Analyse numérique

Francis Filbet

Analyse numérique

Algorithme et étude mathématique

2e édition

DUNOD

Illustration de couverture : © vege - Fotolia.com

ISBN 978-2-10-059910-3

TABLE DES MATIÈRES

AVANT-PROPOS

L'objet de cet ouvrage est de fournir au lecteur une présentation et une analyse des méthodes numériques les plus couramment utilisées pour la résolution de problèmes de mathématiques appliquées. Chaque chapitre fait l'objet d'une introduction et d'exemples puisés dans des applications à divers domaines des mathématiques. Il s'agit ici d'illustrer l'utilisation d'outils mathématiques pour des problèmes actuels (modèles économiques, physiques, algorithme de Google, etc).

Cet ouvrage s'adresse tout particulièrement aux étudiants de Licence 3, Master ou préparation à l'agrégation ainsi qu'aux élèves des écoles d'Ingénieurs intéressés par des problèmes issus de la physique, de la biologie ou encore de l'économie, et qui souhaitent avoir une idée des méthodes de bases utilisées aujourd'hui et être en mesure de les mettre en œuvre sur ordinateur.

Nous avons volontiers glissé plusieurs rappels théoriques qui devraient aider le lecteur à mieux appréhender les difficultés mathématiques de certaines démonstrations. Notons toutefois que certains résultats sont présentés dans un cadre simplifié de manière à ne pas noyer le lecteur dans des difficultés techniques.

L'ensemble des sujets abordés ici à déjà largement été traité dans la littérature et ce cours doit beaucoup à d'autres ouvrages et cours dispensés en France, comme les cours d'analyse numérique de Jean-Marie Thomas, Michel Pierre et Abderrahmane Bendali.

Lors de la rédaction de cet ouvrage, de nombreux collègues et étudiants ont bien voulu me faire part de leurs avis, commentaires, suggestions. À ce propos, je tiens tout particulièrement à exprimer ma reconnaissance à Jean-Pierre Bourgade, Nicolas Crouseilles, Thomas Lepoutre, Céline Parzani, Miguel Rodrigues, qui m'ont fait part de leurs conseils éclairés et pour certains ont bien voulu effectuer une relecture de ce document.

Je dédie ce livre à Raphaël et Flavien.

LES SYSTÈMES LINÉAIRES 1

La résolution d'un système linéaire algébrique est au cœur de la plupart des calculs en analyse numérique. Il paraît donc naturel de débuter un cours de calcul scientifique par là. Ici, nous décrivons les algorithmes de résolution les plus populaires qui sont appliqués à des systèmes généraux. Nous considérons le problème suivant : trouver le vecteur x solution de

$$A\,x \;=\; b,$$

où A est une matrice carrée et b un vecteur donné à coefficients réels ou complexes. La discrétisation d'équations différentielles ordinaires ou d'équations aux dérivées partielles, la modélisation de problèmes en physique, chimie ou économie conduit souvent à la résolution de systèmes linéaires de grande taille avec plusieurs milliers d'inconnues et il devient pratiquement impossible de résoudre ces systèmes d'équations sans l'aide d'un ordinateur. Il s'agit alors de trouver des algorithmes de résolution efficaces où le nombre d'opérations et donc le temps de calcul, n'est pas prohibitif. C'est un problème classique mais difficile en analyse numérique.

L'objectif de ce chapitre est de proposer différentes méthodes numériques de résolution des systèmes linéaires et de sensibiliser le lecteur à l'importance du choix de la méthode en fonction des propriétés du système. Nous distinguerons deux types de méthodes : *les méthodes directes* où nous calculons exactement la solution et *les méthodes itératives* où nous calculons une solution approchée.

1.1 EXEMPLE D'UN SYSTÈME LINÉAIRE

Commençons par présenter un problème provenant de l'économie, où nous sommes amenés à la résolution d'un système linéaire.

Exemple d'une analyse de l'offre et de la demande

Imaginons que plusieurs artisans décident de coopérer pour fabriquer différents produits utiles à chacun d'eux. Afin d'éviter le coût du stockage des produits fabriqués, nous recherchons une situation d'équilibre entre l'offre et la demande. Pour cela, considérons n artisans, sachant que chacun fabrique un produit spécifique. Ces artisans ont choisi de coopérer, c'est-à-dire qu'ils souhaitent adapter leur production, d'une part, à leurs propres besoins déterminés par la quantité de produit nécessaire aux autres artisans pour qu'ils puissent fabriquer leur propre produit et, d'autre part, aux besoins du marché.

Puisque chaque artisan fabrique un produit différent, nous notons par $i \in \{1, \ldots, n\}$ le produit fabriqué par un artisan, x_i désigne le nombre total de produits i fabriqués par un artisan et b_i la demande du marché en ce produit. D'autre part $(c_{i,j})_{1 \leqslant i,j \leqslant n}$ exprime la quantité de produit i nécessaire à la confection d'une unité de produit j. En supposant que la relation qui lie les différents produits est linéaire, nous recherchons l'équilibre entre les besoins et la production, c'est-à-dire que nous imposons que la quantité de produit i fabriquée soit égale à la somme des besoins des autres artisans en produit i pour fabriquer leur propre produit et des besoins du marché en i

$$x_i \;=\; \sum_{j=1}^{n} C_{i,j}\, x_j \;+\; b_i, \quad i = 1, \ldots, n$$

ou encore

$$x \;=\; C\, x \;+\; b,$$

où la matrice C est formée par les coefficients $(c_{i,j})_{1 \leqslant i,j \leqslant n}$ tandis que le vecteur b correspond à la demande du marché $b := (b_1, \ldots, b_n)^T$. Par conséquent, la production totale $x = (x_1, \ldots, x_n)^T$ est la solution du système linéaire $A\,x = b$, où la matrice A est donnée par $A = I_n - C$ et I_n est la matrice identité composée de 1 sur sa diagonale et de 0 ailleurs.

Cette modélisation pose plusieurs difficultés mathématiques. Tout d'abord nous nous interrogeons sur les propriétés de la matrice A permettant d'assurer que ce problème a bien une solution. Il vient ensuite la question du calcul de cette solution. En pratique, lorsque le nombre d'artisans considérés devient grand, il n'est plus possible de calculer l'inverse de la matrice A, il faudra donc fournir des algorithmes qui ne nécessitent pas l'inversion de cette matrice. C'est l'objectif de ce premier chapitre.

1.2 Rappels sur les matrices

Avant de s'intéresser à la résolution numérique de systèmes linéaires, présentons quelques rappels d'algèbre linéaire.

Par la suite, nous noterons $\mathscr{M}_{m,n}(\mathbb{K})$ l'ensemble des matrices à m lignes, n colonnes et à coefficient dans le corps des réels $\mathbb{K} = \mathbb{R}$ ou des complexes $\mathbb{K} = \mathbb{C}$. Pour un vecteur colonne $u \in \mathbb{K}^n$, la valeur u_i avec $1 \leqslant i \leqslant n$, désigne la i-ème composante dans la base canonique de $\mathbb{K}^n$ du vecteur u. Nous appelons *adjoint* du vecteur colonne u, le vecteur ligne $u^\star$ de $\mathbb{K}^n$ tel que $u^\star = (\overline{u}_1, \ldots, \overline{u}_n)$ où $\overline{u}_i$ désigne le conjugué de u_i et transposé de u est le vecteur ligne $u^T = (u_1, \ldots, u_n)$.

Rappelons également le produit matrice-vecteur. Soient A une matrice à m lignes et n colonnes et u un vecteur de $\mathbb{K}^n$, nous définissons $v = A\,u \in \mathbb{K}^m$ le vecteur dont les composantes sont données par

$$v_i = \sum_{j=1}^{n} a_{i,j}\, u_j, \quad i = 1, \ldots, m$$

et pour B une matrice à n lignes et p colonnes, le produit $A\,B$ fournit une matrice $C \in \mathscr{M}_{m,p}$ donnée par

$$c_{i,j} = \sum_{k=1}^{n} a_{i,k}\, b_{k,j}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, p.$$

Introduisons également la notion de matrice adjointe et transposée.

Définition 1.1

Soit $A \in \mathscr{M}_{m,n}(\mathbb{K})$, la matrice adjointe de A, notée $A^\star$, à n lignes et m colonnes, est donnée par $A^\star = (a^\star_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant m}}$ et $a^\star_{i,j} = \overline{a}_{j,i}$, pour $1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant m$.

La matrice transposée de A, notée A^T, à n lignes et m colonnes, est donnée par $A^T = (a^T_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant m}}$ et $a^T_{i,j} = a_{j,i}$, pour $1 \leqslant i \leqslant n$ et $1 \leqslant j \leqslant m$.

Dans la base canonique de $\mathbb{K}^n$, définissons le produit scalaire ou produit hermitien entre deux vecteurs u et $v \in \mathbb{K}^n$, le scalaire de $\mathbb{K}$ donné par

$$\langle u, v\rangle = u^\star\, v = \sum_{i=1}^{n} u_i^\star\, v_i.$$

1.2.1 Cas des matrices carrées

Considérons maintenant le cas particulier des matrices carrées, pour lesquelles le nombre de lignes est égal au nombre de colonnes. Nous rappelons les définitions suivantes :

Définition 1.2

Une matrice carrée $A \in \mathscr{M}_{n,n}(\mathbb{K})$ est dite inversible, ou régulière, ou encore non singulière, s'il existe une matrice $B \in \mathscr{M}_{n,n}(\mathbb{K})$ telle que $A\,B = B\,A = I_n$. Dans ce cas, la matrice B est unique et s'appelle la matrice inverse de A, elle est notée A^{-1}.

Définition 1.3

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$, alors

- A est une matrice normale si $A^\star\, A = A\, A^\star$.
- A est une matrice unitaire si $A^\star\, A = A\, A^\star = I_n$, où I_n désigne la matrice identité. Dans le cas où $\mathbb{K} = \mathbb{R}$, nous parlons de matrice orthogonale et $A^T\, A = A\, A^T = I_n$, c'est-à-dire $A^T = A^{-1}$.
- A est une matrice hermitienne si $A^\star = A$. Dans le cas où $\mathbb{K} = \mathbb{R}$, nous parlons de matrice symétrique et $A^T = A$.

Il est alors facile de faire le lien entre les matrices hermitiennes et normales.

Proposition 1.4

Toute matrice hermitienne est une matrice normale.

Comme nous l'avons vu en début de chapitre, l'objectif de cette partie est de mettre au point des algorithmes de résolution numérique pour un système de la forme

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

où $A \in \mathscr{M}_{n,n}(\mathbb{K})$ et $b \in \mathbb{K}^n$ sont donnés et $x \in \mathbb{K}^n$ est l'inconnue. Nous donnons d'abord une condition nécessaire et suffisante sur la matrice A pour que ce système admette une solution unique. Le système linéaire (1.1) a une solution unique dès que la matrice A est inversible et cette solution est donnée par $x \;=\; A^{-1}\,b$. Il est alors important de connaître quelques propriétés des matrices inversibles [16].

Proposition 1.5

Soient A, $B \in \mathscr{M}_{n,n}(\mathbb{K})$ inversibles, nous avons alors

- pour tout $\alpha \in \mathbb{K}^*$, $(\alpha\,A)^{-1} \;=\; \dfrac{1}{\alpha}\,A^{-1}$.
- $(A\,B)^{-1} \;=\; B^{-1}\,A^{-1}$.
- $(A^\star)^{-1} = (A^{-1})^\star$.

Énonçons quelques propriétés des matrices inversibles, qui sont autant d'outils permettant de s'assurer que le problème (1.1) admet bien une solution [16].

Théorème 1.6 ***Théorème des matrices inversibles***

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$, les propositions suivantes sont équivalentes :

- A est inversible,
- le déterminant de A est non nul,
- le rang de A est égal à n, c'est-à-dire que les n vecteurs colonnes forment une famille libre,
- le système homogène $A\,x \;=\; 0$ a pour unique solution $x \;=\; 0$,
- pour tout b dans $\mathbb{K}^n$, le système linéaire $A\,x \;=\; b$ a exactement une solution.

Pour conclure cette partie, rappelons qu'une valeur propre de A est donnée par $\lambda \in \mathbb{K}$ telle que $\det(A \;-\; \lambda\,I_n) \;=\; 0$. Ainsi, il existe au moins un vecteur v non nul, dit *vecteur propre*, vérifiant $A\,v \;=\; \lambda\,v$. Définissons alors le spectre d'une matrice.

Définition 1.7

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$. Nous appelons spectre de A l'ensemble des valeurs propres de A :

$$\mathrm{Sp}(A) = \{\lambda \in \mathbb{K}; \quad \exists v \in \mathbb{K}^n \quad v \neq 0 \quad Av = \lambda v\}$$

Nous appelons rayon spectral de A le nombre réel positif $\rho(A)$ tel que

$$\rho(A) = \max\{|\lambda|, \lambda \in \mathbb{C}, v \in \mathbb{C} \setminus \{0\} Av = \lambda v\}$$

À partir de la définition d'une valeur propre, nous vérifions facilement qu'une matrice est inversible si et seulement si 0 n'est pas valeur propre.

La conséquence de ces propriétés est que l'ensemble des matrices carrées inversibles forme un groupe, appelé le groupe linéaire et noté habituellement $Gl_n(\mathbb{K})$. En général, « presque toutes » les matrices sont inversibles. Sur le corps $\mathbb{K}$, cela peut être formulé de façon plus précise : l'ensemble des matrices non inversibles, considéré comme sous-ensemble de $\mathscr{M}_{n,n}(\mathbb{K})$, est un ensemble négligeable, c'est-à-dire de mesure de Lebesgue nulle. Intuitivement, cela signifie que si vous choisissez au hasard une matrice carrée à coefficients réels, la probabilité pour qu'elle soit non inversible est égale à zéro. La raison est que des matrices non inversibles peuvent être considérées comme racines d'une fonction polynôme donnée par le déterminant [16].

Par la suite, nous nous intéresserons au calcul proprement dit de la solution de (1.1). Proposons d'abord un premier algorithme naïf qui permet de calculer la solution d'un système linéaire ; cet algorithme est dû à Cramer[1] [23].

Théorème 1.8 ***Résolution de Cramer***

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice inversible. Alors la solution du système $Ax = b$ est donnée par $x_i = \det(A_i)/\det(A)$, pour tout $i = 1, \ldots, n$, où A_i est la matrice A pour laquelle la i-ème colonne est remplacée par le vecteur b.

Cette méthode bien que très élégante est très coûteuse puisqu'elle nécessite plus de $n!$ opérations où $n!$ est la factorielle de n, donnée par $n! = n \times (n-1) \times \ldots \times 2 \times 1$. Elle n'est donc jamais utilisée en pratique sauf en dimension $n = 2$. Cet algorithme revient à calculer explicitement la matrice A^{-1}. Hélas, ce calcul est souvent long et fastidieux, même pour un ordinateur, c'est pourquoi nous avons recours à des algorithmes de résolution exacte d'une complexité moindre (nous parlons alors d'une méthode directe), ou des méthodes itératives qui consistent à construire une suite de solutions approchées qui converge vers la solution exacte. Avant de présenter de tels algorithmes, nous introduirons des matrices dont la structure particulière est bien adaptée à la résolution du système linéaire (1.1). Puis nous verrons comment se ramener à ces cas particuliers.

1. En référence au mathématicien suisse Gabriel Cramer (1704-1752).

1.2.2 Quelques matrices particulières

Cette partie constitue la base théorique permettant de mettre au point des méthodes pour la résolution exacte de (1.1).

- *Matrices triangulaires*

Nous introduisons la notion de matrice triangulaire et montrons que pour les matrices de ce type, la résolution du système linéaire (1.1) devient très facile.

Définition 1.9

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$,

- la matrice A est une matrice triangulaire inférieure si $A = (a_{i,j})_{1 \leqslant i,j \leqslant n}$ avec $a_{i,j} = 0, \quad 1 \leqslant i < j \leqslant n$.
- A est une matrice triangulaire supérieure lorsque $a_{i,j} = 0, 1 \leqslant j < i \leqslant n$.
- A est une matrice diagonale dès lors que $a_{i,j} = 0$ pour $i \neq j$.

Nous vérifions d'abord que l'ensemble des matrices triangulaires supérieures (resp. inférieures) est stable par la somme et le produit. En outre, nous avons le résultat suivant.

Proposition 1.10

Soit $L \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice triangulaire inférieure inversible, ce qui signifie que tous les éléments diagonaux sont non nuls. Alors L^{-1} est aussi une matrice triangulaire inférieure.
Soit U une matrice triangulaire supérieure inversible. Alors U^{-1} est aussi une matrice triangulaire supérieure.

Les matrices triangulaires jouent un rôle important en analyse numérique car elles sont facilement inversibles ou du moins, nous pouvons facilement trouver la solution $x \in \mathbb{K}^n$ du système linéaire $A\,x = b$. En effet, considérons le cas d'une matrice triangulaire supérieure, alors la solution x se calcule par un algorithme dit de remontée. Nous observons d'abord qu'une matrice triangulaire inversible a tous ses éléments diagonaux non nuls, c'est pourquoi nous pouvons écrire $x_n = b_n/a_{n,n}$ puis pour tout $i = n-1, n-2, \ldots, 1$, nous pouvons calculer x_i de la manière suivante :

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

Exemple

Soit $A \in \mathscr{M}_{3,3}(\mathbb{R})$ donnée par

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 2 \\ 0 & 0 & 3 \end{pmatrix},$$

nous cherchons une solution de $Ax = b$ avec $b = (3,4,3)^T$. En commençant par la dernière ligne, nous trouvons $x_3 = 1$. Puis en injectant cette valeur dans l'avant-dernière équation, cela donne $x_2 = 1$. Finalement, connaissant x_2 et x_3, la première équation donne directement $x_1 = 1$.

Nous allons voir par la suite qu'une méthode directe calcule la solution exacte de (1.1) et consiste le plus souvent à trouver un système triangulaire équivalent.

C'est pourquoi nous énonçons le Théorème de Shur [7] qui sera justement utile pour rendre certaines matrices triangulaires par simple changement de base.

Théorème 1.11 ***Théorème de Shur***

Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice quelconque. Alors il existe une matrice unitaire $U \in \mathcal{M}_{n,n}(\mathbb{K})$, c'est-à-dire $U^\star U = I_n$ et une matrice $T \in \mathcal{M}_{n,n}(\mathbb{K})$ triangulaire dont la diagonale est composée par l'ensemble des valeurs propres de A telles que $T = U^\star A U$.

- *Matrices hermitiennes*

Dans le cas des matrices hermitiennes, ce dernier résultat peut être sensiblement amélioré.

Corollaire 1.12

Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice hermitienne. Alors il existe une matrice unitaire $U \in \mathcal{M}_{n,n}(\mathbb{K})$ et une matrice diagonale $D \in \mathcal{M}_{n,n}(\mathbb{K})$ dont la diagonale est composée par l'ensemble des valeurs propres de A, telles que $D = U^\star A U$.

Nous rappelons enfin quelques propriétés de base des matrices hermitiennes, définies positives. Nous introduisons d'abord la notion de sous-matrice principale.

Définition 1.13

Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$ une matrice quelconque. Nous appelons sous-matrice principale d'ordre i (pour $1 \leqslant i \leqslant n$) de A, la matrice $A_i \in \mathcal{M}_{i,i}(\mathbb{K})$ obtenue en ne gardant que les i premières lignes et les i premières colonnes de A.

Définition 1.14

Soit $A \in \mathcal{M}_{n,n}(\mathbb{K})$, nous disons que A est positive si pour tout $x \in \mathbb{K}^n$, elle vérifie $x^T A x \geqslant 0$ et A est définie positive si pour tout $x \in \mathbb{K}^n \setminus \{0\}$, elle vérifie $x^T A x > 0$ et $x^T A x = 0$ implique $x = 0$.

Nous avons alors les résultats suivants dont la preuve est laissée en exercice.

Proposition 1.15

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice hermitienne définie positive. Alors elle vérifie

- toute sous-matrice principale A_i, $1 \leqslant i \leqslant n$ est hermitienne définie positive ;
- tous les coefficients diagonaux de A sont des réels strictement positifs ;
- A est diagonalisable et ses valeurs propres sont strictement positives ;
- le déterminant de A est strictement positif, c'est-à-dire que A est inversible ;
- il existe une constante $\alpha > 0$, telle que $x^\star A\, x \geqslant \alpha \|x\|^2$, pour n'importe qu'elle norme vectorielle $\|.\|$.

- *Matrices de permutations*

Par la suite, nous utiliserons souvent les matrices de permutations pour réordonner les lignes ou les colonnes d'une matrice quelconque.

Définition 1.16

Une matrice de permutations est une matrice carrée qui ne possède que des 0 et des 1 comme coefficients, telle qu'il y ait un seul 1 par ligne et par colonne.

Une matrice de permutations vérifie alors

Proposition 1.17

Soient σ et τ deux permutations des indices $\{1, \ldots, n\}$, nous avons alors

- la matrice de permutations P_σ correspondant à la permutation σ s'écrit comme

$$(P_\sigma)_{i,j} = \delta_{i,\sigma(j)} = \begin{cases} 1, & \text{si } i = \sigma(j), \\ 0, & \text{sinon.} \end{cases}$$

- $P_\sigma\, P_\tau = P_{\sigma\circ\tau}$.
- $(P_\sigma)^T\, P_\sigma = I_n$, c'est une matrice orthogonale.

1.2.3 Conditionnement de matrices

Avant de décrire des algorithmes de résolution de systèmes linéaires (1.1), nous mettons en évidence une difficulté supplémentaire : la sensibilité de la solution par rapport aux perturbations des données $b \in \mathbb{K}^n$ ou des coefficients de la matrice A. En effet, considérons par exemple la matrice de Hilbert donnée par

$$a_{i,j} = \frac{1}{i + j - 1}, \quad i, j = 1, \ldots, n.$$

Nous avons pour $n = 4$,

$$\begin{pmatrix} 1 & 1/2 & 1/3 & 1/4 \\ 1/2 & 1/3 & 1/4 & 1/5 \\ 1/3 & 1/4 & 1/5 & 1/6 \\ 1/4 & 1/5 & 1/6 & 1/7 \end{pmatrix}.$$

Ainsi, pour un vecteur donné $b \in \mathbb{R}^4$ tel que

$$b^T = \left(\frac{25}{12}, \frac{77}{60}, \frac{57}{60}, \frac{319}{420}\right) \simeq (2,08\,,\ 1,28\,,\ 0,95\,,\ 0,76)\,,$$

nous calculons $x = A^{-1}\,b$ et vérifions facilement que $x^T = (1,1,1,1)$. Ensuite, si nous modifions légèrement la source $\tilde{b}^T = (2{,}1\,,\ 1{,}3\,,\ 1\,,\ 0{,}8)$, la solution $\tilde{x}$ du système $\tilde{x} = A^{-1}\tilde{b}$ est donnée par $\tilde{x}^T = (5{,}6\,,\ -48\,,\ 114\,,\ -70)$, ce qui est très loin de la solution de départ x.

En définitive, nous constatons qu'une petite perturbation de la donnée b conduit à une grande modification de la solution x, ce qui engendre des instabilités lors de la résolution du système.

- *Norme matricielle*

Pour mesurer l'ampleur de cette instabilité, nous introduisons la notion de *norme matricielle*. En effet, l'ensemble $\mathscr{M}_{n,n}(\mathbb{K})$ peut être considéré comme étant un $\mathbb{K}$-espace vectoriel muni d'une norme $\|.\|$.

Définition 1.18

Nous appelons *norme matricielle* toute application $\|.\|$ de $\mathscr{M}_{n,n}(\mathbb{K})$ à valeur dans $\mathbb{R}^+ := [0, +\infty[$ qui vérifie les propriétés suivantes :

- pour toute matrice $A \in \mathscr{M}_{n,n}(\mathbb{K})$, $\|A\| = 0 \Rightarrow A = 0_{\mathbb{K}^{n\times n}}$,
- pour toute matrice $A \in \mathscr{M}_{n,n}(\mathbb{K})$, et pour tout $\alpha \in \mathbb{K}$, $\|\alpha A\| = |\alpha|\,\|A\|$,
- pour toutes matrices A et $B \in \mathscr{M}_{n,n}(\mathbb{K})$, $\|A + B\| \leqslant \|A\| + \|B\|$, c'est l'inégalité triangulaire,
- pour toutes matrices A et $B \in \mathscr{M}_{n,n}(\mathbb{K})$, $\|A\,B\| \leqslant \|A\|\,.\,\|B\|$.

Notons bien que définir une norme matricielle requiert une condition supplémentaire par rapport à la définition d'une norme vectorielle (la dernière propriété de la définition). Il n'est donc pas évident *a priori* de pouvoir construire une telle application seulement à partir d'une norme vectorielle. À titre d'exemple la norme de Froebenus $\|.\|_F$ donnée par

$$\|A\|_F = \left(\sum_{i,j=1}^{n} |a_{i,j}|^2\right)^{1/2} \tag{1.2}$$

est bien une norme matricielle. Les trois premières propriétés sont connues puisque $\|.\|_F$ est une norme vectorielle de $\mathbb{K}^{n^2}$. Il reste à démontrer que pour toutes matrices A et $B \in \mathscr{M}_{n,n}(\mathbb{K})$, $\|A\,B\|_F \leqslant \|A\|_F\,\|B\|_F$. En effet,

$$\|A\,B\|_F^2 = \|C\|_F^2 = \sum_{i,j=1}^{n} c_{i,j}^2 = \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} a_{i,k}\,b_{k,j} \right)^2 .$$

En utilisant l'inégalité de Cauchy-Schwarz, nous avons

$$\begin{aligned} \|A\,B\|_F^2 \leqslant \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} a_{i,k}^2 \right) \left(\sum_{k=1}^{n} b_{k,j}^2 \right) &= \left(\sum_{i,k=1}^{n} a_{i,k}^2 \right) \left(\sum_{k,j=1}^{n} b_{k,j}^2 \right), \\ &= \|A\|_F^2\,\|B\|_F^2. \end{aligned}$$

Pour construire une norme matricielle à partir d'une norme vectorielle quelconque, nous introduisons la définition suivante :

Définition 1.19

Soit $\|.\|$ une norme vectorielle sur $\mathbb{K}^n$. Nous définissons la norme matricielle $\|.\|$ subordonnée à la norme vectorielle $\|.\|$ comme étant l'application donnée par

$$A \in \mathscr{M}_{n,n}(\mathbb{K}) \mapsto \|A\| := \sup_{\substack{v \in \mathbb{K}^n \\ v \neq 0}} \frac{\|A\,v\|}{\|v\|}.$$

Nous vérifions facilement que cette application définit bien une norme matricielle.

Par exemple pour $1 \leqslant p \leqslant \infty$, nous savons que l'application qui à $v \in \mathbb{K}^n$ fait correspondre le réel positif

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

ou $\|v\|_\infty = \max_{1 \leqslant i \leqslant n} |v_i|$ pour $p = \infty$, est une norme vectorielle sur $\mathbb{K}^n$. Posons alors pour $p \in [1, +\infty]$ et $A \in \mathscr{M}_{n,n}(\mathbb{K})$

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

qui est une norme matricielle subordonnée à la norme vectorielle $\|.\|_p$. En général nous ne pouvons pas calculer $\|A\|_p$ directement en fonction de $(a_{i,j})_{1 \leqslant i,j \leqslant n}$ sauf pour $p = 1$ et $p = \infty$.

Proposition 1.20 ***Caractérisation de*** $\|.\|_1$ ***et*** $\|.\|_\infty$

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice quelconque. Alors nous avons

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

et

$$\|A\|_\infty = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{i,j}|.$$

DÉMONSTRATION. Soit $v \in \mathbb{K}^n$. D'une part, nous avons

$$\begin{aligned} \|A\,v\|_1 &= \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{i,j}\, v_j \right| \leqslant \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{i,j}|\,|v_j|, \\ &\leqslant \sum_{j=1}^{n} \left(\sum_{i=1}^{n} |a_{i,j}| \right) |v_j| \leqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}| \|v\|_1. \end{aligned}$$

D'où

$$\|A\|_1 \leqslant \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|.$$

D'autre part, nous montrons qu'il existe $w \in \mathbb{K}^n$ tel que $\|w\|_1 = 1$ et

$$\|A\,w\|_1 = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|.$$

En effet, choisissons $j_0 \in \{1, \ldots, n\}$ tel que

$$\sum_{i=1}^{n} |a_{i,j_0}| = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|,$$

puis prenons $w \in \mathbb{K}^n$ tel que $w_i = 0$ pour $i \neq j_0$ et $w_{j_0} = 1$. Alors

$$\|A\,w\|_1 = \sum_{i=1}^{n} |(A\,w)_i| = \sum_{i=1}^{n} |a_{i,j_0}| = \max_{1\leqslant j\leqslant n} \sum_{i=1}^{n} |a_{i,j}|.$$

Par un raisonnement analogue, nous montrons le résultat pour la norme matricielle $\|.\|_\infty$. □

Dans le cas particulier de la norme subordonnée à la norme euclidienne de $\mathbb{K}^n$ définie pour $A \in \mathscr{M}_{n,n}(\mathbb{K})$ par

$$\|A\|_2 \;=\; \sup_{\substack{v\in\mathbb{K}^n\\ v\neq 0}} \frac{\|A\,v\|_2}{\|v\|_2},$$

nous démontrons le résultat suivant.

Proposition 1.21

Pour $A \in \mathscr{M}_{n,n}(\mathbb{K})$, nous avons $\|A\|_2 \;=\; \sqrt{\rho(A^\star\,A)} \;=\; \sqrt{\rho(A\,A^\star)}$.

DÉMONSTRATION. Prenons $\mathbb{K} = \mathbb{C}$ ou $\mathbb{R}$, par définition de la norme subordonnée à la norme euclidienne, il vient

$$\begin{aligned}\|A\|_2^2 \;&=\; \sup_{\substack{v\in\mathbb{K}^n\\ v\neq 0}} \frac{\|A\,v\|_2^2}{\|v\|_2^2} \;=\; \sup_{\substack{v\in\mathbb{K}^n\\ v\neq 0}} \frac{(A\,v)^\star\,(A\,v)}{v^\star\,v},\\ &=\; \sup_{\substack{v\in\mathbb{K}^n\\ v\neq 0}} \frac{v^\star\,A^\star\,A\,v}{v^\star\,v}.\end{aligned}$$

Or, nous vérifions que

$$(A^\star\,A)^\star \;=\; A^\star\,(A^\star)^\star \;=\; A^\star\,A.$$

Ainsi, d'après le Corollaire 1.12 puisque la matrice $A^\star\,A$ est hermitienne, elle est diagonalisable et il existe $U \in \mathscr{M}_{n,n}(\mathbb{K})$ unitaire telle que $U\,A^\star\,A\,U = \mathrm{diag}(\mu_k)$, où $\mathrm{diag}(\mu_k)$ représente une matrice diagonale formée à partir des valeurs μ_k sur la diagonale. Les scalaires $(\mu_k)_{1\leqslant k\leqslant n}$ sont les valeurs propres de la matrice hermitienne $A^\star\,A$. Notons que $\mu_k \geqslant 0$ puisque de la relation $A^\star\,Ap_k = \mu_k\,p_k$, nous déduisons que $(A\,p_k)^\star\,Ap_k = \mu_k\,p_k^\star\,p_k$, c'est-à-dire $\mu_k = \|A\,p_k\|_2^2/\|p_k\|_2^2 \geqslant 0$. Il vient ensuite

$$\sup_{\substack{v\in\mathbb{K}^n\\ v\neq 0}} \frac{v^\star\,A^\star\,A\,v}{v^\star\,v} \;=\; \sup_{\substack{v\in\mathbb{K}^n\\ v\neq 0}} \frac{v^\star\,U^\star\,U\,A^\star\,A\,U^\star\,U\,v}{v^\star\,U^\star\,U\,v}.$$

Puis, comme U est inversible, nous avons par changement de variable $w = U\,v$

$$\|A\|_2^2 \;=\; \sup_{\substack{w\in\mathbb{K}^n\\ w\neq 0}} \frac{w^\star\,U\,A^\star\,A\,U^\star\,w}{w^\star\,w} \;=\; \sup_{\substack{w\in\mathbb{K}^n\\ w\neq 0}} \frac{\sum_{k=1}^n \mu_k|w_k|^2}{\sum_{k=1}^n |w_k|^2},$$

où les μ_k sont les valeurs propres de $A^\star\,A$. Enfin, en prenant le vecteur de la base canonique dont toutes les composantes sont nulles exceptée la k_0-ème qui correspond à la plus grande valeur propre μ_{k_0} en module, nous obtenons $\|A\|_2^2 = \rho(A\,A^\star)$. □

• *Conditionnement d'une matrice*

Essayons maintenant de comprendre de manière plus générale le phénomène d'instabilité par rapport à la donnée $b \in \mathbb{K}^n$. Soient $A \in \mathscr{M}_{n,n}(\mathbb{K})$ inversible et b un vecteur de $\mathbb{K}^n$, non nul. Nous désignons par x la solution du système $A\,x = b$ et pour une perturbation δb du vecteur b, le vecteur $x+\delta x$ est la solution de $A\,(x+\delta x) = b+\delta b$.

Pour une norme vectorielle $\|.\|$ de $\mathbb{K}^n$, nous cherchons à contrôler l'erreur relative $\|\delta x\| \,/\, \|x\|$ en fonction de l'erreur relative $\|\delta b\| \,/\, \|b\|$ et de la norme matricielle subordonnée $\|A\|$.

Par linéarité et puisque A est inversible, nous avons d'une part $A\,\delta x = \delta b \Rightarrow \|\delta x\| \leqslant \|A^{-1}\|\,\|\delta b\|$ et d'autre part $A\,x = b$ implique que $\|b\| \leqslant \|A\|\,\|x\|$, ou encore de manière équivalente puisque b est non nul

$$\frac{1}{\|x\|} \leqslant \|A\|\,\frac{1}{\|b\|}.$$

Ainsi, nous obtenons l'estimation suivante :

$$\frac{\|\delta x\|}{\|x\|} \leqslant \|A^{-1}\|\,\|A\|\,\frac{\|\delta b\|}{\|b\|}$$

et proposons la définition suivante :

Définition 1.22

Nous appelons conditionnement de la matrice A relativement à la norme matricielle $\|.\|$ subordonnée à la norme vectorielle $\|.\|$, le nombre

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

Observons bien que le conditionnement sert à mesurer la sensibilité du système aux perturbations de b et de A. Il permet d'assurer le contrôle de $\|\delta x\|$ en fonction de la perturbation $\|\delta b\|$.

Définition 1.23

Soit $A \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice inversible. Nous disons que

- un système linéaire est bien conditionné, si $\mathrm{cond}(A)$ n'est pas trop grand par rapport à un, ce qui correspond au conditionnement de l'identité ;
- un système linéaire est mal conditionné, si $\mathrm{cond}(A)$ est grand par rapport à un.

Vérifions sur l'exemple précédent la cohérence de cette définition. Nous avons pour la norme $\|.\|_1$, $\mathrm{cond}_1(A) = \|A\|_1\,\|A^{-1}\|_1 \simeq 28\,375 \gg 1$ ou pour la norme $\|.\|_2$, $\mathrm{cond}_2(A) = \|A\|_2\,\|A^{-1}\|_2 \simeq 15\,514 \gg 1$. Les valeurs du conditionnement de la matrice A semblent être assez élevées indépendamment de la norme choisie.

Pour conclure cette partie, nous donnons quelques propriétés sur le conditionnement.

Proposition 1.24

Soient $A \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice carrée inversible et $\|.\|$ une norme matricielle subordonnée à la norme vectorielle $\|.\|$. Alors nous avons

- $\text{cond}(A^{-1}) = \text{cond}(A)$.
- Soit $\alpha \in \mathbb{K} \setminus \{0\}$, $\text{cond}(\alpha\, A) = \text{cond}(A)$.
- $\text{cond}(I_n) = 1$.
- $\text{cond}(A) \geqslant 1$.

L'inconvénient de la définition du conditionnement est qu'il fait apparaître $\|A^{-1}\|$, lequel n'est pas facile à calculer d'autant plus que nous ne connaissons pas la forme explicite de la matrice A^{-1}. Dans le cas particulier d'une matrice hermitienne et pour la norme matricielle $\|.\|_2$, nous avons néanmoins le résultat suivant qui ne nécessite pas le calcul de A^{-1} mais seulement la connaissance des valeurs propres de A.

Proposition 1.25

Soit A une matrice hermitienne ($A^\star = A$). Alors $\|A\|_2 = \rho(A)$. De plus, si A est une matrice hermitienne inversible et $(\lambda_i)_{1\leqslant i\leqslant n}$ ses valeurs propres. Alors,

$$\text{cond}_2(A) = \frac{\max\{|\lambda_i|, i = 1, \ldots, n\}}{\min\{|\lambda_i|, i = 1, \ldots, n\}}.$$

DÉMONSTRATION. Appliquons la Proposition 1.21 et puisque A est hermitienne, nous montrons que $\|A\|_2^2 = \rho(A)^2$.
Supposons ensuite que A est une matrice hermitienne inversible. Pour calculer $\text{cond}_2(A)$, il suffit de remarquer que $1/\lambda_i$, où $\lambda_i \neq 0$, est valeur propre de A^{-1} et donc en appliquant le résultat précédent à A^{-1} (qui est également une matrice hermitienne), nous avons

$$\|A^{-1}\|_2 = \rho(A^{-1}) = \frac{1}{\min\{|\lambda_i|, i = 1, \ldots, n\}}.$$

Nous en déduisons le résultat

$$\text{cond}_2(A) = \|A\|_2\, \|A^{-1}\|_2 = \frac{\max\{|\lambda_i|, i = 1, \ldots, n\}}{\min\{|\lambda_i|, i = 1, \ldots, n\}}. \qquad \square$$

• *Préconditionnement d'un système linéaire*

Pour remédier au problème du mauvais conditionnement d'une matrice, nous pouvons appliquer une méthode de préconditionnement. En effet, en vue de résoudre $Ax = b$ nous multiplions ce système d'équation à gauche par une matrice inversible P, il vient alors $P\,A\,x = P\,b$, avec P choisie de manière à ce que la matrice $P\,A$ soit mieux conditionnée que A (dans le cas le plus favorable, nous aurions $P = A^{-1}$). Cependant, il n'y a pas de méthode standard pour trouver la matrice P, le plus souvent nous chercherons une matrice à la fois facile à inverser et « assez proche » de A^{-1}.

Présentons maintenant les deux types de méthodes pour la résolution d'un système linéaire : les méthodes directes et les méthodes itératives.

1.3 MÉTHODES DIRECTES

1.3.1 Méthodologie générale

Soient $A \in \mathscr{M}_{n,n}(\mathbb{K})$ une matrice inversible et un vecteur $b \in \mathbb{K}^n$, nous recherchons $x \in \mathbb{K}^n$ solution de $A\,x = b$. Pour cela, construisons des matrices M et $N \in \mathscr{M}_{n,n}(\mathbb{K})$ telles que $A = M\,N$, où M est facile à inverser (triangulaire ou unitaire) et N triangulaire. Le système s'écrit alors $N\,x = M^{-1}\,b$, que nous résolvons en deux étapes de la manière suivante :

$$\begin{cases} \text{trouver } y \in \mathbb{K}^n \text{ tel que } M\,y = b, \\ \\ \text{trouver } x \in \mathbb{K}^n \text{ tel que } N\,x = y. \end{cases}$$

1.3.2 Méthode de Gauss avec et sans pivot

Les méthodes directes permettent de calculer la solution exacte du problème (1.1) en un nombre fini d'étapes (en l'absence d'erreurs d'arrondi). La méthode directe la plus classique est la méthode d'élimination de Gauss ou Gauss-Jordan, qui consiste à décomposer la matrice A comme le produit $L\,U$ où L est une matrice triangulaire inférieure et U une matrice triangulaire supérieure[1].

L'élimination de Gauss ou l'élimination de Gauss-Jordan est un algorithme d'algèbre linéaire pour déterminer les solutions d'un système d'équations linéaires, pour déterminer le rang d'une matrice ou pour calculer l'inverse d'une matrice carrée inversible.

1. Cette méthode fut nommée d'après Carl Friedrich Gauss, mathématicien allemand (1777-1855) surnommé le prince des mathématiciens. Cependant, si l'on se réfère au livre chinois « Les neuf chapitres sur l'art du calcul » sa naissance remonte à la dynastie Han au premier siècle de notre ère. Elle constitue le huitième chapitre de ce livre sous le titre de la « disposition rectangulaire » et est présentée au moyen de dix-huit exercices [6].