00001x_2 = 8,00001 \end{cases}$$

dont la solution est donnée par $x_1 = 1$ et $x_2 = 1$.

Si par contre on considère le système linéaire (1.4) obtenu en "perturbant" très légèrement le système (1.3) :

$$(1.4) \quad \begin{cases} 2x_1 + 6x_2 = 8 \\ 2x_1 + 5,99999x_2 = 8,00002 \end{cases}$$

on trouve $x_1 = 10$ et $x_2 = -2$!

Si on note par A , b et x la matrice, le second membre et la solution correspondant au système linéaire (1.3), alors on a :

$$\begin{aligned} \|\delta x\| &= 9 \\ \|\delta A\| + \|\delta b\| &= 2 \cdot 10^{-5} \\ \|\delta x\| &= 4,5 \cdot 10^5 (\|\delta A\| + \|\delta b\|) \end{aligned}$$

où $\|\delta x\| = \max_{1 \leq i \leq 2} |(\delta x)_i|$ et $\|\delta A\| = \max_{1 \leq i \leq 2} \sum_{j=1}^2 |(\delta A)_{ij}|$.

D'où la nécessité d'introduire la notion de *conditionnement* d'une matrice, vis-à-vis de la résolution des systèmes linéaires. Cela fera l'objet de la section 1.5 ci-dessous ; on verra que la matrice A du système linéaire (1.3) est en fait "*mal conditionnée*". Pour une matrice A "*bien conditionnée*" vis-à-vis de la résolution des systèmes linéaires, une "petite" perturbation de A ou du second membre b engendre une "petite perturbation" de la solution du système linéaire $Ax = b$.

Exemple 5 : (Notion de conditionnement d'un problème de valeurs propres)

Soient ε un réel et A_ε la matrice donnée par

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

Si $\varepsilon = 0$, alors les valeurs propres de A_0 , notées $\lambda_1, \dots, \lambda_4$, sont toutes nulles.

Si $\varepsilon = 10^{-4}$, alors les valeurs propres de A_ε sont solutions de l'équation

$$\det(A_\varepsilon - \lambda I) = \lambda^4 - \varepsilon = 0$$

d'où $|\lambda_i|^4 = 10^{-4}$, $1 \leq i \leq 4$. Ainsi $|\lambda_i| = 10^{-1}$. On a alors

$$\|\delta A_0\| = 10^{-4} \text{ et } \delta|\lambda_i| = 10^{-1} = 1000\|\delta A_0\|$$

D'où la nécessité d'introduire la notion de *conditionnement* d'une matrice vis-à-vis de la recherche de valeurs propres. On verra alors, lorsqu'on étudiera les problèmes de valeurs propres, que la matrice A_0 ci-dessus est *mal conditionnée* vis-à-vis de la recherche des valeurs propres ■

Exemple 6 : (Méthode des différences finies)

On considère l'équation de la chaleur dans le cas stationnaire avec conditions aux limites homogènes :

$$(1.5) \quad \begin{cases} -\Delta u = f & \text{dans } \Omega =]0, a[\times]0, b[\subset \mathbb{R}^2 \\ u|_{\partial\Omega} = 0 \end{cases}$$

On subdivise l'intervalle $]0, a[$ en utilisant $N + 2$ points et l'intervalle $]0, b[$ en utilisant $M + 2$ points comme suit :

$$x_i = ih, i = 0, \dots, N + 1 \quad ; \quad h = \delta x = \frac{a}{N + 1}$$

$$y_j = jk, j = 0, \dots, M + 1 \quad ; \quad k = \delta y = \frac{b}{M + 1}$$

Pour chercher une approximation des quantités $u(x_i, y_j)$, $1 \leq i \leq N$ et $1 \leq j \leq M$, où u est solution de l'équation (1.5), on remplace les dérivées partielles de u par des taux d'accroissement (rapports de différences finies) en utilisant la formule de Taylor-Lagrange :

$$(1.6) \quad \begin{aligned} u(x + h, y + k) &= u(x, y) + h \frac{\partial u}{\partial x}(x, y) + k \frac{\partial u}{\partial y}(x, y) + \\ &+ \frac{1}{2} \left(h^2 \frac{\partial^2 u}{\partial x^2}(x, y) + 2hk \frac{\partial^2 u}{\partial x \partial y}(x, y) + k^2 \frac{\partial^2 u}{\partial y^2}(x, y) \right) + \\ &+ O(|h^3| + |k^3|) \end{aligned}$$

Pour $x = x_i$, $y = y_j$ et en choisissant $k = 0$ dans (1.6), on obtient :

$$(1.7) \quad u(x_{i+1}, y_j) = u_{i,j} + h \frac{\partial u}{\partial x}(x_i, y_j) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_i, y_j) + O(|h|^3)$$

où $u_{i,j} = u(x_i, y_j)$.

De même, pour $x = x_i$, $y = y_j$ et en choisissant $k = 0$ dans (1.6), on obtient en remplaçant h par $-h$:

$$(1.8) \quad u(x_{i-1}, y_j) = u_{i,j} - h \frac{\partial u}{\partial x}(x_i, y_j) + \frac{h^2}{2} \frac{\partial^2 u}{\partial x^2}(x_i, y_j) + O(|h|^3)$$

La somme des deux équations (1.7) et (1.8) donne

$$u_{i+1,j} + u_{i-1,j} = 2u_{i,j} + h^2 \frac{\partial^2 u}{\partial x^2}(x_i, y_j) + O(|h|^3)$$

et par suite, on a

$$(1.9) \quad \frac{\partial^2 u}{\partial x^2}(x_i, y_j) = \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2} + O(|h|)$$

En suivant la même démarche que ci-dessus, on obtient :

$$(1.10) \quad \frac{\partial^2 u}{\partial y^2}(x_i, y_j) = \frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{k^2} + O(|k|)$$

Utilisant (1.9) et (1.10) et en négligeant les termes en $|h|$ et $|k|$, on obtient :

$$\begin{aligned} \Delta u(x_i, y_j) &\equiv \frac{\partial^2 u}{\partial x^2}(x_i, y_j) + \frac{\partial^2 u}{\partial y^2}(x_i, y_j) \\ &\simeq \frac{u_{i+1,j} + u_{i-1,j} - 2u_{i,j}}{h^2} + \frac{u_{i,j+1} + u_{i,j-1} - 2u_{i,j}}{k^2} \end{aligned}$$

L'approximation de Δu ci-dessus dans l'équation $-\Delta u = f$ au point (x_i, y_j) , pour $1 \leq i \leq N$ et $1 \leq j \leq M$, et les conditions aux limites imposées sur $\partial\Omega$ conduisent à

un système linéaire de $N \times M$ équations et à $N \times M$ inconnues. Si on range ces inconnues dans un vecteur u comme suit :

$$u = (u_{1,1}, \dots, u_{1,M}, u_{2,1}, \dots, u_{2,M}, \dots, u_{N,1}, \dots, u_{N,M})^T$$

alors, u est solution du système linéaire $Au = b$, où la matrice A à $N \times M$ lignes et $N \times M$ colonnes et le second membre $b \in \mathbb{R}^{N \times M}$ sont donnés par

$$A = \begin{pmatrix} d & e & & c & & & & & & \\ e & d & e & & \ddots & & & & & \\ & \ddots & \ddots & \ddots & & \ddots & & & & \\ & & \ddots & \ddots & \ddots & & \ddots & & & \\ c & & & \ddots & \ddots & \ddots & & \ddots & & \\ & c & \dots & \dots & e & d & e & \dots & \dots & c \\ & & \ddots & \bigcirc & \ddots & \ddots & \ddots & & & c \\ & & & \ddots & \ddots & \ddots & \ddots & & & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & & & & \ddots & \ddots & \ddots & \ddots & e \\ & & & & & c & 0 \dots & \dots 0 & e & d \\ & & & & & & \text{(M-2) colonnes} & & & \end{pmatrix} ; b = \begin{pmatrix} f(x_1, y_1) \\ \vdots \\ f(x_1, y_M) \\ f(x_2, y_1) \\ \vdots \\ f(x_2, y_M) \\ \vdots \\ f(x_N, y_1) \\ \vdots \\ f(x_N, y_M) \end{pmatrix}$$

où, nous avons noté :

$$d = \frac{-2}{h^2} - \frac{2}{k^2} ; \quad c = \frac{1}{h^2} ; \quad e = \frac{1}{k^2}$$

Dans le cas où $N \times M$ est un nombre très grand, il est important d'utiliser des méthodes de résolutions qui sont "économiques" du point de vue du nombre d'opérations élémentaires nécessaires et de la place mémoire utilisée pour la résolution du système linéaire obtenu au moyen d'un ordinateur.

Remarque : Nombre d'opérations nécessaire par la méthode de *Cramer*.

Soit $A = (a_{ij})_{1 \leq i, j \leq n}$ une matrice carrée d'ordre n , $b \in \mathbb{R}^n$ et $x \in \mathbb{R}^n$ solution du système linéaire $Ax = b$. Utilisant la méthode de *Cramer*, chaque composante x_i , $1 \leq i \leq n$, de la solution x est donnée par

$$x_i = \frac{1}{\det A} \det A_i$$

où A_i est la matrice obtenue en remplaçant la i -ème colonne de A par le second membre b , et où $\det A$ est donné par

$$\det A = \sum_{\sigma \in S_n} \varepsilon(\sigma) \prod_{i=1}^n a_{\sigma(i)i}$$

où S_n est l'ensemble des permutations de $\{1, 2, \dots, n\}$, de cardinal $\text{card}(S_n) = n!$.

Le calcul de $\det A$ nécessite

$$n!(n-1)[\text{multiplication}] + (n-1)[\text{addition}]$$

soit de l'ordre de $n.n!$ opérations élémentaires, lorsque n est "assez grand".

Le calcul de la solution x par la méthode de *Cramer*, nécessite alors de l'ordre de

$$(n+1)(n.n!)[\text{multiplication}] + n[\text{division}]$$

soit de l'ordre de $n^2 \cdot n!$ opérations élémentaires, lorsque n est "assez grand".

Si on utilise un ordinateur qui effectue 10^6 opérations élémentaires par seconde, on peut calculer qu'il nous faut de l'ordre de 200 années pour trouver la solution d'un système linéaire d'ordre $n = 18$ par la méthode de *Cramer*!

Pour $n = 50$, il faut $2,4 \cdot 10^{54}$ années de calcul! (pour le même type d'ordinateurs).

D'où la nécessité de développer des méthodes de résolution de systèmes linéaires plus économiques; et l'on verra dans la suite que ces mêmes systèmes seront résolus, avec les méthodes qui seront présentées, en un temps beaucoup plus réduit. ■

Remarque : Soit $A = (a_{ij})_{1 \leq i, j \leq n}$ une matrice triangulaire supérieure d'ordre n :

$$A = \begin{pmatrix} a_{11} & \cdots & \cdots & a_{1n} \\ & a_{22} & \cdots & a_{2n} \\ & & \ddots & \vdots \\ & & & a_{nn} \end{pmatrix}$$

vérifiant $\prod_{i=1}^n a_{ii} \neq 0$ (A est inversible).

Pour résoudre le système linéaire $Ax = b$ (triangulaire supérieur), on utilise la méthode dite *méthode de remontée*, qui consiste à "remonter les équations" comme suit :

$$x_n = b_n / a_{nn}$$

$$x_i = (b_i - \sum_{j=i+1}^n a_{ij}x_j) / a_{ii}, \quad i = n-1, n-2, \dots, 1$$

Cette méthode nécessite, pour chaque composante x_i , $1 \leq i \leq n$

$$\begin{aligned} & 1[\text{division}] + (n-i)[\text{multiplication}] + (n-i-1)[\text{addition}] + 1[\text{addition}] \\ & = (2(n-i) + 1)[\text{opérations élémentaires}] \end{aligned}$$

Ce qui donne, pour toutes les composantes de x :

$$\sum_{i=1}^n (2(n-i) + 1) = \sum_{j=0}^{n-1} 2j + n = \frac{2(n-1)n}{2} + n = n^2$$

opérations élémentaires. Par exemple, pour $n = 50$ il faut effectuer de l'ordre de 2500 opérations élémentaires, soit 0,0025 secondes sur un ordinateur qui fait 10^6 opérations élémentaires par seconde.

La majorité des méthodes directes de résolution de systèmes linéaires consistent à se ramener à des systèmes triangulaires (supérieurs ou inférieurs).

1.2 Principales notations et définitions

On utilise dans toute la suite les notations indiquées dans cette section.

Soit $(e_i)_{1 \leq i \leq n}$ la base canonique de $\mathbb{R}^n$ ou de $\mathbb{C}^n$ ($\mathbb{C}$ étant le corps des nombres complexes).

Notations pour les vecteurs :

Pour un vecteur x de $\mathbb{R}^n$ ou de $\mathbf{C}^n$, on utilise les notations suivantes :

$$x_i, 1 \leq i \leq n \quad : \quad i - \text{ème composante de } x$$

$$x = \sum_{i=1}^n x_i e_i \quad ; \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

$$x^T = (x_1, x_2, \dots, x_n) \quad : \quad \text{transposé de } x$$

$$\bar{x} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_n \end{pmatrix} \quad : \quad \text{conjugué de } x$$

$$x^* = \bar{x}^T = (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n) \quad : \quad \text{transconjugué de } x$$

Notations pour les matrices :

Soit $A = (a_{ij})_{1 \leq i, j \leq n}$ une matrice à n lignes et n colonnes, à coefficients réels ou complexes (i : indice de ligne ; j : indice de colonne).

On désigne par :

$$A^T = (a_{ij}^t)_{1 \leq i, j \leq n} = (a_{ji})_{1 \leq i, j \leq n} \quad , \quad \text{la transposée de } A$$

$$\bar{A} = (\bar{a}_{ij})_{1 \leq i, j \leq n} \quad , \quad \text{la conjuguée de } A$$

$$A^* = \bar{A}^T \quad , \quad \text{la transconjuguée de } A$$

et par $\mathcal{M}_{n,m}(K)$ l'ensemble des matrices à n lignes et m colonnes et à coefficients dans K ($K = \mathbb{R}$ ou $K = \mathbf{C}$). Lorsque $m = n$ on utilisera la notation $\mathcal{M}_n(K)$ au lieu de $\mathcal{M}_{n,n}(K)$.

La matrice identité de $\mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbf{C})$) est notée I .

Produit scalaire de vecteurs :

Le produit scalaire de $\mathbb{R}^n$ est défini pour x et y dans $\mathbb{R}^n$ par :

$$(x, y) = \sum_{i=1}^n x_i y_i = x^T \cdot y$$

et vérifie

$$\forall \lambda \in \mathbb{R}, \forall x \in \mathbb{R}^n, \forall y \in \mathbb{R}^n \quad : \quad \begin{cases} (\lambda x, y) = \lambda(x, y) \\ (x, \lambda y) = \lambda(x, y) \end{cases}$$

Le produit hermitien de $\mathbf{C}^n$ est défini pour x et y dans $\mathbf{C}^n$ par :

$$(x, y) = \sum_{i=1}^n x_i \bar{y}_i = y^* \cdot x$$

et vérifie

$$\forall \lambda \in \mathbf{C}, \forall x \in \mathbf{C}^n, \forall y \in \mathbf{C}^n \quad : \quad \begin{cases} (\lambda x, y) = \lambda(x, y) \\ (x, \lambda y) = \bar{\lambda}(x, y) \end{cases}$$

La norme euclidienne de $\mathbb{R}^n$ et celle de $\mathbb{C}^n$ sont définies comme suit :

$$\|x\|_2 = (x, x)^{1/2} = \begin{cases} \left(\sum_{i=1}^n x_i^2 \right)^{1/2} & \text{si } x \in \mathbb{R}^n \\ \left(\sum_{i=1}^n x_i \bar{x}_i \right)^{1/2} = \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} & \text{si } x \in \mathbb{C}^n \end{cases}$$

où $|z|$ désigne le module d'un nombre complexe z .

Rappelons que deux vecteurs x et y de $\mathbb{R}^n$ (ou $\mathbb{C}^n$) sont dit orthogonaux lorsque $(x, y) = 0$.

1.3 Rappels et compléments sur les matrices

Propriétés 1.1

1) Soient $x \in \mathbb{R}^n$, $y \in \mathbb{R}^n$ et $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathcal{M}_n(\mathbb{R})$. Alors, on a :

$$(Ax, y) = (x, A^T y)$$

2) Soient $x \in \mathbb{C}^n$, $y \in \mathbb{C}^n$ et $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathcal{M}_n(\mathbb{C})$. Alors, on a :

$$(Ax, y) = (x, A^* y)$$

Démonstration :

1) On a, pour tout x, y dans $\mathbb{R}^n$:

$$(Ax, y) = \sum_{i=1}^n (Ax)_i y_i = \sum_{i=1}^n \left(\sum_{j=1}^n a_{ij} x_j \right) y_i = \sum_{j=1}^n \sum_{i=1}^n a_{ij} y_i x_j$$

Utilisant le fait que $a_{ij} = a_{ji}^t$ on obtient

$$(Ax, y) = \sum_{j=1}^n (A^T y)_j x_j = (A^T y, x)$$

le résultat découle alors de la symétrie du produit scalaire de $\mathbb{R}^n$.

2) Le résultat s'obtient en suivant la même démarche que celle utilisée dans 1). ■

Définition 1.1

1) Soit $A \in \mathcal{M}_n(\mathbb{R})$. La matrice A est dite *symétrique* si $A^T = A$.

2) Soit $A \in \mathcal{M}_n(\mathbb{C})$. La matrice A est dite *hermitienne* si $A^* = A$. ■

Propriétés 1.2

1) Si $A \in \mathcal{M}_n(\mathbb{R})$ est *symétrique*, alors on a :

$$\begin{cases} \forall (i, j) \in \{1, \dots, n\}^2, & a_{ij} = a_{ji} \\ \text{et} \\ \forall (x, y) \in \mathbb{R}^n \times \mathbb{R}^n, & (Ax, y) = (x, Ay) \end{cases}$$

2) Si $A \in \mathcal{M}_n(\mathbb{C})$ est *hermitienne*, alors on a :

$$\begin{cases} \forall (i, j) \in \{1, \dots, n\}^2, & a_{ij} = \bar{a}_{ji} \\ \text{et} \\ \forall (x, y) \in \mathbb{C}^n \times \mathbb{C}^n, & (Ax, y) = (x, Ay) \end{cases}$$

Démonstration : Le résultat découle immédiatement de la propriété 1.1. ■

Définition 1.2

Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) est dite *inversible*, s'il existe une matrice $B \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) telle que $AB = BA = I$.

La matrice B est alors unique et est notée A^{-1} . ■

Propriétés 1.3

Si $A \in \mathcal{M}_n(\mathbb{R})$ est inversible, alors A^T est inversible et on a :

$$(A^T)^{-1} = (A^{-1})^T$$

Démonstration :

La matrice A étant inversible, alors on a $AA^{-1} = A^{-1}A = I$. Et par suite $(AA^{-1})^T = (A^{-1}A)^T = I^T$. D'où $(A^{-1})^T A^T = A^T (A^{-1})^T = I$. Ce qui prouve que $(A^T)^{-1} = (A^{-1})^T$. ■

Définition 1.3

1) Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ est dite *orthogonale*, si elle vérifie

$$AA^T = A^T A = I.$$

2) Une matrice $A \in \mathcal{M}_n(\mathbb{C})$ est dite *unitaire*, si elle vérifie

$$AA^* = A^* A = I.$$

■

Propriétés 1.4

1) Si $A \in \mathcal{M}_n(\mathbb{R})$ est orthogonale, alors A est inversible et $A^{-1} = A^T$. De plus, on a :

$$\text{pour tout } x \in \mathbb{R}^n, \quad \|Ax\|_2 = \|x\|_2$$

2) Si $U \in \mathcal{M}_n(\mathbb{C})$ est unitaire, alors U est inversible et $U^{-1} = U^*$. De plus, on a :

$$\text{pour tout } x \in \mathbb{C}^n, \quad \|Ux\|_2 = \|x\|_2$$

d'où $\|U\|_2 = 1$.

Démonstration :

Nous allons donner la démonstration de 2), celle de 1) lui étant analogue.

Soit U une matrice unitaire. Alors il est clair, d'après la définition 1.3, que U est inversible et que $U^{-1} = U^*$.

Par ailleurs, on a pour tout $x \in \mathbb{C}^n$:

$$\|Ux\|_2^2 = (Ux, Ux) = (x, U^* U x)$$

et puisque U est unitaire, on obtient

$$\|Ux\|_2^2 = (x, x) = \|x\|_2^2.$$

Ce qui prouve le résultat souhaité. ■

Remarque : Si $A \in \mathcal{M}_n(\mathbb{R})$ est orthogonale, alors les colonnes de A forment une base orthonormée de $\mathbb{R}^n$. En effet si l'on note $C_i \in \mathbb{R}^n$, $1 \leq i \leq n$, la i -ème colonne de A , alors on obtient pour i et j compris entre 1 et n :

$$C_i^T \cdot C_j = \delta_{ij} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{sinon} \end{cases}$$

■

Définition 1.4

1) Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) est dite triangulaire supérieure, si

$$a_{ij} = 0 \quad \text{pour tout } i > j, \quad 1 \leq i, j \leq n$$

2) Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) est dite triangulaire inférieure, si

$$a_{ij} = 0 \quad \text{pour tout } i < j, \quad 1 \leq i, j \leq n$$

3) Une matrice $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) est dite diagonale, si

$$a_{ij} = 0 \quad \text{pour tout } i \neq j, \quad 1 \leq i, j \leq n$$

On utilisera alors la notation $A = \text{diag}(a_{ii})_{1 \leq i \leq n}$.

■

Propriétés 1.5

Soit $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$).

1) Si A est triangulaire inférieure et triangulaire supérieure, alors A est diagonale.

2) Si A est triangulaire supérieure (respectivement inférieure) et inversible, alors A^{-1} est triangulaire supérieure (respectivement inférieure).

3) Si A est triangulaire supérieure (respectivement inférieure), alors A^T (ou A^*) est triangulaire inférieure (respectivement supérieure).

4) Si A est triangulaire supérieure (ou inférieure), alors le déterminant de A est le produit des coefficients diagonaux de A :

$$\det A = \prod_{i=1}^n a_{ii}$$

5) Si A est triangulaire supérieure (ou inférieure), alors on a :

$$A \text{ est inversible} \iff a_{ii} \neq 0 \text{ pour tout } i \in \{1, 2, \dots, n\}$$

6) Soit $B \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$). Si A et B sont triangulaires inférieures (respectivement supérieures), alors la matrice produit AB est triangulaire inférieure (respectivement supérieure).

Démonstration : À faire en exercice.

■

Définition 1.5

Soit $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) une matrice symétrique (ou hermitienne). Alors, A est dite définie positive, si elle vérifie les deux propriétés suivantes :

$$(1.11) \quad \forall x \in \mathbb{R}^n \text{ (ou } \mathbb{C}^n), \quad (Ax, x) \geq 0$$

$$(1.12) \quad \text{pour } x \in \mathbb{R}^n \text{ (ou } \mathbb{C}^n), \text{ on a : } (Ax, x) = 0 \iff x = 0$$

Dans le cas où A vérifie uniquement (1.11), on dit qu'elle est semi-définie positive.

■

Exemples :

- 1) Si $A \in \mathcal{M}_n(\mathbf{C})$, alors la matrice AA^* est semi-définie positive.
- 2) Si $A \in \mathcal{M}_n(\mathbf{C})$ est inversible, alors la matrice AA^* est définie positive. ■

Définition 1.6

Soient $\lambda \in \mathbf{C}$ et $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbf{C})$). Alors, λ est appelée valeur propre de A , si elle est racine du polynôme caractéristique $P_A(\lambda)$ de A :

$$P_A(\lambda) \equiv \det(A - \lambda I) = 0$$

■

Propriétés 1.6

Soit $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbf{C})$) et soit λ une valeur propre de A . Alors, on a

- i) La matrice $A - \lambda I$ est non inversible.
- ii) Il existe un vecteur $x \in \mathbb{R}^n$ (ou $\mathbf{C}^n$) non nul tel que $Ax = \lambda x$.

Un tel vecteur x est appelé vecteur propre de A associé à la valeur propre λ .

Démonstration : À faire en exercice. ■

L'existence de valeurs propres d'une matrice symétrique (ou hermitienne) est donnée par la proposition suivante :

Proposition 1.1

1) Si $A \in \mathcal{M}_n(\mathbb{R})$ est symétrique, alors il existe une matrice orthogonale $O \in \mathcal{M}_n(\mathbb{R})$ telle que la matrice $D = O^T A O$ soit diagonale réelle.

Si de plus A est définie positive, alors la diagonale de D est strictement positive ($d_{ii} > 0$, $i = 1, 2, \dots, n$).

2) Si $A \in \mathcal{M}_n(\mathbf{C})$ est hermitienne, alors il existe une matrice unitaire $U \in \mathcal{M}_n(\mathbf{C})$ telle que la matrice $D = U^* A U$ soit diagonale réelle.

Si de plus A est définie positive, alors la diagonale de D est strictement positive ($d_{ii} > 0$, $i = 1, 2, \dots, n$).

Démonstration : À faire en exercice. ■**Commentaires :**

Soient $A \in \mathcal{M}_n(\mathbb{R})$ symétrique et O la matrice orthogonale définie à la proposition 1.1 ci-dessus. Notons O_i , $1 \leq i \leq n$, les n colonnes de O et rappelons que les $(O_i)_{1 \leq i \leq n}$ forment une base orthonormée de $\mathbb{R}^n$.

Soit $D = O^T A O = \text{diag}(d_{ii})_{1 \leq i \leq n}$, alors les coefficients d_{ii} de D sont exactement les valeurs propres de A . En effet, on a

$$OD = OO^T A O = A O$$

puisque O est une matrice orthogonale. Ainsi $A O_i = d_{ii} O_i$, $1 \leq i \leq n$. Ce qui prouve que d_{ii} est une valeur propre de A .

Ainsi les coefficients de la matrice diagonale D sont donnés par $d_{ii} = \lambda_i$, où les λ_i , $1 \leq i \leq n$, sont les valeurs propres de A .

De plus, les $(O_i)_{1 \leq i \leq n}$ forment une base de vecteurs propres associés aux valeurs propres λ_i , $1 \leq i \leq n$.

La matrice O est appelée matrice de passage de la base canonique de $\mathbb{R}^n$ à la base de vecteurs propres de la matrice A . ■

Corollaire 1.1

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Alors, A est définie positive si et seulement si toutes ses valeurs propres sont strictement positives.

Démonstration : À faire en exercice. ■

Définition 1.7

Soient $A \in \mathcal{M}_n(\mathbb{R})$ et λ une valeur propre de A .

La multiplicité de λ est la multiplicité de λ comme racine du polynôme caractéristique $P_A(\lambda)$ de A . ■

Remarque : Lorsque A est diagonalisable, la multiplicité de λ est égale à la dimension du sous-espace propre associé à λ , c'est à dire la dimension de $\ker(A - \lambda I)$, où $\ker(M) = \{x \in \mathbb{R}^n \text{ tels que } Mx = 0\}$ pour $M \in \mathcal{M}_n(\mathbb{R})$. C'est le cas, par exemple, d'une matrice symétrique. ■

Propriétés 1.7 (Jordanisation)

Soit $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) et soit ε un réel strictement positif donné. Alors, il existe $S \in \mathcal{M}_n(\mathbb{C})$ inversible telle que la matrice $B = S^{-1}AS$ soit de la forme de Jordan :

$$(1.13) \quad B = \begin{pmatrix} \boxed{J_1} & & & & \\ & \boxed{J_2} & & \circ & \\ & & \boxed{\ddots} & & \\ & \circ & & \boxed{\ddots} & \\ & & & & \boxed{J_p} \end{pmatrix}$$

où les blocs de Jordan J_i , $1 \leq i \leq p \leq n$ sont triangulaires supérieurs et sont donnés par

$$J_i = \begin{pmatrix} \lambda_i & \varepsilon & & \\ & \lambda_i & \ddots & \circ \\ & & \ddots & \ddots \\ \circ & & & \ddots & \varepsilon \\ & & & & \lambda_i \end{pmatrix}$$

où les λ_i , $1 \leq i \leq p \leq n$, sont les valeurs propres de la matrice A .

Dans le cas où $\varepsilon = 1$, on dit que la matrice B donnée par (1.13) est obtenue par jordanisation de A .

Démonstration : Admise.

Remarque : Il faut noter qu'une valeur propre λ_i apparaissant dans le bloc J_i peut apparaître dans un autre bloc J_j pour $j \neq i$. ■

Définition 1.8

Soient $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) et λ_i , $1 \leq i \leq n$, les n valeurs propres de A comptées avec leur ordre de multiplicité.

On appelle rayon spectral de A , le réel positif :

$$\varrho(A) = \max_{1 \leq i \leq n} |\lambda_i|$$

■

Lemme 1.1

Soit $C \in \mathcal{M}_n(\mathbf{C})$ une matrice hermitienne. Alors on a

$$\varrho(C) = \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{|(Cv, v)|}{(v, v)}$$

Démonstration :

Soit λ une valeur propre de C telle que $|\lambda| = \varrho(C)$ et $v_0 \in \mathbf{C}^n \setminus \{0\}$ un vecteur propre qui lui est associé. On a :

$$(1.14) \quad \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{|(Cv, v)|}{\|v\|_2^2} \geq \frac{|(Cv_0, v_0)|}{\|v_0\|_2^2} = |\lambda| = \varrho(C)$$

Par ailleurs, d'après la proposition 1.1, il existe une matrice U unitaire telle que la matrice $D = U^*CU$ soit diagonale. De plus, la diagonale de D est formée des valeurs propres de C . On a alors :

$$\begin{aligned} \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{|(Cv, v)|}{\|v\|_2^2} &= \sup_{w \in \mathbf{C}^n \setminus \{0\}} \frac{|(CUw, Uw)|}{\|w\|_2^2} \quad (\text{où } v = Uw ; \|Uw\|_2 = \|w\|_2) \\ &= \sup_{w \in \mathbf{C}^n \setminus \{0\}} \frac{|(U^*CUw, w)|}{\|w\|_2^2} \\ &= (\text{puisque } D = U^*CU = \text{diag}(\lambda_i)) \\ &= \sup_{w \in \mathbf{C}^n \setminus \{0\}} \frac{|\sum_{i=1}^n \lambda_i |w_i|^2|}{\|w\|_2^2} \end{aligned}$$

où les λ_i , $1 \leq i \leq n$, sont les valeurs propres de C . Il s'en suit que

$$\sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{|(Cv, v)|}{\|v\|_2^2} \leq \sup_{w \in \mathbf{C}^n \setminus \{0\}} \frac{\sum_{i=1}^n |\lambda_i| |w_i|^2}{\|w\|_2^2} \leq \varrho(C)$$

ce qui prouve, grâce à (1.14), le résultat souhaité. ■

1.4 Normes vectorielles et normes matricielles

1.4.1 Normes vectorielles

Définition 1.9

Une norme vectorielle est une application de $\mathbb{R}^n$ (ou $\mathbf{C}^n$) dans $\mathbb{R}_+$, notée $\|\cdot\|$, qui vérifie les propriétés suivantes :

$$\begin{aligned} \|x\| &= 0 \Leftrightarrow x = 0, & \text{pour tout } x \in \mathbb{R}^n \text{ (ou } \mathbf{C}^n) \\ \|x + y\| &\leq \|x\| + \|y\|, & \text{pour tout } x, y \in \mathbb{R}^n \text{ (ou } \mathbf{C}^n) \\ \|\lambda x\| &= |\lambda| \|x\|, & \text{pour tout } \lambda \in \mathbb{R} \text{ (ou } \mathbf{C}) \text{ et } x \in \mathbb{R}^n \text{ (ou } \mathbf{C}^n) \end{aligned}$$

Dans le cas où l'application $\|\cdot\|$ ne vérifie que les deux dernières propriétés ci-dessus, on l'appelle semi-norme. ■

Exemples : Normes vectorielles classiques

Soit $x = (x_i)_{1 \leq i \leq n} \in \mathbb{R}^n$ (ou $\mathbb{C}^n$). Les trois normes suivantes sont les plus couramment utilisées en pratique :

$$\|x\|_1 = \sum_{i=1}^n |x_i| \quad (\text{norme indice 1})$$

$$\|x\|_2 = \left(\sum_{i=1}^n |x_i|^2 \right)^{1/2} \quad (\text{norme euclidienne})$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i| \quad (\text{norme indice l'infini})$$

Il est facile de vérifier que ces applications sont effectivement des normes. Plus généralement, pour tout réel $p \geq 1$, l'application $\|\cdot\|_p$ définie, pour $x \in \mathbb{R}^n$ (ou $\mathbb{C}^n$), par

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}$$

définit une norme sur $\mathbb{R}^n$ (ou $\mathbb{C}^n$) (à vérifier en exercice). ■

Propriétés 1.8

Soit $p \in [1, +\infty]$. Alors les normes $\|\cdot\|_p$ sont équivalentes, c'est à dire :

$$\forall (p, q) \in [1, +\infty]^2, \exists \alpha > 0, \exists \beta > 0 \text{ tels que}$$

$$\forall x \in \mathbb{C}^n, \quad \alpha \|x\|_p \leq \|x\|_q \leq \beta \|x\|_p$$

Démonstration : À faire en exercice. ■

1.4.2 Normes matricielles**Définition 1.10**

On appelle norme matricielle, toute application $\|\cdot\|$ de $\mathcal{M}_n(\mathbb{C})$ dans $\mathbb{R}_+$ qui vérifie les propriétés suivantes :

- i) Pour $A \in \mathcal{M}_n(\mathbb{C})$ on a : $\|A\| = 0 \Leftrightarrow A = O_n$
- ii) $\forall (A, B) \in (\mathcal{M}_n(\mathbb{C}))^2, \quad \|A + B\| \leq \|A\| + \|B\|$
- iii) $\forall \lambda \in \mathbb{C}, \forall A \in \mathcal{M}_n(\mathbb{C}), \quad \|\lambda A\| = |\lambda| \|A\|$
- iv) $\forall (A, B) \in (\mathcal{M}_n(\mathbb{C}))^2, \quad \|AB\| \leq \|A\| \|B\|$

où O_n désigne la matrice nulle de $\mathcal{M}_n(\mathbb{C})$. ■

Exemple : L'application $\|\cdot\|$ définie, pour $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathcal{M}_n(\mathbb{C})$, par

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

est une norme matricielle, appelée la norme de Schur. ■

Notation : On utilisera, dans toute la suite, la notation $\|\cdot\|$ pour désigner une norme matricielle, lorsqu'il n'y a pas de risque de confusion.

1.4.3 Normes matricielles subordonnées

Définition 1.11

Soit $A \in \mathcal{M}_n(\mathbf{C})$. On dit que l'application $A \mapsto \|A\|$ est une norme matricielle subordonnée à la norme vectorielle $v \mapsto \|v\|$, si

$$(1.15) \quad \|A\| = \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{\|Av\|}{\|v\|}$$

■

Remarque : On peut vérifier que l'application $\|\cdot\|$ définie par (1.15) définit effectivement une norme matricielle. ■

Remarque : Toute norme matricielle subordonnée $\|\cdot\|$ vérifie

$$\forall v \in \mathbf{C}^n, \forall A \in \mathcal{M}_n(\mathbf{C}), \quad \|Av\| \leq \|A\| \|v\|$$

et $\|I\| = 1$. ■

Exemples :

1) Cas d'une norme matricielle non subordonnée :

La norme matricielle $\|\cdot\|$ définie, pour $A = (a_{ij})_{1 \leq i, j \leq n} \in \mathcal{M}_n(\mathbb{R})$ ($n \geq 2$), par

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

n'est pas une norme matricielle subordonnée, puisque $\|I\| = \sqrt{n}$ ($\neq 1$).

2) Normes matricielles subordonnées indice p :

$$\|A\|_p = \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{\|Av\|_p}{\|v\|_p}, \quad 1 \leq p \leq \infty$$

Les normes matricielles subordonnées les plus couramment utilisées dans la pratique, sont celles qui correspondent à $p = 1$, $p = 2$ et $p = \infty$. ■

Proposition 1.2

Soit $A \in \mathcal{M}_n(\mathbf{C})$. Alors on a :

$$(1.16) \quad \|A\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$$

$$(1.17) \quad \|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| = \|A^*\|_1$$

$$(1.18) \quad \|A\|_2 = \sqrt{\varrho(A^*A)} = \|A^*\|_2 = \sqrt{\varrho(AA^*)}$$

Démonstration :

i) Prouvons (1.16). On a, pour $A \in \mathcal{M}_n(\mathbf{C})$:

$$\|A\|_1 \equiv \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{\|Av\|_1}{\|v\|_1}$$

avec $\|v\|_1 = \sum_{i=1}^n |v_i|$, et

$$\begin{aligned} \|Av\|_1 &\equiv \sum_{i=1}^n |(Av)_i| = \sum_{i=1}^n \left| \sum_{k=1}^n a_{ik} v_k \right| \\ &\leq \sum_{i=1}^n \sum_{k=1}^n |a_{ik}| |v_k| \\ &\leq \sum_{k=1}^n \left[\left(\sum_{i=1}^n |a_{ik}| \right) |v_k| \right] \end{aligned}$$

Utilisant le fait que pour tout $k \in \{1, 2, \dots, n\}$, on a

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

on obtient

$$\|Av\|_1 \leq \left(\max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| \right) \underbrace{\sum_{k=1}^n |v_k|}_{=\|v\|_1}$$

Ainsi, on a

$$\forall v \in \mathbf{C}^n \setminus \{0\}, \quad \frac{\|Av\|_1}{\|v\|_1} \leq \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$$

d'où

$$(1.19) \quad \|A\|_1 \leq \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}|$$

Prouvons l'inégalité inverse. Soit $j_o \in \{1, 2, \dots, n\}$ tel que

$$(1.20) \quad \max_{1 \leq j \leq n} \sum_{i=1}^n |a_{ij}| = \sum_{i=1}^n |a_{ij_o}|$$

et soit e_{j_o} le j_o -ème vecteur de la base canonique de $\mathbf{C}^n$. Alors, on a

$$(1.21) \quad \|A\|_1 \equiv \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{\|Av\|_1}{\|v\|_1} \geq \frac{\|Ae_{j_o}\|_1}{\|e_{j_o}\|_1}$$

Or, $\|Ae_{j_o}\|_1 = \sum_{i=1}^n |a_{ij_o}|$. Il s'en suit, d'après (1.20) et (1.21), que

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

ce qui prouve, grâce à (1.19), l'égalité (1.16)

ii) Prouvons (1.17). On a, pour tout $v \in \mathbf{C}^n$:

$$\|Av\|_\infty \equiv \max_{1 \leq i \leq n} |(Av)_i| = \max_{1 \leq i \leq n} \left| \sum_{j=1}^n a_{ij} v_j \right| \leq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| |v_j|$$

et par suite

$$\forall v \in \mathbf{C}^n, \quad \|Av\|_\infty \leq \left(\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| \right) \|v\|_\infty$$

Il en découle que

$$(1.22) \quad \|A\|_\infty \equiv \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{\|Av\|_\infty}{\|v\|_\infty} \leq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$$

Montrons l'inégalité inverse. Soit $i_o \in \{1, 2, \dots, n\}$ tel que

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

Si $a_{i_o j} = 0$ pour tout $j \in \{1, 2, \dots, n\}$, alors $A \equiv O_n$, puisque

$$\max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}| = \sum_{j=1}^n |a_{i_o j}| = 0,$$

et $\|A\|_\infty = 0$. L'égalité (1.17) est alors évidente. Dans la suite on supposera qu'il existe au moins un indice j tel que $a_{i_o j} \neq 0$.

Soit $\tilde{v} \in \mathbf{C}^n$ le vecteur dont chaque composante $\tilde{v}_j$, $1 \leq j \leq n$, est donnée par :

$$\tilde{v}_j = \begin{cases} \frac{\bar{a}_{i_o j}}{|a_{i_o j}|} & \text{si } a_{i_o j} \neq 0 \\ 0 & \text{si } a_{i_o j} = 0 \end{cases}$$

On a alors $\|\tilde{v}\|_\infty = 1$ et pour $1 \leq k \leq n$, on a

$$(A\tilde{v})_k = \sum_{j=1}^n a_{kj} \tilde{v}_j = \sum_{j=1}^n a_{kj} \frac{\bar{a}_{i_o j}}{|a_{i_o j}|}$$

et par suite

$$\begin{aligned} \|A\tilde{v}\|_\infty &\equiv \max_{1 \leq k \leq n} |(A\tilde{v})_k| \geq |(A\tilde{v})_{i_o}| \\ &\geq \left| \sum_{j=1}^n a_{i_o j} \frac{\bar{a}_{i_o j}}{|a_{i_o j}|} \right| = \left| \sum_{j=1}^n |a_{i_o j}| \right| = \sum_{j=1}^n |a_{i_o j}| \end{aligned}$$

Il s'en suit, puisque $\|\tilde{v}\|_\infty = 1$, que

$$\frac{\|A\tilde{v}\|_\infty}{\|\tilde{v}\|_\infty} \geq \sum_{j=1}^n |a_{i_o j}| = \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$$

d'où

$$\|A\|_\infty \geq \frac{\|A\tilde{v}\|_\infty}{\|\tilde{v}\|_\infty} \geq \max_{1 \leq i \leq n} \sum_{j=1}^n |a_{ij}|$$

Ce qui prouve, grâce à (1.22), la première égalité de (1.17).

La deuxième égalité $\|A\|_\infty = \|A^*\|$ de (1.17) découle immédiatement de la définition de A^* et de (1.16).

iii) **Prouvons** (1.18). Commençons par démontrer l'égalité $\|A\|_2 = \sqrt{\varrho(A^*A)}$. On a

$$(1.23) \quad \begin{aligned} \|A\|_2^2 &= \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{\|Av\|_2^2}{\|v\|_2^2} = \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{(Av, Av)}{(v, v)} \\ &= \sup_{v \in \mathbf{C}^n \setminus \{0\}} \frac{(v, A^*Av)}{(v, v)} \end{aligned}$$

Utilisant le lemme 1.1 et le fait que la matrice A^*A est hermitienne et semi-définie positive, on obtient, grâce à (1.23), que

$$\|A\|_2^2 = \varrho(A^*A).$$

Ce qui prouve la première égalité de (1.18).

Montrons que $\|A\|_2 = \|A^*\|_2$. On a, d'après ce qui précède, $\|A^*\|_2 = \sqrt{\varrho(AA^*)}$. Il suffit alors de prouver que $\varrho(A^*A) = \varrho(AA^*)$.

1er cas : $\varrho(A^*A) \neq 0$.

Soit $\lambda \neq 0$ une valeur propre de A^*A . On a alors

$$(1.24) \quad \exists v \in \mathbf{C}^n \setminus \{0\}; \quad A^*Av = \lambda v$$

ou encore $AA^*(Av) = \lambda(Av)$. Il s'en suit, puisque Av est non nul¹, que λ est une valeur propre de AA^* . Ainsi, toute valeur propre non nulle de A^*A est une valeur propre de AA^* .

Réciproquement, en utilisant la même démarche, on montre que toute valeur propre non nulle de AA^* est une valeur propre de A^*A . Ce qui prouve que $\varrho(A^*A) = \varrho(AA^*)$.

2ème cas : $\varrho(A^*A) = 0$.

Dans ce cas, toutes les valeurs propres de A^*A sont nulles. Et par suite, comme A^*A est hermitienne² alors $A^*A = O_n$. On a alors :

$$\forall v \in \mathbf{C}^n \setminus \{0\}, \quad (A^*Av, v) = 0.$$

d'où $\|Av\|_2^2 = (Av, Av) = (A^*Av, v) = 0$. Ainsi, pour tout $v \in \mathbf{C}^n \setminus \{0\}$, $Av = 0$. D'où nécessairement $A = O_n$ et donc $A^* = O_n$. Il s'en suit que $\varrho(A^*A) = \varrho(AA^*)$. ■

Proposition 1.3

Soit $A \in \mathcal{M}_n(\mathbf{C})$ une matrice hermitienne. Alors, on a

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

Démonstration :

Il suffit d'utiliser (1.18) et la proposition 1.1 qui permet de voir que les valeurs propres de A^2 sont les carrés des valeurs propres de A , lorsque A est hermitienne. ■

Théorème 1.1

1) Soit $\|\cdot\|$ une norme matricielle, subordonnée ou non, quelconque. Alors on a :

$$\forall A \in \mathcal{M}_n(\mathbf{C}), \quad \varrho(A) \leq \|A\|.$$

2) Soit $\varepsilon > 0$ et soit $A \in \mathcal{M}_n(\mathbf{C})$. Alors, il existe une norme matricielle subordonnée notée $\|\cdot\|_{A,\varepsilon}$ telle que

$$\|A\|_{A,\varepsilon} \leq \varrho(A) + \varepsilon$$

¹Si $Av = 0$, alors, d'après (1.24), $\lambda v = 0$. Ce qui contredit le fait que $\lambda \neq 0$ et $v \neq 0$.

²Si une matrice M est telle que toutes ses valeurs propres sont nulles, alors M n'est pas nécessairement nulle sauf si elle est hermitienne. Prendre par exemple $M = \begin{pmatrix} 0 & 0 \\ 10^6 & 0 \end{pmatrix}$

Démonstration :

1) Soit λ une valeur propre de A telle que $|\lambda| = \varrho(A)$ et $u \in \mathbf{C}^n \setminus \{0\}$ un vecteur propre qui lui est associé. Soit $v \in \mathbf{C}^n$ un vecteur tel que la matrice uv^t ne soit pas nulle³. On a alors :

$$\varrho(A) \|uv^t\| = \|\lambda uv^t\| = \|Auv^t\| \leq \|A\| \|uv^t\|,$$

d'après la propriété iv) des normes matricielles. Ainsi, on a $\varrho(A) \leq \|A\|$.

2) Soit $\varepsilon > 0$ et soit $A \in \mathcal{M}_n(\mathbf{C})$. Il existe alors, grâce à la propriété 1.7, une matrice inversible $S \in \mathcal{M}_n(\mathbf{C})$ telle que la matrice $B = S^{-1}AS$ soit de la forme de Jordan :

$$B = \begin{pmatrix} \boxed{J_1} & & & & \\ & \boxed{J_2} & & \circ & \\ & & \boxed{\ddots} & & \\ & \circ & & \boxed{\ddots} & \\ & & & & \boxed{J_p} \end{pmatrix}$$

où les blocs de Jordan J_i , $1 \leq i \leq p$, sont donnés par

$$J_i = \begin{pmatrix} \lambda_i & \varepsilon & & & \\ & \lambda_i & \ddots & \circ & \\ & & \ddots & \ddots & \\ & \circ & & \ddots & \varepsilon \\ & & & & \lambda_i \end{pmatrix}$$

où les λ_i , $1 \leq i \leq p \leq n$, sont les valeurs propres de la matrice A .

Il s'en suit que

$$(1.25) \quad \|B\|_1 = \max_{1 \leq j \leq n} \sum_{i=1}^n |B_{ij}| \leq \max_{1 \leq i \leq n} |\lambda_i| + \varepsilon \leq \varrho(A) + \varepsilon$$

Posons, pour tout $C \in \mathcal{M}_n(\mathbf{C})$, $\|C\|_{A,\varepsilon} = \|S^{-1}CS\|_1$. Alors, il est clair que $\|\cdot\|_{A,\varepsilon}$ définit une norme matricielle sur $\mathcal{M}_n(\mathbf{C})$, et vérifie

$$\|A\|_{A,\varepsilon} \leq \varrho(A) + \varepsilon$$

Montrons que c'est une norme matricielle subordonnée. Pour cela, soit $C \in \mathcal{M}_n(\mathbf{C})$, alors on a

$$\begin{aligned} \|C\|_{A,\varepsilon} &= \sup_{v \neq 0} \frac{\|S^{-1}CSv\|_1}{\|v\|_1} \\ &= \text{(en posant } w = Sv) \\ &= \sup_{w \neq 0} \frac{\|S^{-1}Cw\|_1}{\|S^{-1}w\|_1} \end{aligned}$$

³Car sinon, si pour tout $v \in \mathbf{C}^n$ on a $uv^t = O$, alors en choisissant $v = (1, 0, \dots, 0)^t$ on obtient

$$uv^t = \begin{pmatrix} u_1 & 0 & \cdot & 0 \\ u_2 & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ u_n & 0 & \cdot & 0 \end{pmatrix} = O$$

ce qui ne peut pas se produire, puisque $u \neq 0$.

En posant $N(w) = \|S^{-1}w\|_1$, qui définit une norme vectorielle, on obtient

$$\|C\|_{A,\varepsilon} = \sup_{w \neq 0} \frac{N(Cw)}{N(w)}$$

Ce qui prouve que $\|\cdot\|_{A,\varepsilon}$ définit une norme matricielle subordonnée à la norme vectorielle $N(\cdot)$. ■

Théorème 1.2

1) Soit $\|\cdot\|$ une norme matricielle quelconque, subordonnée ou non, et soit $B \in \mathcal{M}_n(\mathbf{C})$ telle que $\|B\| < 1$. Alors la matrice $I - B$ est inversible et vérifie

$$(1.26) \quad \|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}$$

$$(1.27) \quad (I - B)^{-1} = \sum_{k=0}^{+\infty} B^k$$

2) Si une matrice de la forme $(I - B)$ est non inversible, alors pour toute norme matricielle $\|\cdot\|$, subordonnée ou non, on a $\|B\| \geq 1$.

Démonstration :

1) Posons $S_N = \sum_{k=0}^N B^k$. Alors on a :

$$(1.28) \quad S_N(I - B) = \sum_{k=0}^N B^k - \sum_{k=0}^N B^{k+1} = I - B^{N+1}$$

De même, on a

$$(1.29) \quad (I - B)S_N = I - B^{N+1}$$

La suite S_N est convergente dans $(\mathcal{M}_n(\mathbf{C}), \|\cdot\|)$ ⁴, puisque l'on a

$$\forall N, \forall M, N > M, \quad \|S_N - S_M\| = \left\| \sum_{k=M+1}^N B^k \right\| \leq \sum_{k=M+1}^N \|B\|^k$$

et que $\|B\| < 1$ (une suite géométrique de raison strictement inférieure à 1 est convergente et donc de Cauchy).

Notons S la limite de S_N . Il s'en suit, grâce à (1.28) et (1.29), que

$$S(I - B) = (I - B)S = I$$

Ainsi, $I - B$ est inversible et

$$(I - B)^{-1} = S = \sum_{k=0}^{+\infty} B^k$$

Par ailleurs, on a

$$\|S_N\| \leq \sum_{k=0}^N \|B\|^k = \frac{1 - \|B\|^{N+1}}{1 - \|B\|}$$

En passant à la limite, lorsque $N \rightarrow +\infty$, on obtient

$$\|S\| = \|(I - B)^{-1}\| \leq \frac{1}{1 - \|B\|}$$

Ce qui prouve (1.26) et achève la démonstration.

2) Si $I - B$ est non inversible, alors 1 est une valeur propre de B , et par suite on a $\|B\| \geq 1$, puisque $\varrho(B) \geq 1$ et d'après le théorème 1.1 on a $\|B\| \geq \varrho(B)$. ■

⁴ $(\mathcal{M}_n(\mathbf{C}), \|\cdot\|)$ est un espace vectoriel normé.

Commentaires :

Soit $\|\cdot\|$ une norme matricielle subordonnée et soit $A \in \mathcal{M}_n(\mathbf{C})$ inversible. Alors, la boule ouverte $\mathcal{B}(A, \frac{1}{\|A^{-1}\|})$ de $\mathcal{M}_n(\mathbf{C})$, de centre A et de rayon $\frac{1}{\|A^{-1}\|}$:

$$\mathcal{B}(A, \frac{1}{\|A^{-1}\|}) = \left\{ M \in \mathcal{M}_n(\mathbf{C}) ; \|M - A\| < \frac{1}{\|A^{-1}\|} \right\}$$

est contenue dans l'ensemble des matrices inversibles.

En effet, Soit $C \in \mathcal{B}(A, \frac{1}{\|A^{-1}\|})$, alors on a

$$\|A^{-1}(A - C)\| \leq \|A^{-1}\| \|A - C\| < 1$$

Il s'en suit, grâce au théorème 1.2, que $A^{-1}C$ est inversible, donc C est inversible.

L'ensemble des matrices inversibles est donc une partie ouverte de l'espace vectoriel normé $(\mathcal{M}_n(\mathbf{C}), \|\cdot\|)$. ■

1.5 Conditionnement d'un système linéaire

On a vu au §1.1 (exemple 4), que si les données d'un système linéaire ne sont connues qu'à 10^{-5} près, la solution peut être entachée d'une erreur $4,5 \cdot 10^5$ fois supérieure à l'erreur sur les données.

Nous donnons au théorème 1.5 ci-dessous une majoration de l'erreur commise sur la solution d'un système linéaire en fonction d'une erreur éventuelle sur les données. Pour cela, on se donne une matrice inversible A , un vecteur b et leurs perturbations respectives δA et δb , et on va comparer les solutions exactes x et $(x + \delta x)$ des systèmes linéaires

$$(1.30) \quad Ax = b$$

$$(1.31) \quad (A + \delta A)(x + \delta x) = b + \delta b$$

Définition 1.12

Soit $\|\cdot\|$ une norme matricielle subordonnée et A une matrice inversible. Le nombre

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

est appelé le conditionnement de la matrice A , relativement à la norme matricielle considérée. ■

Remarque : Les conditionnements de matrice utilisés dans la pratique correspondent aux trois normes matricielles subordonnées introduites au paragraphe 1.4.3. On les note

$$\text{cond}_p(A) = \|A\|_p \|A^{-1}\|_p, \quad \text{pour } p = 1, 2, \infty.$$

On considèrera, dans toute la suite de ce chapitre, une norme vectorielle $\|\cdot\|$ et la norme matricielle $\|\cdot\|$ qui lui est subordonnée.

Théorème 1.3

Soient A une matrice inversible dans $\mathcal{M}_n(\mathbf{C})$, b et δb deux éléments de $\mathbf{C}^n$ avec $b \neq 0$, et soient x et $x + \delta x$ les solutions des systèmes linéaires :

$$Ax = b \quad , \quad A(x + \delta x) = b + \delta b.$$

On a alors :

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

Démonstration :

Utilisant le fait que x et $x + \delta x$ sont solutions des systèmes linéaires $Ax = b$ et $A(x + \delta x) = b + \delta b$, on obtient :

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

Et par suite, puisque $\|b\| \leq \|A\| \|x\|$, on a :

$$\|\delta x\| \|b\| \leq \|A\| \|x\| \|A^{-1}\| \|\delta b\| = \text{cond}(A) \|x\| \|\delta b\|.$$

D'où finalement :

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

■

Théorème 1.4

Soient A et δA deux matrices de $\mathcal{M}_n(\mathbf{C})$ avec A inversible et $b \in \mathbf{C}^n$. Soient x et $x + \delta x$ les solutions des systèmes linéaires :

$$Ax = b \quad , \quad (A + \delta A)(x + \delta x) = b.$$

On a alors :

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \text{cond}(A) \frac{\|\delta A\|}{\|A\|}$$

Démonstration :

On a $\delta x = -A^{-1}\delta A(x + \delta x)$. D'où :

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

Il s'en suit que

$$\frac{\|\delta x\|}{\|x + \delta x\|} \leq \|A^{-1}\| \|\delta A\| = \text{cond}(A) \frac{\|\delta A\|}{\|A\|},$$

ce qui achève la démonstration. ■

Théorème 1.5

Soient $A \in \mathcal{M}_n(\mathbf{C})$ inversible et $b \in \mathbf{C}^n$. On considère les systèmes linéaires (1.30) et (1.31).

Si

$$\|\delta A\| < \frac{1}{\|A^{-1}\|}$$

alors on a

$$(1.32) \quad \frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \frac{1}{1 - \|\delta A\| \|A^{-1}\|} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

Démonstration :

Soient x la solution de (1.30) et $(x + \delta x)$ la solution de (1.31). On a alors :

$$(1.33) \quad A(I + A^{-1}\delta A)\delta x = \delta b - (\delta A)x$$

Or $\|A^{-1}\delta A\| < 1$, puisque par hypothèse $\|\delta A\| < \frac{1}{\|A^{-1}\|}$, et par suite, grâce au théorème 1.2, la matrice $I + A^{-1}\delta A$ est inversible et

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|}$$

Ainsi, on peut déduire de (1.33) que

$$\delta x = (I + A^{-1}\delta A)^{-1} A^{-1}(\delta b - (\delta A)x)$$

et que :

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \frac{1}{1 - \|A^{-1}\delta A\|} \|A^{-1}\| \left(\frac{\|\delta b\|}{\|x\|} + \|\delta A\| \right) \\ &\leq \|A^{-1}\| \|A\| \frac{1}{1 - \|A^{-1}\| \|\delta A\|} \left(\frac{\|\delta b\|}{\|A\| \|x\|} + \frac{\|\delta A\|}{\|A\|} \right) \\ &\leq (\text{puisque } \|b\| \leq \|A\| \|x\|) \\ &\leq \text{cond}(A) \frac{1}{1 - \|A^{-1}\| \|\delta A\|} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right) \end{aligned}$$

Ce qui prouve (1.32) et achève la démonstration. ■

Propriétés 1.9

Soit $A \in \mathcal{M}_n(\mathbf{C})$ inversible. On a

- 1) $\text{cond}(A) \geq 1$.
- 2) Pour tout $\alpha \in \mathbf{C} \setminus \{0\}$, $\text{cond}(\alpha A) = \text{cond}(A)$.
- 3) $\text{cond}(A) = \text{cond}(A^{-1})$.
- 4) Soit $U \in \mathcal{M}_n(\mathbf{C})$ une matrice unitaire. Alors, $\text{cond}_2(U) = 1$, et $\text{cond}_2(UA) = \text{cond}_2(A)$

Démonstration :

1) Il suffit d'utiliser le fait que pour toute norme matricielle subordonnée $\|\cdot\|$, on a $\|I\| = 1$.

2) La preuve de cette propriété est immédiate.

3) La preuve de cette propriété est immédiate.

4) Soit U une matrice unitaire. Alors, d'après la propriété 1.4, on a $\|U\|_2 = \|U^{-1}\|_2 = 1$, et par suite $\text{cond}_2(U) = 1$.

Par ailleurs, utilisant le fait que $U^{-1} = U^*$, on obtient

$$\|AU\|_2 \equiv \sup_{v \neq 0} \frac{\|AUv\|_2}{\|v\|_2} = \sup_{w \neq 0} \frac{\|Aw\|_2}{\|U^*w\|_2} = \|A\|_2$$

puisque $\|U^*w\|_2 = \|w\|_2$. De même, on démontre que

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

Ce qui prouve que

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

En suivant la même démarche que ci-dessus, on montre que $\text{cond}_2(UA) = \text{cond}_2(A)$. ■

Théorème 1.6

Soit $A \in \mathcal{M}_n(\mathbf{C})$ une matrice hermitienne et inversible. Alors on a

$$\text{cond}_2(A) = \frac{|\lambda_n|}{|\lambda_1|}$$

où $|\lambda_n| = \max_{1 \leq i \leq n} |\lambda_i|$ et $|\lambda_1| = \min_{1 \leq i \leq n} |\lambda_i|$; les λ_i étant les valeurs propres de A .

Démonstration :

Soit A une matrice hermitienne. Alors, d'après la proposition 1.3, on a $\|A\|_2 = \varrho(A) = |\lambda_n|$ et $\|A^{-1}\|_2 = \varrho(A^{-1}) = \frac{1}{|\lambda_1|}$. Ce qui prouve que $\text{cond}_2(A) = \frac{|\lambda_n|}{|\lambda_1|}$. ■

Remarque : La majoration d'erreur (1.32) donnée au théorème 1.5 :

$$(1.34) \quad \frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \frac{1}{1 - \|\delta A\| \|A^{-1}\|} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

est optimale.

En effet, soit A une matrice symétrique et inversible et soient $\lambda_1, \dots, \lambda_n$ ses valeurs propres, avec $|\lambda_n| = \sup |\lambda_i|$ et $|\lambda_1| = \inf |\lambda_i|$. Soient v_n un vecteur propre normé associé à λ_n et v_1 un vecteur propre normé associé à λ_1 .

On considère le système linéaire $Ax = b$, où $b = \lambda_n v_n$, dont la solution est trivialement $x = v_n$, et le système linéaire $(A + \delta A)(x + \delta x) = b + \delta b$, avec $\delta A = O_n$ et $\delta b = t v_1$, où $t \neq 0$. Ainsi, on obtient $A \delta x = t v_1$, puisque $Ax = b$. Il s'en suit que $\delta x = \frac{t}{\lambda_1} v_1$. D'où

$$\frac{\|\delta x\|}{\|x\|} = \frac{|t|}{|\lambda_1|} \frac{\|v_1\|}{\|v_n\|} = \frac{|t|}{|\lambda_1|}$$

puisque $\|v_n\| = \|v_1\| = 1$. De même, on a :

$$\frac{\|\delta b\|}{\|b\|} = \frac{|t|}{|\lambda_n|} \frac{\|v_1\|}{\|v_n\|} = \frac{|t|}{|\lambda_n|}$$

et par suite, pour $\|\cdot\| = \|\cdot\|_2$, on a

$$\text{cond}_2(A) \frac{\|\delta b\|}{\|b\|} = \frac{|\lambda_n|}{|\lambda_1|} \frac{|t|}{|\lambda_n|} = \frac{|t|}{|\lambda_1|} = \frac{\|\delta x\|}{\|x\|}$$

car $\text{cond}_2(A) = \frac{|\lambda_n|}{|\lambda_1|}$.

Ce qui prouve que l'inégalité (1.34) est une égalité dans ce cas.

On déduit de ce qui précède que l'inégalité (1.34) est une inégalité optimale (on ne peut pas l'améliorer puisque l'égalité est réalisée dans certains cas). ■