Chapitre 3

Résolution de systèmes d'équations algébriques

Sommaire

Les systèmes d'équations algébriques jouent un rôle important dans les sciences d'ingénierie et peuvent être classés en deux grandes familles : les systèmes linéaires et les systèmes non linéaires.
La taille considérable de ces systèmes, exige la mise en place de méthodes efficaces qui minimisent le temps de calcul et l'espace mémoire requis pour leur résolution sur ordinateur.
Dans ce chapitre, nous allons aborder les principales méthodes de résolution de systèmes linéaires, à savoir la méthode d'élimination de Gauss et la décomposition LU. L'effet des erreurs dues à l'arithmétique flottante sera également étudié et nous introduirons le concept de conditionnement d'une matrice. Nous verrons ensuite, comment linéariser les systèmes non linéaires afin de pouvoir appliquer les diverses techniques de résolution des systèmes linéaires.

1 Introduction

• 1.1 - Systèmes linéaires

Un système d'équations linéaires est un ensemble d'équations dont les inconnues notées $x_1, x_2, \cdots, x_n$, sont en relations linéaires à travers chacune de ses équations

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n = b_3 \\ \qquad\qquad \vdots \qquad\qquad \vdots \qquad\quad \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n = b_n \end{cases} \tag{3.1}$$

La notation matricielle du système (3.1) est plus pratique du fait de sa compacité, et s'écrit comme suit

$$A\vec{x} = \vec{b}, \tag{3.2}$$

où la matrice A et les vecteurs $\vec{x}$, $\vec{b}$ sont définis par

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}, \quad \vec{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \text{et} \quad \vec{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Ici la matrice A et le vecteur *membre de droite* $\vec{b}$ sont connus, et la seule inconnue est le vecteur $\vec{x}$.

Remarque(s) :

- Nous traitons des matrices non singulières ou inversibles (c'est-à-dire dont la matrice inverse existe). Ce qui assure l'existence d'une solution
- Ainsi, la solution de l'équation (3.2) peut s'écrire $\vec{x} = A^{-1}\vec{b}$. Sauf que calculer l'inverse d'une matrice, A^{-1}, est plus difficile et plus long que la résolution.

Question: Comment résoudre de manière automatique et à des difficultés moindre de tel systèmes ?

Exemple 1

Considérons le système linéaire suivant, présenté sous ses formes matricielle et d'équations linéaires respectivement

$$\begin{bmatrix} 2 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 11 \end{bmatrix} \longleftrightarrow \begin{array}{l} 2x_1 + 3x_2 = 8 \\ 3x_1 + 4x_2 = 11 \end{array}$$

On utilise la méthode classique de résolution consistant à éliminer les équations une à une par *substitution successive*. En effet, en isolant x_1 de la première équation, on a

$$x_1 = \frac{8 - 3x_2}{2}. \tag{3.3}$$

Et si l'on substitue dans la deuxième, on obtient l'équation suivante

$$3\left(\frac{8 - 3x_2}{2}\right) + 4x_2 = 11,$$

dont la solution est facile à obtenir $x_2 = 2$; d'où en retournant à l'équation (3.3), on en tire $x_1 = 1$.

Bien qu'il est théoriquement possible d'étendre la substitution successive à des systèmes de grande taille, il est cependant difficile de la transcrire sous forme d'algorithme afin d'être programmé dans un langage informatique.

Question: Existent-t-ils de types de systèmes linéaires faciles à résoudre, et ce, même s'ils sont de grande taille ?

Définition 1.1

La matrice $A = (a_{ij})_{1 \leq i,j \leq n}$ est dite diagonale, si ses entrées sont nulles en dehors de sa diagonale (i.e. $a_{ij} = 0,\ i \neq j$). Une telle matrice est de la forme

$$A = \begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 \\ 0 & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & a_{nn} \end{bmatrix}$$

Exemple 2

Intéressons nous maintenant au système linéaire suivant

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 9 \end{bmatrix}$$

dont la solution est facile à obtenir : Il suffit de considérer séparément chaque ligne pour obtenir le vecteur solution $\vec{x} = [2 \quad 1 \quad 3]^t$. On voit tout de suite comment résoudre le cas général suivant

$$\begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 \\ 0 & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \dots & \dots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ \vdots \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ \vdots \\ \vdots \\ b_n \end{bmatrix}$$

sous les conditions $a_{ii} \neq 0, i = 1, \cdots, n$ qui assurent l'unicité de la solution. La solution est ontenue en considérant pour chaque ligne $i = 1, 2, \cdots, n$, la valeur $x_i = \frac{b_i}{a_{ii}}$.

Note: Les systèmes diagonaux (dont la matrice A n'a de coefficients non nuls que sur la diagonale) sont faciles à résoudre.

Définition 1.2

La matrice $A = (a_{ij})_{1\leq i,j\leq n}$ est dite triangulaire inférieure (ou supérieure) si tous les a_{ij} (ou tous les a_{ji}) sont nuls pour $i < j$. La matrice triangulaire inférieure A, a la forme type suivante

$$A = \begin{bmatrix} a_{11} & 0 & \dots & \dots & 0 \\ a_{21} & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ a_{n1} & \dots & \dots & a_{n(n-1)} & a_{nn} \end{bmatrix}$$

Note: Une matrice triangulaire supérieure est tout simplement la transposée d'une matrice triangulaire inférieure.

Exemple 3

Considérons encore l'exemple suivant, où la partie supérieure de la matrice est nulle

$$\begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 14 \end{bmatrix}$$

La résolution de ce système n'est pas compliquée car il suffit de faire une première descente pour obtenir

$$x_1 = \frac{9}{3} = 3.$$

À la deuxième descente, on calcul avec l'aide de x_1 la valeur

$$x_2 = \frac{7 - (1)(3)}{2}.$$

Et la dernière valeur est obtenue grâce à x_1 et x_2 par la relation suivante

$$x_3 = \frac{14 - (3)(3) - (2)(2)}{1} = 1.$$

On peut également généraliser, dans le cas d'un système de taille n, la résolution de ce système à matrice triangulaire inférieure. La solution est déterminée par les relations génériques suivantes

$$x_1 = \frac{b_1}{a_{11}}, \quad x_i = \frac{\left(b_i - \sum_{k=1}^{i-1} a_{ik}x_k\right)}{a_{ii}}, \; i = 2, 3, \cdots, n. \tag{3.4}$$

Dans le cas d'un système à matrice triangulaire supérieure, la solution est donnée comme suit

$$x_n = \frac{b_n}{a_{nn}}, \quad x_i = \frac{\left(b_i - \sum_{k=i+1}^{n} a_{ik}x_k\right)}{a_{ii}}, \; i = n-1, n-2, \cdots, 2, 1. \tag{3.5}$$

Note: Une autre réponse à la question précédente est que les systèmes triangulaires sont également faciles à résoudre. Il suffit en effet de commencer par l'équation qui se trouve à la pointe du triangle (la première pour une matrice triangulaire inférieure et la dernière pour une matrice triangulaire supérieure) et de résoudre une à une les équations. On parle de descente triangulaire ou de remontée triangulaire, selon le cas.

Note: Pour résoudre un système linéaire quelconque, on se ramenera toujours à un système triangulaire plutôt qu'à un système diagonal car ce dernier est rarement rencontré en pratique et exige plus de travail pour s'y ramener.

Remarque(s) :

- Les équations (3.4) et (3.5) sont valides si les a_{ii} sont tous non nuls. Dans le cas contraire, la matrice A n'est pas inversible et, donc, le système $A\vec{x} = \vec{b}$ n'a pas une solution unique.
- Le déterminant de la matrice triangulaire inférieure A (ou supérieure), est égale au produit des éléments de la diagonale :

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

Question: Comment ramener un système linéaire quelconque à un système triangulaire sans changer la solution du système de départ ?

1.2 - Opérations élémentaires sur les lignes

La question à répondre ici est, comment transformer le système linéaire

$$A\vec{x} = \vec{b} \tag{3.6}$$

en un système triangulaire sans en modifier la solution. La réponse est de multiplier à gauche de chaque côté du système (3.6) par une matrice inversible W. C'est à dire, on s'intéressera au système triangulaire suivant

$$WA\vec{x} = W\vec{b} \tag{3.7}$$

Remarque(s) : Les systèmes (3.6) et (3.7) ont la même solution $\vec{x}$. Ce résultat n'est plus vrai si la matrice W n'est pas inversible, i.e. à partir du système (3.7) on ne pourra pas revenir au système (3.6) si la matrice W^{-1} n'existe pas.

Exemple 4

Nous avons vu que la solution du système linéaire

$$\begin{bmatrix} 3 & 0 & 0 \\ 1 & 2 & 0 \\ 3 & 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 7 \\ 14 \end{bmatrix}$$

est le vecteur $\vec{x} = [3 \quad 2 \quad 1]^t$. D'autre part, si on multiplie ce système par la matrice inversible

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 0 \\ 1 & 2 & 3 \end{bmatrix}$$

on obtient le nouveau système

$$\begin{bmatrix} 3 & 0 & 0 \\ 5 & 4 & 0 \\ 14 & 10 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 9 \\ 23 \\ 65 \end{bmatrix}$$

dont la solution est toujours $\vec{x} = [3 \quad 2 \quad 1]^t$.

Question: On réitère la question précédente : Comment obtenir la matrice W ? En d'autres termes, comment transformer un système linéaire quelconque en un système triangulaire ?

Note: Pour transformer un système quelconque en un système triangulaire, il suffit d'utiliser trois opérations élémentaires sur les lignes de la matrice. Ces trois opérations élémentaires correspondent à trois différentes types de matrices W. En notant $\vec{l_i}$ la ligne i de la matrice A du système de départ, on a :

1. L'opération qui consiste à remplacer la ligne i par un multiple d'elle-même notée $(\vec{l_i} \longleftarrow \lambda\vec{l_i})$;
2. L'opération qui consiste à intervertir la ligne i et la ligne j représentée par $(\vec{l_i} \longleftrightarrow \vec{l_j})$;
3. L'opération qui est de remplacer la ligne i par la ligne i plus un multiple de la ligne j notée par $(\vec{l_i} \longleftarrow \vec{l_i} + \lambda\vec{l_j})$.

Chacune de ces opérations élémentaires est équivalente à multiplier le système (3.6) par une matrice inversible.

- **1.2.1 - Multiplication d'une ligne par un scalaire $(\vec{l_i} \longleftarrow \lambda\vec{l_i})$ et matrice équivalente**

L'opération $(\vec{l_i} \longleftarrow \lambda\vec{l_i})$, consistant à remplacer la ligne l_i par un multiple d'elle-même, équivaut à multiplier le système linéaire (3.6) par une matrice diagonale inversible W dont tous les éléments diagonaux sont 1, sauf l'élément à la position (i, i), qui vaut λ.

Note: On obtient W à partir de la matrice identité en remplaçant l'élément à la position (i, i) par λ.

Remarque(s) :

- Dans ce cas, le déterminant de la matrice diagonale W est λ. La matrice est donc inversible si $\lambda \neq 0$.
- La matrice inverse de W, est celle diagonale dont tous les éléments diagonaux sont 1, sauf l'élément à la position (i, i), qui vaut $\frac{1}{\lambda}$.

Exemple 5

Soit le système suivant dont la solution est donnée par $\vec{x} = [1 \quad 1 \quad 1]^t$

$$\begin{bmatrix} 3 & 1 & 2 \\ 6 & 4 & 1 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 11 \\ 10 \end{bmatrix} \tag{3.8}$$

Si l'on souhaite effectuer l'opération $(\vec{l_2} \longleftarrow 3\vec{l_2})$ sur le système (3.8), c'est à dire multiplier la ligne 2 par un facteur 3, alors cela revient à multiplier ce système par la matrice

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Et l'on obtient ainsi le nouveau système suivant

$$\begin{bmatrix} 3 & 1 & 2 \\ 18 & 12 & 3 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 33 \\ 10 \end{bmatrix}$$

dont la solution reste la même que celle du système (3.8) de départ puisque la matrice W est inversible car $det(W) \neq 0$:

$$det(W) = 1 \times 3 \times 1 = 3.$$

- **1.2.2 - Permutation de deux lignes $(\vec{l_i} \longleftrightarrow \vec{l_j})$ et matrice équivalente**

L'opération $(\vec{l_i} \longleftrightarrow \vec{l_j})$ qui est à intervertir les lignes i et j, est équivalente à la multiplication du système (3.6) par une matrice inversible W, obtenue en permutant les lignes i et j de la matrice identité.

Note: On obtient la matrice W à partir de la matrice identité en permutant les lignes i et j.

Remarque(s) :

- Le déterminant de la matrice W est -1. Lorsque l'on permute deux lignes, le déterminant du système de départ change de signe.
- L'inverse de la matrice W est elle même.

Exemple 6

Si l'on souhaite effectuer l'opération $(\vec{l}_2 \longleftrightarrow \vec{l}_3)$ sur le système (3.8) de l'exemple précédent, c'est à dire intervertir la ligne 2 et la ligne 3, il suffit de multiplier ce système par la matrice

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

Ainsi, on obtient le nouveau système suivant dont la solution n'a pas changé (car la matrice W est inversible)

$$\begin{bmatrix} 3 & 1 & 2 \\ 5 & 4 & 1 \\ 6 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 10 \\ 11 \end{bmatrix}$$

- **1.2.3 - L'opération $\vec{l}_i \longleftarrow \vec{l}_i + \lambda \vec{l}_j$ et matrice équivalente**

L'opération élémentaire $\vec{l}_i \longleftarrow \vec{l}_i + \lambda \vec{l}_j$ consiste à remplacer la ligne i par elle même plus un multiple de la ligne j. Elle est équivalente à multiplier le système de départ par une matrice inversible W qui vaut 1 sur toute la diagonale et 0 partout ailleurs, sauf l'élément à la position (i, j) qui vaut λ.

Note: On obtient la matrice W à partir de la matrice identité en ajoutant λ à la position (i, j).

Remarque(s) :

- On peut montrer facilement que le déterminant de la matrice W est 1.
- La matrice W est inversible et son inverse est obtenue en remplacant λ par $-\lambda$.

Exemple 7

En souhaitant, dans le système (3.8), effectuer l'opération $\vec{l}_2 \longleftarrow \vec{l}_2 - 2\vec{l}_1$ qui est de remplacer la deuxième ligne par la deuxième ligne ($i = 2$) moins deux fois ($\lambda = -2$) la première ligne ($j = 1$), il suffit de multiplier ce système par la matrice inversible

$$W = \begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

On obtient le nouveau système

$$\begin{bmatrix} 3 & 1 & 2 \\ 0 & 2 & -3 \\ 5 & 4 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -1 \\ 10 \end{bmatrix}$$

Note: La stratégie de résolution qu'on va adopter, pour transformer un système quelconque en système triangulaire, sera donc basée sur ces trois opérations élémentaires qui seront effectuées à travers deux démarches appelées **méthodes directes** : la méthode d'élimination de Gauss et la décomposition LU.
En pratique, on ne multiplie jamais explicitement les systèmes considérés par les différentes matrices W, car ce serait trop long. Il faut cependant garder en tête que les opérations effectuées sont équivalentes à ces multiplications.

2 Élimination de Gauss

Méthode : La méthode d'élimination de Gauss consiste à éliminer tous les termes sous la diagonale de la matrice A en utilisant systématiquement les opérations élémentaires précédentes. Ces opérations permettent d'introduire des zéros sous la diagonale de la matrice A et procurent ainsi un système triangulaire supérieur.

Note: La validité de la méthode de Gauss repose donc sur le fait que ces opérations élémentaires équivalent à multiplier le système de départ par une matrice inversible.

Définition 2.1

La matrice augmentée du système linéaire (3.1) est la matrice de dimension n sur $n+1$ que l'on obtient en ajoutant le membre de droite $\vec{b}$ à la matrice A, c'est-à-dire :

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & \dots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} & b_n \end{array}\right]$$

Note: Puisque les opérations élémentaires doivent être effectuées à la fois sur les lignes de la matrice A et sur celles du vecteur $\vec{b}$, cette notation sera très utile.

Exemple 8

Considérons de système linéaire suivant dont l'objectif est de déterminer la solution $\vec{x} = [x_1\ x_2\ x_3]^t$, après application de la méthode d'élimination de Gauss :

$$\underbrace{\begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix}}_{\text{A}} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 10 \\ 26 \\ 35 \end{bmatrix}}_{\vec{b}}$$

Sachant que les opérations élémentaires doivent être effectuées à la fois sur les lignes de la matrice A et sur celles du vecteur $\vec{b}$, alors on considère la matrice augmentée suivante :

$$\left[\begin{array}{ccc|c} \boxed{2} & 1 & 2 & 10 \\ 6 & 4 & 0 & 26 \\ 8 & 5 & 1 & 35 \end{array}\right] \tag{3.9}$$

Les opérations élémentaires nécessaires, pour éliminer les termes non nuls sous la diagonale de la première colonne de la matrice (3.9), sont les suivantes

$$\vec{l}_2 \longleftarrow \vec{l}_2 - (6/\boxed{2})\vec{l}_1$$
$$\vec{l}_3 \longleftarrow \vec{l}_3 - (8/\boxed{2})\vec{l}_1$$

Note: Il est à noter que l'on divise par 2 (l'élément a_{11} de la matrice A) les coefficients qui multiplient la ligne 1. On dit alors que l'élément 2 est le pivot.

On obtient, en effectuant les opérations indiquées préalablement, la nouvelle matrice

$$\left[\begin{array}{ccc|c} 2 & 1 & 2 & 10 \\ 0 & \boxed{1} & -6 & -4 \\ 0 & 1 & -7 & -5 \end{array}\right] \tag{3.10}$$

Pour produire une matrice triangulaire supérieure, il suffit maintenant d'introduire des 0 sous la diagonale de la deuxième colonne de la matrice (3.10). Et pour cela, il suffit d'effectuer l'opération élémentaire suivante

$$\vec{l}_3 \longleftarrow \vec{l}_3 - (1/\boxed{1})\vec{l}_2$$

Note: Ici, le pivot est l'élément 1 de la matrice (3.10) puisque maintenant $a_{22} = 1$.

On obtient donc, en effectuant cette opération élémentaire, la nouvelle matrice

$$\left[\begin{array}{ccc|c} 2 & 1 & 2 & 10 \\ 0 & 1 & -6 & -4 \\ 0 & 0 & -1 & -1 \end{array}\right] \tag{3.11}$$

La matrice triangulaire supérieure (3.11) ainsi obtenue, il reste qu'à faire la remontée triangulaire de l'algorithme (3.5) pour obtenir la solution. On a :

$$x_3 = \frac{-1}{-1} = 1$$

ensuite, on calcule x_2 comme suit

$$x_2 = \frac{-4 - (-6)(1)}{1} = 2$$

et enfin, on a x_1

$$x_1 = \frac{10 - (1)(2) - (2)(1)}{2} = 3.$$

On a construit le système triangulaire (3.11) en effectuant des opérations élémentaires directement sur les lignes de la matrice. En désignant la matrice triangulaire obtenue par U, les opérations effectuées pour obtenir U sont équivalentes à multiplier le système de départ par une suite de matrices inversibles T_i, $i = 1, 2, 3$. On a ainsi

$$U = T_3 T_2 T_1 A. \tag{3.12}$$

Les matrices T_i correspondent aux différentes opérations effectuées sur les lignes de la matrice A mais aussi sur celles du vecteur $\vec{b}$. Explicitement, on a eu la succession de matrices générées par chacune des opérations élémentaires

$$\vec{l}_2 \longleftarrow \vec{l}_2 - (6/\boxed{2})\vec{l}_1 \iff T_1 = \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\vec{l}_3 \longleftarrow \vec{l}_3 - (8/\boxed{2})\vec{l}_1 \iff T_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix}$$

$$\vec{l}_3 \longleftarrow \vec{l}_3 - (1/\boxed{1})\vec{l}_2 \iff T_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix}$$

Ainsi, l'égalité matricielle (3.12) peut s'écrire sans équivoque par

$$\begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -6 \\ 0 & 0 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -4 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix}$$

Si l'on poursuit le raisonnement, on a également

$$A = T_1^{-1} T_2^{-1} T_3^{-1} U$$

Puisque l'on sait inverser les matrices T_i, on a immédiatement que

$$\begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{T_1^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & 0 & 1 \end{bmatrix}}_{T_2^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}}_{T_3^{-1}} \begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -6 \\ 0 & 0 & -1 \end{bmatrix}$$

ou encore

$$\begin{bmatrix} 2 & 1 & 2 \\ 6 & 4 & 0 \\ 8 & 5 & 1 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 4 & 1 & 1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 2 & 1 & 2 \\ 0 & 1 & -6 \\ 0 & 0 & -1 \end{bmatrix}}_{U}$$

Note:

- On remarque que les coefficients de la matrice triangulaire inférieure sont ceux qui ont permis d'éliminer les termes non nuls sous la diagonale de la matrice A.
- Tout cela revient à décomposer la matrice A en un produit d'une matrice triangulaire inférieure, notée L, et d'une matrice triangulaire supérieure U. C'est ce que l'on appelle une décomposition LU.

Remarque(s) : Le déterminant de la matrice de départ est le même que celui de la matrice triangulaire (3.11) puisqu'on n'a effectué que des opérations de la forme $\vec{l_i} \longleftarrow \vec{l_i} + \lambda\vec{l_j}$, ce qui revient à multiplier le système de départ par une matrice dont le déterminant est 1. On a donc

$$det(A) = (2)(1)(-1) = -2$$

soit le produit des termes diagonaux de la matrice (3.11). Pour être plus précis

$$det(A) = det\left(T_1^{-1}\right) det\left(T_2^{-1}\right) det\left(T_3^{-1}\right) det\,(U) = (1)(1)(1)[(2)(1)(-1)]$$

puisque le déterminant des trois matrices T_i est 1.

Remarque(s) : La méthode d'élimination de Gauss revient à factoriser la matrice A en un produit de deux matrices triangulaires L et U seulement dans le cas où aucune permutation de lignes n'est effectuée.

3 Décomposition LU

Méthode : Si on réussit à exprimer la matrice A en un produit de deux matrices triangulaires L et U, alors le système $A\vec{x} = \vec{b}$ peut s'écrire comme suit

$$A\vec{x} = LU\vec{x} = \vec{b}.$$

Ainsi en posant $U\vec{x} = \vec{y}$, la résolution du système linéaire $A\vec{x} = \vec{b}$ se fait alors en deux étapes

$$\begin{aligned} L\vec{y} &= \vec{b} \\ U\vec{x} &= \vec{y} \end{aligned} \tag{3.13}$$

qui sont deux systèmes triangulaires.

Remarque(s) :

- On utilise d'abord une descente triangulaire sur la matrice L pour obtenir $\vec{y}$;
- ensuite, une remontée triangulaire sur la matrice U pour obtenir la solution recherchée $\vec{x}$.

Note: Il est important de souligner que la décomposition LU n'est pas unique.

Exemple 9

Pour illustrer la non-unicité de la décomposition LU, il suffit de vérifier les deux égalités suivantes

$$\begin{bmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 6 & -3 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -4 & 0 \\ 6 & 0 & 4 \end{bmatrix} \begin{bmatrix} 1 & -0,5 & -0,5 \\ 0 & 1 & -0,5 \\ 0 & 0 & 1 \end{bmatrix}$$

et

$$\begin{bmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 6 & -3 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & -1 \\ 0 & -4 & 2 \\ 0 & 0 & 4 \end{bmatrix}$$

3.1 - Décomposition de Crout

La décomposition LU n'étant pas unique, il faut faire au préalable des choix arbitraires. Le choix le plus populaire consiste à imposer que la matrice U ait des 1 sur sa diagonale. C'est la décomposition de Crout.

Méthode : La décomposition (ou factorisation) de Crout d'une matrice A consiste à déterminer les coefficients l_{ij} et u_{ij} des matrices triangulaires inférieure L et supérieure U tels que $A = LU$. Les éléments de la diagonale de U sont composés de 1. Par exemple si A est une matrice de dimension 4 sur 4, on a

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{U} \tag{3.14}$$

On procédera par identification, pour déterminer les $4^2 = 16$ (n^2 si A est une matrice de dimension n sur n) coefficients inconnus des matrices L et U. En effet, effectuant le produit des matrices L et U et se servant des différents coefficients a_{ij} on détermine les coefficients l_{ij} et u_{ij} suivant cet ordre :

1. Pour les coefficients de la première colonne de L, faire le produit des lignes de L par la première colonne de U

$$l_{11} = a_{11}, \quad l_{21} = a_{21}, \quad l_{31} = a_{31}, \quad l_{41} = a_{41}. \tag{3.15}$$

2. Produit de la première ligne de L par les colonnes de U fournit la première ligne de U, pourvu que $l_{11} \neq 0$,

$$u_{12} = \frac{a_{12}}{l_{11}}, \quad u_{13} = \frac{a_{13}}{l_{11}}, \quad u_{14} = \frac{a_{14}}{l_{11}}, \tag{3.16}$$

3. Produit des lignes de L par la deuxième colonne de U détermine la deuxième colonne de L

$$l_{22} = a_{22} - l_{21}u_{12}, \quad l_{32} = a_{32} - l_{31}u_{12}, \quad l_{42} = a_{42} - l_{41}u_{12}. \tag{3.17}$$

4. Le produit de la deuxième ligne de L par les colonnes de U fournit d'avoir la deuxième ligne de U

$$u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}}, \quad u_{24} = \frac{a_{24} - l_{21}u_{14}}{l_{22}}. \tag{3.18}$$

5. Par le produit des lignes de L par la troisième colonne de U, on a la troisième colonne de L

$$l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}, \quad l_{43} = a_{43} - l_{41}u_{13} - l_{42}u_{23}. \tag{3.19}$$

6. Le produit de la troisième ligne de L par la quatrième colonne de U, on obtient l'élément

$$u_{34} = \frac{a_{34} - l_{31}u_{14} - l_{32}u_{24}}{l_{33}}. \tag{3.20}$$

7. Le produit de la quatrième ligne de L par la quatrième colonne de U, fournit dernier coefficient de L

$$l_{44} = a_{44} - l_{41}u_{14} - l_{42}u_{24} - l_{43}u_{34}. \tag{3.21}$$

Remarque(s) : La décomposition (3.14) ne fonctionne que si les pivots l_{ii} sont tous non nuls. Ce n'est pas toujours le cas et il est possible qu'il faille permuter deux lignes pour éviter cette situation (voir la sous-section 3.2), tout comme pour l'élimination de Gauss. Le coefficient l_{ii} est encore appelé pivot.

Définition 3.1

La *notation compacte* de la décomposition LU, d'une matrice de dimension 4 sur 4, est la matrice de coefficients

$$\begin{bmatrix} l_{11} & u_{12} & u_{13} & u_{14} \\ l_{21} & l_{22} & u_{23} & u_{24} \\ l_{31} & l_{32} & l_{33} & u_{34} \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix} \tag{3.22}$$

Remarque(s) :

- La matrice initiale A est tout simplement détruite.
- Les coefficients 1 sur la diagonale de la matrice U ne sont pas indiqués explicitement dans (3.22), mais doivent tout de même être pris en compte.
- De façon plus rigoureuse, la notation compacte revient à mettre en mémoire la matrice

$$L + U - I$$

et à détruire la matrice A.

Exemple 10

Soit le système suivant, dont on veut résoudre en effectuant d'abord la décomposition de Crout de la matrice A :

$$\begin{bmatrix} 3 & -1 & 2 \\ 1 & 2 & 3 \\ 2 & -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 12 \\ 11 \\ 2 \end{bmatrix}$$

Afin d'illustrer la notation compacte, on remplacera au fur et à mesure les coefficients a_{ij} par les coefficients l_{ij} ou u_{ij} ; les cases, accompagnées de la couleur, soulignent que l'élément a_{ij} correspondant a été détruit.

1. Pour les coefficients de la première colonne de L, prendre tout simplement la première colonne de A :

$$\begin{bmatrix} \boxed{3} & -1 & 2 \\ \boxed{1} & 2 & 3 \\ \boxed{2} & -2 & -1 \end{bmatrix}$$

2. Première ligne de U : Le pivot de la première ligne est 3. On divise donc la première ligne de A par 3

$$\begin{bmatrix} \boxed{3} & \boxed{-\frac{1}{3}} & \boxed{\frac{2}{3}} \\ \boxed{1} & 2 & 3 \\ \boxed{2} & -2 & -1 \end{bmatrix}$$

3. En utilisant la relation (3.17), on obtient la deuxième colonne de la matrice L

$$l_{22} = a_{22} - l_{21}u_{12} = 2 - (1)(-\frac{1}{3}) = \frac{7}{3}, \quad l_{32} = a_{32} - l_{31}u_{12} = -2 - (2)(-\frac{1}{3}) = -\frac{4}{3}$$

La matrice devient alors

$$\begin{bmatrix} \boxed{3} & \boxed{-\frac{1}{3}} & \boxed{\frac{2}{3}} \\ \boxed{1} & \boxed{\frac{7}{3}} & 3 \\ \boxed{2} & \boxed{-\frac{4}{3}} & -1 \end{bmatrix}$$

4. En utilisant la relation (3.18), on obtient la deuxième ligne de la matrice U

$$u_{32} = \frac{a_{23} - l_{21}u_{13}}{l_{22}} = \frac{3 - (1)(\frac{2}{3})}{\frac{7}{3}} = 1$$

La matrice compacte devient

$$\begin{bmatrix} \boxed{3} & \boxed{-\frac{1}{3}} & \boxed{\frac{2}{3}} \\ \boxed{1} & \boxed{\frac{7}{3}} & \boxed{1} \\ \boxed{2} & \boxed{-\frac{4}{3}} & -1 \end{bmatrix}$$

5. D'après la relation (3.19), on a le calcul de l_{33} qui donne

$$l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23} = -1 - (2)(\frac{2}{3}) - (-\frac{4}{3})(1) = -1.$$

La matrice compacte est donc

$$\begin{bmatrix} \boxed{3} & \boxed{-\frac{1}{3}} & \boxed{\frac{2}{3}} \\ \boxed{1} & \boxed{\frac{7}{3}} & \boxed{1} \\ \boxed{2} & \boxed{-\frac{4}{3}} & \boxed{-1} \end{bmatrix}$$

À la suite de cette décomposition LU, la matrice de départ A (maintenant détruite) vérifie nécessairement :

$$A = \underbrace{\begin{bmatrix} 3 & 0 & 0 \\ 1 & \frac{7}{3} & 0 \\ 2 & -\frac{4}{3} & -1 \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & -\frac{1}{3} & \frac{2}{3} \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}}_{U}$$

Pour la résolution, on effectue une descente triangulaire sur $L\vec{y} = \vec{b}$ pour obtenir

$$\begin{aligned} y_1 &= \frac{12}{3} &&= 4 \\ y_2 &= \frac{11-(1)(4)}{\frac{7}{3}} &&= 3 \\ y_3 &= \frac{2-(2)(4)-(-\frac{4}{3})(3)}{(-1)} &&= 2 \end{aligned}$$

Et une remontée triangulaire sur le système $U\vec{x} = \vec{y}$ pour obtenir la solution recherchée

$$\begin{aligned} x_3 &= 2 \\ x_2 &= 3-(1)(2) &&= 1 \\ x_1 &= 4-\left(-\frac{1}{3}\right)(1)-\left(\frac{2}{3}\right)(2) &&= 3 \end{aligned}$$

La solution recherchée est donc $\vec{x} = [3 \quad 1 \quad 2]^t$.

De façon générale, on a l'algorithme de décomposition de Crout détaillé dans le manuel [1] du cours.

• 3.2 - Décomposition LU et permutation de lignes

La décomposition LU exige que les pivots l_{ii} soient non nuls, au cas contraire, il faut faire une permutation de lignes. De plus, contrairement à la méthode d'élimination de Gauss, la décomposition LU n'utilise le terme de droite $\vec{b}$ qu'à la toute fin, au moment de la descente triangulaire $L\vec{y} = \vec{b}$. On doit ainsi garder la trace de ces permutations afin de les appliquer sur $\vec{b}$. À cette fin, on introduit un vecteur $\vec{O}$ dit de permutation qui contient tout simplement la numérotation des équations.

Remarque(s) : Dans une décomposition LU, la permutation de lignes s'effectue toujours après le calcul de chaque colonne de L. On place en position de pivot le plus grand terme en valeur absolue de cette colonne (sous le pivot actuel), pour des raisons de précision que nous verrons plus loin.

Exemple 11

On veut maintenant résoudre le système linéaire suivant, en effectuant d'abord une décompistion LU de la matrice A :

$$\begin{bmatrix} 0 & 2 & 1 \\ 1 & 0 & 0 \\ 3 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 5 \\ -1 \\ -2 \end{bmatrix} \tag{3.23}$$

Avant de commencer, on définit le vecteur $\vec{O}$ qui indique la numérotation des équations du système (3.23)

$$\vec{O} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

1. André Fortin : Analyse numérique pour ingénieurs, Presses Internationales Polytechnique, Montréal, Québec (2016).

1. Première colonne de L : Puisqu'il s'agit de la première colonne de A, on a

$$\begin{bmatrix} \boxed{0} & 2 & 1 \\ \boxed{1} & 0 & 0 \\ \boxed{3} & 0 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

où le vecteur de permutation n'a pas été modifié.
Sachant qu'on a un pivot nul, alors on effectue la permutation $(l_1 \longleftrightarrow l_3)$. On aurait pu aussi permuter les lignes 1 et 2, mais on choisit immédiatement le plus grand pivot possible (en valeur absolue). Le vecteur de permutation est alors modifié

$$\begin{bmatrix} \boxed{3} & 0 & 1 \\ \boxed{1} & 0 & 0 \\ \boxed{0} & 2 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

2. Première ligne de U : Il suffit de diviser cette ligne par le nouveau pivot 3

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{\frac{1}{3}} \\ \boxed{1} & 0 & 0 \\ \boxed{0} & 2 & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

3. Deuxième colonne de L : De la relation (3.17), on tire

$$l_{22} = a_{22} - l_{21}u_{12} = 0 - (1)(0) = 0, \quad l_{32} = a_{32} - l_{31}u_{12} = 2 - (0)(0) = 2$$

On a maintenant

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{\frac{1}{3}} \\ \boxed{1} & \boxed{0} & 0 \\ \boxed{0} & \boxed{2} & 1 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

On a encore un pivot nul, qui oblige à intervertir les lignes 2 et 3 et à modifier le vecteur $\vec{O}$ en conséquence $(l_2 \longleftrightarrow l_3)$

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{\frac{1}{3}} \\ \boxed{0} & \boxed{2} & 1 \\ \boxed{1} & \boxed{0} & 0 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

4. Le calcul de l'élément u_{23} est fait à partir de la relation (3.18) et on a

$$u_{23} = \frac{a_{23} - l_{21}u_{13}}{l_{22}} = \frac{1 - (0)(\frac{1}{3})}{2} = \frac{1}{2}$$

et la matrice compacte devient

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{\frac{1}{3}} \\ \boxed{0} & \boxed{2} & \boxed{\frac{1}{2}} \\ \boxed{1} & \boxed{0} & 0 \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

5. Pour le calcul de l'élément l_{33}, on utilise la relation (3.19) et on obtient

$$l_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23} = 0 - (1)\left(\frac{1}{3}\right) - (0)\left(\frac{1}{2}\right) = -\frac{1}{3}$$

La décomposition LU de la matrice A est donc

$$\begin{bmatrix} \boxed{3} & \boxed{0} & \boxed{\frac{1}{3}} \\ \boxed{0} & \boxed{2} & \boxed{\frac{1}{2}} \\ \boxed{1} & \boxed{0} & \boxed{-\frac{1}{3}} \end{bmatrix} \quad \vec{O} = \begin{bmatrix} 3 \\ 1 \\ 2 \end{bmatrix}$$

Remarque(s) : Il faut toutefois remarquer que le produit LU donne

$$\underbrace{\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & -\frac{1}{3} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} 1 & 0 & \frac{1}{3} \\ 0 & 1 & \frac{1}{2} \\ 0 & 0 & 1 \end{bmatrix}}_{U} = \begin{bmatrix} 3 & 0 & 1 \\ 0 & 2 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$

c'est-à-dire *la matrice A permutée suivant le vecteur $\vec{O}$*.

Pour passer maintenant à la résolution du système (3.23) on commence d'abord, compte tenu du vecteur $\vec{O}$, par le système suivant

$$L\vec{y} = \begin{bmatrix} -2 \\ 5 \\ -1 \end{bmatrix}$$

À noter l'ordre des valeurs dans le membre de droite. La descente triangulaire donne $\vec{y} = \begin{bmatrix} -\frac{2}{3} & \frac{5}{2} & 1 \end{bmatrix}^t$.
Il suffit maintenant d'effectuer la remontée triangulaire au système suivant

$$U\vec{x} = \begin{bmatrix} -\frac{2}{3} \\ \frac{5}{2} \\ 1 \end{bmatrix}$$

qui nous donne la solution finale $\vec{x} = \begin{bmatrix} -1 & 2 & 1 \end{bmatrix}^t$.

Remarque(s) : Le déterminant de la matrice A dans cet exemple est donné par

$$det(A) = (-1)(-1)\left[(3)(2)\left(-\frac{1}{3}\right)\right] = -2.$$

Ayant effectué deux permutations de lignes, le déterminant a deux fois changé de signe.

Cela nous amène au théorème suivant.

Théorème 3.1

On peut calculer le déterminant d'une matrice A à l'aide de la méthode de décomposition LU de Crout par

$$det(A) = (-1)^N \prod_{i=1}^{n} l_{ii} \tag{3.24}$$

où N est le nombre de fois où on a interverti deux lignes.

• 3.3 - Factorisation de Choleski

Si la matrice A est symétrique, on peut mettre seulement en mémoire sa moitié inférieure plus sa diagonale principale. Ce qui permettra de réduire l'espace mémoire nécessaire à la résolution du système linéaire $A\vec{x} = \vec{b}$. De plus à la résolution de ce système par une décomposition LU, la matrice triangulaire supérieure U peut être remplacée par la matrice transposée de L pour donner la décomposition $A = LL^t$. C'est ce qu'on appelle la factorisation de Choleski.

Méthode : La décomposition de Choleski de la matrice A consiste à déterminer les coefficients l_{ij} de la matrice L, tel que $A = LL^t$. Par exemple si A est une matrice de dimension 4 sur 4, on a

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}}_{L} \underbrace{\begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}}_{L^t}$$

Par le produit des matrices L et L^t et se servant des coefficients a_{ij}, on détermine les coefficients l_{ij} comme suit

1. Pour les coefficients de la première colonne de L, on a

$$l_{11} = \sqrt{a_{11}}, \quad l_{21} = \frac{a_{21}}{l_{11}}, \quad l_{31} = \frac{a_{31}}{l_{11}}.$$

2. Pour la deuxième colonne de L, on a

$$l_{22} = \sqrt{a_{22} - l_{21}^2}, \quad l_{32} = \frac{a_{32} - l_{31}l_{21}}{l_{22}}.$$

3. On a le dernier coefficient de L qui est

$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2}.$$

Dans le cas générale, on a l'algorithme de Factorisation de Choleski détaillé dans le manuel [2] du cours.

Remarque(s) : Tout comme nous l'avons fait pour la décomposition LU, on peut remplacer, au fur et à mesure que les calculs progressent, les éléments utilisés de la matrice A par l'élément correspondant de L.

Exemple 12

Considérons la matrice symétrique A définie par

$$A = \begin{bmatrix} 4 & * & * \\ 6 & 10 & * \\ 2 & 5 & 14 \end{bmatrix}$$

où la partie supérieure de la matrice est remplacée par des $(*)$ simplement pour indiquer que cette partie de la matrice n'est pas mise en mémoire et ne servira aucunement dans les calculs. En suivant l'algorithme précédent, on a :

- La première colonne de A

$$l_{11} = \sqrt{a_{11}} = 2, \quad l_{21} = \frac{a_{21}}{l_{11}} = 3, \quad l_{31} = \frac{a_{31}}{l_{11}} = 1.$$

Sous forme compacte, on a

$$\begin{bmatrix} \boxed{2} & * & * \\ \boxed{3} & 10 & * \\ \boxed{1} & 5 & 14 \end{bmatrix}$$

- Pour la deuxième colonne

$$l_{22} = \sqrt{10 - 3^2} = 1, \quad l_{32} = \frac{5 - 1 \times 3}{1} = 2.$$

ce qui donne la forme compacte

$$\begin{bmatrix} \boxed{2} & * & * \\ \boxed{3} & \boxed{1} & * \\ \boxed{1} & \boxed{2} & 14 \end{bmatrix}$$

- Enfin pour le dernier élément

$$l_{33} = \sqrt{14 - 1^2 - 2^2} = 3$$

et on a la factorisation sous forme compacte

$$\begin{bmatrix} \boxed{2} & * & * \\ \boxed{3} & \boxed{1} & * \\ \boxed{1} & \boxed{2} & \boxed{3} \end{bmatrix}$$

Ainsi, on a la décomposition suivante de la matrice A

$$A = \begin{bmatrix} 4 & * & * \\ 6 & 10 & * \\ 2 & 5 & 14 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 0 \\ 3 & 1 & 0 \\ 1 & 2 & 3 \end{bmatrix} \begin{bmatrix} 2 & 3 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{bmatrix}$$

Remarque(s) :

- La factorisation de Choleski n'est pas unique. On a par exemple $A = LL^t = (-L)(-L)^t$ qui sont deux factorisa-

2. André Fortin : Analyse numérique pour ingénieurs, Presses Internationales Polytechnique, Montréal, Québec (2016).

tions différentes. On peut s'assurer de l'unicité en imposant $l_{ii} > 0$, ce qui revient au choix naturel de prendre la valeur positive de la racine carrée lors du calcul de l_{ii}.

- le déterminant de A est donné par

$$det(A) = det(L) \times det(L^t) = (det(L))^2 = \prod_{i=1}^{n} l_{ii}^2$$

- La factorisation de Choleski ne peut être menée à terme qu'avec une matrice symétrique définie positive.

Définition 3.2

Une matrice symétrique A est dite définie positive, si elle vérifie la condition $A\vec{x} \cdot \vec{x} > 0, \forall\, \vec{x} \neq \vec{0}$.

Note: A est définie positive si le produit scalaire de tout vecteur $\vec{x}$ avec $A\vec{x}$ est strictement positif.

Question: Peut-t-on caractériser les matrices symétriques et définies positives ?

Théorème 3.2

Les énoncés suivants sont équivalents :

1. A est une matrice symétrique et définie positive ;
2. Toutes les valeurs propres de A sont réelles et strictement positives ;
3. Le déterminant des sous-matrices principales de A est strictement positif ;
4. Il existe une factorisation de Choleski $A = LL^t$.

Remarque(s) :

- Les critères du théorème sont difficiles à vérifier au préalable et il semble bien que la meilleure façon de s'assurer qu'une matrice est définie positive est d'y appliquer l'algorithme de factorisation de Choleski.
- Dans le cas d'une matrice symétrique non définie positive, on doit recourir à la factorisation LU classique.

Proposition 3.1

Soit A une matrice vérifiant

- A est symétrique et $a_{ii} > 0$,
- A est à diagonale strictement dominante, c'est-à-dire que,

$$|a_{ii}| > \sum_{j=1}^{n} |a_{ij}|,\ i = 1, \cdots, n,$$

alors A admet une décomposition de Cholesky.

Note: La proposition ne dit rien dans le cas où la matrice ne vérifie pas les conditions ; elle est donc une condition suffisante mais pas nécessaire.

4 Effets de l'arithmétique flottante

Jusqu'ici, nous n'avons effectué les opérations qu'à l'échelle humaine, c'est à dire en utilisant l'arithmétique exacte. Il est donc important de voir si l'arithmétique flottante, utilisée par les ordinateurs, a une influence sur ces calculs. Et nous allons voir que certaines matrices sont très sensibles aux effets de l'arithmétique flottante (appelées matrices mal conditionnées) et d'autres, très peu. Pour s'en convaincre, examinons l'exemple suivant :

Exemple 13

Soit le système suivant :

$$\begin{bmatrix} 1 & 2 \\ 1,1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 10 \\ 10,4 \end{bmatrix}$$

dont la solution exacte est $\vec{x} = [4 \quad 3]^t$. Si l'on remplace le terme $1,1$ de la matrice par $1,05$, la nouvelle solution exacte devient $\vec{x} = [8 \quad 1]^t$.

Remarque(s) : Cet exemple démontre qu'une petite modification sur un terme de la matrice peut entraîner une grande modification de la solution exacte. En pratique, l'arithmétique flottante provoque inévitablement de petites modifications de chaque terme de la matrice et de sa décomposition LU. Il est alors tout à fait possible que ces petites erreurs aient d'importantes répercussions sur la solution et, donc, que les résultats numériques soient très éloignés de la solution exacte

Remarque(s) : En arithmétique flottante à m chiffres dans la mantisse, on doit effectuer chaque opération arithmétique en représentant les opérandes en notation flottante et en arrondissant le résultat de l'opération au m^e chiffre de la mantisse (voir chapitre 1).

Exemple 14

Considérons le système linéaire suivant

$$\begin{bmatrix} 0,0003 & 3,0000 \\ 1,0000 & 1,0000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2,0001 \\ 1,0000 \end{bmatrix}$$

dont la solution exacte est $\vec{x} = \left[\frac{1}{3} \quad \frac{2}{3}\right]^t$. Il s'agit maintenant d'effectuer la décomposition LU en arithmétique flottante à 4 chiffres. Et pour cela, il faut d'abord effectuer la notation flottante à 4 chiffres du système linéaire

$$\begin{bmatrix} 0,3000 \times 10^{-3} & 0,3000 \times 10^{1} \\ 0,1000 \times 10^{1} & 0,1000 \times 10^{1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0,2000 \times 10^{1} \\ 0,1000 \times 10^{1} \end{bmatrix}$$

où on peut remarquer que le 1 de l'élément $2,0001$ disparaît. Ainsi la décomposition LU donne dans ce cas

$$LU = \begin{bmatrix} 0,3000 \times 10^{-3} & 0 \\ 0,1000 \times 10^{1} & -0,9999 \times 10^{4} \end{bmatrix} \begin{bmatrix} 0,1000 \times 10^{1} & 0,1000 \times 10^{5} \\ 0 & 0,1000 \times 10^{1} \end{bmatrix}$$

Remarque(s) : On peut remarquer que le terme u_{12} est très grand puisque le pivot $0,0003$ est presque nul.

En effectuant la descente triangulaire sur le système linéaire $L\vec{y} = \vec{b}$, on obtient

$$y_1 = fl\left(\frac{0,2000 \times 10^{1}}{0,3000 \times 10^{-3}}\right) = 0,6667 \times 10^{4}, \quad y_2 = fl\left(\frac{1 - 6667}{-9999}\right) = 0,6667.$$

Puis la remontée triangulaire du système $U\vec{x} = \vec{y}$ donne enfin la solution

$$x_2 = 0,6667, \quad x_1 = 6667 - (10000)(0,6667) = 0.$$

Remarque(s) : Si l'on compare ce résultat avec la solution exacte $\left[\frac{1}{3} \quad \frac{2}{3}\right]^t$, on constate une variation importante de la valeur de x_1. On imagine aisément ce qui peut se produire avec un système de plus grande taille.

Question: Comment peut-on limiter ou même éviter de tels dégâts ?

Une première possibilité consiste à utiliser plus de chiffres dans la mantisse, mais cela n'est pas toujours possible. Dans cet exemple, la source des ennuis est la division par un pivot presque nul remarquée précédemment. Et une solution de rechange consiste donc à permuter les lignes même si le pivot n'est pas parfaitement nul.

En reprenant dans notre exemple ce système linéaire avec une permutation de ligne, on aura

$$\begin{bmatrix} 1,0000 & 1,0000 \\ 0,0003 & 3,0000 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1,0000 \\ 2,0001 \end{bmatrix}$$

Cette fois, la décomposition LU en arithmétique flottante à 4 chiffres donne alors

$$LU = \begin{bmatrix} 0,1000 \times 10^{1} & 0,000 \\ 0,3000 \times 10^{-3} & 3,000 \end{bmatrix} \begin{bmatrix} 1,000 & 1,000 \\ 0,000 & 1,000 \end{bmatrix}$$

Ainsi la descente triangulaire donne $\vec{y} = [1,000 \quad 0,6666]^t$, alors que par la remontée triangulaire nous avons la solution $\vec{x} = [0,3333 \quad 0,6667]^t$ qui est très près de la solution exacte.

Remarque(s) : Une excellente stratégie de recherche du pivot consiste, une fois la i^e colonne de L calculée, à placer en position de pivot le plus grand terme en valeur absolue de cette colonne. Cette recherche ne tient compte que des lignes situées sous le pivot actuel.

Exemple 15

Lire les exemples 3.19 et 3.20 du livre [a] qui illustrent comment la stratégie du pivot et la mise à l'échelle améliorent la précision de la solution du système, respectivement !

a. André Fortin : Analyse numérique pour ingénieurs, Presses Internationales Polytechnique, Montréal, Québec (2016).

Définition 4.1

La mise à l'échelle consiste à diviser chaque ligne du système linéaire par le plus grand terme (en valeur absolue) de la ligne correspondante de la matrice A. *On ne tient pas compte du terme de droite $\vec{b}$ pour déterminer le plus grand terme de chaque ligne*.

5 Conditionnement d'une matrice

La section précédente illustre clairement la sensibilité qu'ont certains systèmes linéaires aux erreurs dues à l'arithmétique flottante. Ici, nous allons essayer de déterminer cette sensibilité en mesurant l'erreur liée aux systèmes linéaires. Cela nous amène à introduire la notion de métrique ou norme, permettant de mesurer l'écart entre une solution numérique et une solution exacte.

Définition 5.1

Une norme vectorielle est une application de $\mathbb{R}^n$ dans $\mathbb{R}$ ($\mathbb{R}$ désigne l'ensemble des réels) qui associe à un vecteur $\vec{x}$, un scalaire noté $\|\vec{x}\|$ qui vérifie les trois propriétés suivantes :

1. La norme d'un vecteur est toujours strictement positive, sauf si le vecteur a toutes ses composantes nulles
$$\|\vec{x}\| > 0, \text{ sauf si } \vec{x} = \vec{0}$$
2. Si α est un scalaire, alors
$$\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|$$
où $|\alpha|$ est la valeur absolue de α.
3. L'inégalité triangulaire est toujours vérifiée entre deux vecteurs $\vec{x}$ et $\vec{y}$ quelconques
$$\|\vec{x} + \vec{y}\| \leq \|\vec{x}\| + \|\vec{y}\|$$

Note: Toute application vérifiant ces trois propriétés est une norme vectorielle dont les plus connues sont la norme euclidienne, la norme l_1 et la norme l_∞.

Définition 5.2

La norme euclidienne d'un vecteur $\vec{x}$ est notée $\|\vec{x}\|_2$ et est définie par

$$\|\vec{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Tandis que les normes l_1 et l_∞ d'un vecteur $\vec{x}$, notées respectivement $\|\vec{x}\|_1$ et $\|\vec{x}\|_\infty$, sont définies par

$$\|\vec{x}\|_1 = \sum_{i=1}^{n} |x_i|, \quad \|\vec{x}\|_\infty = \max_{1\leq i\leq n} |x_i|$$

Exemple 16

Soit le vecteur $\vec{x} = [1 \quad -3 \quad -8]^t$. Sa norme euclidienne, l_1 et l_∞ sont définies par

$$\begin{aligned} \|\vec{x}\|_1 &= 1+3+8 &&= 12 \\ \|\vec{x}\|_\infty &= \max(1,3,8) &&= 8 \\ \|\vec{x}\|_2 &= \sqrt{1+9+64} &&= \sqrt{74} \end{aligned}$$

Définition 5.3

Une norme matricielle est une application qui associe à une matrice A un scalaire noté $\|A\|$ vérifiant les propriétés :

1. La norme d'une matrice est toujours strictement positive, sauf si la matrice a toutes ses composantes nulles

$$\|A\| > 0, \text{ sauf si } A = 0 \tag{3.25}$$

2. Si α est un scalaire, alors

$$\|\alpha A\| = |\alpha|\|A\| \tag{3.26}$$

3. L'inégalité triangulaire est toujours vérifiée entre deux matrices A et B quelconques, c'est-à-dire

$$\|A+B\| \leq \|A\| + \|B\| \tag{3.27}$$

4. Une quatrième propriété est nécessaire pour les matrices

$$\|AB\| \leq \|A\|\|B\| \tag{3.28}$$

Théorème 5.1

Pour toute norme vectorielle $\|\cdot\|_*$, l'application $\|A\|$ définie pour toute matrice A par

$$\|A\| = \sup_{\|\vec{x}\|_*=1} \|A\vec{x}\|_*$$

est une norme matricielle (car vérifiant les quatre propriétés (3.25)-(3.28)).

Note: Les normes matricielles induites par les normes vectorielles $\|\cdot\|_1$, $\|\cdot\|_\infty$ et $\|\cdot\|_2$ sont définies par

$$\begin{aligned} \|A\|_1 &= \sup_{\|\vec{x}\|_1=1} \|A\vec{x}\|_1 &&= \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \\ \|A\|_\infty &= \sup_{\|\vec{x}\|_\infty=1} \|A\vec{x}\|_\infty &&= \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}| \\ \|A\|_2 &= \sup_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2 &&= \left(\rho(A^tA)\right)^{\frac{1}{2}} \end{aligned}$$

où ρ désigne le rayon spectral, c'est-à-dire la plus grande valeur propre.

Remarque(s) : La norme suivante, dite de Frobenius, n'est induite par aucune norme vectorielle

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} a_{ij}^2}$$

Exemple 17

Soit la matrice A suivante

$$A = \begin{bmatrix} 1 & -2 & 5 \\ -3 & 1 & -5 \\ 1 & -9 & 0 \end{bmatrix}$$

Les différentes normes matricielles précédentes de la matrice A prennent alors les valeurs suivantes

$$\begin{aligned} \|A\|_1 &= \max(5, 12, 10) &&= 12 \\ \|A\|_\infty &= \max(8, 9, 10) &&= 10 \\ \|A\|_F &= \sqrt{1+4+25+9+1+25+1+81} &&= \sqrt{147} \end{aligned}$$

Définition 5.4

Une norme vectorielle et une norme matricielle sont dites *compatibles* si la condition

$$\|A\vec{x}\| \leq \|A\|\|\vec{x}\| \tag{3.29}$$

est valide quels que soient la matrice A et le vecteur $\vec{x}$.

Remarque(s) :

- Toutes les normes matricielles induites sont compatibles avec leurs normes vectorielles

$$\|A\vec{x}\|_1 \leq \|A\|_1\|\vec{x}\|_1, \quad \|A\vec{x}\|_\infty \leq \|A\|_\infty\|\vec{x}\|_\infty, \quad \|A\vec{x}\|_2 \leq \|A\|_2\|\vec{x}\|_2$$

- La norme de Frobenius est compatible avec la norme euclidienne

$$\|A\vec{x}\|_2 \leq \|A\|_F\|\vec{x}\|_2$$

Note: La norme matricielle induite par la norme euclidienne est rarement utilisée sauf dans un cadre théorique. On préférera la norme l_1 ou la norme l_∞, plus faciles et demandant moins d'effort de calcul.

Exemple 18

Considérons de nouveau le vecteur $\vec{x} = [1 \quad -3 \quad -8]^t$ et la matrice A suivante

$$A = \begin{bmatrix} 1 & -2 & 5 \\ -3 & 1 & -5 \\ 1 & -9 & 0 \end{bmatrix}$$

Ainsi le produit $A\vec{x}$ donne le vecteur $[-33 \quad 34 \quad 28]$ et donc

$$\|A\vec{x}\|_1 = 95 \quad \|A\vec{x}\|_\infty = 34 \quad \|A\vec{x}\|_2 = \sqrt{3029}$$

L'inégalité (3.29) devient, respectivement aux calculs précédents, ce-ci

$$\begin{aligned} 95 &\leq (12)(12) && \text{en norme } l_1 \\ 34 &\leq (10)(8) && \text{en norme } l_\infty \\ \sqrt{3029} &\leq (\sqrt{147})(\sqrt{74}) && \text{en norme euclidienne} \end{aligned}$$

Définition 5.5

Le conditionnement d'une matrice (noté $cond\ A$) est défini par le produit des normes de A et de son inverse

$$cond\ A = \|A\|\|A^{-1}\|$$

Note: Le conditionnement dépend de la norme matricielle utilisée et la norme $\|\cdot\|_\infty$ sera souvent utilisée. Quelle que soit la norme matricielle utilisée, le conditionnement d'une matrice A est un nombre supérieur ou égal à 1 :

$$1 \leq cond\ A < \infty$$

Si l'on revient au système linéaire $A\vec{x} = \vec{b}$ dont la solution exacte est $\vec{x}$, on sait que l'effet de l'arithmétique flottante est qu'une solution approximative $\vec{x}^*$ est obtenue. Il est donc souhaitable (même si ce n'est pas toujours le cas) que ces deux vecteurs soient près l'un de l'autre, c'est-à-dire que la norme de l'erreur $\vec{e} = \vec{x} - \vec{x}^*$ soit petite.

Le vecteur $\vec{r} = \vec{b} - A\vec{x}^* = A\vec{x} - A\vec{x}^* = A(\vec{x} - \vec{x}^*) = A\vec{e}$ est appelé **résidu** associé à la solution approximative $\vec{x}^*$.

Si l'on utilise des normes vectorielles et matricielles compatibles, en tenant en compte que $\vec{r} = A\vec{e}$ et $\vec{e} = A^{-1}\vec{r}$, on a

$$\|\vec{e}\| \leq \|A^{-1}\|\|\vec{r}\|, \quad \|\vec{r}\| \leq \|A\|\|\vec{e}\| \tag{3.30}$$

En regroupant ainsi les deux inégalités dans (3.30), on obtient

$$\frac{\|\vec{r}\|}{\|A\|} \leq \|\vec{e}\| \leq \|A^{-1}\|\|\vec{r}\| \tag{3.31}$$

Par ailleurs, en refaisant le même raisonnement avec les égalités $A\vec{x} = \vec{b}$ et $\vec{x} = A^{-1}\vec{b}$, on obtient

$$\frac{\|\vec{b}\|}{\|A\|} \leq \|\vec{x}\| \leq \|A^{-1}\|\|\vec{b}\| \quad \longrightarrow \quad \frac{1}{\|A^{-1}\|\|\vec{b}\|} \leq \frac{1}{\|\vec{x}\|} \leq \frac{\|A\|}{\|\vec{b}\|} \tag{3.32}$$

En multipliant les inégalités (3.31) et (3.32), on obtient le résultat fondamental

$$\frac{1}{cond\ A}\frac{\|\vec{r}\|}{\|\vec{b}\|} \leq \frac{\|\vec{e}\|}{\|\vec{x}\|} \leq cond\ A\frac{\|\vec{r}\|}{\|\vec{b}\|}$$

qu'on reformule dans le théorème suivant :

Théorème 5.2

Soit A une matrice de dimension n sur n et $\vec{b} \in \mathbb{R}^n$. Si le vecteur $\vec{x} \in \mathbb{R}^n$ est la solution du système linéaire $A\vec{x} = \vec{b}$ et $\vec{x}^* \in \mathbb{R}^n$ une solution approximative, alors on a

$$\frac{1}{cond\ A}\frac{\|\vec{r}\|}{\|\vec{b}\|} \leq \frac{\|\vec{e}\|}{\|\vec{x}\|} \leq cond\ A\frac{\|\vec{r}\|}{\|\vec{b}\|} \tag{3.33}$$

Remarque(s) : Plusieurs remarques s'imposent pour bien comprendre l'inégalité (3.33)

1. Le terme du milieu représente l'erreur relative entre la solution exacte $\vec{x}$ et la solution approximative $\vec{x}^*$.
2. Si le conditionnement de la matrice A est près de 1, l'erreur relative est coincée entre deux valeurs très près l'une de l'autre. Si la norme du résidu est petite, l'erreur relative est également petite et la précision de la solution approximative a toutes les chances d'être satisfaisante.
3. Par contre, si le conditionnement de la matrice A est grand, la valeur de l'erreur relative est quelque part entre 0 et un nombre possiblement très grand. *Il est donc à craindre que l'erreur relative soit alors grande, donc que la solution approximative soit de faible précision et même, dans certains cas, complètement fausse*.
4. *Même si la norme du résidu est petite, il est possible que l'erreur relative liée à la solution approximative soit quand même très grande*.
5. Plus le conditionnement de la matrice A est grand, plus on doit être attentif à l'algorithme de résolution utilisé.
6. Il importe de rappeler que, même si une matrice est bien conditionnée, un mauvais algorithme de résolution peut conduire à des résultats erronés.

Note: Le calcul de l'inverse d'une matrice étant numériquement coûteux, il n'est pas utile de calculer le conditionnement. On peut cependent déterminer une borne inférieure facilement computable de celui-ci à partir de (3.33)

$$cond\ A \geq \max\left\{\frac{\|\vec{e}\|\|\vec{b}\|}{\|\vec{x}\|\|\vec{r}\|}, \frac{\|\vec{x}\|\|\vec{r}\|}{\|\vec{e}\|\|\vec{b}\|}\right\} = \max\left\{\frac{\|\vec{x}-\vec{x}^*\|}{\|\vec{x}\|}\frac{\|\vec{b}\|}{\|\vec{b}-A\vec{x}^*\|}, \frac{\|\vec{x}\|}{\|\vec{x}-\vec{x}^*\|}\frac{\|\vec{b}-A\vec{x}^*\|}{\|\vec{b}\|}\right\}$$

On peut obtenir une autre inégalité qui illustre le rôle du conditionnement. Lorsque l'on résout sur ordinateur le système linéaire

$$A\vec{x} = \vec{b}$$

où la représentation des nombres n'est pas toujours exacte, on résout en fait

$$(A+E)\vec{x}^* = \vec{b}. \tag{3.34}$$

La matrice E représente une perturbation du système initial due par exemple aux erreurs de représentation sur ordinateur des coefficients de la matrice A et $\vec{x}^*$ est la solution du système perturbé (3.34). On a alors

$$\vec{x} = A^{-1}\vec{b} = A^{-1}\left((A+E)\vec{x}^*\right) = \left(I + A^{-1}E\right)\vec{x}^* = \vec{x}^* + A^{-1}E\vec{x}^*$$

On en conclut que $\vec{x} - \vec{x}^*$. En vertu de la relation (3.28) et de la compatibilité matricielle, on a

$$\|\vec{x}-\vec{x}^*\| \leq \|A^{-1}\|\|E\|\|\vec{x}^*\| = \frac{\|A\|\|A^{-1}\|\|E\|\|\vec{x}^*\|}{\|A\|}$$

qui conduit au théorème suivant.

Théorème 5.3

Soit A une matrice de dimension n sur n et $\vec{b} \in \mathbb{R}^n$. Si le vecteur $\vec{x} \in \mathbb{R}^n$ est la solution du système linéaire $A\vec{x} = \vec{b}$ et $\vec{x}^* \in \mathbb{R}^n$ une solution approximative, alors on a

$$\frac{\|\vec{x}-\vec{x}^*\|}{\|\vec{x}^*\|} \leq cond\ A\frac{\|E\|}{\|A\|} \tag{3.35}$$

Remarque(s) : Les remarques suivantes permettent de bien mesurer la portée de l'inégalité (3.35).

1. Le terme de gauche est une approximation de l'erreur relative entre la solution exacte et la solution du système perturbé. (On devrait avoir $\|\vec{x}\|$ au dénominateur pour représenter vraiment l'erreur relative.)
2. Le terme de droite est en quelque sorte l'erreur relative liée aux coefficients de la matrice A multipliée par le conditionnement de A.
3. Si $cond\ A$ est petit, une petite perturbation sur la matrice A entraîne une petite perturbation sur la solution $\vec{x}$.
4. Par contre, si $cond\ A$ est grand, une petite perturbation sur la matrice A pourrait résulter en une très grande perturbation sur la solution du système. Il est par conséquent possible que les résultats numériques soient peu précis et même, dans certains cas, complètement faux.

Note: La matrice E étant la perturbation de A, par définition de la précision machine ε_m et de la norme l_∞, on a

$$\|E\|_\infty \leq \varepsilon_m\|A\|_\infty$$

On peut ainsi réécrire la relation (3.35) sous la forme suivante

$$\frac{\|\vec{x}-\vec{x}^*\|_\infty}{\|\vec{x}^*\|_\infty} \leq \varepsilon_m cond\ A = \varepsilon_m\|A\|_\infty\|A^{-1}\|_\infty$$

Plus le conditionnement est élevé, plus la précision machine ε_m doit être petite (si la simple précision est insuffisante, on recourt à la double précision).

Exemple 19

Nous avons déjà considéré la matrice

$$A = \begin{bmatrix} 0,0003 & 3,0 \\ 1,0 & 1,0 \end{bmatrix}$$

dont l'inverse est

$$A^{-1} = \begin{bmatrix} -0,333367 & 1,0001 \times 10^{0} \\ 0,333367 & 1,0001 \times 10^{-4} \end{bmatrix}$$

On a ainsi un conditionnement d'environ 4, ce qui est relativement faible.

Nous avons vu que la résolution d'un système linéaire à l'aide de cette matrice, sans effectuer de permutation de lignes, aboutit à de mauvais résultats.

Cela démontre bien qu'un algorithme mal choisi (la décomposition LU sans permutation de lignes dans ce cas) peut se révéler inefficace, et ce, même si la matrice est bien conditionnée.

6 Systèmes non linéaires

Dans cette section, nous examinons les systèmes non linéaires et nous montrons comment les résoudre à l'aide d'une suite de problèmes linéaires, auxquels on peut appliquer diverses techniques de résolution comme la décomposition LU.

Les méthodes de résolution des systèmes non linéaires sont nombreuses (presque toutes les méthodes du chapitre 2 peuvent être généralisées aux systèmes non linéaires). Nous ne présentons ici que la méthode la plus importante et la plus utilisée en pratique, soit *la méthode de Newton*.

L'application de cette méthode à un système de deux équations non linéaires est suffisante pour illustrer le cas général.

Méthode : Le problème consiste à trouver le vecteur $\vec{x} = [x_1 \;\; x_2]^t$, vérifiant les 2 équations non linéaires suivantes

$$\begin{cases} f_1(x_1, x_2) = 0 \\ f_2(x_1, x_2) = 0 \end{cases} \tag{3.36}$$

où f_1 et f_2 sont deux fonctions de deux variables que nous supposons différentiables.

En révisant le développement de la méthode de Newton pour une équation non linéaire (voir le chapitre 2), on considère une approximation initiale (x_1^0, x_2^0) de la solution du système (3.36) qu'on cherchera à corriger.

Note: Cette approximation initiale est cruciale et doit toujours être choisie avec soin.

- Par un développement de Taylor en deux variables pour chacune des équations du système (3.36) (où les termes d'ordre supérieur sont négligés), on détermine une correction $\delta\vec{x} = (\delta x_1, \delta x_2)$ à (x_1^0, x_2^0) en résolvant le système

$$\begin{aligned} \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0)\delta x_1 + \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0)\delta x_2 &= -f_1(x_1^0, x_2^0) \\ \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0)\delta x_1 + \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0)\delta x_2 &= -f_2(x_1^0, x_2^0) \end{aligned} \tag{3.37}$$

  où sous la forme matricielle, on écrira le système comme suit

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x_1^0, x_2^0) & \frac{\partial f_1}{\partial x_2}(x_1^0, x_2^0) \\ \frac{\partial f_2}{\partial x_1}(x_1^0, x_2^0) & \frac{\partial f_2}{\partial x_2}(x_1^0, x_2^0) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} f_1(x_1^0, x_2^0) \\ f_2(x_1^0, x_2^0) \end{bmatrix} \tag{3.38}$$

- On construit ainsi une nouvelle approximation de la solution du système non linéaire (3.36) en posant

$$\begin{aligned} x_1^1 &= x_1^0 + \delta x_1 \\ x_2^1 &= x_2^0 + \delta x_2 \end{aligned}$$

- On cherchera par la suite à corriger (x_1^1, x_2^1) d'une nouvelle quantité $\delta\vec{x}$ en réitérant le procédé ci-dessus

Remarque(s) : Le système linéaire (3.38) s'écrit également sous une forme plus compacte

$$J(x_1^0, x_2^0)\vec{\delta x} = -\vec{R}(x_1^0, x_2^0) \tag{3.39}$$

où $J(x_1^0, x_2^0)$ désigne la matrice des dérivées partielles ou *matrice jacobienne* évaluée au point (x_1^0, x_2^0), $\vec{\delta x}$ est le vecteur des corrections relatives à chaque variable et où $-\vec{R}(x_1^0, x_2^0)$ est *le vecteur résidu* évalué en (x_1^0, x_2^0). Le déterminant de la matrice jacobienne est appelé *le jacobien*. Le jacobien doit bien entendu être différent de 0 pour que la matrice jacobienne soit inversible.

De manière plus générale, si le problème à résoudre consiste à trouver le ou les vecteurs $\vec{x} = [x_1 \;\; x_2 \;\; x_3 \;\; \cdots \;\; x_n]^t$ vérifiant les n équations non linéaires suivantes

$$\begin{cases} f_1(x_1, x_2, x_3, \cdots, x_n) = 0 \\ f_2(x_1, x_2, x_3, \cdots, x_n) = 0 \\ f_3(x_1, x_2, x_3, \cdots, x_n) = 0 \\ \qquad\qquad \vdots \qquad\qquad \vdots \\ f_n(x_1, x_2, x_3, \cdots, x_n) = 0 \end{cases} \tag{3.40}$$

on a la matrice jacobienne, évaluée au point $\vec{x}^i = (x_1, x_2, x_3, \cdots, x_n)$, qui est définie par

$$J(\vec{x}^i) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\vec{x}^i) & \frac{\partial f_1}{\partial x_2}(\vec{x}^i) & \cdots & \frac{\partial f_1}{\partial x_n}(\vec{x}^i) \\ \frac{\partial f_2}{\partial x_1}(\vec{x}^i) & \frac{\partial f_2}{\partial x_2}(\vec{x}^i) & \cdots & \frac{\partial f_2}{\partial x_n}(\vec{x}^i) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(\vec{x}^i) & \frac{\partial f_n}{\partial x_2}(\vec{x}^i) & \cdots & \frac{\partial f_n}{\partial x_n}(\vec{x}^i) \end{bmatrix}$$

Le vecteur résidu $\vec{R}(\vec{x}^i)$ évalué en $\vec{x}^i$ ainsi que le vecteur des corrections $\vec{\delta x}$ sont donnés par

$$\vec{R}(\vec{x}^i) = \begin{bmatrix} f_1(\vec{x}^i) \\ f_2(\vec{x}^i) \\ \vdots \\ f_n(\vec{x}^i) \end{bmatrix} \qquad \vec{\delta x} = \begin{bmatrix} \delta x_1 \\ \delta x_2 \\ \vdots \\ \delta x_n \end{bmatrix}$$

On arrive ainsi à l'algorithme général suivant :

Algorithme 6.1: Méthode de Newton appliquée aux systèmes [a] :

1. Étant donné ε_a, un critère d'arrêt
2. Étant donné N, le nombre maximal d'itérations
3. Étant donné $\vec{x}^0 = [x_1^0 \;\; x_2^0 \;\; x_3^0 \;\; \cdots \;\; x_n^0]^t$, une approximation initiale de la solution du système
4. Résoudre le système linéaire :

$$J(\vec{x}^i)\vec{\delta x} = -\vec{R}(\vec{x}^i) \tag{3.41}$$

et poser

$$\vec{x}^{i+1} = \vec{x}^i + \vec{\delta x}$$

5. Si $\frac{\|\vec{x}^i\|}{\|\vec{x}^{i+1}\|} < \varepsilon_a$ et $\|\vec{R}(\vec{x}^{i+1})\| \leq \varepsilon_a$:
   - convergence atteinte
   - écrire la solution $\vec{x}^{i+1}$
   - arrêt
6. Si le nombre maximal d'itérations N est atteint :
   - convergence non atteinte en N itérations
   - arrêt
7. Retour à l'étape 4

a. André Fortin : Analyse numérique pour ingénieurs, Presses Internationales Polytechnique, Montréal, Québec (2016).

Exemple 20

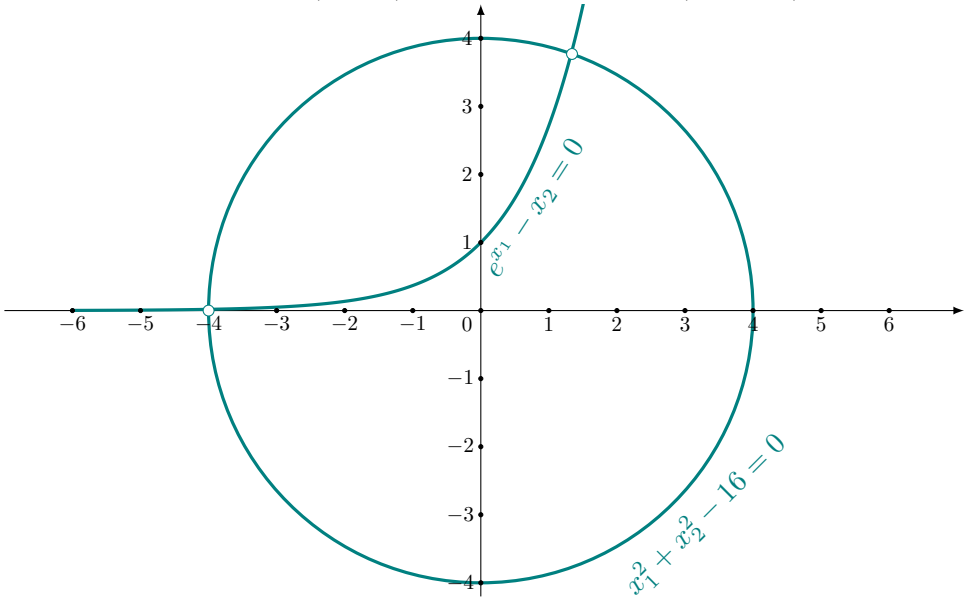
On cherche à trouver l'intersection de la courbe $x_2 = e^{x_1}$ et du cercle de rayon 4 centré à l'origine d'équation $x_1^2 + x_2^2 = 16$. L'intersection de ces courbes est une solution de

$$e^{x_1} - x_2 = 0$$
$$x_1^2 + x_2^2 - 16 = 0$$

La première étape consiste à calculer la matrice jacobienne de dimension deux $J(x_1, x_2)$

$$J(x_1, x_2) = \begin{bmatrix} e^{x_1} & -1 \\ 2x_1 & 2x_2 \end{bmatrix}$$

Un graphique de ces deux courbes montre qu'il y a deux solutions à ce problème non linéaire (voir la figure ci-dessous). La première solution se trouve près du point $(-4 \quad 0)$ et la deuxième, près de $(2,8 \quad 2,8)$.



Représentation des courbes d'équations $x_1^2 + x_2^2 - 16 = 0$ et $e^{x_1} - x_2 = 0$

Prenons le vecteur $\vec{x}_0 = [2,8 \quad 2,8]^t$ comme approximation initiale de la solution de ce système non linéaire.

1. À la première, le système (3.41) devient

$$\begin{bmatrix} e^{2,8} & -1 \\ 2(2,8) & 2(2,8) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} e^{2,8} - 2,8 \\ (2,8)^2 + (2,8)^2 - 16 \end{bmatrix} \longleftrightarrow \begin{bmatrix} 16,445 & -1 \\ 5,6 & 5,6 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 13,645 \\ -0,3200 \end{bmatrix}$$

   dont la solution est $\vec{\delta x} = [-0,77890 \quad 0,83604]^t$. La nouvelle approximation de la solution est donc

$$x_1^1 = x_1^0 + \delta x_1 = 2,8 - 0,77890 = 2,0211$$
$$x_2^1 = x_2^0 + \delta x_2 = 2,8 + 0,83604 = 3,63604$$

2. On effectue une deuxième itération à partir de $[2,0211 \quad 3,63604]^t$. Le système (3.41) devient alors

$$\begin{bmatrix} e^{2,0211} & -1 \\ 2(2,0211) & 2(3,63604) \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} e^{2,0211} - 3,63604 \\ (2,0211)^2 + (3,63604)^2 - 16 \end{bmatrix} \longleftrightarrow \begin{bmatrix} 7,5466 & -1 \\ 4,0422 & 7,2721 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 3,9106 \\ 1,3056 \end{bmatrix}$$

   dont la solution est $\vec{\delta x} = [-0,5048 \quad 0,10106]^t$. On a maintenant

$$x_1^2 = x_1^1 + \delta x_1 = 2,0211 - 0,50480 = 1,5163$$
$$x_2^2 = x_2^1 + \delta x_2 = 3,63604 + 0,10106 = 3,7371$$

3. À la troisième itération, on doit résoudre

$$\begin{bmatrix} 4,5554 & -1 \\ 3,0326 & 7,4742 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 0,81824 \\ 0,26508 \end{bmatrix}$$

   ce qui entraîne que $\vec{\delta x} = [-0,17208 \quad 0,034355]^t$. La nouvelle solution est

$$x_1^3 = x_1^2 + \delta x_1 = 1,5163 - 0,17208 = 1,3442$$
$$x_2^3 = x_2^2 + \delta x_2 = 3,7371 + 0,034355 = 3,7715$$

4. À la quatrième itération, on a le système linéaire suivant à résoudre

$$\begin{bmatrix} 3,8351 & -1 \\ 2,6884 & 7,5430 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 0,063617 \\ 0,031086 \end{bmatrix}$$

ce qui entraîne que $\vec{\delta x} = [-0,0161616 \quad 0,0163847]^t$. La nouvelle approximation de la solution est

$$x_1^4 = x_1^3 + \delta x_1 = 1,3442 - 0,0161616 = 1,3280$$
$$x_2^4 = x_2^3 + \delta x_2 = 3,7715 + 0,0163847 = 3,7731$$

5. À l'itération 5 et à partir de $[1,3280 \quad 3,7731]^t$, on doit résoudre

$$\begin{bmatrix} 3,7735 & -1 \\ 2,6560 & 7,5463 \end{bmatrix} \begin{bmatrix} \delta x_1 \\ \delta x_2 \end{bmatrix} = - \begin{bmatrix} 0,34886 \times 10^{-3} \\ 0,16946 \times 10^{-3} \end{bmatrix}$$

dont la solution est $\vec{\delta x} = [9,03 \times 10^{-5} \quad 9,25 \times 10^{-6}]^t$. La solution du système non linéaire devient

$$x_1^5 = x_1^4 + \delta x_1 = 1,3281$$
$$x_2^5 = x_2^4 + \delta x_2 = 3,7731$$

On déduit la convergence de l'algorithme de Newton du fait que les modules de $\vec{\delta x}$ et de $\vec{R}$ diminuent avec les itérations.

Remarque(s) :

1. La convergence de la méthode de Newton dépend de l'approximation initiale $\vec{x}^0$ de la solution. *Un mauvais choix de $\vec{x}^0$ peut résulter en un algorithme divergent*.
2. On peut démontrer que, lorsqu'il y a convergence de l'algorithme, cette convergence est généralement quadratique dans le sens suivant

$$\|\vec{x} - \vec{x}^{i+1}\| \simeq c\|\vec{x} - \vec{x}^i\|^2 \tag{3.42}$$

ce qui équivaut, en posant $\vec{e}^i = \vec{x} - \vec{x}^i$

$$\|\vec{e}^{i+1}\| \simeq c\|\vec{e}^i\|^2 \tag{3.43}$$

Cela signifie que la norme de l'erreur à l'itération $i+1$ est approximativement égale à une constante c multipliée par le carré de la norme de l'erreur à l'étape i. L'analogie est évidente avec le cas d'une seule équation non linéaire étudié au chapitre 2. En effet, on peut écrire les deux algorithmes sous la forme

$$x_{i+1} = x_i - (f'(x_i))^{-1} f(x_i) \quad \text{en dimension } 1$$
$$\vec{x}^{i+1} = \vec{x}^i - \left(J(\vec{x}^i)\right)^{-1} \vec{R}(\vec{x}^i) \quad \text{en dimension } n$$

3. La convergence quadratique est perdue si la matrice jacobienne est singulière au point $\vec{x}$, solution du système non linéaire. *Encore une fois, ce comportement est analogue au cas d'une seule équation où la méthode de Newton perd sa convergence quadratique si la racine est de multiplicité plus grande que 1 (où $f'(r) = 0$)*.

Exercices suggérés du manuel !

- Exercices [a] suggérés : 1-8, 10a), 11-14, 16, 21 26-28, 30, 31, 34, 36, 37.
- Exercices fortement suggérés : 2, 8, 10a), 16.

a. André Fortin : Analyse numérique pour ingénieurs, Presses Internationales Polytechnique 2016.

À Retenir !

Je dois pouvoir répondre aux questions [a] suivantes :

1. Je sais résoudre par une méthode d'élimination de Gauss un système linéaire et définir les matrices équivalentes aux opérations élémentaires.
2. Je sais faire une décomposition et une résolution par LU et Cholesky.
3. Je distingue Cholesky et Crout.

4. Je sais reconnaitre des matrices permettant d'appliquer ou d'exclure Cholesky.
5. Je comprends la notion de conditionnement d'une matrice.
6. Je connais le théorème sur le conditionnement.
7. Je sais appliquer le théorème de conditionnement pour caractériser une solution et son erreur relative.
8. Je sais construire une borne inférieure du conditionnement et je comprends ses limitations.
9. Je sais calculer une matrice Jacobienne.
10. Je sais utiliser la méthode de Newton.

a. André Fortin : Analyse numérique pour ingénieurs, Presses Internationales Polytechnique, Montréal, Québec (2016), pages 97-169.