République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Analyse**

Par

Afaf Zoubiri

Titre :

Effet des Erreurs d'arrondi sur les systèmes linéaires

Membres du Comité d'Examen :

Dr.	Dakhia Ghania	UMKB	Président
Dr.	Rajah Faouzia	UMKB	Encadreur
Dr.	Benaba Fadhila	UMKB	Examinateur

Juin 2018

Dédicace

Je dédie ce modeste travail à :

Ma mère que dieu la garde dans son vaste paradis.

A l'homme de ma vie, mon exemple éternel, mon soutient moral et source de joie et de bonheur : mon père.

Aux personnes dont j'ai bien aimé la présence dans ce jour, à tous mes frères et mes soeurs, je dédie ce travail pour leurs conseils, aides et encouragements.

REMERCIEMENTS

Nous tenons tout d'abord à remercier Dieux le tout puissant et miséricordieux, qui nous a donné la force et la patience d'accomplir ce modeste travail.

En second lieu, nous tenons à remercier notre encadreur **Rajah Faouzia**, son précieux conseil et son aide durant toute la période du travail.

Nos vifs remerciements vont également aux membres du jury pour l'intérêt qu'ils ont accepté d'examiner notre travail et de l'enrichir par leur proposition.

Nous tenons également à remercier toutes les personnes qui ont participé de près ou de loin à la réalisation de ce travail.

Je veux également remercier tout les membres de ma famille pour leurs encouragements.

Sans oublie le grand remerciement à ma soeur l'aînée **Latifa** qu'elle ma donnée le courage et la force durant toute mes études, "du primaire jusqu'à l'universitaire".

Enfin, je remercie tous mes amis(es).

Table des matières

Table des figures

Introduction

A cause des effets des erreurs d'arrondi pendant l'exécution numérique d'un problème, donc on cherche toujours à trouver des méthodes plus efficaces et plus performantes.

A cette raison, dans ce mémoire nous faisons appel à des méthodes numériques ayant des temps de calcul acceptables et de nombre d'opérations de l'ordre de n^3 pour résoudre un système linéaire $Ax = b$, où A est une matrice carrée d'ordre n.

Comme on a cité précédemment, notre objectif est donc de chercher à élaborer des méthodes numériques qui n'amplifient pas trop les erreurs d'arrondi au cours des calculs. Intuitivement, nous imaginons bien que cette amplification est d'autant plus grand que le nombre d'opérations est important (addition, multiplication, etc. . .). Donc, nous cherchons à réduire le nombre d'opérations au maximum et par conséquence le temps de calcul.

Ce travail que nous présentons ici est composé de trois chapitres.

Dans le premier chapitre, On va appliquer tout d'abord les méthodes directes pour résoudre un système linéaire de grande taille. Il est connu que ces méthodes ont l'inconvénient de nécessiter une assez grande place mémoire car elle nécessite le stockage de toutes les matrices en mémoire ; en plus, elles ont des effets négatifs des erreurs d'arrondi au cours de calcul numérique. On introduit aussi le concept de conditionnement d'une matrice et à la fin on fait une comparaison entre les méthodes directes étudiées.

Dans le deuxième chapitre, on va présenter les méthodes itératives les plus connues (Jacobi, Gauss-Seidel, Relaxation) et on va étudier les performances de chaque méthode

itérative. Ces méthodes génèrent une suite itérative de vecteurs $\left\{x^k\right\}; k \geq 0$ qui sont des approximations de la solution exacte. Théoriquement, ces méthodes fournissent une réponse approximative en un nombre infini d'opérations, mais pratiquement, on essaye de trouver la solution approchée par un nombre d'itérations relativement petite.

Dans le dernier chapitre, on va appliquer l'une des méthodes de résolution d'un système linéaire dans le domaine d'électronique et plus précisément dans les circuits électriques.

Chapitre 1

Accumulation des erreurs d'arrondi dans les calculs numériques

Dans ce chapitre, on essaye de démontrer l'effet négatif des erreurs d'arrondi au cours de calcul numérique, on expose quelques méthodes de résolution des systèmes linéaires (algébriques et numériques directes) afin de faire une petite comparaison entre ces méthodes pour choisir la plus performante. Pour aboutir à notre but, il est naturel de donner au début une idée générale sur les erreurs d'arrondi.

Les erreurs numériques ce sont des erreurs associées au système de numération. Elles sont dues au fait qu'un ordinateur ne peut prendre en considération qu'un nombre fini de chiffres.
Il est connu qu'il ya plusieurs façons de mesurer l'erreur entre une valeur approchée x^* et une valeur exacte x.

Définition (1.1) : *Soit x un nombre réel et x^*une approximation de x.*

L'erreur absolu $\triangle x$ est définie par $\triangle x = \mid x - x^* \mid$.
L'erreur relative est $\dfrac{\mid x - x^* \mid}{\mid x \mid} = \dfrac{\triangle x}{\mid x \mid}$.
Le pourcentage d'erreur est l'erreur relative multipliée par 100.

1.1 Systèmes linéaires

Un système linéaire de type $(n \times p)$; $n, p \geq 1$ est un ensemble de n équations linéaires à p inconnues de la forme :

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1p}x_p = b_1 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{np}x_p = b_n. \end{cases} \tag{1.1}$$

Les coefficients a_{ij} et les secondes membres b_i sont des éléments donnés de $\mathbb{R}$. Les inconnues $x_1, x_2, ..., x_p$ sont à chercher dans $\mathbb{R}$.

Le système homogène associé à 1.1 est la système obtenu en remplaçant les b_i par 0.

Une solution de 1.1 est un $p-$uplet $(x_1, x_2, ..., x_p) \in \mathbb{R}^n$ qui vérifier simultanément les n équations de 1.1.

Un système est impossible ou incompatible s'il n'admet pas de solution ; un système est possible ou compatible, s'il admet une ou plusieurs solutions .

Dans ce chapitre, nous ne traiterons que des systèmes linéaires carrés d'ordre n à coefficients réels. autrement dit $A \in M_n(\mathbb{R})$ et $b \in \mathbb{R}^n$; alors : $Ax = b$; $x \in \mathbb{R}^n$.
Dans ce cas, on est assuré de l'existence et l'unicité de la solution si A est inversible ; c'est-à-dire : $det(A) \neq 0$.

1.2 Méthodes directes

Une méthode directe de résolution d'un système est une méthode qui donne exactement la solution x et on l'utilise lorsque $n \leq 100$. Dans cette section, on va considérer trois méthodes directes (Cramer, Gauss et Cholesky).

1.2.1 Méthode de Cramer

Le principe est de résoudre le système $Ax = b$ par la formule $x_i = \dfrac{\det A_i}{\det A}$; avec x_i est la ième composante du vecteur x et A_i est la matrice A, où la ième colonne est remplacée par b.

En effet, soient $(c_1, c_2, \ldots, c_n)$ les n vecteurs colonnes de la matrice A associée au système linéaire, $b = (b_1, b_2, \ldots, b_n) \in \mathbb{R}^n$ le vecteur de second membre et $(x_1, x_2, \ldots, x_n)$ est l'unique solution de ce système. Alors :

$$x_i = \frac{\det(c_1, \ldots, c_{i-1}, b, c_{i+1}, \ldots, c_n)}{\det A}; \text{ pour } i = 1, 2, \ldots, n.$$

1.2.2 Méthode de Gauss

Le principe de cette méthode est de transformer le système $Ax = b$ à un système triangulaire supérieur. Donc, on se ramène à la résolution de ce dernier. Cette méthode est associée à la factorisation $A = LU$ de la matrice A avec, L triangulaire inférieure et U triangulaire supérieure. On peut résoudre le système $Ax = b$ en résolvant successivement les deux systèmes triangulaires $Ly = b$ puis $Ux = y$.

On considère le système linéaire $Ax = b$ et on pose : $b^{(1)} = b$ et $A^{(1)} = A = \left(a_{ij}^{(1)}\right)_{1 \leq i,j \leq n}$; le système s'écrit alors $A^{(1)}x = b^{(1)}$. Puisque A est inversible, On échange la première ligne de $A^{(1)}$ avec une autre (supposant que $a_{11}^{(1)} \neq 0$) ; le nombre $a_{11}^{(1)}$ est le premier pivot de l'élimination de Gauss.

Pour $i = 2, ..., n$, on multiplie la première équation $A^{(1)}x = b^{(1)}$ par $g_{i1} = \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}$ et on retranche l'équation obtenue a la ième équation du système. La ième ligne L_i devient donc $L_i - g_{i1}L_1$, on obtient alors un nouveau système $A^{(2)}x = b^{(2)}$; avec :

$$\begin{cases} a_{1j}^{(2)} = a_{1j}^{(1)}; \quad j = 1, \ldots, n \text{ et } b_1^{(2)} = b_1^{(1)} \\ a_{i1}^{(2)} = 0; \quad i = 2, \ldots, n \\ a_{ij}^{(2)} = a_{ij}^{(1)} - g_{i1} a_{1j}^{(1)} \text{ et } b_i^{(2)} = b_i^{(1)} - g_{i1} b_1^{(1)}; \ i, j = 2, \ldots, n. \end{cases}$$

La matrice $A^{(2)}$ est la vecteur $b^{(2)}$ sont donc de la forme :

$$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}, \ b^{(2)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix}.$$

Après K étapes de la même manière, on obtient $A^{(k)}x = b^{(k)}$, avec :

$$A^{(k)} = \begin{pmatrix} a_{11}^{(1)} & & \cdots & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \ddots & \ddots & \vdots & & \vdots \\ 0 & \ldots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}.$$

Supposant maintenant que le K ième pivot de l'élimination de Gauss est non nul ($a_{k,k}^{(k)} \neq 0$). Par le même principe que l'étape 1 et en utilisant $g_{ik} = \dfrac{a_{ik}^{(1)}}{a_{kk}^{(1)}}$ pour $i > k$, on obtient alors : $A^{(k+1)}x = b^{(k+1)}$. Finalement, le système $A^{(n)}x = b^{(n)}$ obtenu est triangulaire supérieure, avec :

$$A^{(n)} = \begin{pmatrix} a_{11}^{(1)} & \cdots & \cdots & a_{1,n}^{(1)} \\ 0 & a_{22}^{(2)} & \cdots & a_{1,n}^{(2)} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & a_{nn}^{(n)} \end{pmatrix}$$

On peut calculer les composantes de x en "*remontant*" x_n à la composante x_1 :

$$\begin{cases} x_n &= \dfrac{b^{(n)}}{a_{nn}^{(n)}} \\ x_i &= \dfrac{1}{a_{ii}^{(n)}}(b^{(i)} - \sum\limits_{j=i+1}^{n} a_{ij}x_j);\ i = n-1, \ldots, 1. \end{cases}$$

1.2.3 Méthode de Cholesky

La méthode de Cholesky est une alternative à l'élimination de Gauss qui s'applique au matrice symétrique et définie positive.

Définition (1.2) : *$A \in M_n(\mathbb{R})$ est dite symétrique définie positive si :*

1) $A = A^T$; où A^T désigne le transposée $a_{ij} = a_{ji}$ (A est symétrique).

2) $x^t Ax > 0$ pour tout $x \in \mathbb{R}^n$.

3) $x^t Ax = 0 \Longleftrightarrow x = 0$.

Théorème (1.1) *(Décomposition de Cholesky) : Si $A \in M_n(\mathbb{R})$ est une matrice symétrique définie positive ; alors, il existe une unique matrice triangulaire inférieure à valeurs diagonale positives ; notée R ; telle que : $A = RR^t$.* $\Diamond$

Preuve : On veut démontrer par récurrence. Si $n = 1$: c'est évident parce que $A = (a_{11})$, avec $a_{11} > 0$ et la seule solution est $R = \sqrt{a_{11}}$. On suppose que la propriété est vraie pour n et on démontre qu'elle est vraie pour $(n+1)$. Soit A une matrice symétrique définie positive d'ordre $(n+1)$; il faut montrer l'existence et l'unicité d'une matrice triangulaire inférieure R à coefficients diagonaux positives d'ordre $(n+1)$, telle que $A = RR^t$.

On peut écrire $A = \begin{pmatrix} A_n & L_n^T \\ L_n & a \end{pmatrix}$ et chercher R sous la forme $R = \begin{pmatrix} R_n & 0 \\ T_n & a_{11} \end{pmatrix}$; .où

- A_n est la sous matrice principale d'ordre n de A.
- R_n est triangulaire inférieure à coefficients diagonaux strictement positifs d'ordre n.
- L_n et T_n sont deux vecteurs lignes de taille n.
- a et a_{11} sont deux réels strictement positifs.

On doit prouver l'existence et l'unicité de R_n, T_n et a_{11}. Avec les notations précédentes :

$$RR^T = A = \begin{pmatrix} R_n & 0 \\ T_n & a_{11} \end{pmatrix} \begin{pmatrix} R_n^T & T_n^T \\ 0 & a_{11} \end{pmatrix} = \begin{pmatrix} A_n & L_n^T \\ L_n & a \end{pmatrix}$$

ce qui est équivalent à :

$$\begin{cases} A_n &= R_n R_n^T & (1) \\ L_n &= T_n R_n^T & (2) \\ a &= T_n T_n^T + a_{11}^2. & (3) \end{cases}$$

Puisque A_n est symétrique définie positive, L'hypothèse de récurrence assure que l'égalité (1) admet une solution R_n unique triangulaire inférieure à coefficients diagonaux > 0; cette matrice est inversible. L'égalité (2) donne alors $T_n = L_n(R_n^T)^{-1}$ de façon unique. Il reste l'égalité (3), qui s'écrit : $a_{11}^2 = a - T_n T_n^T$ (T_n est maintenant connu). Pour l'instant, on peut considérer que a_{11} est l'une des deux racines carrées complexes du scalaire $a - T_n T_n^T$. Avec un tel a_{11} provisoire, on a effectivement $A_n = R_n R_n^T$. On sait que $det(A) > 0$.

On a : $det(R) = (\det R_n)a_{11}$ donc : $\det(A) = \det(RR^T) = (\det(R))^2 = (\det(R_n))^2 a_{11}^2$ .Ainsi $\lambda^2 > 0$; .ceci prouve que a_{11} peut être choisi dans $\mathbb{R}^{*+}$ et ce d'une façon unique.

De cette manière, la résolution du système $Ax = b$ est équivalent à la résolution : d'un système triangulaire inférieur $Ry = b$ et un système triangulaire supérieur $R^T x = y$; où les élément de R sont donnés par : en commençant par les éléments de la première colonne :

$$\bullet \quad r_{i1} = \frac{a_{i1}}{r_{11}}; \text{ pour } i = 2 \text{ à } n, \tag{1.2}$$

ensuite les éléments diagonaux :

$$\bullet \quad r_{kk} = \sqrt{a_{kk} - \sum_{j=1}^{k-1} r_{kj}} \tag{1.3}$$

et enfin les autres éléments de R

$$\bullet \ r_{ik} = \frac{a_{ik} - \sum_{j=1}^{k-1} r_{ij} r_{kj}}{r_{kk}}; \text{ pour } i = k+1 \text{ à } n. \quad \blacklozenge \tag{1.4}$$

1.3 Complexité de calcule

Dans cette section, on présente ; en détaille ; le coût de calcul de chaque méthode qu'on a exposé précédemment.

1.3.1 Coût de l'algorithme de Gauss

Nous allons intéresser au calcul de nombre d'opérations nécessaires pendant l'exécution de la méthode de Gauss et la résolution du système $A^{(n)}x = b^{(n)}$.

Soit $A \in M_n(\mathbb{R})$ une matrice carrée inversible. A l'étape K de l'élimination de Gauss, on doit faire $(n-k)$ divisions pour calculer les g_{ik}, puis nous avons $(n-k)^2$ coefficients $a_{ij}^{(k+1)}$ à calculer et $(n-k)$ coefficients $b_i^{(k+1)}$, cela nécessite donc $(n-k)\ (n-k+1)$ multiplications puis $(n-k)\ (n-k+1)$ soustractions. Au total, on effectue :

$$\begin{aligned}
\sum_{k=1}^{n-1}(n-k) + 2\sum_{k=1}^{n-1}(n-k)(n-k+1) &= \sum_{k=1}^{n-1} k + 2\sum_{k=1}^{n-1} k(k+1) \\
&= \sum_{k=1}^{n-1} k + 2\sum_{k=1}^{n-1} k + \sum_{k=1}^{n-1} k^2 \\
&= 3\sum_{k=1}^{n-1} k + \sum_{k=1}^{n-1} k^2 \\
&= 3\frac{n(n-1)}{2} + \frac{n(n-1)(2n-1)}{6} \\
&= \frac{4n^3 + 3n^2 - 7n}{6}
\end{aligned}$$

opérations et la résolution d'un système triangulaire coûte n^2 opérations. Donc, au totale, la résolution du système $Ax = b$ par cette méthode coûtera :

$$G(n) \quad = n^2 + \frac{4n^3 + 3n^2 - 7n}{6} = \frac{4n^3 + 3n^2 - 7n}{6}$$

opérations. Lorsque n tend vers l'infini on a $G(n) \simeq \frac{2n^3}{3}$ opérations.

1.3.2 Coût de la méthode de Cholesky

Pour suivre le nombre d'opérations nécessaires au cours du calcule de R; la première colonne de cette dernière nécessite N opérations. Calculons, si $n = 1, \ldots, N-1$; pour la colonne $(n+1)$, le nombre d'opérations par ligne est $(2n+1)$, car le calcul de $r_{n+1,n+1}$ par la formule 1.3 nécessite n multiplications, n soustractions et une extraction de racine et le calcule de $r_{i,n+1}$ par la formule 1.4 nécessite n multiplications, n soustractions et une divisions, alors se fait en $(2n+1)$ opérations.

Comme les calculs se font des lignes, $n+1$ *à* N (car $r_{i,n+1} = 0$ pour $i \leq n$), le nombre d'opérations pour calculer la $(n+1)-$ième colonne est donc $(2n+1)(N-n)$. on en déduit que le nombre d'opérations N_R nécessaires au calcul de R est :

$$\begin{aligned} N_R \quad &= \sum_{n=0}^{N-1}(2n+1)(N-n) = 2N\sum_{n=0}^{N-1} n - 2\sum_{n=0}^{N-1} n^2 + N\sum_{n=0}^{N-1} 1 - \sum_{n=0}^{N-1} n \\ &= N^2(N-1) - \frac{2N(N+1)(2N+1)}{6} + N^2 - \frac{N(N-2)}{2} \\ &= (2N-1)\frac{N(N-1)}{2} + N^2 - \frac{2N(N+1)(2N+1)}{6} \\ &= \frac{N(2N^2+3N+1)}{6} = \frac{N^3}{3} + \frac{N^2}{2} + \frac{N}{6} \end{aligned}$$

opérations. Pour la résolution du système, au niveau de la ligne 1 ; le calcule de $y_1 = \frac{b_1}{r_{11}}$ s'effectue en une opération. Pour les lignes $n = 2$ à N, le calcul de $y_n = \frac{(b_n - \sum_{i=1}^{n-1} r_{i,n}x_i)}{r_{n,n}}$ s'effectue en $(n-1)$ multiplications, $(n-2)$ additions, une soustraction et une division$=$ $2n+1$ opérations. Le calcul de $y(Ry = b)$ s'effectue donc en $N_1 = \sum_{n=1}^{N}(2n-1) = N(N+1) - N = N^2$. On peut calculer d'une manière similaire le nombre d'opérations pour l'étape

de remontée $(R^T x = y)$; $N_2 = N^2$. Donc le nombre totale d'opérations pour calculer x; solution du système ; est $N_C = N_R + N_1 + N_2 = \dfrac{n^3}{3}$ opérations lorsque n tend vers l'infini.

1.3.3 Coût de la méthode de Cramer

La complexité de cette méthode, c'est-à-dire : le nombre d'opérations nécessaires pour calculer la solution est :

- N divisions.
- $(N+1)$ déterminants à calculer.
- $N!$ opérations pour calculer un déterminant.

Donc, la complexité vaut $(N+1)N! + N$

1.4 Conditionnement d'une matrice

On a vu précédemment que l'accumulation des erreurs d'arrondi au cours de la résolution d'un système linéaire donne une solution dans quelques cas est incomparable à la solution exacte. Dans ce qui suit, on introduit une notion importantes ; *conditionnement* ; qui quantifier la sensibilité de la solution d'un système linéaire $Ax = b$ lorsqu'on perturbe les données A et b.

Définition (1.3) : *Le conditionnement d'une matrice (noté cond$A$) est défini par :*

$$condA = \| A \| \| A^{-1} \|, \tag{1.5}$$

il s'agit simplement du produit de la norme de A et de la norme de son inverse. Ce nombre mesure la «sensibilité» de la solution par rapport aux données du problème.

Proposition (1.2) : *Soient $\mathbb{R}^n$ et $M_n(\mathbb{R})$ deux espaces muni d'une norme vectorielle et matricielle respectivement.*

1) Soit $A \in M_n(\mathbb{R})$ une matrice inversible, alors cond$(A) \geq 1$.

2) Soit $A \in M_n(\mathbb{R})$ une matrice inversible et $\alpha \in \mathbb{R}^*$, alors $cond(\alpha A) = cond(A)$.

3) Soient A et $B \in M_n(\mathbb{R})$ des matrices inversibles, alors $cond(AB) \leq cond(A).cond(B)$. $\Diamond$

Preuve : 1) Soit $\| \, . \, \|$ une norme matricielle dans $M_n(\mathbb{R})$, donc topologiquement, elle vérifie toutes les propriétés d'une norme et on veut démontrer que le conditionnement est un nombre supérieur ou égal â 1. En effet, si I désigne la matrice identité ; on a : $\| A \| = \| AI \| \leq \| A \| \| I \|$. Lorsqu'on divise les deux membres de l'inégalité par $\|A\|$, on obtient $\| A \| \geq 1$. De cette façon, on conclut que : $1 \leq \| I \| = \| AA^{-1} \| \leq \| A \| \| A^{-1} \|$ et donc : $1 \leq cond(A) \leq \infty$.

2) Par définition, on a :

$$\begin{aligned} cond(\alpha A) &= \| \alpha A \| \| (\alpha A)^{-1} \| = | \alpha | \| A \| \frac{1}{| \alpha |} \| A^{-1} \| \\ &= \| A \| \| A^{-1} \| = cond(A). \end{aligned}$$

3) Soient A et B deux matrices inversibles, alors algébriquement AB est une matrice inversible ; donc, il est possible d'écrire :

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

1.4.1 Estimation théorique de l'erreur a priori

Dans les parties suivantes, Considérons le système linéaire $Ax = b$, et on note par x la solution exacte et par x^* la solution approchée. D'une façon générale, ces deux vecteurs sont près l'une de l'autre, c'est -à-dire : la norme de l'erreur $\| e \| = \| x - x^* \|$ est petite. Ce n'est pas toujours le cas.

Premier cas : b est perturbé

Théorème (1.3) : *Soit $A \in M_n(\mathbb{R})$ inversible et $b \in \mathbb{R}^n$; tels que : $Ax = b$ et $A(x + \triangle x) = b + \triangle b$ avec $x \neq 0$. Alors, on a :* $\frac{1}{cond(A)} \frac{\| \triangle b \|}{\| b \|} \leq \frac{\| \triangle x \|}{\| x \|} \leq cond(A) \frac{\| \triangle b \|}{\| b \|}$. $\Diamond$

Preuve : Définissons le résidu par : $\triangle b = b - Ax^*$, on a alors : $\triangle b = b - Ax^* = Ax - Ax^* = A(x - x^*) = Ae$, donc $e = A^{-1}\triangle b$. Si en tenant compte les normes vectorielles et matricielles compatibles, on trouve :

$$\| e \| \leq \| A^{-1} \| \| \triangle b \| . \tag{1.6}$$

De la même manière et puisque $Ae = \triangle b$: $\| \triangle b \| \leq \| A \| \| e \|$ qu'on peut l'écrire aussi

$$\frac{\| \triangle b \|}{\| A \|} \leq \| e \| . \tag{1.7}$$

De 1.6 et 1.7, on obtient :

$$\frac{\| \triangle b \|}{\| A \|} \leq \| e \| \leq \| A^{-1} \| \| \triangle b \| \tag{1.8}$$

Par le même raisonnement appliqué sur $Ax = b$ et $x = A^{-1}b$, on trouve :

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

alors :

$$\frac{1}{\| A^{-1} \| \| b \|} \leq \frac{1}{\| x \|} \leq \frac{\| A \|}{\| b \|}. \tag{1.9}$$

En multipliant 1.8 et 1.9 terme à terme, on obtient immédiatement le résultat. ♦

Remarque : *Il est important de remarquer que si le conditionnement de la matrice A est très proche de 1, l'erreur relative est entre deux valeurs très près l'une de l'autre. Si la quantité $\| \triangle b \|$ est petite, alors l'erreur relative est également petite ; donc l'approche de la solution est très précise. Il importe de rappeler que, même si une matrice est bien conditionnée, un mauvais algorithme de résolution peut conduire à des résultats erronés.*

Deuxième cas : A est perturbée

Le théorème suivant donne immédiatement le majorant de l'erreur relative si on fait une

perturbation sur A.

Théorème (1.4) : *Soit $A \in M_n(\mathbb{R})$ inversible, $b \in \mathbb{R}^n$ et $\triangle A \in M_n(\mathbb{R})$; tel que : $\| A^{-1} \|\| \triangle A \| < 1$. Alors $A + \triangle A$ est inversible, de plus, si on suppose $Ax = b$ et $(A + \triangle A)x^* = b$ avec $x \neq 0$, alors :*

$$\frac{\| x - x^* \|}{\| x \|} \leq cond(A) \frac{\| \triangle A \|}{\| A \|}. \quad \Diamond$$

Preuve : Lorsqu'on effectue la résolution d'un système linéaire $Ax = b$ sur un ordinateur ne peut faire notre calculs qu'approximativement. A cette raison, la résolution effectif se fait par : $(A + \triangle A)x^* = b$, où la matrice $\triangle A$ représente une perturbation du système initial et x^* est la solution du système perturbé. On a :

$$\begin{aligned} x &= A^{-1}b = A^{-1}[(A + \triangle A)x^*] \\ &= (I + A^{-1}\triangle A)x^* = x^* + A^{-1}\triangle A x^*, \end{aligned}$$

ce qui est équivalent à : $x - x^* = A^{-1}\triangle A x^*$. On introduit la norme, on trouve :

$$\| x - x^* \| \leq \| A^{-1} \|\| \triangle A \|\| x^* \| = \frac{\| A \|\| A^{-1} \|\| \triangle A \|\| x^* \|}{\| A \|}.$$

Donc : $\dfrac{\| x - x^* \|}{\| x^* \|} \leq cond(A) \dfrac{\| \triangle A \|}{\| A \|}$. $\blacklozenge$

Troisième cas : A et b sont perturbés :

Le théorème qu'on va exposer présente la valeur maximale de l'erreur relative de la solution x si on perturbe les deux donnés A et b.

Théorème (1.5) : *Soit $A \in M_n(\mathbb{R})$ inversible. $b \in \mathbb{R}^n$ et $\triangle A \in M_n(\mathbb{R})$ vérifiant $\| A^{-1} \|\| \triangle A \| < 1$, si l'on suppose que : $Ax = b$ et $(A + \triangle A)(x + \triangle x) = b + \triangle b$, avec*

$x \neq 0$, *alors on a :*

$$\frac{\| \triangle A \|}{\| x \|} \leq \frac{cond(A)}{1 - \frac{\|\triangle A\|}{\|A\|}} (\frac{\| \triangle b \|}{\| b \|} + \frac{\| \triangle A \|}{\| A \|}). \quad \Diamond$$

1.4.2 Estimation théorique de l'erreur a posteriori

Maintenant, on va estimer ; en fonction du conditionnement ; l'erreur commise sur la solution du système linéaire $Ax = b$. Soit x la solution exacte et y la solution obtenue par la machine. On pose le résidu $r = Ay - b$ et on obtient le résultat suivant :

Théorème (1.6) : *Avec les notations précédentes, si* $x \neq 0$, *alors :*

$$\| y - x \| \leq cond(A) \frac{\| r \|}{\| b \|} \| x \| . \quad \Diamond$$

Preuve : On a : $r = Ay - b = Ay - Ax = A(y - x)$. Par une prémultiplication par A^{-1}, on trouve : $y - x = A^{-1}r$. En norme : $\| y - x \| \leq \| A^{-1} \| \| r \|$ ce qui est équivalent à : $\| y - x \| \leq cond(A) \frac{\| r \|}{\| A \|}$. Car $Ax = b$, on a alors : $\| A \| \geq \frac{\| b \|}{\| x \|}$, Ce qui implique : $\frac{1}{\| A \|} \leq \frac{\| b \|}{\| x \|}$ par la substitution directe, on trouve directement le résultat. ♦

Cette majoration n'est pas très facile à utiliser car le conditionnement est en générale inconnu. Soit C une approximation de A^{-1} et $R = AC - I_n$, le résultat suivant, nous permet de calculer l'erreur relative d'une manière plus simple.

Théorème (1.7) : Avec les notations précédentes, si $\| R \| \leq 1$, alors :

$$\| y - c \| \leq \frac{\| r \| \| C \|}{1 - \| R \|}. \quad \Diamond$$

Dans le cas où la matrice est symétrique, on peut diminuer l'espace mémoire nécessaire à la résolution d'un système linéaire (le cas de Cholesky). Dans un premier temps, la matrice étant symétrique. On peut se limiter à ne mettre en mémoire que la moitié inférieure de la matrice d'un coté et d'un autre coté, si $A \in M_n(\mathbb{R})$ est inversible, la résolution du système

$Ax = b$ par la méthode de Gauss demande $\frac{2n^3}{3}$ opérations et dans le cas d'une matrice symétrique définie positive la méthode de Cholesky demande $\frac{n^3}{3}$ opérations et la méthode de Cramer demande $(n+1)n! + n$ opérations. A titre d'exemple pour $N = 10$, la méthode de Gauss nécessite 700 opérations, la méthode de Cholesky 350 et la méthode de Cramer 40000000 opérations, cette dernière méthode est donc proscrire.

Chapitre 2

Performance des méthodes itératives

Si le système $Ax = b$ est de dimension grand, les méthodes directes qu'on a étudié dans le chapitre précédent n'est plus applicable parce qu'elles demandent souvent trop d'espace mémoire et trop de calculs. Il faut utiliser des méthodes plus performantes et plus efficaces. Ces méthodes connues sous le nom : *"méthodes itératives"*, le principe général d'une méthode itérative pour résoudre un système linéaire est de générer une suite de vecteurs converge vers la solution exacte $A^{-1}b$.

2.1 Ecriture matricielle d'une méthode itérative

Pour aboutir à notre but, le système $Ax = b$, sera écrire sous une forme équivalente permettant de voir la solution comme un point fixe d'une certaine fonction

$$Ax = b \Longleftrightarrow Bx + c = x, \tag{2.1}$$

avec $B \in M_n(\mathbb{R})$ et $c \in \mathbb{R}^n$ bien choisis ; c'est-à-dire : $I - B$ inversible et $c = (I - B)A^{-1}b$. Par exemple, si $A = M - N$ pour deux matrice $M, N \in M_n(\mathbb{R})$ avec M inversible, on peut choisir $B = M^{-1}N$ et $c = M^{-1}b$. Dans la suite on suppose toujours que $B \in M_n(\mathbb{R})$ et $c \in \mathbb{R}^n$.

On se donne alors un vecteur $x^{(0)} \in \mathbb{R}^n$et on construit une suite de vecteurs $x^{(k)} \in \mathbb{R}^n$ à l'aide du schéma itératif :

$$x^{(k+1)} = Bx^{(k)} + c;\ k = 1, 2, \ldots \tag{2.2}$$

Si la suite $(x^{(k)})_{k\epsilon N}$ est convergente, alors elle converge vers la solution du système $A^{-1}b$. En effet, si la limite x^* existe, donc ce dernier est un point fixe de la fonction $x \to Bx + c$; c'est-à-dire : $x^* = Bx^* + c$ qui est équivalent à $Ax^* = b$ d'après 2.1. Pratiquement, une méthode itérative de la forme 2.2 nécessite la donnée d'un point de départ $x^{(0)}$ et d'une précision sur la solution que l'on cherche à calculer. On calcule ensuite les itérés $x^{(k)}$, $k \geq 1$. En utilisant la formule 2.2 Jusqu'à ce que le résidu $b- Ax^{(k)}$ soit plus petit que la précision proposée.

2.2 Convergence d'une méthode itérative

Une méthode itérative 2.2 de résolution d'un système linéaire $Ax = b$ est dite convergente si pour toute valeur initiale $x^{(0)}$, on a : $\lim\limits_{k\to+\infty} x^{(k)} = A^{-1}b$.

Lemme (2.1) : *Si la méthode itérative 2.2 est convergente et si on note $x = A^{-1}b$ est la solution exacte du système, on a alors : $x^{(k)} - x = B^k(x^{(0)} - x)$.* $\Diamond$

Preuve : On a vu précédemment que : $c = (I_n - B)A^{-1}b = (I_n - B)x$, c'est-à-dire : $x^{(k+1)} = Bx^{(k)} + (I_n - B)x$ ce qui équivalent à : $x^{(k+1)} - x = B^k(x^{(k)} - x)$; par récurrence finie, on obtient immédiatement le résultat. $\blacklozenge$

Il est clair que : $x^{(k)} - x$ représente l'erreur à la *k-ième* étape, alors cette formule nous permet d'estimer cette erreur en fonction de l'erreur initiale.

Le résultat suivant présente les critères nécessaires pour tester la convergence d'une méthode itérative 2.2.

Théorème (2.2) : *Les assertions suivantes sont équivalentes :*

i) La méthode itérative 2.2 est convergente.

ii) $\rho(B) < 1$*, où* $\rho(B)$ *désigne le rayon spectral de la matrice* B*, c'est-à-dire : le maximum des modules des valeurs propres de* B*.*

iii) Il existe une norme matricielle $\| \, . \, \|$ *sur* $M_n(\mathbb{R})$ *subordonnée à une norme vectorielle sur* $\mathbb{R}^n$*; telle que :* $\| B \| < 1$. $\Diamond$

Preuve : Au début, on démontre tout d'abord que si la méthode itérative converge, *alors* $\rho(B) < 1$. On a :

$$x^{(k+1)} = Bx^{(k)} + c \text{ et } x = Rx + c. \tag{2.3}$$

La suite (méthode itérative) converge, c'est-à-dire :

$$\begin{aligned}
\lim_{k\to+\infty}(x^{(k+1)} - x) = 0 &\iff \lim_{k\to+\infty}(Rx^{(k)} + c - (Rx + c)) = 0 \\
&\iff \lim_{k\to+\infty} R(x^{(k)} - x) = 0 \iff \lim_{k\to+\infty} R(Rx^{(k-1)} + c - (Rx + c) = 0 \\
&\iff \lim_{k\to+\infty} R^2(Rx^{(k-1)} - x) = 0 \\
&\iff \lim_{k\to+\infty} R^3(Rx^{(k-2)} - x) = 0 \\
&\iff \lim_{k\to+\infty} R^{(k+1)}(x^{(0)} - x) = 0 \iff \lim_{k\to+\infty} R^{(k+1)} = 0 \iff p(B) < 1.
\end{aligned}$$

On veut démontrer que : $p(B) < 1 \iff \lim_{k\to+\infty} R^{(k+1)} = 0$.

Si $\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ sont les valeurs propres de R, donc : $\exists\ P \in M_n(R)$; telle que :

$$PRP^{-1} = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} = D.$$

En multipliant cette dernière matrice $(k+1)$ fois, on obtient :

$$PR^{(k+1)}P^{-1} = D^{(k+1)} = \begin{pmatrix} \lambda_1^{(k+1)} & 0 & \cdots & 0 \\ 0 & \lambda_2^{(k+1)} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n^{(k+1)} \end{pmatrix}.$$

On passe à la limite :

$$\begin{aligned} \lim_{k\to+\infty} PR^{(k+1)}P^{-1} &= \lim_{k\to+\infty} D^{(k+1)} = 0 \\ &\iff P \lim_{k\to+\infty} R^{(k+1)}P^{-1} = \lim_{k\to+\infty} D^{(k+1)} = 0 \\ &\iff \lim_{k\to+\infty} R^{(k+1)} = 0. \end{aligned}$$

Maintenant, on démontre l'implication inverse : $\rho(B) < 1 \Rightarrow$ la méthode itérative converge, c'est-à-dire : $\lim_{k\to+\infty} x^{(k+1)} = x$. On a :

$$\begin{aligned} x^{(k+1)} &= Rx^{(k)} + c \\ &= R(Rx^{(k-1)} + c) + c \\ &= R^2(Rx^{(k-2)} + c) + Rc + c \\ &\vdots \\ &= R^{(k+1)}x^{(0)} + R^{(k)}c + R^{(k-1)}c + \cdots + R^2c + Rc + c \\ &= R^{(k+1)}x^{(0)} + (R^{(k)} + R^{(k-1)} + \cdots + R^2 + R + I)c. \end{aligned}$$

Puisque on a : $\rho(B) < 1 \iff \lim_{k\to+\infty} R^{(k+1)} = 0$, on trouve immédiatement :

$$\begin{aligned} \lim_{k\to+\infty} x^{(k+1)} &= \lim_{k\to+\infty} R^{(k+1)} + (I + R + R^2 + R^3 + \cdots)c \\ &= (I - R)^{-1}c = x. \end{aligned}$$

Alors, la suite (méthode itérative) converge. ♦

Dans ce qui suit, on essaye de donner une idée générale sur la vitesse de convergence

que l'on utilise pour mesurer l'efficacité d'une méthode itérative. L'égalité donnée précédemment $x^{(k)} - x = B^k(x^{(0)} - x)$ implique que la norme des puissances de la matrice B joue un rôle très important dans la vitesse de convergence d'une méthode itérative. Nous définissons ici les outils permettant de comparer les vitesses de convergence de différentes méthodes itératives.

Définition (2.1) : Soient $\| \, . \, \|$ une norme matricielle sur $M_n(\mathbb{R})$ est k un entier ; tel que : $\| \, B^k \, \| < 1$. Dans le cas où le schéma itératif 2.2 converge, on appelle taux moyen de convergence associé à la norme $\| \, . \, \|$ pour k itérations, le nombre positif : $R_k(B) = -\ln([\| \, B \, \|]^{\frac{1}{K}})$.

Si en tenant compte ce nombre positif, on peut maintenant faire une comparaison entre les méthodes itératives.

Définition (2.2) : *Considérons deux méthodes itératives convergentes :*

1) $x^{(k+1)} = B_1 x^{(k)} + c_1$ *;* $k = 1, 2, \ldots$

2) $x^{(k+1)} = B_2 x^{(k)} + c_2$; $k = 1, 2, \ldots$

Soit k un entier, tel que : $\| \, B_1^k \, \| < 1$ *et* $\| \, B_2^k \, \| < 1$. *On dit que (1) est plus rapide que (2) relativement à la norme* $\| \, . \, \|$ *si* $R_k(B_1) \geq R_k(B_2)$.

En pratique le calcule des $R_k(B)$ est trop couteux car il nécessite l'évaluation des B^k. On préfère donc estimer le taux asymptotique de convergence.

Définition (2.3) : *Le taux asymptotique de convergence est le nombre :*
$R_\infty(B) = \lim R_k(B) = -\ln(\rho(B))$.

Avec les notions précédentes, on peut dire qu'une méthode itérative est plus rapide si son taux asymptotique de convergence est grand ; c'est-à-dire : si $\rho(B)$ est petit.

2.3 Méthodes itératives classiques

On considère un système linéaire $Ax = b$ avec A inversible. La première étape est de décomposer la matrice A sous la forme $A = M - N$ où M est une matrice inversible d'une part et d'autre part les systèmes de matrice M sont *‹‹faciles››* à résoudre ; pour assurer ça, on peut prendre M diagonale ou triangulaire. Le système $Ax = b$ s'écrit alors : $Mx = Nx + b$, c'est-à-dire : $x = Bx + c$, avec $B = M^{-1}N$ et $c = M^{-1}b$ et on considère le schéma itératif associé :

$$x \in \mathbb{R}^n,\ Mx^{(k+1)} = Nx^{(k)} + b.$$

Nous allons considérer trois méthodes classiques : Jacobi, Gauss-Seidel et relaxation. Le point de départ de chacune de ces méthodes est l'unique décomposition de la matrice $A = (a_{ij})_{1\leq i,j\leq n}$ sous la forme $A = D - E - F$, avec :

- $D = (d_{ij})_{1\leq i,j\leq n}$ diagonale, telle que : $d_{ii} = a_{ii}$ et $d_{ij} = 0$ pour $i \neq j$.
- $E = (e_{ij})_{1\leq i,j\leq n}$ triangulaire inférieure stricte, telle que : $e_{ij} = -a_{ij}$ si $i \geq j$ et $e_{ij} = 0$ si $i \leq j$.
- $F = (f_{ij})_{1\leq i,j\leq n}$ triangulaire supérieure stricte, telle que : $f_{ij} = -a_{ij}$ si $i \leq j$ et $f_{ij} = 0$ si $i \geq j$.

Dans les sous sections suivantes, on présente que chaque méthode itérative basée sur le choix des deux matrices de la décomposition M et N ; Comme on remarque aussi que la méthode de Gauss-Seidel est un cas particulier de la méthode de relaxation.

2.3.1 Méthode de Jacobi

On a toujours un système linéaire $Ax = b$ avec A inversible. On pose : $A = M - N$ avec $M = D$ inversible et $N = F + E$. Le schéma itératif s'écrit alors :

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

où $B_J = D^{-1}(E+F)$ est la matrice de Jacobi associée à A.

Complexité arithmétique

On veut estimer le nombre d'opérations nécessaires pour calculer récursivement $x^{(k+1)}$ à partir de x^k. On a : $Dx^{(k+1)} = (E+F)x^{(k)} + b$, donc la $ième$ composante est $(Dx^{(k+1)})_i = ((E+F)x^k)_i + b_i$; c'est-à-dire :

$$a_{ii}x_i^{(k+1)} = -\sum_{j=1(j\neq i)}^{n} a_{ij}x_j^{(k)} + b_i \Longleftrightarrow x_i^{(k+1)} = \frac{1}{a_{ii}}[(b_i - \sum_{j=1(j\neq i)}^{n} a_{ij}x_j^{(k)})].$$

Pour calculer $x_i^{(k+1)}$ à partir de $x_i^{(k)}$, on a besoin de $(n-1)$ multiplications, $(n-1)$ additions et une division, donc $(2n-1)$ opérations. Par conséquent, il faut exécuter $n(2n-1)$ opérations pour calculer $x^{(k+1)}$ à partir de $x^{(k)}$; et pour k itérations, on aura besoin de $kn(2n-1)$ opérations. A titre d'exemple le processus de calculer : $k = 100$ itérations de la méthode de Jacobi coûte approximativement $2kn^2 = 2 \times 10^8$ opérations.

D'après le théorème principal de convergence, il est facile de déduire que la méthode de Jacobi converge si et seulement si $\rho(B_J) < 1$.

2.3.2 Méthode de Gauss-Seidel

Sous les mêmes hypothèses, on pose : $A = M - N$. Dans cette méthode, on peut prendre $M = D - E$ inversible et $N = F$. Le schéma itératif s'écrit alors :

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

Complexité arithmétique

Pour trouver la complexité de cette méthode pour calculer $x^{(k+1)}$ à partir de $x^{(k)}$, on a : $(D-E)x^{(k+1)} = Fx^{(k)} + b$, donc pour tout $i = 1, \ldots, n$; $((D-E)x^{(k+1)})_i = (Fx^{(k)})_i + b_i$

c'est-à-dire :

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

qu'on peut l'écrire encore :

$$x_1^{(k+1)} = \frac{1}{a_{11}}[b_1 - \sum_{j=2}^{n} a_{1j}x_j^{(k)}]$$

et

$$x_i^{(k+1)} = \frac{1}{a_{ii}}[b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}];\ i = 2, \ldots, n.$$

La complexité arithmétique de la méthode de Gauss-Seidel est la même que celle de la méthode de Jacobi. Il est facile de remarquer que la méthode de Gauss-Seidel est plus économique de la mémoire de la machine parce qu'on peut écraser la valeur de $x_i^{(k)}$ et ne stocker au cours des calculs qu'un seul vecteur de taille n suivant : $(x_1^{(k+1)} \cdots x_i^{(k+1)} x_{i+1}^{(k)} \cdots x_n^{(k)})^T$ au lieu de deux vecteurs pour la méthode de Jacobi. D'après le même argument, on peut dire que : La méthode de Gauss-Seidel converge si et seulement si ; $\rho(B_{cs}) < 1; B_{cs} = (D-E)^{-1}F$.

2.3.3 Méthode de relaxation

On considère toujours un système linéaire $Ax = b$ avec A inversible et w est un paramètre réel non nul. On pose $A = M - N$ avec $M = \frac{1}{w}(D - wE)$ inversible et $N = (\frac{1-w}{w})D + F$. Donc l'écriture matricielle de cette méthode prend la forme :

$$\frac{1}{w}(D - wE)x^{(k+1)} = ((\frac{1-w}{w})D + F)x^{(k)} + b,$$

ce qui est équivalent à :

$$x^{(k+1)} = (D - wE)^{-1}[(1-w)D + wF]x^{(k)} + w(D - wE)^{-1}b,$$

où $B_R(w) = (D - wE)^{-1}[(1 - w)D + wF]$ s'appelle la matrice de relaxation associée à A et w est le facteur de relaxation.

• si $w < 1$, on parle de sous-relaxation.

• si $w = 1$, on retrouve la méthode de Gauss-Seidel.

• si $w > 1$, on parle de sur-relaxation.

On se basant sur le théorème de convergence pour conclure que la méthode de relaxation converge $\Leftrightarrow \rho(B_R(w)) < 1$.

2.4 Résultats particuliers de convergence

On s'intéresse tout d'abord au cas des matrices symétriques définies positives. Cette définition signifie que : $\forall x \in R^n$; $x \neq 0 : x^t Ax > 0$.

Théorème (2.3) : *Soit A une matrice symétrique définie positive et écrivons $A = M - N$, avec M inversible et $M^T + N$ définie positive. Alors, la méthode itérative $x^{(0)} \in K^n$: $x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b$ converge.* $\Diamond$

Corollaire (2.4) : *Si A est une matrice symétrique définie positive, alors la méthode de Gauss-Seidel converge.* $\Diamond$

Preuve : Pour la méthode de Gauss-Seidel, on a choisit : $M = D - E$ et $N = F$. Comme données, on a : M est inversible, A est supposée définie positive, de plus $M^T + N = D - E^T + F$. Car A est symétrique, on a alors $E^T = F$; c'est -à-dire : $M^T + N = D$. Cette dernière matrice est donc définie positive, parce que pour tout $i = 1, \ldots, n$; $\langle De_i, e_i \rangle = a_{ii}$ et $a_{ii} > 0$ puisque A est définie positive. D'après le théorème précédent, on peut conclure immédiatement le résultat. ⧫

Dans ce qui suit, on traite le cas où les matrices sont à diagonale strictement dominante.

Définition (2.4) : Une matrice $A = (a_{ij})_{1 \leq i,j \leq n}$ est dite à diagonale strictement dominante si :

$$\forall i = 1, \ldots, n \quad \mid a_{ii} \mid > \sum_{j=1, j \neq i}^{n} \mid a_{ij} \mid .$$

Pour démontrer que $\rho(B_J) < 1$ et $\rho(B_{CS}) < 1$ pour les matrices à diagonale strictement dominante, il faut exposer le lemme suivant :

Lemme (2.5) : *Le rayon spectral d'une matrice A vérifie : $\rho(A) \leq \| A \|$ pour n'importe quelle norme matricielle utilisée.* $\Diamond$

Théorème (2.6) : Si A est une matrice à diagonale strictement dominante, alors A est inversible et les méthodes de Jacobi et de Gauss-Seidel convergent. $\Diamond$

Preuve : Pour démontrer que la méthode de Jacobi converge, il suffit de montrer que $\rho(B_J) = \rho(D^{-1}(E+F) < 1$. Si on utilise les normes vectorielles et matricielles compatibles $\| \, . \, \|_\infty$et par l'application directe du lemme précédent, il est clair que : $\rho(B_J) = \rho(D^{-1}(E+F) < \| \ B_J \ \|_\infty < 1$. Puisque : $\rho(B_J) < 1$, donc Jacobi converge. On utilise les mêmes arguments pour démontrer la convergence de la méthode de Gauss-Seidel. $\blacklozenge$

Dans ce dernier paragraphe, on parle de l'implémentation des méthodes de Jacobi et de Gauss-Seidel et de leurs variantes. D'après ce qui précède, une méthode itérative de la forme 2.2 peuvent également s'écrire : $x^{(k+1)} = x^{(k)} + M^{-1} r^{(k)}$; $k \geq 0$, où le vecteur $r^{(k)} = b - Ax^{(k)}$ est le résidu à l'étape k . En tenant compte cette dernière écriture pour proposer l'algorithme suivant :

- *Initialisation : le vecteur $x^{(0)}$ donné.*
- *Calcul du résidu.*
- *Résolution du système linéaire ayant M pour matrice et le résidu comme second membre.*
- *Mise à jour de l'approximation de la solution.*
- *Répéter jusqu'à ce que la norme du résidu soit très petite par rapport à une précision bien définie au début.*

L'exécution des étapes de cet algorithme pour un système d'ordre n, nécessite $2n$ additions et n^2 multiplications pour le calcul du résidu, n divisions (pour la méthode de Jacobi) ou $\frac{n(n-1)}{2}$ additions, $\frac{n(n-1)}{2}$ multiplications et n divisions (pour la méthode de Gauss-Seidel) et la résolution du système linéaire associé à M coûte n additions pour la mise à jour de la solution approchée, $(n-1)$ additions, n multiplications. Enfin, on peut conclure que la complexité de cet algorithme est d'ordre $\frac{1}{2}n^2$ additions et $\frac{3}{2}n^2$ multiplications s'avère donc très favorable par rapport à celui des méthodes directes du chapitre 1 si le nombre d'itérations à effectuer reste petit devant n. L'efficacité des méthodes itératives non seulement au niveau de la complexité mais aussi au niveau de la mémoire de l'ordinateur.

Chapitre 3

Application

Dans ce dernier chapitre, on donne une application sur les circuits électriques. Le but de cette application est de présenter la manière de calculer les courants de branches par la résolution d'un système linéaire en utilisant la méthode de mailles.

Pour faciliter l'étude, il est naturel de commencer par une définition simple d'un circuit électrique.

Un circuit électrique est un ensemble simple ou complexe de conducteurs et de composantes électriques ou électroniques parcours par un courant électrique.

3.1 Courants de mailles

Dans la littérature, on trouve deux techniques puissantes pour faire l'analyse de circuits ; la première est la technique des courants de mailles et la deuxième est la technique des tensions de noeud. Dans ce mémoire, on va faire notre étude par l'utilisation de la première technique.

Avant de procéder, il faut définir qu'une maille dans un circuit est constituée d'éléments connectés entre eux sous forme d'une boucle fermée.

Pour démontrer la méthode, on commence en premier avec un circuit simple dont on fait l'analyse avec les lois de kirchhoff. Ces lois sont également connues aussi sous les noms : loi des noeuds et loi de mailles.

Première Loi de kirchhoff ou loi des noeuds

Le premier type de loi est caractérisé par le fait que : la somme des courants entrant au noeud d'un réseau est égale à la somme des courants sortant de ce noeud comme l'indique l'exemple suivant :

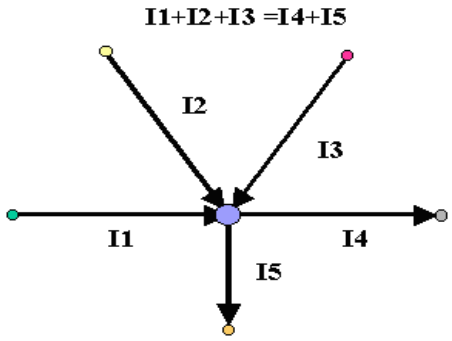


FIG. 3.1 – loi des noeuds appliquée à un exemple de circuit

Deuxième loi de kirchhoff ou loi de mailles

Dans le deuxième type de loi, on trouve que la somme algébrique des tensions aux bornes des différentes branches d'une maille est égale à zéro. Pour faciliter l'apprentissage, on applique la loi de mailles sur un exemple simple d'un circuit.

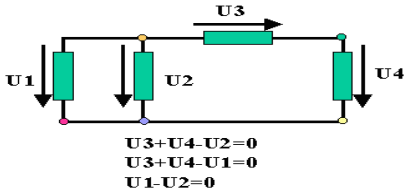


FIG. 3.2 – loi des maille appliquée à un exemple de circuit

3.2 Méthodes de mailles

Maintenant, on s'intéresse à la méthode de mailles. On peut résumer le principe de cette méthode en trois points suivants :

1. Ignorer la maille qui a le plus de branches communes avec les autres et attribuer un courant à chacune des $(N-1)$ mailles restantes.

2. Appliquer la loi de kirchhoff sur les tensions à chacune des mailles et exprimer les tensions en fonction des courants dans les mailles.

3. Résoudre le système de $(N-1)$ équations obtenues pour trouver les tensions I_i.

On peut alors déterminer tout tensions dans le circuit à partir des tensions I_i.

Illustration

Pour la méthode des courants de mailles, il faut définir tout d'abord un courant qui circule dans la maille, dans le sens horaire.

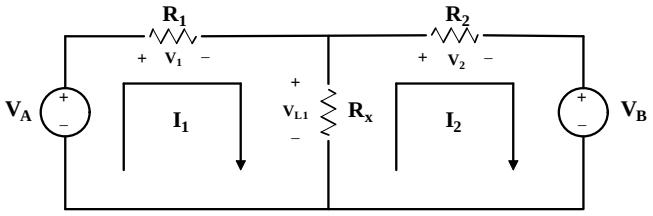


FIG. 3.3 – exemple pour courants de maille

Ensuite, on va suivre le sens des courants afin d'écrire l'équation de la tension dans la maille *(selon la loi de kirchhoff)*. Le courant dans la résistance R_X est la différence entre les deux courants : $I_1 - I_2$.

On a pour la maille 1 :

$$V_1 + V_{L1} = V_A,$$

avec

$$V_1 = R_1 I_1 \text{ et } V_{L1} = R_X(I_1 - I_2).$$

On en déduit alors :

$$(R_1 + R_X)I_1 - R_X I_2 = V_A.$$

Pour la maille 2, on a :

$$-R_X I_1 + (R_2 + R_X)I_2 = -V_B.$$

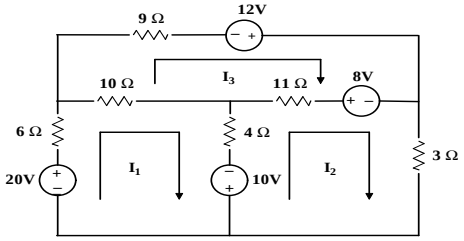
Enfin, on obtient un système linéaire de deux équations à deux inconnues suivant :

$$\begin{cases} (R_1 + R_X)I_1 - R_X I_2 = V_A \\ -R_X I_1 + (R_2 + R_X)I_2 = -V_B \end{cases}$$

Donc, on peut écrire les équations précédentes sous forme matricielle et les résoudre :

$$\begin{pmatrix} (R_1 + R_X) & -R_X \\ -R_X & (R_2 + R_X) \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \end{pmatrix} = \begin{pmatrix} V_A \\ -V_B \end{pmatrix}.$$

Dans la figure suivante, on donne un circuit et on veut calculer dans lequel les courants de branche.



On a trois mailles dans ce circuit, donc, il faut utiliser trois courants de mailles. Lorsqu'on applique la loi de Kirchhoff des tensions aux trois mailles, on trouve :

$$\begin{aligned}
\text{Maille 1}: &\quad 6I_1 + 10(I_1 - I_3) + 4(I_1 - I_2) = 20 + 10.\\
\text{Maille 2}: &\quad 4(I_2 - I_1) + 11(I_2 - I_3) + 3I_2 = -10 - 8.\\
\text{Maille 3}: &\quad 9I_3 + 11(I_3 - I_2) + 10(I_3 - I_1) = 12 + 8.
\end{aligned}$$

En simplifiant les équations, on obtient un système linéaire de trois équations à trois inconnues de la forme $Ax = b$:

$$\begin{pmatrix} 20 & -4 & -10 \\ -4 & 18 & -11 \\ -10 & 11 & 30 \end{pmatrix} \begin{pmatrix} I_1 \\ I_2 \\ I_3 \end{pmatrix} = \begin{pmatrix} 30 \\ -18 \\ 20 \end{pmatrix}.$$

Pour la résolution de ce système, on applique la méthode de Gauss étape par étape :

1. Tout d'abord, on pose : $A^{(1)} = A$ et $b^{(1)} = b$.

2. Ensuite, lorsqu'on applique la première étape de cette méthode, on obtient :

$$A^{(2)} = \begin{pmatrix} 20 & -4 & -10 \\ 0 & \frac{86}{5} & -13 \\ 0 & 9 & 25 \end{pmatrix} \text{ et } b^{(2)} = \begin{pmatrix} 30 \\ 24 \\ 45 \end{pmatrix}.$$

3. Enfin, on trouve un système triangulaire supérieur suivant :

$$A^{(3)} = \begin{pmatrix} 20 & -4 & -10 \\ 0 & \frac{86}{5} & -13 \\ 0 & 0 & \frac{2735}{86} \end{pmatrix} \text{ et } b^{(3)} = \begin{pmatrix} 30 \\ 24 \\ \frac{2530}{86} \end{pmatrix}.$$

Finalement, on utilise la méthode de substitutions successives pour résoudre ce système, on obtient les résultats suivants : $I_1 \simeq 2.364$, $I_2 \simeq 2.07$ et $I_3 \simeq 0.9$.

Conclusion

Dans ce travaille, on a intéressé à faire une comparaison entre les méthodes directes et les méthodes indirectes de résolution d'un système linéaire $Ax = b$.

D'après ce qu'on a vu, on démontre que les méthodes directes (Cramer, Cholesky et Gauss) sont des méthodes de résolution qui ont un coût de calcul en n^3.

Dans le cas des systèmes linéaires de grandes tailles, il peut être impossible de stocker en mémoire centrale l'ensemble des coefficients ; donc les méthodes directe n'est plus applicable.

Dans ce cas là, il faut utiliser des méthodes itératives de résolution beaucoup moin économique en mémoire et plus rapides parce qu'elles sont des suites itératives convergent vers la solution exacte du système.

Ce travaille est achevé par le choix d'une application en électronique ; exactement sur les circuits électrique. Le but de cette application est de présenter la manière de calculer les courants de branches par la résolution d'un système linéaire.

Bibliographie

[1] **André Fortin**, *Analyse numérique pour ingénieurs.troisième édition.*

[2] **Gloria Faccanoni**, *Analyse numérique, Recueil d'exercices corrigés et aide mémoire.*

[3] **Guillaume legendre**, *Méthodes numériques, Introduction à l'analyse numériques et au calcul scientifique, Cours 2009-2010.*

[4] **Mazen SAAD**, *Analyse numérique.*

[5] **Tahar Neffati**, *Electricité générale, Analyse et synthèse des circuits.2*^{ème} *édition.*

[6] **Thomas Gluzeau**, *Analyse numérique, Ecole National Supérieure d'Ingénieurs de Limoges, 16 rue d'atlantis, Parc ester technopole, 87068 Limoges CEDEX,cluzeau@ensil.unilim.fr.*