Ministère de l'enseignement supérieur
et la recherche scientique
Université Ibn Khaldoun de Tiaret
Faculté des sciences appliquées

Polycopié du cours

Méthodes Numériques

Réalisé par Dr.Becheikh Mostefa

Destiné aux étudiants 2^{eme} année S.T

Année Universitaire 2019-2020

Table des matières

Préface **3**

1 Résolution des équations non linéaire **4**
1.1 Introduction . . . 4
1.2 Méthode de dichotomie ou bissection . . . 4
1.3 Méthode du point fixe . . . 6
1.4 Méthode de Newton-Raphson . . . 9

2 Interpolation Polynomiales **14**
2.1 Introduction . . . 14
2.2 Méthode directe . . . 14
2.3 Méthode d'interpolation de Lagrange . . . 15
2.4 L'erreur dans l'interpolation polynomiales . . . 17
2.5 Méthode d'interpolation de Newton . . . 18

3 Intégration Numérique **22**
3.1 Introduction . . . 22
3.2 La règle du trapèze . . . 23
3.3 Les règles de Simpson . . . 26
3.3.1 Règle Simpson $\frac{1}{3}$ . . . 26
3.3.2 Règle de Simpson $\frac{3}{8}$ . . . 28
3.4 Quadrature de Gauss . . . 30
3.4.1 Introduction . . . 30
3.4.2 Quadrature de Gauss pour l'intervalle $[-1, 1]$ . . . 30
3.4.3 Quadrature de Gauss pour l'intervalle $[a, b]$ . . . 33

4 Équations différentielles Ordinaires **34**
4.1 Introduction . . . 34
4.2 Méthode d'Euler . . . 34
4.3 Méthode d'Euler améliorée : Méthode de Heun . . . 37
4.4 Méthode de la série de Taylor . . . 38
4.5 Méthodes de Runge-Kutta (RK) . . . 39
4.5.1 Méthodes de Runge-Kutta d'ordre 2 , RK2 . . . 40
4.5.2 Méthodes de Runge-Kutta d'ordre 4 RK4 . . . 41

5 Méthodes de résolutions directes des systèmes linéaires **43**
5.1 Introduction . . . 43
5.2 Méthode d'élimination de Gauss . . . 44
5.2.1 Choix de la ligne de pivot : pivot partiel . . . 46
5.2.2 Coût de de la méthode l'élimination de Gauss . . . 47
5.3 Factorisation **LU** . . . 48

5.3.1 Factorisation de Doolittle 48
5.3.2 Factorisation de Crout 49
5.3.3 Factorisation de Cholesky 51
5.3.4 Algorithme de Thomas pour les systèmes tridiagonaux 54

6 Méthodes Itératives pour systèmes linéaires 57
6.1 Introduction 57
6.2 Méthodes de Jacobi et Gauss-Seidel 57
6.3 Étude de la convergence 59
6.4 Méthode de relaxation 62

A A 64

Préface

L'analyse numérique est la branche des mathématiques qui traite de la résolution de problèmes mathématiques avec les ordinateurs dans la recherche de résultats numériques, par opposition aux réponses symboliques. Ces solutions numériques fournissent des résultats adaptées aux applications pratiques et réelles. En revanche, les solutions symboliques sont exactes, mais peuvent être difficiles ou impossibles à trouver . L'analyse numérique est une branche appliquée des mathématiques. elle cherche des solutions aux problèmes pratiques. L'aspect pratique de celui-ci conduit à un désir de solutions qui sont 'assez bonnes'. Mais assez bon est plus difficile à définir. La valeur et la pertinence des solutions numériques sont limitées par deux concepts en général .Premièrement, nous nous préoccupons de l'efficacité ou de la quantité de travail nécessaire pour parvenir à une solution. Deuxièmement, nous nous préoccupons de la précision de notre solution, ou de notre proximité avec la vraie valeur. Nous allons examiner chacun de ces concepts tour à tour et apprendre à les appliquer à l'analyse numérique.
Ce polycopiée est un support au cours d'analyse numérique en deuxième année d'une Licence sciences appliquées . Il aborde : la recherche de racines d'une fonction réelle de variable réelle, l'interpolation polynomiale, l'intégration numériques, l'intégration d'équations différentielles et la résolution de systèmes linéaires. Les notions supposées connues correspondent au programme des cours de mathématiques (Analyse mathématique des fonctions réelles d'une variable réelle et Algèbre Linéaire) et Informatiques (Initiation à l'algorithmique et aux langages Matlab et octave) de la première année de Licence. L'objet de ce polycopiée est de proposer une explication succincte des concepts vu en cours. De nombreux livres, parfois très fournis, existent. Ici on a cherché, compte tenu des contraintes de volume horaire, des acquis des étudiants à la première année et des exigences pour la suite du cursus, à dégager les points clés permettant de structurer le travail personnel de l'étudiant voire de faciliter la lecture d'autres ouvrages. Ce polycopiée ne dispense pas des séances de cours et de TD ni de prendre des notes complémentaires. Il est d'ailleurs important d'apprendre le cours au fur et à mesure et de comprendre les nombreux exercices corrigés proposé.

Chapitre 1

Résolution des équations non linéaire

1.1 Introduction

Ce chapitre se concentre sur la solution numérique des équations à une seule variable, qui apparaissent sous la forme générale.

$$f(x) = 0 \tag{1.1}$$

Graphiquement, une solution (ou racine) de $f(x) = 0$ fait référence au point d'intersection de $f(x)$ avec l'axe des x. Par conséquent, selon la position du graphe de $f(x)$ par rapport à l'axe des x, l'équation (1.1) peut avoir une solution unique, plusieurs solutions ou aucune solution. La racine d'une l'équation peut parfois être déterminée analytiquement, ce qui donne une solution exacte.Dans la plupart des situations, cela n'est pas possible et il faut faire recourt aux méthodes numériques.

1.2 Méthode de dichotomie ou bissection

La méthode de dichotomie (ou méthode de la bissection) repose directement sur le théorème des valeurs intermédiaire (A.1).

Son principe est le suivant :Soit une fonction f continue sur un intervalle $[a, b]$ tel que $f(a).f(b) < 0$, On pose $a_1 = a$, $b_1 = b$, on note $p_1 = \frac{1}{2}(a_1 + b_1)$ le milieu de l'intervalle de départ et on évalue la fonction f en ce point. Si $f(p_1) = 0$, le point p_1 est le zéro de f et le problème est résolu. Sinon,voir si $f(a_1).f(p_1)) < 0$, alors le zéro p est contenu dans l'intervalle $]a_1; p_1[$,alors qu'il appartient à $]p_1, b_1[$ si $f(p_1).f(b_1) < 0$. On réitère ensuite ce processus sur l'intervalle $[a_2, b_2]$, avec $a_2 = a_1$ et $b_2 = p_1$ dans le premier cas, ou $a_2 = p_1$ et $b_2 = b_1$ dans le second, et ainsi de suite.

De cette manière, on construit de manière récurrente trois suites $(a_k)_{k\in\mathbb{N}}$, $(b_k)_{k\in\mathbb{N}}$ et $(p_k)_{k\in\mathbb{N}}$ telles que $a_1 = a$,$b_1 = b$ et vérifiant , pour entier naturel k

- $p_k = \frac{1}{2}(a_k + b_k)$
- $a_{k+1} = a_k$ et $b_{k+1} = p_k$ si $f(a_k).f(p_k) < 0$
- $a_{k+1} = p_k$ et $b_{k+1} = b_k$ si $f(p_k).f(b_k) < 0$

La figure (1.2) illustre la mise en œuvre de la méthode. Pour le test d'arrêt nous pouvons choisir une tolérance $\epsilon > 0$ et générer $p_1, p_2, \ldots, p_n$ jusqu'à ce que l'une des conditions suivantes soit remplie :

$$\begin{aligned} |p_k - p_{k-1}| &< \epsilon \\ \frac{|p_k - p_{k-1}|}{|p_k|} &< \epsilon, \qquad |p_k| \neq 0 \\ |f(p_k)| &< \epsilon \end{aligned} \tag{1.2}$$

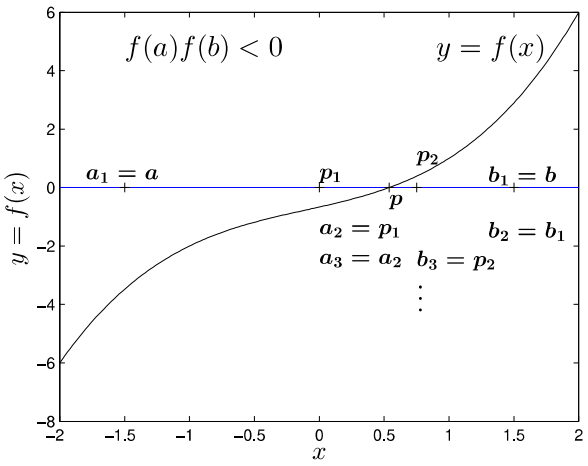


FIGURE 1.1 – Construction des premiers itérés de la méthode de dichotomie

Exemple 1.1 *Utilisez la méthode de bissection pour trouver une solution, avec une précision de* 10^{-5} *, pour le problème suivant*

$$3x - e^x = 0 \quad pour \quad 1 \leq x \leq 2$$

TABLE 1.1 – Résolution de $3x - e^x = 0$ par la méthode de dichotomie

k	a_k	b_k	p_k	$\lvert p_k - p_{k-1}\rvert$
1	1.000000	2.000000	1.500000	
2	1.500000	2.000000	1.750000	0.250000
3	1.500000	1.750000	1.625000	0.125000
4	1.500000	1.625000	1.562500	0.062500
5	1.500000	1.562500	1.531250	0.031250
6	1.500000	1.531250	1.515625	0.015625
7	1.500000	1.515625	1.507812	0.007812
8	1.507812	1.515625	1.511718	0.003906
9	1.511718	1.515625	1.513671	0.001953
10	1.511718	1.513671	1.512695	0.000976
11	1.511718	1.512695	1.512207	0.000488
12	1.511718	1.512207	1.511962	0.000244
13	1.511962	1.512207	1.512084	0.000122
14	1.512084	1.512207	1.512145	0.000061
15	1.512084	1.512145	1.512115	0.000030
16	1.512115	1.512145	1.512130	0.000015
17	1.512130	1.512145	1.512138	0.000007

$p_{17} = 1.512138$ est une approximation à la racine p avec une erreur $< 10^{-5}$

Théoreme 1.1 *Supposons que* $f \in \mathbf{C}[a,b]$*et* $f(a).f(b) < 0$*. La méthode de bissection génère une séquence* $\{p_n\}_{n=1}^{\infty}$ *se rapprochant d'une racine* p *de* f *avec*

$$|p_n - p| \leq \frac{b-a}{2^n} \quad avec\ n \geq 1 \tag{1.3}$$

Démonstration : Pour $n \geq 1$ on peut vérifier

$$b_n - a_n = \frac{b-a}{2^{n-1}} \quad \text{et} \quad p \in [a_n, b_n]$$

Puisque $p_n = \frac{1}{2}(a_n + b_n)$ pour tout $n \geq 1$ il s'ensuit que

$$|p_n - p| \leq \frac{1}{2}(b_n - a_n) = \frac{b-a}{2^n}$$

Pour établir une borne du nombre d'itérations, n qui satisfait (1.2), on résout simplement

$$|p - p_n| \leq \frac{b-a}{2^n} \leq \epsilon$$

qui donne

$$n \geq \frac{\ln(b-a) - \ln(\epsilon)}{\ln 2} \tag{1.4}$$

Exemple 1.2 *Pour l'exemple (1.2) trouver une borne n du nombre d'itération qui assure une erreur $\leq \epsilon = 10^{-5}$*

On applique l'équation (1.4) on aura

$$n \geq \frac{\ln(2-1) - \ln(10^{-5})}{\ln 2} = 16.61$$

Puisque le nombre $n \in \mathbb{N}$, $n = 17$ assure une erreur $\leq 10^{-5}$.

Il est important de noter que le théorème (1.1) ne donne qu'une borne pour l'erreur d'approximation et que l'erreur réelle est beaucoup plus petite : dans l'exemple(1.2) l'erreur réelle est $|p - p_{17}| = 3.815.10^{-6}$.

Remarque : La méthode de dichotomie présente plusieurs avantages.La convergence est certaine dès lors que les conditions précédentes sont remplies, la programmation est trop simple,le calcul de f n'a pas besoin d'être très précis,parce que nous utilisons en fait le signe de f et non sa valeur.Le désavantage est que la convergence est assez lente si une grande précision est demandée.

1.3 Méthode du point fixe

Définition 1.1 *Le nombre p est un point fixe pour une fonction g donnée si*

$$p = g(p) \tag{1.5}$$

l'équation (1.5) a une interprétation géométrique simple, $(p, g(p))$ est le point d'intersection de la courbe $y = g(x)$ et la première bissectrice $y = x$.

Exemple 1.3 *Déterminez les points fixes de la fonction $g(x) = (x^2 - 3)/2$*

$$p = g(p) \Leftrightarrow p = (p^2 - 3)/2 \Leftrightarrow p = -1, p = 3$$

On a deux point fixe $p = -1$ et $p = 3$ qui sont montrés dans la figure 1.2

Dans cette section, nous considérons le problème de la recherche de solutions aux problèmes de points fixes et la connexion entre les problèmes de points fixes et les problèmes de recherche de racines que nous souhaitent résoudre. Les problèmes de recherche de racines et les problèmes de points fixes sont des classes équivalentes dans le sens suivant :

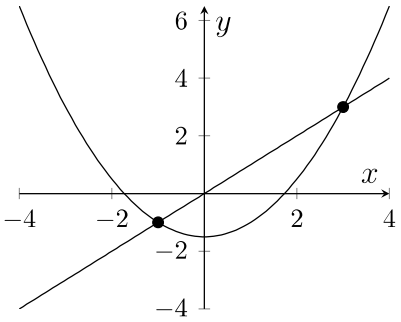


FIGURE 1.2 – Points fixes de $g(x) = (x^2 - 3)/2$

Résoudre $f(x) = 0$, revient à définir une fonctions $g(x)$ tel que le point fixe p de $g(x)$ vérifie $f(p) = 0$.Par exemple si

$$g(x) = x + f(x) \text{ ou } g(x) = x + 5f(x) \text{ ou.....}$$

On voit bien si $p = g(p)$ alors $f(p) = 0$ et inversement si $f(p) = 0$ alors $g(p) = p$.

L'itération du point fixe :

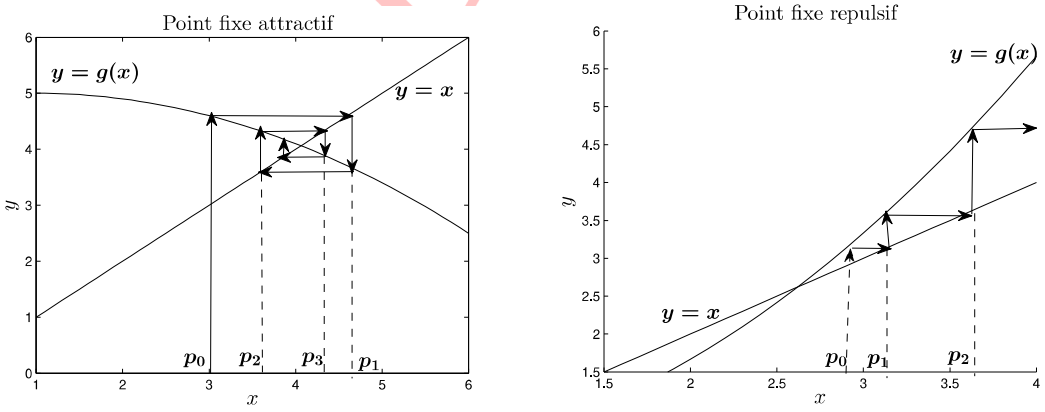
Pour approximer le point fixe d'une fonction g, nous choisissons une valeur initiale p_0 et générons la séquence $\{p_n\}_{n=0}^{\infty}$ en mettant

$$p_n = g(p_{n-1}), \quad \text{pour } n \geq 1. \tag{1.6}$$

Si la séquence (1.6) converge vers p et g est continue, alors

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g(\lim_{n\to\infty} p_{n-1}) = g(p) \tag{1.7}$$

L'ordonné $g(p_0)$ est la nouvelle abscisse p_1 ; le point $(p_1, 0)$ se déduit du point $(0, g(p_0))$ par une symétrie orthogonale par rapport à la droite $y = x$.En générant la suite (1.7),on peut rencontrer deux dispositions de la suite $p_0, p_1, ...$, correspondant soit à un point fixe attractif (on s'approche du point fixe) ou un point fixe répulsif (on s'éloigne du point fixe),comme nous le montre la figure(1.3).



(a) Convergence (b) Divergence

FIGURE 1.3 – Itération du point fixe

Théoreme 1.2 *Théorie de l'existence et de la convergence de l'itération du point fixe*
Soit $g \in C([a, b])$ avec $a \leq g(x) \leq b$ pour tout $x \in [a, b]$:

1. *g a au moins un point fixe $p \in [a, b]$*
2. *S'il existe une valeur $\kappa < 1$ tel que*

$$|g(x) - g(y)| \leq \kappa |x - y| \tag{1.8}$$

pour tout x et y dans $[a, b]$, alors

(a) p est unique

(b) L'itération $p_{n+1} = g(p_n)$ converge vers p pour toute valeur initiale $p_0 \in [a, b]$

(c) L'erreur estimée

$$|p - p_n| \leq \frac{\kappa^n}{1 - \kappa} |p_1 - p_0| \tag{1.9}$$

3. *Si la dérivée de g est continue sur $[a, b]$ avec*

$$\max_{x \in [a,b]} |g'(x)| = \kappa < 1 \tag{1.10}$$

(a) p est unique

(b) L'itération $p_{n+1} = g(p_n)$ converge vers p pour toute valeur initiale $p_0 \in [a, b]$

(c) L'erreur estimée

$$|p - p_n| \leq \frac{\kappa^n}{1 - \kappa} |p_1 - p_0| \tag{1.11}$$

(d) on a la limite

$$\lim_{n \to \infty} \frac{p - p_{n+1}}{p - p_n} = g'(p)$$

Démonstration : Définissant $h(x) = g(x) - x$ alors

$$h(b) = g(b) - b \leq 0$$

et

$$h(a) = g(a) - a \geq 0$$

Par conséquent, le théorème des valeurs intermédiaire (A.1) h a une racine $p \in [a, b]$; ainsi $h(p) = 0$, qui implique que $p = g(p)$ qui démontre (1).
Supposons maintenant que (1.8) est vérifié,et qu'un deuxième point fixe existe,$q \in [a, b]$.Ainsi on a

$$p = g(p)$$

et

$$q = g(q)$$

de sorte (de (1.8))

$$|p - q| = |g(p) - g(q)| \leq \kappa |p - q|$$

qui implique

$$|p - q|(1 - \kappa) \leq 0$$

Puisque $0 < \kappa < 1$, la seule façon que $|p - q| \leq 0$ soit vrai est $p = q$,ce qui implique que le point fixe p est unique d'où (2) est prouvé.
Considérons maintenant l'itération (1.6) et la définition du point fixe (1.5). Si nous soustrayons et prenons des valeurs absolues, nous obtenons

$$|p - p_{n+1}| = |g(p) - g(p_n)| \leq \kappa |p - p_n| \tag{1.12}$$

Maintenant écrivant l'erreur $e_n = |p - p_n|$ et l'équation (1.12) devient $e_{n+1} \leq \kappa e_n$. Par la récursivité on aura $e_n \leq \kappa^n e_0$, d'où il résulte que $e_n \to 0$ quand $n \to \infty$;par conséquent, l'itération converge. Cela prouve (2b)
Finalement ,nous notons que

$$|p - p_0| = |p - g(p_0) + p_1 - p_0| \leq |g(p) - g(p_0)| + |p_1 - p_0| \leq \kappa|p - p_0| + |p_1 - p_0|$$

D'où il résulte que

$$|p - p_0| \leq \frac{1}{1-\kappa}|p_1 - p_0|$$

Pour que

$$e_n \leq \kappa^n e_0 \leq \frac{\kappa^n}{1-\kappa}|p_1 - p_0|$$

ce qui prouve (2c).
Nous notons maintenant que (1.10) implique (1.8), de sorte que (3a)-(3c) suivent exactement (2a)-(2c). Il ne reste plus qu'à prouver (3d).
Nous avons, de (1.6) et (1.7), et du théorème (A.2) que

$$p - p_{n+1} = g(p) - g(p_n) = g'(\xi_n)(p - p_n)$$

D'ou

$$\frac{p - p_{n+1}}{p - p_n} = g'(\xi_n) \to g'(p)$$

puisque $\xi_n \to p$ est forcé par la convergence de p_n vers p

Exemple 1.4 *La fonction $f(x) = e^x - x - 2$ à une racine $p \in [1, 2]$. Utiliser la méthode du point fixe pour calculer cette racine avec un test d'arrêt $|x_{k+1} - x_k| \leq 10^{-5}$.*

On a $f(1).f(2) = (-0.2817).(3.3890) < 0$, donc il existe $p \in [1, 2]$ tel que $f(p) = 0$. Pour pouvoir utiliser l'itération du point fixe,on tire une équation $x = g(x)$ de $f(x) = 0$.

1.
$$e^x - x - 2 = 0 \Leftrightarrow x = e^x - 2 = g_1(x)$$
$$|g'(x)| = |e^x| > 1 \quad \forall x \in [1, 2] \text{ l'itération du point fixe } x_{k+1} = g_1(x_k) \text{ diverge}$$

2.
$$e^x - x - 2 = 0 \Leftrightarrow e^x = x + 2 \Rightarrow x = \ln(x + 2) = g_2(x)$$
$$|g'(x)| = \left|\frac{1}{x+2}\right| < 1 \quad \forall x \in [1, 2] \text{ l'itération du point fixe } x_{k+1} = g_2(x_k) \text{ converge}$$

Pour $x_0 = 1$
$p_{10} = 1.146196$ est une bonne approximation à solution de $f(x) = 0$ dans l'intervalle $[1, 2]$

1.4 Méthode de Newton-Raphson

La méthode la plus courante pour résoudre des équations non linéaires est la méthode d'itération de Newton-Raphson (N-R). Un gros avantage est que si elle converge,elle converge plus rapidement que les méthodes discutées précédemment. La convergence est d'ordre 2 , c'est-à-dire que l'erreur à chaque étape est proportionnelle au carré de l'erreur à l'étape précédente. Il existe deux bonnes dérivations de la méthode de (N-R), une géométrique et une analytique. Nous allons discuter des deux, en commençant par la dérivation géométrique.

TABLE 1.2 – L'itération du point fixe

k	$x_{k+1} = g_1(x_k)$	$x_{k+1} = g_2(x_k)$
0	2.48168	1.252762
1	9.96145	1.179504
2	21191.52	1.156725
3	Inf	1.149535
4	Inf	1.147254
5	Inf	1.146530
6	Inf	1.146300
7	Inf	1.146227
8	Inf	1.146204
9	Inf	1.146196

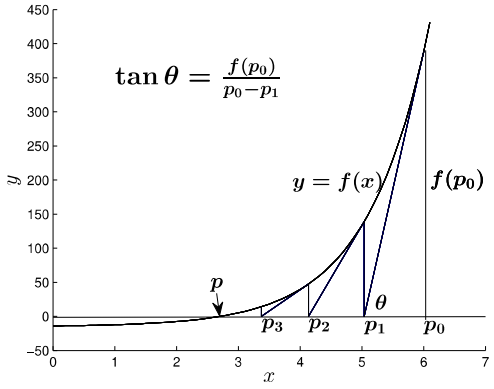


FIGURE 1.4 – Itération de Newton-Raphson

Considérez la figure (1.4). Nous souhaitons trouver une racine,de $y = f(x)$, étant donné une estimation initiale p_0 .Comment pouvons-nous améliorer cette estimation pour obtenir une meilleure approximation ?L'idée fondamentale de la méthode de (N-R) est d'utiliser l'approximation de la ligne tangente à la fonction au point $(p_0, f(p_0))$ et au lieu de chercher l'intersection de la fonction f avec l'axe des x ,on cherche le point d'intersection de cette droite tangente avec l'axe des x.

$$f'(p_0) = \tan\theta = \frac{f(p_0)}{p_0 - p_1}$$

d'ou

$$p_1 = p_0 - \frac{f(p_0)}{f'(p_0)} \tag{1.13}$$

Maintenant, continuons le processus avec une autre ligne droite tangente en $(p_1, f(p_1))$ pour obtenir

$$p_2 = p_1 - \frac{f(p_1)}{f'(p_1)}$$

ou, généralement

$$p_{k+1} = p_k - \frac{f(p_k)}{f'(p_k)} \tag{1.14}$$

La deuxième dérivation de la méthode de (N-R) dépend du théorème de Taylor (A.3). Soit $f(x) \in C^2 I \subset \mathbb{R}$.

$$f(x) = f(p_0) + f'(p_0)(x - p_0) + \frac{f''(\xi)}{2}(x - p_0)^2, \quad \xi \in [x, p_0]$$

Pour p racine de f ,on a $f(p) = 0$,d'où

$$0 = f(p_0) + f'(p_0)(p - p_0) + \frac{f''(\xi)}{2}(p - p_0)^2$$

$$p = p_0 - \frac{f(p_0)}{f'(p_0)} - \frac{f''(\xi)}{2f'(p_0)}$$

Si la valeur approximative p_0 est suffisamment proche de p, Le terme

$$R_2 = \frac{f''(\xi)}{2f'(p_0)}$$

peut être négligé.L'équation restante ne donne pas la valeur exacte p,mais une solution approximative améliorée par rapport à la valeur de départ p_0.Poursuivant ainsi, à partir d'une approximation p_n,une approximation améliorée p_{n+1} peut être obtenu par la règle d'itération suivante

$$p_{k+1} = p_k - \frac{f(p_k)}{f'(p_k)} = \phi(p_k) \quad k \geq 0 \tag{1.15}$$

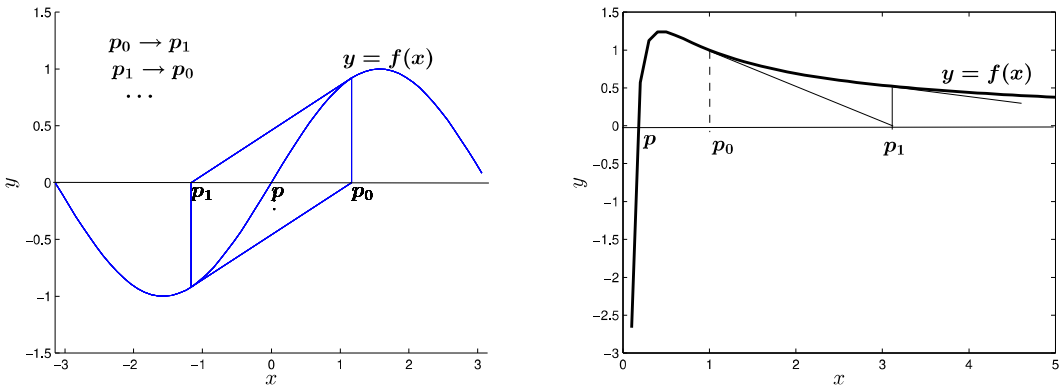
Il faut examiner à quelle condition une valeur initiale p_0 doit répondre,pour créer une séquence d'itération $\{p_n\}_{n=0}^{\infty}$ convergente. La convergence est assurée selon le théorème du point fixe (1.2),si $|\phi(x)| \leq \kappa < 1$ pour tout $x \in I$.

$$\phi(x) = x - \frac{f(x)}{f'(x)} \quad \phi(x)' = \frac{f(x).f''(x)}{(f'(x))^2}$$

Soit la condition suffisante pour la convergence

$$\left|\frac{f(x).f''(x)}{(f'(x))^2}\right| \leq \kappa < 1 \tag{1.16}$$

La méthode d'itération de Newton peut échouer si la condition (1.16) est violée. La figure (1.5) illustre deux cas de divergence. **Estimation des erreurs :** Dans la méthode (N-R), le test



(a) (b)

FIGURE 1.5 – Divergence de l'itération de Newton-Raphson

d'arrêt généralement utilisée est $|p_{k+1} - p_k| < \epsilon$

Mais l'estimation de l'erreur est plus complexe. Pour cela, on suppose que la fonction f est deux fois différentiable dans I et que

$$|f'(x)| \geq k > 0 \quad \text{et} \quad |f''(x)| \leq K \quad (x \in I)$$

Le développement de Taylor de f autour de p est

$$f(x) = f(p) + (x-p)f'(\xi) = (x-p)f'(\xi) \quad (\xi \in [x,p])$$

Pour $x = p_{n+1}$ on a

$$\begin{aligned} f(p_{n+1}) &= f(p) + (p_{n+1}-p)f'(\xi) \quad (\xi \in [p_{n+1},p]) \\ |f(p_{n+1})| &= |p_{n+1}-p||f'(\xi)| \\ &\geq |p_{n+1}-p|k \end{aligned} \tag{1.17}$$

Le développement de Taylor de f autour de $x = p_n$ est

$$\begin{aligned} f(x) &= f(p_n) + (x-p_n)f'(p_n) \\ &+ \frac{1}{2}(x-p_n)^2 f''(\xi') \quad (\xi' \in [x,p_n]) \end{aligned}$$

Pour $x = p_{n+1}$

$$\begin{aligned} f(p_{n+1}) &= f(p_n) + (p_{n+1}-p_n)f'(p_n) \\ &+ \frac{1}{2}(p_{n+1}-p_n)^2 f''(\xi') \quad (\xi' \in [x,p_n]) \\ &= \frac{1}{2}(p_{n+1}-p_n)^2 f''(\xi') \\ |f(p_{n+1})| &= \frac{1}{2}|p_{n+1}-p_n|^2 f''(\xi') \\ &\leq \frac{1}{2}|p_{n+1}-p_n|^2 K \end{aligned} \tag{1.18}$$

Les inégalités (1.17) et (1.18) résultent

$$|p_{n+1}-p|k \leq |f(p_{n+1})| \leq \frac{1}{2}|p_{n+1}-p_n|^2 K$$

Soit

$$|p_{n+1}-p| \leq \frac{K}{2k}|p_{n+1}-p_n|^2 \tag{1.19}$$

L'estimation d'erreur (1.19) est appelée estimation a posteriori.
Une autre estimation approximative peut être décrite : supposons que p_n diffère de la racine p d'une petite quantité ε_n,ainsi $p_n = p + \varepsilon_n$ et $p_{n+1} = p + \varepsilon_{n+1}$.

$$p + \varepsilon_{n+1} = p + \varepsilon_n - \frac{f(p+\varepsilon_n)}{f'(p+\varepsilon_n)}$$

$$\begin{aligned} \varepsilon_{n+1} &= \varepsilon_n - \frac{f(p) + \varepsilon_n f'(p) + \frac{1}{2}\varepsilon_n^2 f''(p) + \ldots}{f'(p) + \varepsilon f''(p) + \ldots} \\ &= \varepsilon_n - \frac{\varepsilon_n f'(p) + \frac{1}{2}\varepsilon_n^2 f''(p) + \ldots}{f'(p) + \varepsilon f''(p) + \ldots} \\ &= \frac{\varepsilon_n^2 f''(p)}{2\left[f'(p) + \varepsilon f''(p)\right]} = \frac{1}{2}\varepsilon_n^2 \frac{f''(p)}{f'(p)} \end{aligned} \tag{1.20}$$

(1.20) est obtenue en négligeant les puissances supérieures de ε_n.

Exemple 1.5 *La fonction* $f(x) = x \ln x - 1$ *a un zéro dans l'intervalle* $[1.5, 2]$ *; Nous recherchons l'estimation a posteriori de la règle d'itération de Newton-Raphson.*

Solution : La fonction $f'(x) = \ln x + 1$ est strictement croissante sur $[1.5, 2]$ alors sa valeur minimale est $f'(1.5)$. Donc

$$|f'(x)| \geq \ln 1.5 + 1 > 1.4 = k$$

La fonction $f''(x) = 1/x$ est strictement décroissante sur $[1.5, 2]$ alors sa valeur maximale est $f''(1.5)$.Donc

$$f''(x) \leq \frac{1}{1.5} < 0.7 = K$$

Il en résulte une estimation a posteriori

$$|p_{n+1} - p| \leq 0.25.|p_{n+1} - p_n|^2$$

Exemple 1.6 *Calculer la racine de la fonction* $f(x) = e^x - x^2 + 2x - 2$ *dans l'intervalle* $[0, 1]$*, avec une précision* $\epsilon = 5.10^{-6}$.

Chapitre 2

Interpolation Polynomiales

2.1 Introduction

Le problème est le suivant : à partir d'une fonction $f(x)$ connue seulement en $(n+1)$ points de la forme $(x_i, f(x_i))$ pour $\{i = 0, 1, 2, \ldots, n\}$, peut-on construire une approximation de $f(x)$, et ce, pour tout x ? Les x_i sont appelés abscisses ou nœuds d'interpolation tandis que les couples $(x_i, f(x_i))$ pour $\{i = 0, 1, 2, \ldots, n\}$,sont les points de collocation ou points d'interpolation et peuvent provenir de données expérimentales ou d'une table. En d'autres termes, si l'on ne connait que les points de collocation $(x_i, f(x_i))$ d'une fonction, peut-on obtenir une approximation de $f(x)$ pour une valeur de x différente des x_i.

Si x se situe en dehors de cet intervalle, le problème est appelé extrapolation. Le problème de l'interpolation polynomiale consiste à trouver un polynôme $P_n(x)$ de degré inférieur ou égal à n, tel que $P_n(x_i) = f(x_i)$ pour $\{i = 0, 1, 2, \ldots, n\}$ À travers deux points distincts, nous pouvons construire un polynôme unique de degré 1 (Fig 2.1 a). À travers trois points distincts, nous pouvons construire un polynôme unique de degré 2 (Fig 2.1 b) ou un polynôme unique de degré 1 (si les points sont alignés). À travers quatre points distincts nous pouvons construire un polynôme unique de degré 3 (Fig 2.1 c) etc.... En général, par $n+1$ points distincts, nous pouvons construire un polynôme unique de degré $\leqslant n$.

Théoreme 2.1 *: Étant donné $n+1$ points distincts $x_0, .., x_n$ et $n+1$ valeurs correspondantes $y_0, ..., y_n$, il existe un unique polynôme $P_n(x)$ de degré $\leq n$ tel que $P_n(x_i) = y_i$ pour $i = 0, ..., n$.*

2.2 Méthode directe

Une manière apparemment simple de résoudre ce problème est d'écrire le polynôme sous la forme :

$$P_n(x) = \sum_{k=0}^{n} c_k p_k(x) \qquad p_k = x^k \qquad (k = 0, 1 \ldots, n) \tag{2.1}$$

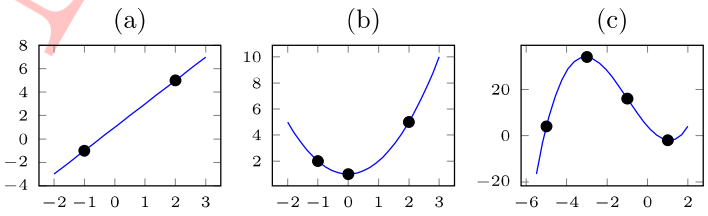


FIGURE 2.1 – Graphes de polynôme de degrés 1 ,2 et 3

Le problème d'interpolation conduit alors au système d'équations linéaires dont les constantes c_i à déterminer.

$$P_n(x_i) = f(x_i) \tag{2.2}$$

$$\begin{pmatrix} x_0^0 & x_0^1 & \dots & x_0^n \\ x_1^0 & x_1^1 & \dots & x_1^n \\ \vdots & \vdots & \vdots & \vdots \\ x_n^0 & x_n^1 & \dots & x_n^n \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \tag{2.3}$$

Résoudre un tel système d'équations à $(n+1)$ inconnues n'est pas une tache triviale. Cette méthode pour trouver le polynôme $P_n(x)$ n'est donc pas une bonne méthode en pratique. Dans la suite on va étudier des méthodes plus astucieuses pour construire le polynôme. $P_n(x_i)$.

2.3 Méthode d'interpolation de Lagrange

Dans ce cas, toutes les fonctions $p_k(x)$ sont des polynômes de degré n. Qui seront notées $L_k(x)$

Définition 2.1 *Le polynôme suivant est appelé polynôme de Lagrange*

$$L_k(x) = \prod_{i=0, i\neq k}^{n} \frac{x - x_i}{x_k - x_i} \quad (k = 0, 1 \dots, n) \tag{2.4}$$

Pour satisfaire (2.2) les polynômes de Lagrange ont donc la propriété :

$$L_k(x_i) = \delta_{ki} = \begin{cases} 1 : i = k \\ 0 : i \neq k \end{cases} \tag{2.5}$$

Exemple 2.1 *À titre d'illustration, nous considérons un problème d'interpolation avec les points suivants :* $x_0 = -2$ *;* $x_1 = 0$ *et* $x_2 = 1$.

$$L_0(x) = \prod_{i=1}^{2} \frac{x - x_i}{x_0 - x_i} = \frac{x(x-1)}{(-2-0)(-2-1)} = \frac{x(x-1)}{6}$$

$$L_1(x) = \prod_{i=0, i\neq 1}^{2} \frac{x - x_i}{x_1 - x_i} = \frac{(x+2)(x-1)}{(0+2)(0-1)} = \frac{(x+2)(x-1)}{2}$$

$$L_2(x) = \prod_{i=0, i\neq 2}^{2} \frac{x - x_i}{x_2 - x_i} = \frac{(x+2)x}{(1+2)(1-0)} = \frac{(x+2)x}{3}$$

Aux points d'interpolation x_i , la propriété (2.5) s'applique au polynôme d'interpolation. Et puisque $P_n(x_i) = f(x_i)$ alors $c_i = f(x_i)$ et le polynôme d'interpolation de Lagrange s'écrit :

$$\boxed{P_n(x) = \sum_{k=0}^{n} f(x_k) L_k(x)} \tag{2.6}$$

Exemple 2.2 *Nous recherchons le polynôme d'approximation pour la fonction* $y = \sin x$ *qui passe par les trois points* $(0,0)$,$(\pi/2,1)$ *et* $(\pi,0)$.

Les polynômes de Lagrange sont :

$$L_0(x) = \prod_{i=1}^{2} \frac{x - x_i}{x_0 - x_i} = \frac{(x-\pi/2)(x-\pi)}{\pi^2/2}$$

$$L_1(x) = \prod_{i=0, i\neq 1}^{2} \frac{x - x_i}{x_1 - x_i} = \frac{x(x-\pi)}{(-\pi^2/4)}$$

$$L_2(x) = \prod_{i=0, i\neq 2}^{2} \frac{x - x_i}{x_2 - x_i} = \frac{x(x-\pi/2)}{(\pi^2/2)}$$

Et le polynôme d'interpolation de Lagrange est :

$$P_2(x) = 0.\frac{(x-\pi/2)(x-\pi)}{\pi^2/2} + 1.\frac{x(x-\pi)}{(-\pi^2/4)} + 0.\frac{x(x-\pi/2)}{(\pi^2/2)} = \frac{x(x-\pi)}{(-\pi^2/4)}$$

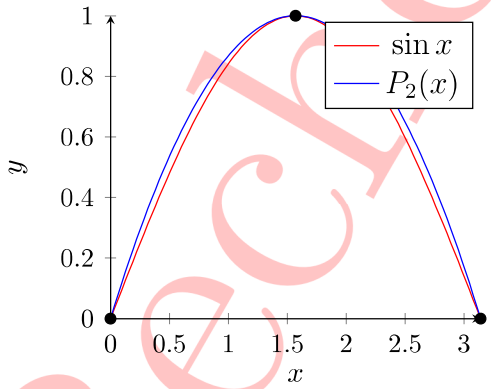


FIGURE 2.2 – Graphes $\sin x$ et $P_2(x)$

Exemple 2.3 *Nous recherchons le polynôme d'approximation pour la fonction* $y = \sin x$ *qui passe par les quatre points* $(0,0)$,$(\pi/2,1)$,$(\pi,0)$ *et* $(3\pi/2,-1)$.

Les polynômes de Lagrange sont :

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-\pi/2)(x-\pi)(x-3\pi/2)}{-3\pi^3/4}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{x(x-\pi)(x-3\pi/2)}{\pi^3/4}$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{x(x-\pi/2)(x-3\pi/2)}{-\pi^3/4}$$

$$L_3(x) = \frac{(x_3-x_0)(x_3-x_1)(x_3-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{x(x-\pi/2)(x-\pi)}{3\pi^3/4}$$

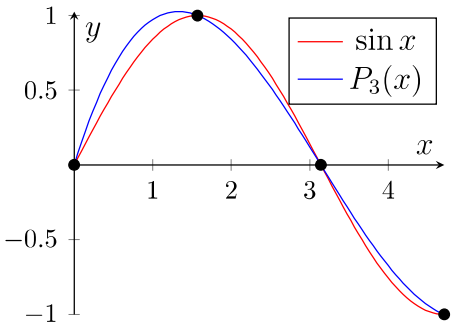


FIGURE 2.3 – Graphes $\sin x$ et $P_3(x)$

Puisque $f(x_0) = f(x_2) = 0$ on a besoin seulement de $L_1(x)$ et $L_3(x)$ pour écrire le polynôme d'interpolation de Lagrange

$$P_3(x) = 1.\frac{x(x-\pi)(x-3\pi/2)}{(\pi^3/4)} - 1.\frac{x(x-\pi/2)(x-\pi)}{(3\pi^3/4)} = \frac{8x(x-\pi)(x-2\pi)}{3\pi^3}$$

Exemple 2.4 *Nous recherchons le polynôme d'interpolation de Lagrange qui passe par les points* $(-1,1)$,$(1,-1)$,$(2,4)$ *et* $(5,1)$.

Les polynômes de Lagrange sont :

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = \frac{(x-1)(x-2)(x-5)}{36}$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{(x+1)(x-2)(x-5)}{8}$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = \frac{(x+1)(x-1)(x-5)}{9}$$

$$L_3(x) = \frac{(x_3-x_0)(x_3-x_1)(x_3-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{(x+1)(x-1)(x-2)}{72}$$

Et le polynôme d'interpolation de Lagrange est :

$$P_3(x) = -\frac{(x-1)(x-2)(x-5)}{36} - \frac{(x+1)(x-2)(x-5)}{8}$$
$$-4\frac{(x+1)(x-1)(x-5)}{9} + \frac{(x+1)(x-1)(x-2)}{72}$$

2.4 L'erreur dans l'interpolation polynomiales

L'interpolation permet, à partir d'un certain nombre de données sur les valeurs d'une fonction, de faire l'approximation de $f(x)$ en tout point x. Toutefois, cette opération entraîne une erreur d'interpolation qu'il convient d'étudier en détail

Théoreme 2.2 *Soit $x_0 < x_1 < x_2 \ldots < x_n$ les abscisses des points de collocation. On suppose que la fonction $f(x)$ est définie dans l'intervalle $[x_0, x_n]$et qu'elle est $(n+1)$ fois dérivable dans $]x_0, x_n[$. Alors, pour tout x dans l'intervalle $[x_0, x_n]$, il existe $\xi(x)$appartenant à l'intervalle $]x_0, x_n[$ tel que l'erreur $E_n(x)$ est donnée par :*

$$E_n(x) = f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\prod_{i=0}^{n}(x - x_i) \tag{2.7}$$

Exemple 2.5 *Estimez l'erreur dans l'exemple 2.3*

$$\mid E_3(x) \mid = \mid \sin x - P_3(x) \mid = \left|\frac{\sin^{(n+1)}\xi(x))}{(n+1)!}\right| \mid (x-0)(x-\pi/2)(x-\pi)(x-3\pi/2) \mid$$

$\xi(x) \in [0, 3\pi/2]$ alors $\left|\sin(\xi(x))\right| \leqslant 1.$ et $(n+1)! = 4! = 24$
ainsi :

$$\mid \sin x \ - P_3(x) \mid \leqslant \frac{1}{24} \mid x(x-\pi/2)(x-\pi)(x-3\pi/2) \mid$$

Pour chaque x donné on peut trouver une borne de l'erreur. Et sur l'intervalle entier on utilise le graphe de $y = x(x-\pi/2)(x-\pi)(x-3\pi/2)$:

$$\mid x(x-\pi/2)(x-\pi)(x-3\pi/2) \mid \leqslant 6.088068199$$

Ainsi la borne d'erreur est :

$$\mid \sin x \ - P_3(x) \mid \leqslant \frac{1}{24} \mid x(x-\pi/2)(x-\pi)(x-3\pi/2) \mid \leqslant \frac{1}{24} 6.088068199 = 0.25367$$

2.5 Méthode d'interpolation de Newton

La méthode de Lagrange permet de trouver le polynôme d'interpolation mais elle n'est pas la plus efficace d'un point de vue pratique. En effet, pour calculer le polynôme d'interpolation d'un ensemble de $(n+1)$ points on doit calculer les $(n+1)$ polynômes $L_0, L_1, \ldots, L_n$.Si ensuite on ajoute un point d'interpolation, on doit calculer les $(n+2)$ polynômes $L'_0, L'_1, \ldots, L'_{(n+1)}$ qui diffèrent tous des $(n+1)$ calculés précédemment (comparez exemple 2.2 et 2.3). La méthode de Newton est basée sur le choix d'autre fonctions de base de sort à ce que l'ajout d'un point comporte juste l'ajout d'une fonction c.à.d. un polynôme . Dans la méthode de Newton les polynômes de base sont :

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= (x - x_0) \\ &\vdots \\ p_n(x) &= (x - x_0)(x - x_1)\ldots(x - x_{n-1}) \end{aligned} \tag{2.8}$$

La forme générale est :

$$p_k(x) = \prod_{i=0}^{k-1}(x - x_i) \tag{2.9}$$

Pour les points de collocation ces polynômes ont la propriété suivante :

$$p_k(x_i) = 0 \qquad (i < k) \tag{2.10}$$

Cela signifie que $p_0(x)$ est non nul pour tous les nœuds. Le polynôme $p_1(x)$ est nul en x_0 et il est non nul pour les autres nœuds. Le nombre de zéros parmi les points de collocation est toujours en augmentation continue jusqu'à ce que le dernier polynôme $p_n(x)$ est nul pour tous les nœuds de x_0 à x_{n-1}. Et ne diffère de zéro que pour x_n.Cette propriété garantit l'approche

$$P_n(x) = \sum_{k=0}^{n} c_k p_k(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1)\dots(x - x_{n-1}) \tag{2.11}$$

D'où

$$\begin{aligned} P_n(x_0) &= f(x_0) = c_0 \\ P_n(x_1) &= f(x_1) = c_0 + c_1(x_1 - x_0) \\ P_n(x_2) &= f(x_2) = c_0 + c_1(x_2 - x_0) + c_2(x_2 - x_0)(x_2 - x_1) \\ &\vdots \\ P_n(x_n) &= f(x_n) = c_0 + c_1(x_n - x_0) + +c_n(x_n - x_0)(x_n - x_1)(x_n - x_{n-1}) \end{aligned} \tag{2.12}$$

Cela conduit à la détermination des inconnues c_i

$$\begin{aligned} c_0 &= f(x_0) \\ c_1 &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\ c_2 &= \frac{\frac{f(x_2)-f(x_1)}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2 - x_0} \\ &\vdots \end{aligned} \tag{2.13}$$

Nous introduisant maintenant la notion de différence divisée.

Définition 2.2 *: Différences divisées*
Soit $\{(x_i, y_i) \quad i = 0, 1, , n\}$ *un ensemble de* $(n+1)$ *points distincts. La différence divisée d'ordre 0 pour* x_i *est simplement la valeur de* $f(x)$ *en* x_i*.*

$$f[x_i] = f(x_i) \tag{2.14}$$

La différence divisée d'ordre 1 de x_i *et* x_{i+1} *est*

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i} \tag{2.15}$$

La différence divisée d'ordre 2 de x_i *,* x_{i+1} *et* x_{i+2}*.*

$$f[x_i, x_{i+1}, x_{i+2}] = \frac{f[x_{i+1}, x_{i+2}] - f[x_i, x_{i+1}]}{x_{i+2} - x_i} \tag{2.16}$$

La différence divisée d'ordre n *des* $(n+1)$ *points* $\{x_i, , x_{n+1}\}$ *est définie par récurrence en utilisant deux différences divisées d'ordre* $n-1$ *comme suit :*

$$f[x_i, ..., x_{i+n}] = \frac{f[x_{i+1}, ..., x_{i+n}] - f[x_i, ..., x_{i+n-1}]}{x_{i+n} - x_i} \tag{2.17}$$

Pour expliciter le processus récursif, les différences divisées peuvent être calculées en les disposants de la manière suivante dans un tableau :

x_i	$f[x_i]$	$f[xi, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$	....	...
x_0	$f[x_0]$					
		$f[x_0, x_1]$				
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$			
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$		
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$		...	
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$		...
x_3	$f[x_3]$		$f[x_2, x_3, x_4]$		...	
		$f[x_3, x_4]$		$f[x_2, x_3, x_4, x_5]$		
x_4	$f[x_4]$		$f[x_3, x_4, x_5]$			
		$f[x_4, x_5]$				
x_5	$f[x_5]$					

Théoreme 2.3 *Formule de Newton*
Soit $\{(x_i, y_i) \quad i = 0, 1, \ldots, n\}$ *un ensemble de* $(n+1)$ *points distincts. Le polynôme d'interpolation de Newton est donné par :*

$$\boxed{P_n(x) = f[x_0] + \sum_{k=1}^{n} f[x_0, x_1, ..., x_k](x - x_0)...(x - x_{n-1})} \tag{2.18}$$

.

Exemple 2.6 *Trouver le polynôme d'interpolation de Newton qui passe par les points* $(-1, 1)$,$(1, -1)$, $(2, 4)$, *et* $(5, 1)$.

En utilisant la diagonale supérieure ,les coefficients (en gras) du polynôme de Newton sont mis en évidence dans le diagramme suivant

x_i	$f[x_i]$	$f[xi, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
-1	**1**			
		$\frac{-1-1}{1-(-1)} =$ **-1**		
1	-1		$\frac{5-(-1)}{2-(-1)} =$**2**	
		$\frac{4-(-1)}{2-1} = 5$		$\frac{\frac{-3}{2}-2}{5-(-1)} =$**-7/12**
2	4		$\frac{-1-5}{5-1)} = \frac{-3}{2}$	
		$\frac{1-4}{5-2} = -1$		
5	1			

Ainsi, le polynôme d'interpolation de Newton est écrit sous la forme :

$$P_3(x) = 1 - (x+1) + 2(x+1)(x-1) - \frac{7}{12}(x+1)(x-1)(x-2)$$

$$P_3(x) = -\frac{7}{12}x^3 + \frac{19}{6}x^2 - \frac{5}{12}x - \frac{19}{6}$$

On peut aussi utiliser la diagonale inférieure et les coefficients en gras dans le tableau suivant :

x_i	$f[x_i]$	$f[xi, x_{i+1}]$	$f[x_i, x_{i+1}, x_{i+2}]$	$f[x_i, x_{i+1}, x_{i+2}, x_{i+3}]$
-1	1			
		$\frac{-1-1}{1-(-1)} = -1$		
1	-1		$\frac{5-(-1)}{2-(-1)} = 2$	
		$\frac{4-(-1)}{2-1} = 5$		$\frac{\frac{-3}{2}-2}{5-(-1)} =$ **-7/12**
2	4		$\frac{-1-5}{5-1)} =$ **-3/2**	
		$\frac{1-4}{5-2} =$ **-1**		
5	**1**			

Ainsi, le polynôme d'interpolation de Newton est écrit sous la forme :

$$P_3(x) = 1 - (x-5) - \frac{3}{2}(x-5)(x-2) - \frac{7}{12}(x-5)(x-2)(x-1)$$

$$P_3(x) = -\frac{7}{12}x^3 + \frac{19}{6}x^2 - \frac{5}{12}x - \frac{19}{6}$$

Chapitre 3

Intégration Numérique

3.1 Introduction

Dans ce chapitre on va étudier des méthodes pour approcher les intégrales de fonctions.On sait bien qu'il n'est pas toujours possible, pour une fonction arbitraire, de trouver la forme explicite d'une primitive. Dans les méthodes d'intégration, l'intégrale d'une fonction f continue sur un intervalle borné $[a, b]$ est remplacée par une somme finie. Le choix de la subdivision de l'intervalle d'intégration et celui des coefficients qui interviennent dans la somme approchant l'intégrale sont des critères essentiels pour minimiser l'erreur. Ces méthodes se répartissent en deux grandes catégories : les méthodes composées dans lesquelles la fonction est remplacée par un polynôme d'interpolation sur chaque intervalle élémentaire $[x_i, x_{i+1}]$ de la subdivision de $[a, b]$ ($[a, b] = \bigcup_i [x_i, x_{i+1}]$) et les méthodes de Gauss pour lesquelles les points de la subdivision sont imposés.

La méthode de base impliquée dans l'approximation de $\int_a^b f(x)dx$ est appelée quadrature numérique.

Elle utilise la somme $\sum_{i=1}^{n} w_i f(x_i)$ pour l'approximation de $\int_a^b f(x)dx$.

La méthode de quadrature dans cette section est basée sur l'interpolation polynomiales étudiée dans la Chapitre 2. L'idée de base est de sélectionner un ensemble de points $\{x_0, x_1, \ldots, x_n\}$, de l'intervalle $[a, b]$. Puis intégrer le polynôme de Lagrange.
On a :

$$f(x) = P_n(x) + \frac{f^{(n+1)}(\xi(x))}{(n+1)!} \prod_{i=0}^{n} (x - x_i) \tag{3.1}$$

Avec

$$P_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) \tag{3.2}$$

Ainsi :

$$\begin{aligned} \int_a^b f(x)dx &= \int_a^b \sum_{i=0}^{n} f(x_i) L_i(x)dx + \int_a^b \prod_{i=0}^{n} (x - x_i) \frac{f^{(n+1)}(\xi(x))}{(n+1)!} dx \\ &= \sum_{i=0}^{n} a_i f(x_i)dx + \int_a^b \prod_{i=0}^{n} (x - x_i) \frac{f^{(n+1)}(\xi(x))}{(n+1)!} dx \end{aligned} \tag{3.3}$$

Avec
$\xi(x) \in$[a,b] pour chaque x et

$$a_i = \int_a^b L_i(x)dx \quad i = 0, 1, , n. \tag{3.4}$$

et la formule de quadrature est ainsi

$$\int_a^b f(x)dx \approx \sum_{i=0}^{n} a_i f(x_i)dx \tag{3.5}$$

Avec l'erreur donnée par

$$E(f) = \frac{1}{(n+1)!} \int_a^b \prod_{i=0}^{n} (x - x_i) f^{(n+1)}(\xi(x))dx \tag{3.6}$$

3.2 La règle du trapèze

Nous commençons par l'application la plus simple de l'intégration numérique basée sur l'interpolation. Soit $f(x)$ une fonction continue ainsi que sa première dérivée et cela sur l'intervalle $[a, b]$.Soit $x_0 = a\,;\ x_1 = b\,; h = b - a$. Considérons le polynôme interpolant de degré 1 $P_1(x)$ passant par $(x_0, f(x_0))$ et $(x_1, f(x_1))$ ansi que le terme d'erreur

$$f(x) = P_1(x) + E(f) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1) + \frac{(x - x_0)(x - x_1)}{2!} f''(\xi_x) \tag{3.7}$$

Intégrant les deux membre de l'équation (3.7)

$$\int_{x_0}^{x_1} f(x)dx = \int_{x_0}^{x_1} P_1(x)dx + \int_{x_0}^{x_1} E(f)dx \tag{3.8}$$

La première intégrale de droite donne

$$\int_{x_0}^{x_1} P_1(x)dx = f(x_0) \int_{x_0}^{x_1} \frac{x - x_1}{x_0 - x_1} dx + f(x_1) \int_{x_0}^{x_1} \frac{x - x_0}{x_1 - x_0} dx = \frac{h}{2}(f(x_0) + f(x_1)) \tag{3.9}$$

La formule (3.9) calcule l'aire du trapèze(figure 3.1) d'où le nom de règle du trapèze
Le terme de l'erreur est

$$\int_{x_0}^{x_1} E(f)dx = \frac{1}{2!} f''(\xi) \int_{x_0}^{x_1} (x - x_0)(x - x_1) f''(\xi(x))dx \tag{3.10}$$

En utilisant le théorème des valeurs moyenne des intégrales on aura

$$\int_{x_0}^{x_1} E(f)dx = \frac{1}{2!} f''(\xi) \int_{x_0}^{x_1} (x - x_0)(x - x_1)dx = -\frac{h^3}{12} f''(\xi) \tag{3.11}$$

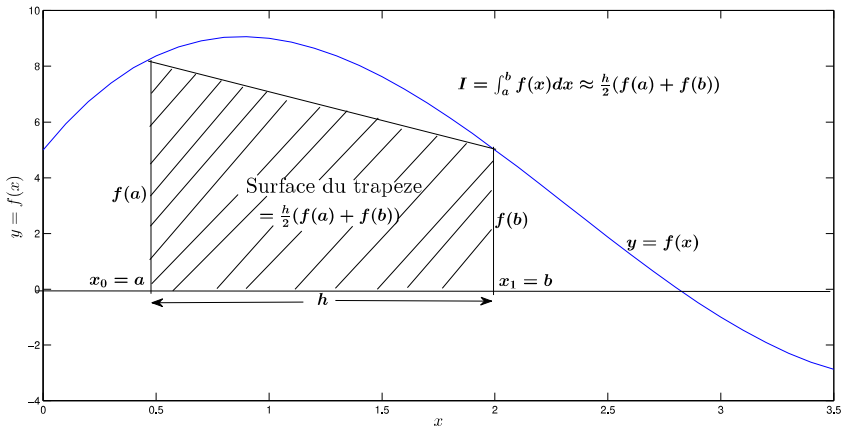


FIGURE 3.1 – l'approximation de l'intégrale par l'aire du trapèze

La règle du Trapèze simple est :

$$\int_a^b f(x)dx \approx \frac{h}{2}(f(b) + f(b))$$

Avec une erreur

$$E(I) = \int_a^b E(f)dx = -\frac{(b-a)^3}{12}f''(\xi); \qquad a < \xi < b$$

.

Exemple 3.1 *Utilisez la règle du trapèze simple pour faire une approximation à*

$$\int_1^2 \ln x dx$$

Et trouvez une limite supérieure pour l'erreur .

$$\int_1^2 \ln x dx \approx \frac{1}{2}(\ln 2 + \ln 1) = 0.3466$$

L'erreur dans la règle du trapèze est $-h^3 f''(\xi)/12$ avec $1 < \xi < 2$ puisque $f''(x) = -1/x^2$:

$$\frac{1}{12\xi^2} \leq \frac{1}{12} \approx 0.0834 \tag{3.12}$$

Ainsi

$$0.3466 - 0.0834 \leq \int_1^2 \ln x dx \leq 0.3466 + 0.0834$$

L'intégrale peut être calculée exactement en utilisant l'intégration par parties :

$$\int_1^2 \ln x dx = x\ln x|_1^2 - x|_1^2 = 2\ln 2 - 1 \approx 0.386294 \tag{3.13}$$

De l'équation (3.11) on voit que l'erreur dépend de h ,l'amélioration de la précision de la méthode du trapèze ,consiste à réduire la valeur de h en décomposant l'intervalle $[a, b]$ on n sous-intervalle de longueur $h = \frac{b-a}{n}$ (Fig 3.2). Les différents points engendrés sont notés x_i pour $i = 0, 1, 2, ..., n$.Les valeurs aux extrémités sont $a = x_0$ et $b = x_n$ .Dans chaque sous intervalle $[x_i, x_{i+1}]$ on applique la règle du trapèze.

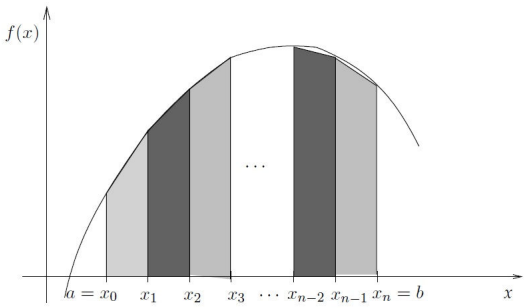


FIGURE 3.2 – décomposition de l'aire en n trapèze

$$\begin{aligned}\int_a^b f(x)dx &= \sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}} f(x)dx \\ &= \frac{h}{2}\left([f(x_0)+f(x_1)]+[f(x_1)+f(x_2)]+,...,[f(x_{n-1}+f(x_n)]\right) - \sum_{i=0}^{m-1}\frac{h^3}{12}f''(c_i) \\ &= \frac{h}{2}\left[f(a)+f(b)+2\sum_{i=1}^{n-1}f(x_i)\right] - \sum_{i=0}^{n-1}\frac{h^3}{12}f''(c_i)\end{aligned} \tag{3.14}$$

En utilisant le théorème des valeur intermédiaires généralisé, le terme d'erreur peut être écrit :

$$\frac{h^3}{12}\sum_{i=0}^{n-1}f''(c_i) = n\frac{h^3}{12}f''(c), \qquad a < c < b \tag{3.15}$$

La règle du Trapèze composée est :

$$\int_a^b f(x)dx \approx \frac{h}{2}\left(f(a)+2f(x_1)+2f(x_2)+,...,2f(x_{n-1})+f(b)\right)$$

Avec une erreur

$$E(I) = -\frac{(b-a)h^2}{12}f''(c); \qquad a < c < b$$

.

Exemple 3.2 *Refaire l'exemple 3.1 en utilisant la règle du trapèze composée avec* $m = 4$ *sous-intervalle,de largueur* $h = \frac{2-1}{4} = 0.25$

$$\int_a^b f(x)dx \approx \frac{0.25}{2}\left[\ln 1 + 2\ln 1.25 + 2\ln 1.5 + 2\ln 1.75 + \ln 2\right] \approx 0.383699$$

L'erreur est tout au plus

$$\frac{(b-a)h^2}{12}\left|f''(c)\right| = \frac{1/16}{12}\frac{1}{c^2} \leq \frac{1}{(16)(12)(1^2)} = 0.052$$

3.3 Les règles de Simpson

La règle du trapèze converge lentement car le remplacement des courbures par des segments de droite est trop grossier. Une amélioration peut être attendue si la courbe est approximée par un polynôme de degré 2 (Simpson 1/3) ou degré 3 (Simpson 3/8).

3.3.1 Règle Simpson $\frac{1}{3}$

En évaluant $\int_a^b f(x))dx$, la règle de Simpson 1/3 utilise un polynôme de second degré pour approximer l'intégrant de $f(x)$. Les trois points nécessaires pour déterminer le polynôme $P_2(x)$ sont choisis comme $x_0 = a, x_1 = (a+b)/2$, et $x_2 = b$ (Fig.3.3).

$$f(x) = P_2(x) + E(f) = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_1}f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2}f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1}f(x_2) + \frac{(x-x_0)(x-x_1)(x-x_2)}{3!}f'''(c_x) \quad (3.16)$$

L'intégration donne

$$\int_{x_0}^{x_2} f(x)dx = \int_{x_0}^{x_2} P_2(x)dx + \int_{x_0}^{x_2} E(f)dx \quad (3.17)$$

avec

$$\int_{x_0}^{x_2} P_2(x)dx = f(x_0)\int_{x_0}^{x_2}\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}dx + f(x_1)\int_{x_0}^{x_2}\frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}dx + f(x_2)\int_{x_0}^{x_2}\frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}dx = \frac{h}{3}(f(x_0)+4f(x_1)+f(x_2)) \quad (3.18)$$

Le terme de l'erreur donne

$$E(I) = \int_{x_0}^{x_2} E(f)dx = -\frac{h^5}{90}f^{(4)}(c) \qquad c \in [x_0, x_2] \quad (3.19)$$

Et cela à condition que $f^{(4)}(x)$ existe et soit continue

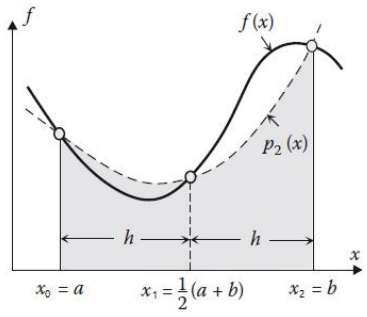


FIGURE 3.3 – l'approximation de l'intégrale par l'aire de $P_2(x)$

Exemple 3.3 *Refaire l'exemple (3.1) en utilisant la règle du Simpson 1/3 .*

ici $h = (2-1)/2 = 0.5$

$$\int_1^2 \ln(x)dx = \frac{0.5}{3}(\ln(1) + 4\ln(1.5) + \ln(2)) = 0.385834$$

L'erreur pour la règle de Simpson est $-h^5/90 f^{(4)}(c)$, avec $1 < c < 2$. Puisque $f^{(4)}(x) = -6/x^4$ l'erreur est au plus

$$\frac{6(0.5)^5}{90c^4} \leq \frac{6(0.5)^5}{90(1)^4} = 0.00208$$

$$0.3466 - 0.00208 \leq \int_1^2 \ln(x)dx \leq 0.3466 + 0.00208$$

Dans la règle de Simpson 1/3 composée, l'intervalle $[a, b]$ est divisé en un nombre pair de sous-intervalle $n = 2m$ définis par $n + 1$ points $a = x_0, x_1, ..., x_n = b$. les points sont également espacés avec $h = (b-a)/n$. Comme trois points sont nécessaires pour construire un polynôme du deuxième degré, le La règle 1/3 de Simpson doit être appliquée à deux sous-intervalles adjacents à la fois. Par exemple, la première application concernera les deux premiers sous-intervalles $[x_0, x_1]$ et $[x_1, x_2]$ de sorte que les trois points correspondant à x_0, x_1 et x_2 soient utilisés pour la construction polynomiale. La prochaine application sera pour $[x_2, x_3]$ et $[x_3, x_4]$, de sorte que x_2, x_3 et x_4 soient utilisés pour la construction polynomiale (Fig. 3.4) Sur l'intervalle $[x_{2i}, x_{2i+2}]$ on applique Simpson 1/3

$$\int_{x_{2i}}^{x_{2i+2}} f(x)dx = \frac{h}{3}[f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2i+2})] - \frac{h^5}{90}f^{(4)}(c_i) \tag{3.20}$$

Sur chaque 2 sous-intervalles la méthode Simpson 1/3 est appliquée, le nombre total d'application est $m = n/2$ ainsi :

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(a) + f(b) + 4\sum_{i=1}^{m} f(x_{2i-1}) + 2\sum_{i=1}^{m-1} f(x_{2i})\right] - \sum_{i=0}^{m-1}\frac{h^5}{90}f^{(4)}(c_i) \tag{3.21}$$

Pour $(b-a) = nh = 2mh$ le terme de l'erreur s'écrit :

$$\sum_{i=0}^{m-1}\frac{h^5}{90}f^{(4)}(c_i) = m\frac{h^5}{90}f^{(4)}(c) = (b-a)\frac{h^4}{180}f^{(4)}(c) \quad a < c < b \tag{3.22}$$

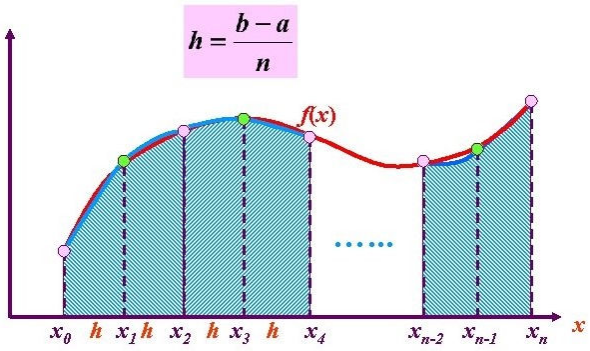


FIGURE 3.4 – La règle de Simpson 1/3 composée

La règle de Simpson 1/3 simple:

$$\int_a^b f(x)dx \approx \frac{h}{3}(f(a) + 4f(\frac{a+b}{2}) + f(b))$$

Avec une erreur

$$E(I) = \frac{h^5}{90}f^{(4)}(c) \quad h = (b-a)/2 \quad a < c < b$$

La règle de Simpson 1/3 composée:

$$\int_a^b f(x)dx \approx \frac{h}{3}\left[f(a) + f(b) + 4\sum_{i=1}^{m} f(x_{2i-1}) + 2\sum_{i=1}^{m-1} f(x_{2i})\right]$$

Avec une erreur

$$E(I) = (b-a)\frac{h^4}{180}f^{(4)}(c) \quad a < c < b$$

3.3.2 Règle de Simpson $\frac{3}{8}$

Si on utilise un polynôme de degré 3 dans l'intervalle $[a, b]$ et passant par les points $(xi, f(xi))$ pour $(i = 0, 1, 2, 3)$, avec $x_0 = a, x_1 = x_0 + h, x_2 = x_0 + 2h$, $x_3 = b$ et $h = (b-a)/3$ on obtient la formule de Simpson 3/8 simple qui s'écrit.

La règle de Simpson 3/8 simple:

$$\int_{x_0}^{x_3} f(x)dx \approx \frac{3h}{8}\left[f(x_0)+3f(x_1)+3f(x_2)+f(x_3)\right]$$

Avec une erreur

$$E(I) = -\frac{3}{80}h^5 f^{(4)}(c) \quad x_0 < c < x_3$$

.

On peut également composer cette méthode en divisant l'intervalle d'intégration $[a,b]$ en $3n$ sous-intervalles de longueur $h=(b-a)/3n$: et en utilisant la formule de Simpson 3/8 simple dans chaque triplet de sous-intervalle. On obtient alors la règle de Simpson composée :

La règle de Simpson 3/8 composé:

$$\int_a^b f(x)dx = \sum_{i=0}^{n-1}\int_{x_{3i}}^{x_{3i+3}} f(x)dx \approx \frac{3h}{8}\sum_{i=0}^{n-1}(f(x_{3i})+3f(x_{3i+1})+3f(x_{3i+2})+f(x_{3i+3}))$$

Avec une erreur

$$E(I) = -\frac{(b-a)}{80}h^4 f^{(4)}(c) \quad a < c < b$$

.

Exercice 2.1 : Soit $I=\int_0^1 e^x dx$ =1.7182818284590 46

1. Faite une approximation de I par la règle du trapèze simple.
2. Faite une approximation de I par la règle du trapèze composée, on prend 5 sous-intervalles.
3. quelle est le nombre de sous-intervalle pour avoir une erreur inférieure à 10^{-3}.

Solution abrégée :

1. Trapèze simple $I=\int_0^1 e^x dx \approx \frac{1}{2}(e^0+e^1)=1,8591409142.$
2. Trapèze composée $I=\int_0^1 e^x dx \approx \frac{0.2}{2}(e^0+2e^{0.2}+2e^{0.4}+2e^{0.6}+2e^{0.8}+e^1)=1.7240056197$
3. le nombre de sous-intervalle pour avoir une erreur inférieure à 10^{-3}.
   $|E| = |(b-a)\frac{h^2}{12}f''(c)| = |\frac{h^2}{12}e^c| \leq |\frac{h^2}{12}e| \leq 10^{-3} \quad h=(1-0)/n \leq 0.06644$ donc $n \geq 15.051$ donc apartir de $n=16$ sous-intervalle on peut avoir une erreur $< 10^{-3}$.

Exercice 2.2 : refaire l'exercice précédent avec la méthode de Simpson $\frac{1}{3}$
Solution abrégée :

1. Simpson simple $h=(1-0)/2=0.5$, $I=\int_0^1 e^x dx \approx \frac{0.5}{3}(e^0+4e_0.5+e1)=1.7188611518.$
2. Simpson composée pour appliquer la règle de Simpson composée il faut que le nombre de sous intervalle doit être pair .Au lieu de 5 prenant 4 sous-intervale de longueur $h=(1-0)/4=0.25$ ainsi
   $I=\int_0^1 e^x dx \approx \frac{0.25}{3}(e^0+4e^{0.25}+2e^{0.5}+4e^{0.75}+e^1)=1.7183188419$

3. le nombre de sous-intervalle pour avoir une erreur inférieure à 10^{-3}.

$$|E| = \left|(b-a)\frac{h^4}{180}f^{(4)}(c)\right| = \left|\frac{h^4}{180}e^c\right| \leq \left|\frac{h^4}{180}e\right| \leq 10^{-3}, \quad h = (1-0)/n \leq 0.507276$$

soit $n \geq 1.97$ donc a partir de $n = 2$ sous-intervalle on peut avoir une erreur $< 10^{-3}$.c.a.d la règle de Simpson simple donne cette précision et on le voit dans le résultat 1 avec 3 chiffres après la virgule exacts.

3.4 Quadrature de Gauss

3.4.1 Introduction

La quadrature de Gauss est un outil très puissant pour approximer les intégrales. Elle est dérivé d'une manière très différente des règles trapézoïdales et Simpson . Avant d'entrer dans les détails de la dérivation de la méthode, ce pourrait être une bonne idée pour donner une idée de la précision de la quadrature de Gauss. Les règles de quadrature sont tous basés sur des valeurs spéciales de poids et d'abscisses (points d'évaluation, communément appelés "Points et Poids de Gauss") qui sont pré-calculés et stockés. Ils sont disponibles dans des tableaux mathématiques standards, et peuvent également être calculés par des codes informatiques standard disponibles. Le tableau (Table 3.1) donne certaines des valeurs pour quelques cas.

3.4.2 Quadrature de Gauss pour l'intervalle $[-1, 1]$

La règle de quadrature est écrite sous la forme.

$$I = \int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{n} w_i^{(n)} f(x_i^{(n)}) \tag{3.23}$$

Le degré de précision d'une formule d'intégration numérique est défini comme étant le degré maximal du polynôme qui est intégré exactement par la formule (3.23) où les poids $w_i^{(n)}$ et les points de Gauss x_i des inconnues à déterminer. pour cela on pose la question suivante : Quels sont les $w_i^{(n)}$ et x_i pour lesquels la quadrature (3.23) soit exacte dans le cas des polynômes de degré le plus élevé possible ?.
Autrement on veut trouver $w_i^{(n)}$ et x_i tels que

$$\int_{-1}^{1} x^k dx = \sum_{i=1}^{n} w_i^{(n)} (x_i^{(n)})^k \qquad k = 0, 1, ..., N \tag{3.24}$$

Lemme 3.1 *Si $N = 2n$, alors il n'y a pas de poids et de points de Gauss tels que* (3.24) *est satisfait pour tout $k = 0, 1, 2, ..., N$.*

Démonstration Soit $\left\{w_i^{(n)}\right\}$ et $\left\{x_i^{(n)}\right\}$ satisfont (3.24) , et supposons que (3.24) est valable pour tout $k = 0, 1, 2, ..., N$, c.a.d la règle de quadrature est exacte pour tous les polynômes de degré inférieur ou égal à $N = 2n$. Définissant

$$L(x) = \prod_{j=1}^{n} (x - x_j^{(n)})^2$$

Ainsi $L(x) \geq 0$ et $\int_{-1}^{1} L(x)dx > 0$ .Cependant

$$\sum_{i=1}^{n} w_i^{(n)} L(x_i^{(n)}) = 0$$

Par conséquent, nous avons une contradiction, car la règle de quadrature (3.24) est censée être exacte pour tous les polynômes de degré $2n$, et c'est clairement pas exact pour le polynôme particulier de degré $2n$ donné par $L(x)$. Par conséquent, notre hypothèse selon laquelle une solution existe doit être fausse, et le lemme est prouvé.

Lemme 3.2 *Soit* $\left\{w_i^{(n)}\right\}$ *un ensemble de poids et* $\left\{x_i^{(n)}\right\}$ *un ensemble de points de Gauss, tels que* (3.24) *est satisfaite pour* $k = 0, 1, 2, ..., N = 2n - 1$. *Alors les poids doivent satisfaire*

$$w_i^{(n)} = \int_{-1}^{1} L_i^{(n)}(x)dx \tag{3.25}$$

Avec

$$L_i^{(n)}(x) = \prod_{\substack{k=1 \\ k\neq i}}^{n} \frac{x - x_k^{(n)}}{x_i^{(n)} - x_k^{(n)}}$$

Démonstration : Notez que les $L_i^{(n)}$ sont de degré $n-1 < 2n-1$, et rappelons que $L_i^{(n)}(x_j) = \delta_{ij}$. Ainsi, le fait que la règle de quadrature soit supposée être exacte pour tous les polynômes de degré $< 2n - 1$ forces

$$\int_{-1}^{1} L_i^{(n)}(x)dx = \sum_{j=1}^{n} w_j^{(n)} L_i^{(n)}(x_j^{(n)}) = w_i^{(n)}$$

et ceci complète la preuve. Une méthode pour trouver la formule de quadrature de Gauss est la méthode des coefficients indéterminés dans laquelle les nœuds et les poids sont traités comme des inconnues et les équations (3.24) sont résolues pour une valeur de k aussi grande que possible. Dans cette section on se limitera à la résolution de (3.24) dans le cas $n = 2$ c.a.d $N = 2n - 1 = 3$ et $k = 0, 1, 2, 3$

k	*Intégrale* $\int_{-1}^{1} x^k dx$	*Quadrature* $\sum_{i=1}^{n} w_i^{(n)}(x_i^{(n)})^k$
0	$\int_{-1}^{1} dx = x\|_{-1}^{1} = 2$	$w_1.1 + w_2.1$
1	$\int_{-1}^{1} xdx = \frac{1}{1}x^2\|_{-1}^{1} = 0$	$w_1x_1 + w_2x_2$
2	$\int_{-1}^{1} x^2dx = \frac{1}{3}x^3\|_{-1}^{1} = \frac{2}{3}$	$w_1x_1^2 + w_2x_2^2$
3	$\int_{-1}^{1} x^3dx = \frac{1}{4}x^4\|_{-1}^{1} = 0$	$w_1x_1^3 + w_2x_2^3$

d'où le système d'équations non linéaire suivant :

$$\begin{cases} w_1.1 + w_2.1 = 2 \\ w_1x_1 + w_2x_2 = 0 \\ w_1x_1^2 + w_2x_2^2 = \frac{2}{3} \\ w_1x_1^3 + w_2x_2^3 = 0 \end{cases} \tag{3.26}$$

Dont une des solutions

$$x_1 = -\frac{\sqrt{3}}{3}, x_2 = \frac{\sqrt{3}}{3}, w_1 = 1, w_2 = 1$$

Le tableau suivant résume les valeurs des x_i et w_i pour $n = 1$ à $n = 5$

TABLE 3.1 – Poids et points de Gauss.

Nombre de points	Poids w_i	Points x_i
1	$w_1 = 2$	$x_1 = 0$
2	$w_1 = 1$	$x_1 = -0.57735027$
	$w_2 = 1$	$x_2 = 0.57735027$
3	$w_1 = 0.5555556$	$x_1 = -0.77459667$
	$w_2 = 0.8888889$	$x_2 = 0$
	$w_3 = 0.5555556$	$x_3 = 0.77459667$
4	$w_1 = 0.3478548$	$x_1 = -0.86113631$
	$w_2 = 0.6521452$	$x_2 = -0.33998104$
	$w_3 = 0.6521452$	$x_3 = 0.33998104$
	$w_4 = 0.3478548$	$x_4 = 0.86113631$
5	$w_1 = 0.2369269$	$x_1 = -0.90617985$
	$w_2 = 0.4786287$	$x_2 = -0.53846931$
	$w_3 = 0.5688889$	$x_3 = 0$
	$w_4 = 0.4786287$	$x_4 = 0.53846931$
	$w_5 = 0.2369269$	$x_5 = 0.90617985$

Si $f(x)$ est un un polynôme de degrés $\leq (2n-1)$ on utilise la formule exacte (3.24) et pour les autres fonctions on utilise la formule approximative (3.23).
Exercice : Calculer

$$I = \int_{-1}^{1} f(x)dx = \int_{-1}^{1} \frac{1}{1+x^2}dx$$

en utilisant la quadrature de Gauss de 1,2 et 3 points et comparer les résultats avec la valeur exacte .

1. : La formule à un point

$$I = \int_{-1}^{1} \frac{1}{1+x^2}dx \approx 2f(0) = 2$$

2. : La formule à 2 points

$$I = \int_{-1}^{1} \frac{1}{1+x^2}dx \approx f\left(\frac{-\sqrt{3}}{3}\right) + f\left(\frac{\sqrt{3}}{3}\right) = 1.5$$

3. : La formule à 3 points

$$I = \int_{-1}^{1} \frac{1}{1+x^2}dx \approx \frac{5}{9}f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9}f(0) + f\left(\sqrt{\frac{3}{5}}\right) = 1.58333$$

La solution exacte est donnée par

$$\int_{-1}^{1} \frac{1}{1+x^2}dx = \tan^{-1}(x) = \frac{\pi}{2} = 1.571$$

Ainsi le plus de points donne le meilleur résultat.

3.4.3 Quadrature de Gauss pour l'intervalle $[a, b]$

Que faire si l'intégrale n'est pas posée sur l'intervalle $[-1, 1]$?.Ce n'est pas un obstacle à l'application de la quadrature de Gauss, car on peut appliquer un simple changement de variable pour réécrire toute intégrale sur $[a, b]$ comme intégrale sur [—1,1].on pose

$$x = \frac{b-a}{2}z + \frac{b+a}{2}$$

Alors

$$\int_a^b f(x)dx = \frac{1}{2}(b-a)\int_{-1}^{1} f(\frac{b-a}{2}z + \frac{b+a}{2})dz = \int_{-1}^{1} g(z)dz$$

Exercice : Considérant l'intégrale $\int_1^3 (x^6 - x^2 \sin 2x)dx = 317.344$ .Comparez ce résultat avec la quadrature de Gauss à 2 et 3 points.

$$x = \frac{3-1}{2}z + \frac{3+1}{2} = z + 2, \qquad dx = dz$$

$$\int_1^3 (x^6 - x^2 \sin 2x)dx = \int_{-1}^{1} ((z+2)^6 - (z+2)^2 \sin(2(z+2))dz = \int_{-1}^{1} g(z)dz$$

1. La quadrature de Gauss à 2 points :

$$\int_1^3 (x^6 - x^2 \sin 2x)dx = g(-0.57773 + 2) + g(0.57773 + 2) = 306.81993$$

2. La quadrature de Gauss à 3 points :

$$\begin{aligned}\int_1^3 (x^6 - x^2 \sin 2x)dx &= 0.55555g(-0.77459 + 2) + 0.88888g(2) \\ &+ 0.55555g(0.77459 + 2) = 317.26415\end{aligned}$$

Chapitre 4

Équations différentielles Ordinaires

4.1 Introduction

Les équations différentielles ordinaires se produisent fréquemment dans les modèles mathématiques qui surviennent dans de nombreuses branches des sciences, de l'ingénierie et de l'économie. Malheureusement il est rare que ces équations aient des solutions analytiques exactes, il est donc courant de rechercher des solutions approximatives au moyen de méthodes numériques. De nos jours, cela peut généralement être réalisé avec une grande précision et avec une limite fiable sur l'erreur entre la solution analytique et son approximation numérique. Dans cette section, nous nous intéresserons à la construction et à l'analyse de méthodes numériques pour les équations différentielles de premier ordre (problème de la condition initiale) de la forme

$$y' = f(x, y(x)), \qquad y(x_0) = y_0 \tag{4.1}$$

4.2 Méthode d'Euler

La méthode d'Euler est la technique d'approximation la plus élémentaire pour résoudre les problèmes de la condition initiale . Bien qu'elle soit rarement utilisée dans la pratique, la simplicité de sa dérivation peut être utilisée pour illustrer les techniques impliquées dans la construction de certaines des techniques les plus avancées, Le but de la méthode d'Euler est d'obtenir des approximations à la solution de (4.1)

$$y' = f(x, y), \qquad a \leq x \leq b, \qquad y(a) = \alpha \tag{4.2}$$

Une approximation continue de la solution $y(x)$ ne sera pas obtenue ; à la place, des approximations de y seront générées à différentes valeurs de x dans l'intervalle $[a, b]$, Nous choisissons les points de maillage également répartis sur l'intervalle $[a, b]$. Cette condition est assurée en choisissant un entier positif N et en calculant la pas $h = (b - a)/N = x_{i+1} - x_i$ ainsi les points de maillage sont :

$$x_i = a + ih, \qquad i = 0, 2, ..., N$$

Soit les $w_i = w(x_i)$ les approximations des $y_i = y(x_i)$ $(i = 0, 1, ...N - 1)$ dont le calcul se fait par

$$w(x_{i+1}) = w(x_i + h) = w(x_i) + hf(x_i, w(x_i)), \qquad i = 0, 1, ..., N - 1 \tag{4.3}$$

L'équation (4.3) définit ce que l'on appelle la méthode d'Euler pour résoudre (approximativement) une équation différentielle du premier ordre . Il existe deux principales dérivations de cette méthode ; l'une est géométrique, l'autre analytique . Nous commençons par la dérivation

géométrie. Considérant la figure (4.2).Elle montre le graphique de la solution exacte $y(t)$ au problème de la condition initiale

$$y' = f(x, y), \qquad y(x_0) = y_0 = w_0. \tag{4.4}$$

Par (x_0, w_0) on trace la droite tangente de pente $f(x_0, w_0)$. et pour $x = x_1$ au lieu de chercher y_1 sur la courbe inconnue $y(x)$, on lit w_1 sur la droite tangente par :

$$w_1 = w(x_1) = w_0 + hf(x_0, w_0) \approx y(x_1) = y(x_0 + h) \tag{4.5}$$

On continue le même raisonnement en lisant w_2 sur la droite passant par (x_1, w_1) et de pente $f(x_1, w_1)$

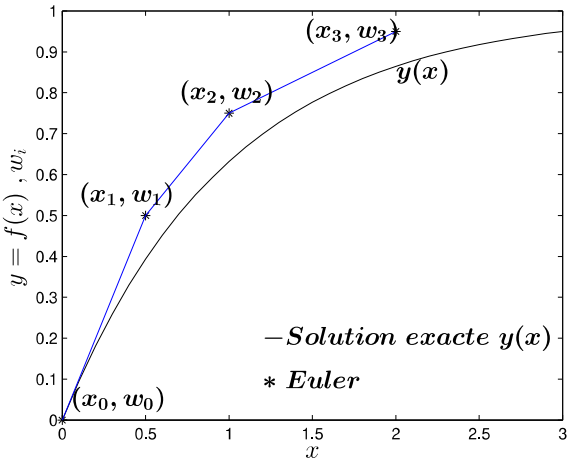


FIGURE 4.1 – Interprétation graphique de la méthode d'Euler

$$w_2 = w_1 + hf(x_1, w_1) \approx y(x_2) \tag{4.6}$$

Ou plus généralement

$$w_{i+1} = w_i + hf(x_i, w_i) \approx y(x_{i+1}) \tag{4.7}$$

La méthode d'Euler est:

$$w_0 = y(a) = \alpha,$$

$$w_{i+1} = w_i + hf(x_i, w_i), \qquad i = 0, 1, ..., N-1 \tag{4.9}$$

La dérivation analytique est basée sur le théorème de Taylor (A.3).En supposant que $y(x)$ est deux fois continuellement différenciable sur $[a, b]$

$$\begin{aligned} y(x+h) &= y(x) + hy'(x) + \frac{h^2}{2}y''(x) + ... + \frac{h^{(n)}}{n!}y^{(n)}(x) + ... \\ &= y(x) + hy'(x) + \frac{h^2}{2!}f''(\xi(x)) \end{aligned} \tag{4.10}$$

$\xi(x)$ est entre x et $x + h$. Mais l'équation différentielle implique que $y'(x) = f(x, y(x))$ nous avons donc

$$y(x+h) = y(x) + hf(x, y(x)) + \frac{h^2}{2}f''(\xi(x)) \tag{4.11}$$

négligeant le terme du reste

$$y(x+h) \approx y(x) + hf(x, y(x)) \tag{4.12}$$

Maintenant pour $x = x_i$, $w_i \approx y(x_i)$ et $w_{i+1} \approx y(x_i + h)$ on aura l'équation d'Euler (4.9)

$$w_{i+1} = w_i + hf(x_i, w_i), \qquad i = 0, 1, ..., N-1$$

L'avantage de cette dérivation est que nous avons (4.11), qui relie la solution exacte à la méthode numérique avec un reste très précis. C'est la clé pour obtenir une erreur estimation pour la méthode d'Euler. La quantité

$$\mathcal{R}_2(x, h) = \frac{h^2}{2} f''(\xi(x)) \tag{4.13}$$

est appelée le résidu pour la méthode d'Euler . Et la quantité

$$E(x, h) = \frac{\mathcal{R}_2(x, h)}{h} \tag{4.14}$$

est appelée erreur de troncature

Exemple 4.1 *Considérant l'équation différentielle suivante*

$$y' + y = 1, \qquad y(0) = 0 \tag{4.15}$$

dont la solution exacte est $y(x) = 1 - e^{-x}$ *.Calculez la valeur approximative à* $y(1)$ *ainsi que l'erreur commise en fonction du pas* h*, étant donné que la valeur exacte de* $y(1) = 1 - e^{-1} =$ 0.63212

h	*Euler*	*Erreur*
0.5	0.75	0.117879
0.2	0.67232	0.0402
0.1	0.65132	0.01920
0.05	0.64151	0.009393
0.025	0.636767	0.004647
0.02	0.63583	0.003709
0.01	0.633967	0.001841

Exemple 4.2 *Considérant l'équation différentielle suivante*

$$y' = y - x^2 + 1, \qquad y(0) = 0.5, \qquad 0 < x < 2 \tag{4.16}$$

Avec $N = 10$*. déterminez les approximations* w_i *et comparez les avec les valeurs exactes donnée par* $y = (x+1)^2 - 0.5e^x$

i	x_i	w_i	$y_i = y(x_i)$	*Erreur*
0	0.0	0.5	0.5	0.0
1	0.2	0.8	0.82930	0.02930
2	0.4	1.15200	1.21409	0.06209
3	0.6	1.55040	1.64894	0.09854
4	0.8	1.98844	2.12723	0.13875
5	1.0	2.45817	2.64086	0.18268
6	1.2	2.94981	3.17994	0.23013
7	1.4	3.45177	3.73240	0.28067
8	1.6	3.95012	4.28348	0.33335
9	1.8	4.42815	4.81517	0.38702
10	2.0	4.86578	5.30547	0.43968

d'où le graphe de la solution exacte et approximative

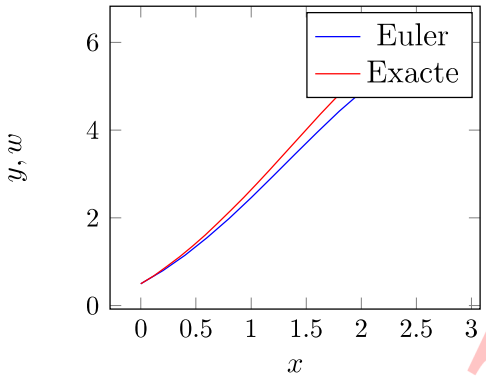


FIGURE 4.2 – L'erreur entre la solution exacte et approximation d'Euler

TABLE 4.1 – Résultats Heun

i	x_i	w_i^*	w_i	$y_i = y(x_i)$	*Erreur*
0	0.0		0.5	0.5	0
1	0.2	0.800000	0.826000	0.82930	3.30e-03
2	0.4	1.183200	1.206920	1.21409	7.17e-03
3	0.6	1.616304	1.637242	1.64894	1.17e-02
4	0.8	2.092691	2.110236	2.12723	1.70e-02
5	1.0	2.604283	2.617688	2.64086	2.32e-02
6	1.2	3.141225	3.149579	3.17994	3.04e-02
7	1.4	3.691495	3.693686	3.73240	3.87e-02
8	1.6	4.240423	4.235097	4.28348	4.84e-02
9	1.8	4.770117	4.755619	4.81517	5.96e-02
10	2.0	5.258742	5.233055	5.30547	7.24e-02

4.3 Méthode d'Euler améliorée : Méthode de Heun

La méthode d'Euler utilise la pente en un point initial (x_i, w_i) . Dans la méthode d'Euler améliorée, nous utilisons la valeur moyenne des pentes au point initial (x_i, w_i) et au point suivant (x_{i+1}, w_{i+1}^*) .Cela améliore l'estimation de la pente sur l'intervalle $[x_i, x_{i+1}]$. Au début on calcule par la méthode d'Euler une première estimation à w_{i+1} appelé w_{i+1}^* puis on utilise cette première estimation pour calculer la valeur finale de w_{i+1} .

La méthode de Heun :

$$w_0 = y(a) = \alpha,$$

$$w_{i+1}^* = w_i + h\left(f(x_i, w_i)\right) \qquad i = 0, 1, ..., N-1$$

$$w_{i+1} = w_i + \frac{h}{2}\left(f(x_i, w_i) + f(x_{i+1}, w_{i+1}^*)\right)$$

.

Exemple 4.3 *Refaire l'exemple (4.2) par la méthode de Heun*

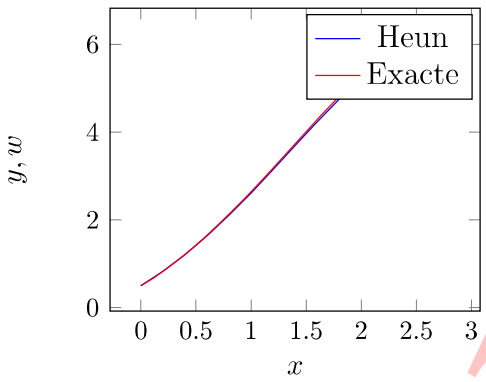


FIGURE 4.3 – L'erreur entre la solution exacte et approximation de Heun

4.4 Méthode de la série de Taylor

Une façon de réduire l'erreur de la méthode d'Euler serait d'inclure des termes d'ordre supérieur de l'expansion de la série Taylor dans la solution. Soit $y(t)$ la solution au problème de la condition initiale

$$y' = f(x, y), \qquad a \leq x \leq b, \qquad y(a) = \alpha$$

est dérivable à l'ordre $n+1$ .Si nous développons la solution $y(x)$, en termes du polynôme de Taylor d'ordre n autour du point x_i et évaluée en x_{i+1}.

$$y(x_{i+1}) = y(x_i + h) = y(x_i) + hy'(x_i) + \frac{h^2}{2}y''(x_i) + ... + \frac{h^n}{n!}y^{(n)}(x_i) + \frac{h^{(n+1)}}{(n+1)!}y^{(n+1)}(\xi_i) \quad (4.17)$$

$$y(x_{i+1}) = y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2}f'(x_i, y(x_i)) + ... + \frac{h^n}{n!}f^{(n-1)}(x_i, y(x_i)) + \frac{h^{(n+1)}}{(n+1)!}f^{(n)}(\xi_i, y(\xi_i)) \quad (4.18)$$

Avec ξ_i appartient à $[x_i, x_{i+1}]$. En négligeant le terme du reste (ξ_i) de la solution exacte (4.17) on obtient la méthode approximative de Taylor d'ordre n

$$\begin{gathered} x_0 = a, \qquad w_0 = y(a) = \alpha \\ T^{(n)}(x_i, w_i) = hf(x_i, w_i) + \tfrac{h^2}{2}f'(x_i, w_i) + ... + \tfrac{h^n}{n!}f^{(n-1)}(x_i, w_i) \\ w_{i+1} = w_i + hT^{(n)}(x_i, w_i) \qquad i = 0, 1, ..., N-1 \end{gathered} \quad (4.19)$$

La difficulté d'appliquer les méthodes de Taylor d'ordre supérieur réside dans la dérivation de la fonction f, car cette fonction dépend des deux variables x et y et y est fonction de x. La règle de la dérivation de la fonction composée est

$$\begin{aligned} f'(x, y(x)) &= \frac{\partial f(x, y(x))}{\partial x}\frac{dx}{dx} + \frac{\partial f(x, y(x))}{\partial y}\frac{dy}{dx} \\ &= \frac{\partial f(x, y(x))}{\partial x} + \frac{\partial f(x, y(x))}{\partial y}f(x, y(x)) \end{aligned} \quad (4.20)$$

La méthode d'Euler est une méthode de Taylor d'ordre 1

Exemple 4.4 *Refaire l'exemple (4.2) par la méthode de Taylor d'ordre 2*

TABLE 4.2 – Résultats Taylor d'ordre 2.

i	x_i	w_i	$y(x_i)$	*Erreur*
0	0.00	0.500000	0.500000	0.00
1	0.20	0.830000	0.829299	7.01e-04
2	0.40	1.215800	1.214088	1.71e-03
3	0.60	1.652076	1.648941	3.14e-03
4	0.80	2.132333	2.127230	5.10e-03
5	1.00	2.648646	2.640859	7.79e-03
6	1.20	3.191348	3.179942	1.14e-02
7	1.40	3.748645	3.732400	1.62e-02
8	1.60	4.306146	4.283484	2.27e-02
9	1.80	4.846299	4.815176	3.11e-02
10	2.00	5.347684	5.305472	4.22e-02

$$y' = y - x^2 + 1 \qquad 0 \leq 2 \qquad y(0) = 0.5$$

$$f'(x, y(x)) = \frac{d}{dt}(y - x^2 + 1) = y' - 2x = y - x^2 + 1 - 2t$$

$$T^{(2)}(x_i, w_i) = f(x_i, w_i) + \frac{h}{2} f'(x_i, w_i) = w_i - x_i^2 + 1 + \frac{h}{2}(w_i - x_i^2 + 1 - 2x_i)$$

Avec

$$w_{i+1} = w_i + hT^{(2)}(x_i, w_i)$$

$$w_0 = 0.5$$

$$w_1 = w_0 + hT^{(2)}(x_0, w_0) = 0.5 + 0.2T^{(2)}(0, 0.5) = 0.83$$

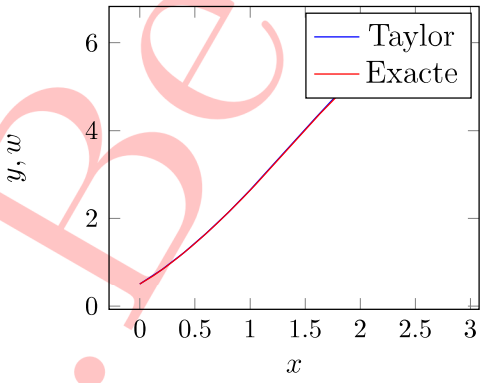


FIGURE 4.4 – L'erreur entre la solution exacte et approximation de Taylor d'ordre 2

4.5 Méthodes de Runge-Kutta (RK)

Il serait avantageux de disposer de méthodes d'ordre de plus en plus élevé tout en évitant les désavantages des méthodes de Taylor, qui nécessitent l'évaluation des dérivées partielles de la fonction $f(x, y)$. Les méthodes Runge-Kutta $(R - K)$ permettent d'obtenir la précision d'une approche de la série Taylor sans nécessiter le calcul des dérivés d'ordre élevé . Dans cette section on développe les méthodes RK d'ordre 2 et 4 Les méthodes de Runge-Kutta sont une famille de techniques numériques explicites à un pas pour résoudre une équation différentielle

du premier ordre. Pour un sous-intervalle (un pas) défini par $[x_i;, x_{i+1}]$; la valeur de w_{i+1} est calculé par :

$$w_{i+1} = w_i + Pente.h \tag{4.21}$$

où *Pente* est une constante. La valeur de la pente dans l'équation (4.21) est obtenue en considérant la pente en plusieurs points du sous-intervalle. Différents types de méthodes Runge-Kutta sont classés selon leur ordre. L'ordre identifie le nombre de points dans le sous-intervalle qui sont utilisés pour déterminer la valeur de la pente dans l'équation (4.21). Les méthodes *RK* du second ordre utilisent les pentes en deux points . Les méthodes du troisième ordre utilisent trois points, etc. L'ordre de la méthode est également lié à l'erreur de troncature globale de chaque méthode. Par exemple, les méthodes *RK* de second ordre sont précises au second ordre , c'est-à-dire qu'il a une erreur de troncature locale de $\mathcal{O}(h^3)$ et globale de $\mathcal{O}(h^2)$. Pour chaque ordre, il existe plusieurs méthodes. Les différences entre les méthodes résident dans l'emplacement des points dans le sous-intervalle considéré .Les méthodes *RK* offrent une solution plus précise que la méthode d'Euler. La précision augmente (c'est-à-dire que l'erreur de troncature diminue) avec l'accroissement de l'ordre .Cependant, a chaque pas ils nécessitent plusieurs évaluations (selon l'ordre) de la fonction pour la dérivée $f(x, y)$.

4.5.1 Méthodes de Runge-Kutta d'ordre 2 , RK2

La forme générale des méthodes Runge-Kutta de second ordre est :

$$w_{i+1} = w_i + (c_1k_1 + c_2k_2)h \tag{4.22}$$

$$\begin{aligned} k_1 &= f(x_i, w_i) \\ k_2 &= f(x_i + a_2h, w_i + b_{21}k_1h) \end{aligned} \tag{4.23}$$

Où c_1, c_2, a_2 et b_{21} sont des constantes. Les valeurs de ces constantes varient avec la méthode spécifique .Les méthodes Runge-Kutta du second ordre peuvent être associées à l'expansion de la série Taylor. Pour l'intervalle défini par $[x_i, x_{i+1}]$ la valeur de $y(x_i)$ est connue, et la valeur de $y(x_{i+1})$ peut être approximée en utilisant le développement de Taylor et l'équation (4.20) :

$$\begin{aligned} y(x_{i+1}) = y(x_i + h) &= y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2}f'(x_i, y(x_i)) + \mathcal{O}(h^3) \\ &= y(x_i) + h\left(f(x_i, y(x_i)) + \frac{h}{2}\frac{\partial f(x_i, y(x_i))}{\partial x} + \frac{h}{2}\frac{\partial f(x_i, y(x_i))}{\partial y}f(x_i, y(x_i))\right) + \mathcal{O}(h^3) \end{aligned} \tag{4.24}$$

Le développement de k_2 suivant $(A.1)$ on posant $\Delta x = a_2h$ et $\Delta y = b_{21}k_1h$.

$$k_2 = f(x_i, w_i) + \frac{\partial f(x_i, w_i)}{\partial x}a_2h + \frac{\partial f(x_i, w_i)}{\partial y}b_{21}hk_1 + \mathcal{O}(h^2) \tag{4.25}$$

Ainsi (4.22) peut s'écrire :

$$w_{i+1} = w_i + hc_1f(x_i, w_i) + c_2hf(x_i, w_i) + \frac{\partial f(x_i, w_i)}{\partial x}a_2h^2 + \frac{\partial f(x_i, w_i)}{\partial y}b_{21}h^2k_1 + \mathcal{O}(h^3) \tag{4.26}$$

Comparant les équations (4.24) et (4.26) en prenant en considération que $y(x_{i+1}) \approx w_{i+1}$ et $y(x_i) \approx w_i$ on aura le système

$(c_1 + c_2) = 1$, $c_2a_2 = \frac{1}{2}$ et $c_2b_{21} = \frac{1}{2}$ .C'est un système de 3 équations à 4 inconnues

$$\begin{cases} c_1 + c_2 = 1 \\ c_2a_2 = \frac{1}{2} \\ c_2b_{21} = \frac{1}{2} \end{cases} \tag{4.27}$$

Le système (4.27) est sous-déterminé ,il y a moins d' équations que d'inconnues ,il n'a donc pas de solution unique. Cela offre une marge de manœuvre qui favorise la mise au point de plusieurs variantes de la méthode de Runge-Kutta. Voici le choix le plus couramment utilisé.

1. La méthode d'Euler modifié ou Heun (section 4.3)

$$c_1 = \frac{1}{2}, c_2 = \frac{1}{2}, a_2 = 1, b_{21} = 1$$

2. La méthode du point milieu est la méthode de Runge-Kutta d'ordre 2 la plus utilisée

$$c_1 = 0, c_2 = 1, a_2 = \frac{1}{2}, b_{21} = \frac{1}{2}$$

$$\begin{aligned} &x_0 = a, \quad w_0 = \alpha \\ &x_{i+1} = x_i + h \\ &k = hf(x_i, w_i) \\ &w_{i+1} = w_i + hf(x_i + \tfrac{h}{2}, w_i + \tfrac{k}{2}) \quad (i = 0, 1, ..., N-1) \end{aligned} \tag{4.28}$$

4.5.2 Méthodes de Runge-Kutta d'ordre 4 RK4

Les méthodes de Runge-Kutta les plus populaires sont celles du quatrième ordre. Comme pour les approches de second ordre, il existe un nombre infini de versions. La forme la plus utilisée ou la forme classique de la méthode RK4 est donnée par :

$$\begin{aligned} &x_0 = a, \quad w_0 = \alpha \\ &k_1 = hf(x_i, w_i) \\ &k_2 = hf(x_i + \tfrac{h}{2}, w_i + \tfrac{k_1}{2}) \\ &k_3 = hf(x_i + \tfrac{h}{2}, w_i + \tfrac{k_2}{2}) \\ &k_4 = hf(x_i + h, w_i + k_3) \\ &w_{i+1} = w_i + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) \quad (i = 0, 1, ..., N-1) \end{aligned} \tag{4.29}$$

Exemple 4.5 *Appliquez les méthodes RK2 et RK4 avec $h = 0.5$ pour approximation de la solution de l'équation différentielle suivante :*

$$y' = xy, \qquad , y(0) = 1, \qquad 0 \leq x \leq 2, \qquad h = 0.5$$

Comparez le résultat numérique avec le résultat analytique donné $y = e^{\frac{x^2}{2}}$

TABLE 4.3 – Résultats Runge-Kutta d'ordre 2

i	x_i	k	w_i	y	*Erreur*
0	0.00	—	1	1.0	0
1	0.50	0.000000	1.125000	1.133148	0.008148
2	1.00	0.281250	1.599609	1.648721	0.049112
3	1.50	0.799805	2.849304	3.080217	0.230913
4	2.00	2.136978	6.277373	7.389056	1.111683

On voit qu'avec un pas $h = 0.5$ qui est assez large ,on a une erreur de relative de 15% .

Malgré que $h = 0.5$ est assez large la méthode de Runge-Kutta d'ordre 4 donne un résultat de grande précision avec une erreur relative $< 0.3\%$

TABLE 4.4 – Résultats Runge-Kutta d'ordre 4

i	x_i	k_1	k_2	k_3	k_4	w_i	$Erreur$
0	0.00	—-	—-	—-	——	1.0	0.0
1	0.50	0.000000	0.250000	0.265625	0.566406	1.133138	0.000010
2	1.00	0.566569	0.956085	1.029119	1.647698	1.648528	0.000194
3	1.50	1.648528	2.575825	2.865605	4.621995	3.077976	0.002241
4	2.00	4.616964	7.406380	8.626750	14.782702	7.366803	0.022253

Chapitre 5

Méthodes de résolutions directes des systèmes linéaires

5.1 Introduction

Dans ce chapitre, nous présentons la solution des systèmes de n équations linéaires à n inconnues. Les systèmes d'équations sont associés à de nombreux problèmes de science et d'ingénierie , ainsi qu'aux applications des mathématiques aux sciences sociales et à l'étude quantitative des problèmes économiques et commerciaux. Un système d'équations algébriques a la forme :

$$\mathcal{S}\begin{cases} a_{11}x_1+ & a_{12}x_2 & +... & +a_{1n}x_n & = b_1 \\ a_{21}x_1+ & a_{22}x_2 & +... & +a_{2n}x_n & = b_2 \\ \vdots & & & & \vdots \\ a_{n1}x_1+ & a_{n2}x_2 & +... & +a_{nn}x_n & = b_n \end{cases} \tag{5.1}$$

Où les coefficients a_{ij} et les constantes b_j sont connus et les x_i représentent les inconnues à déterminer. En notation matricielle, le système est écrit comme :

$$\begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{5.2}$$

Ou simplement $\mathbf{Ax} = \mathbf{b}$. Le système a une solution unique, si la matice A est non singulier c.a.d le déterminant $|A| \neq 0$. Les lignes et les colonnes d'une matrice non singulière sont linéairement indépendantes. Si $|A| = 0$,le système $\mathcal{S}$ peut avoir une infinité de solution ou aucune solution. Il existe deux classes de méthodes pour résoudre un système d'équations linéaires : les méthodes directes et les méthodes itératives. Les caractéristiques communes des méthodes directes sont qu'elles transforment le système original en un système équivalent qui peut être résolu plus facilement en un nombre prédéterminé d'opérations . La transformation s'effectue par l'application de certaines opérations.

Les méthodes itératives ou indirectes utilisent des itérations pour trouver la solution approximative. Le processus d'itération commence par un vecteur initial et génère des approximations successives qui finissent par converger vers la solution réelle. Contrairement aux méthodes directes , le nombre d'opérations requises par les méthodes itératives n'est pas connu à l'avance.

Dans ce chapitre, nous présenterons quelques méthodes directes. Ces méthodes transforment le système d'origine en un système équivalent dans lequel la matrice des coefficients est triangulaire supérieure, triangulaire inférieure ou diagonale.

5.2 Méthode d'élimination de Gauss

La méthode d'élimination de Gauss a pour but de transformer le système $\mathbf{Ax} = \mathbf{b}$ en un système équivalent (ayant la même solution) de la forme $\mathbf{Ux} = \widehat{\mathbf{b}}$, où $\mathbf{U}$ est une matrice triangulaire supérieure et $\widehat{\mathbf{b}}$ est un second membre convenablement modifie. Ce dernier système peut être alors résolu par une méthode de substitution arrière. Il est important de noter que la matrice augmentée $[\mathbf{A}|\mathbf{b}]$ représente complètement le système linéaire $\mathbf{Ax} = \mathbf{b}$. Par conséquent, toutes les modifications doivent être appliquées à la matrice augmentée et non à la seul matrice $\mathbf{A}$.La transformation en forme triangulaire supérieure est réalisée par les opérations suivantes qui ne changent pas la solution du système.

- Multipliez une ligne de la matrice augmentée par une constante non nulle
- Échangez deux rangées de la matrice augmentée
- Multipliez la $i^{ème}$ ligne de la matrice augmentée par une constante $\alpha \neq 0$ et ajoutez le résultat à la $k^{ième}$ ligne.La $i^{ème}$ rangée est appelé la ligne pivot

Maintenant organisons un algorithme d'élimination de Gauss pour le système général $\mathcal{S}$, qui contient n équations à n inconnues. Dans cet algorithme, les données d'origine sont écrasées avec de nouvelles valeurs calculées. Dans la phase d'élimination avant , il y'a $n-1$ étapes . La première de ces étapes utilise la première équation pour produire $n-1$ zéros comme coefficients pour chaque x_1 dans toutes les équations sauf la première. Cela se fait en soustrayant les multiples appropriés de la première équation des autres équations. Dans ce processus, nous nous référons à la première équation comme première équation pivot et à a_11 comme premier élément pivot. Pour chacun les équations $(2 \leq i \leq n)$, nous calculons

$$\begin{cases} a_{ij} \leftarrow a_{ij} - \left(\frac{a_{i1}}{a_{11}}\right) a_{1j} & (1 \leq j \leq n) \\ b_i \leftarrow b_i - \left(\frac{a_{i1}}{a_{11}}\right) b_1 \end{cases} \tag{5.3}$$

Le symbole $\leftarrow$ indique un remplacement.Ainsi le contenu de la case mémoire alloué à l'ancien a_{ij} est remplacé par le nouveau a_{ij} qui devient $a_{ij} - \left(\frac{a_{i1}}{a_{11}}\right) a_{1j}$
Après la première étape la matrice augmenté sera

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ 0 & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ 0 & a_{32} & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n2} & a_{n3} & \dots & a_{nn} & b_n \end{array}\right) \tag{5.4}$$

Maintenant la deuxième équation sera l'équation pivot et le a_{22} est le deuxième élément pivot et le processus continu pour $(3 \leq i \leq n)$

$$\begin{cases} a_{ij} \leftarrow a_{ij} - \left(\frac{a_{i2}}{a_{22}}\right) a_{2j} & (2 \leq j \leq n) \\ b_i \leftarrow b_i - \left(\frac{a_{i2}}{a_{22}}\right) b_2 \end{cases} \tag{5.5}$$

Après la deuxième étape la matrice augmenté sera

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ 0 & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ 0 & 0 & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n3} & \dots & a_{nn} & b_n \end{array}\right) \tag{5.6}$$

A la dernière étape $i = n$

$$\begin{cases} a_{ij} \leftarrow a_{ij} - \left(\frac{a_{n,n-1}}{a_{n-1,n-1}}\right) a_{n-1,j} & (n-1 \leq j \leq n) \\ b_i \leftarrow b_i - \left(\frac{a_{n,n-1}}{a_{n-1,n-1}}\right) b_{n-1} \end{cases} \tag{5.7}$$

$$\left(\begin{array}{ccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & b_1 \\ 0 & a_{22} & a_{23} & \dots & a_{2n} & b_2 \\ 0 & 0 & a_{33} & \dots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{nn} & b_n \end{array}\right) \tag{5.8}$$

Évidemment, nous devons supposer que tous les diviseurs de cet algorithme sont différents de zéro.

La substitution arrière commence par résoudre la $n^{ième}$ équation pour x_n :

$$x_n = \frac{b_n}{a_{nn}}$$

Ensuite, en utilisant la $(n-1)^{ème}$ équation, nous résolvons pour x_{n-1} :

$$x_{n-1} = \frac{1}{a_{n-1,n-1}}(b_{n-1} - a_{n-1,n}x_n)$$

Nous continuons la substitution arrière , en calculant chaque x_i par la formule :

$$x_i = \frac{1}{a_{ii}}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right) \tag{5.9}$$

Exemple 5.1 *En utilisant l'élimination de Gauss, résoudre le système d'équations suivant*

$$\begin{cases} -x_1 + 2x_2 + 3x_3 + x_4 = 3 \\ 2x_1 - 4x_2 + x_3 + 2x_4 = -1 \\ -3x_1 + 8x_2 + 4x_3 - x_4 = 6 \\ x_1 + 4x_2 + 7x_3 - 2x_4 = -4 \end{cases} \tag{5.10}$$

Solution :

La matrice augmentée du système (5.10)

$$\left(\begin{array}{cccc|c} -1 & 2 & 3 & 1 & 3 \\ 2 & -4 & 1 & 2 & -1 \\ -3 & 8 & 4 & -1 & 6 \\ 1 & 4 & 7 & -2 & -4 \end{array}\right)$$

Étant donné que l'élément $(1,1)$ est différente de zéro, nous l'utilisons comme pivot pour éliminer les coefficient directement en dessous.

$$\begin{bmatrix} -1 & 2 & 3 & 1 & 3 \\ 2 & -4 & 1 & 2 & -1 \\ -3 & 8 & 4 & -1 & 6 \\ 1 & 4 & 7 & -2 & -4 \end{bmatrix} \begin{matrix} \\ \hookleftarrow{}^{2}{}_{+} \\ \leftarrow\!\!\rightarrow{}^{-3}{}_{+} \\ \leftarrow\!\!\rightarrow{}^{1}{}_{+} \end{matrix} \rightarrow \begin{bmatrix} -1 & 2 & 3 & 1 & 3 \\ 0 & 0 & 7 & 4 & 5 \\ 0 & 2 & -5 & -4 & -3 \\ 0 & 6 & 10 & -1 & -1 \end{bmatrix} \begin{matrix} \\ \hookleftarrow \\ \hookleftarrow \end{matrix}$$

Permutez les lignes 2 et 3 car $a_{22} = 0$

$$\begin{bmatrix} -1 & 2 & 3 & 1 & 3 \\ 0 & 2 & -5 & -4 & -3 \\ 0 & 0 & 7 & 4 & 5 \\ 0 & 6 & 10 & -1 & -1 \end{bmatrix} \begin{matrix} \\ \\ \\ \longleftarrow +^{-3} \end{matrix} \rightarrow \begin{bmatrix} -1 & 2 & 3 & 1 & 3 \\ 0 & 2 & -5 & -4 & -3 \\ 0 & 0 & 7 & 4 & 5 \\ 0 & 0 & 25 & 11 & 8 \end{bmatrix} \begin{matrix} \\ \\ \\ \longleftarrow +^{\frac{-25}{7}} \end{matrix}$$

La matrice augmentée finale

$$\begin{bmatrix} -1 & 2 & 3 & 1 & 3 \\ 0 & 2 & -5 & -4 & -3 \\ 0 & 0 & 7 & 4 & 5 \\ 0 & 0 & 0 & \frac{-23}{7} & \frac{-69}{7} \end{bmatrix}$$

La solution est ensuite trouvée par substitution inverse comme suit. La dernière ligne donne

$$\frac{-23}{7}x_4 = \frac{-69}{7} \Rightarrow x_4 = 3$$

En remontant d'une ligne à la fois, en utilisant à chaque fois les dernières informations sur les inconnues, nous trouvons.

$$x_3 = \frac{1}{7}(5 - 4x_4) \Rightarrow x_3 = -1$$

$$x_2 = \frac{1}{2}(5x_3 + 4x_4 - 3) \Rightarrow x_2 = 2$$

$$x_1 = 2x_2 + 3x_3 + x_4 - 3 \Rightarrow x_1 = 1$$

5.2.1 Choix de la ligne de pivot : pivot partiel

Pour bien gérer les erreurs d'arrondi tout en traitant de grandes matrices, il est préférable utiliser le pivot partiel. dans la première étape du processus d'élimination, on choisit comme ligne pivot la ligne dans laquelle x_1 a le plus grand coefficient (en valeur absolue).

- Dans chaque ligne i de $\mathbf{A}$, recherchez l'élément avec la plus grande valeur absolue et appelez-la M_i.
- Dans chaque ligne i, trouvez le rapport entre la valeur absolue du coefficient de x_1 et la valeur absolue de M_i, c'est-à-dire

$$r_i = \frac{|a_{i1}|}{|M_i|} \tag{5.11}$$

- Parmi les r_i $(i = 1, 2, \ldots, n)$, choisissez le plus grand. La ligne correspondante à cette valeur maximale est choisie comme ligne pivot. Éliminez x_1 pour obtenir un nouveau système.
- Dans le nouveau système, considérons la sous-matrice de dimension $(n-1)(n-1)$ occupant le coin inférieur droit. Dans cette matrice, utilisez la même logique que précédemment pour choisissez la ligne pivot pour éliminer x_2, et ainsi de suite...

Exemple 5.2 *Utilisez le pivot partiel pour résoudre le système avec la matrice augmentée suivante*

$$[\mathbf{A}|\mathbf{b}] = \left(\begin{array}{ccc|c} -4 & 3 & 5 & 0 \\ 6 & 7 & -3 & 2 \\ 2 & -1 & 1 & 6 \end{array}\right)$$

Solution : Les trois valeurs de r_i sont

$$r_1 = \frac{|-4|}{|5|} = \frac{4}{5}, \qquad r_2 = \frac{|6|}{|7|} = \frac{6}{7}, \qquad r_3 = \frac{|2|}{|2|} = 1$$

Puisque r_3 est le plus grand, c'est la troisième ligne qui sera choisi comme ligne pivot.Permutant la première et la troisième ligne éliminant les coefficient de x_1 on aura la matrice augmentée suivante

$$\left(\begin{array}{ccc|c} 2 & -1 & 1 & 6 \\ 0 & \boxed{\begin{array}{cc} 10 & -6 \\ -5 & 7 \end{array}} & & -16 \\ 0 & & & 12 \end{array}\right)$$

Pour éliminer les coefficients de x_2, considérant la sous-matrice 22 et calculant r_i correspondants,

$$\frac{|10|}{|10|} = 1, \qquad \frac{|-5|}{|7|} = \frac{5}{7}$$

La première ligne de la sous matrice est prise comme ligne pivot. Éliminant les coefficient de x_2 on aura la matrice augmentée finale

$$\left(\begin{array}{ccc|c} 2 & -1 & 1 & 6 \\ 0 & 10 & -6 & -16 \\ 0 & 0 & 4 & 4 \end{array}\right)$$

La substitution arriéré donne : $x_3 = 1, x_2 = -1, x_1 = 2$

5.2.2 Coût de de la méthode l'élimination de Gauss

Étant donné que de nombreux problèmes pratiques en algèbre linéaire impliquent de très grandes matrices, un problème important est le nombre d'opérations qu'un algorithme spécifique nécessite pour gérer une matrice de taille donnée. Dans cette section, nous allons passer en revue les détails de la dérivation du nombre d'opération pour l'élimination de Gauss ainsi que pour la substitution arrière.

a) L'élimination : Supposons que les $(k-1)$ premières étapes d'élimination ont été effectuées et que nous sommes à la $k^{ème}$ étape. Cela signifie que les coefficients de x_k doivent être transformés en zéros dans les $(n-k)$ lignes restantes de la matrice augmentée.Ainsi

- $n-k$ divisions sont nécessaires pour déterminer les multiplicateurs.
- $(n-k)(n-k+1)$ multiplications
- $(n-k)(n-k+1)$ additions

Le processus d'élimination consiste en $(n-1)$ étapes. Le nombre total d'opération N_e est :

$$N_e = \underbrace{\sum_{k=1}^{n-1}(n-k)}_{Division} + \underbrace{\sum_{k=1}^{n-1}(n-k)(n-k+1)}_{Multiplication} + \underbrace{\sum_{k=1}^{n-1}(n-k)(n-k+1)}_{Addition} \tag{5.12}$$

$$N_e = \sum_{s=1}^{n-1} s + 2\sum_{s=1}^{n-1} s(s+1) = 3\sum_{s=1}^{n-1} s + 2\sum_{s=1}^{n-1} s^2 \tag{5.13}$$

En utilisant les identités (A.1) et (A.2)

$$N_e = 3\frac{(n-1)n}{2} + 2\frac{(n-1)n(2n-1)}{6} = \frac{2}{3}n^3 + \mathcal{O}(n^2) \tag{5.14}$$

b) Substitution arrière : Quand la substitution arrière est utilisée pour calculer x_k on a :
— $n-k$ multiplications.
— $(n-k)(n-k+1)$ soustractions
— *une* division
le nombre d'opération N_s

$$\begin{aligned} N_s &= \underbrace{\sum_{k=1}^{n} 1}_{Division} + \underbrace{\sum_{k=1}^{n}(n-k)}_{Multiplication} + \underbrace{\sum_{k=1}^{n}(n-k)}_{Soustraction} \\ &= n + \frac{n(n-1)}{2} + \frac{n(n-1)}{2} = n^2 \end{aligned} \tag{5.15}$$

Pour n large N_e domine N_s alors on peut dire que le nombre d'opération total $N \approx N_e = \frac{2}{3}n^3$

5.3 Factorisation LU

L'algorithme de l'élimination de Gauss développé au (7.2) est faible sur un point : le travail effectué en réduisant **A** à la forme triangulaire supérieure n'est pas exploiter si nous voulons résoudre plusieurs systèmes linéaires ayant la même matrice **A**,mais plusieurs vecteurs du coté droit qui ne sont pas connus a l'avance . Car, il est possible d'avoir une séquence de problèmes de systèmes linéaires dans laquelle le vecteur de droite pour un problème dépend de la solution du problème précédent. Ainsi, nous devrons répéter tout ce travail, qui est, en fait, la majeure partie du travail impliqué dans la solution du problème. Ce que nous allons faire est de montrer que nous pouvons factoriser la matrice **A** en un produit d'une matrice triangulaire inférieure et d'une matrice triangulaire supérieure.

$$\mathbf{A} = \mathbf{LU}$$

Cela nous permet de résoudre des systèmes linéaires en résolvant deux systèmes triangulaires :

$$\mathbf{A}x = b \Rightarrow \mathbf{U}x = y, \quad \text{avec} \quad \mathbf{L}y = b$$

Ainsi, nous résolvons d'abord $\mathbf{L}y = b$ puis $\mathbf{U}x = y$ pour obtenir la solution :

$$\mathbf{U}x = y, \mathbf{L}y = b, \Rightarrow \mathbf{U}x = \mathbf{L}^{-1}b \Leftrightarrow \mathbf{LU}x = b \Leftrightarrow \mathbf{A}x = b$$

Cela montre que nous pouvons résoudre des systèmes linéaires en calculant la factorisation **LU** de A ; ensuite résoudre les systèmes triangulaires inférieur et supérieur appropriés.

5.3.1 Factorisation de Doolittle

Théoreme 5.1 *Si l'élimination de Gauss peut être effectuée sur le système linéaire* $\mathbf{Ax} = \mathbf{b}$ *sans échanger de lignes,alors la matrice* **A** *peut être factorisée en un produit d'une matrice triangulaire inférieure* **L** *et supérieur* **U**. *De plus la matrice* **U** *est celle obtenue par l'algorithme de l'élimination de Gauss et les élément de la diagonale de* **L** *sont des un. cette factorisation est celle de Doolittle*

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & ... & a_{nn} \end{pmatrix}$$

$$\mathbf{U} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & ... & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & ... & a_{2n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & ...0 & a_{nn}^{(n)} \end{pmatrix} \text{ et } \mathbf{L} = \begin{pmatrix} 1 & 0 & ..... & 0 \\ m_{21} & 1 & ..... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ m_{n1} & m_{n2} & ....m_{n,n-1} & 1 \end{pmatrix}$$

avec $m_{ji} = a_{ji}^{(i)}/a_{ii}^{(i)}$

Exemple 5.3 *Résoudre le système suivant en appliquant la factorisation de Doolittle*

$$\begin{cases} x + 2y + 3z = 0 \\ 2x + 2y + 3z = 1 \\ 3x + 3y + 3z = 0 \end{cases}$$

$$\begin{pmatrix} 1 & 2 & 3 \\ 2 & 2 & 3 \\ 3 & 3 & 3 \end{pmatrix} \overset{m_{21}=2, m_{31}=3}{\rightarrow} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -2 & -3 \\ 0 & -3 & -6 \end{pmatrix} \overset{m_{32}=3/2}{\rightarrow} \begin{pmatrix} 1 & 2 & 3 \\ 0 & -2 & -3 \\ 0 & 0 & -3/2 \end{pmatrix}$$

Ainsi

$$\mathbf{U} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -2 & -3 \\ 0 & 0 & -3/2 \end{pmatrix} \text{ et } \mathbf{L} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 3/2 & 1 \end{pmatrix}$$

On résout au début $\mathbf{Ly} = \mathbf{b} = (0, 1, 0)^T$ par substitution avant on aura $\mathbf{y} = (0, 1, -3/2)^T$ Puis on résout $\mathbf{Ux} = \mathbf{y} = (0, 1, -3/2)^T$ par substitution arrière on aura $\mathbf{x} = (1, -2, 1)^T$ Le nombre d'opération est résumé dans la table suivante :

	Addition/Soustraction	*Multiplication*	*Division*
Factorisation	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)(2n-1)}{6}$	$\frac{n(n-1)}{2}$
Substitution avant	$\frac{n(n-1)}{2}$	$\frac{n(n-1)}{2}$	0
Substitution arrière	$\frac{n(n-1)}{2}$	$\frac{n(n-1)}{2}$	n
Total	$\frac{n(n-1)(2n+5)}{6}$	$\frac{n(n-1)(2n+5)}{6}$	$\frac{n(n+1)}{2}$

On voit que le nombre total d'opérations est identique à celui de l'élimination de Gauss. Pour plusieurs côté droit, la substitution avant et arrière doit être répétée le nombre approprié de fois.

5.3.2 Factorisation de Crout

Une approche alternative à celle de Doolittle implique une matrice $\mathbf{U}$ avec 1 sur la diagonale. C'est ce qu'on appelle la décomposition de Crout. Bien qu'il existe certaines différences entre les deux approches, leurs performances sont comparables.L'approche de décomposition de Crout génère $\mathbf{U}$ et $\mathbf{L}$ en balayant la matrice par colonnes et lignes. Il peut être implémenté par les formules suivantes :

$$\begin{aligned} l_{i1} &= a_{i1} \qquad i = 1, 2, ..., n \\ u_{1j} &= \frac{a_{1j}}{l_{11}} \qquad j = 2, 3, ..., n \end{aligned} \tag{5.16}$$

Pour $j = 2, 3, ..., n-1$

$$\begin{aligned} l_{ij} &= a_{ij} - \sum_{k=1}^{j-1} l_{ik} u_{kj} \qquad i = j, j+1, ..., n \\ u_{ij} &= \frac{a_{jk} - \sum_{i=1}^{j-1} l_{ji} u_{ik}}{l_{jj}} \qquad k = j+1, j+2, ..., n \\ l_{nn} &= a_{nn} - \sum_{k=1}^{n-1} l_{nk} u_{kn} \end{aligned} \tag{5.17}$$

Exemple 5.4 *Utilisez la méthode de Crout pour résoudre le système :*

$$\begin{cases} x_1 + x_2 + x_3 + x_4 = 10 \\ 2x_1 + 3x_2 + x_3 + 5x_4 = 31 \\ -x_1 + x_2 - 5x_3 + 3x_4 = -2 \\ 3x_1 + x_2 + 7x_3 - 2x_4 = 18 \end{cases} \tag{5.18}$$

On a

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 3 & 1 & 5 \\ -1 & 1 & -5 & 3 \\ 3 & 1 & 7 & -2 \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} & u_{14} \\ 0 & 1 & u_{23} & u_{24} \\ 0 & 0 & 1 & u_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

— La multiplication de la première ligne de $\mathbf{L}$ par les colonnes de $\mathbf{U}$ donne

$$l_{11} = 1, \quad u_{12} = 1, \quad u_{13} = 1, \quad u_{14} = 1$$

— La multiplication de la deuxième ligne de $\mathbf{L}$ par les colonnes de $\mathbf{U}$ donne

$$l_{21} = 2, \quad l_{22} = 1, \quad u_{23} = -1, \quad u_{24} = 3$$

— La multiplication de la troisième ligne de $\mathbf{L}$ par les colonnes de $\mathbf{U}$ donne

$$l_{31} = -1, \quad l_{32} = 2, \quad l_{33} = -2, \quad u_{34} = 1$$

— La multiplication de la quatrième ligne de $\mathbf{L}$ par les colonnes de $\mathbf{U}$ donne

$$l_{41} = 3, \quad l_{42} = -2, \quad l_{43} = 2, \quad l_{44} = -1$$

d'ou

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 2 & -2 & 0 \\ 3 & -2 & 2 & -1 \end{pmatrix}, \qquad \mathbf{U} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & -1 & 3 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

L'application de la substitution avant à $\mathbf{Ly} = \mathbf{b}$ donne

$$y_1 = 10, \quad y_2 = 11, \quad y_3 = 7, \quad y_4 = 4$$

Et l'application de la substitution arriere à $\mathbf{Ux} = \mathbf{y}$ donne

$$x_1 = 1, \quad x_2 = 2, \quad x_3 = 3, \quad x_4 = 4$$

5.3.3 Factorisation de Cholesky

Il existe un certain nombre de classes de matrices dont les propriétés facilitent la solution des systèmes linéaires. Une classe importante de ce type se compose de matrices symétriques définies positives .

Définition 5.1 *Une $n \times n$ matrice $\mathbf{A}$ est dite symétrique définie positive (SDP) si $\mathbf{A}^T = \mathbf{A}$ et $\mathbf{x}^T\mathbf{A}\mathbf{x} > 0$ pour tous vecteurs $\mathbf{x} \neq 0$.*

Exemple 5.5 *Montrez que la matrice $\mathbf{A} = \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix}$ est symétrique définie positive.*

On voit bien que $\mathbf{A}$ est symétrique car $\mathbf{A} = \mathbf{A}^T$ .Pour montrer que $\mathbf{A}$ est définie positive, on applique la définition.

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 2 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 2x_1^2 + 4x_1x_2 + 5x_2^2 \\ &= 2(x_1 + x_2)^2 + 3x_2^2 > 0 \text{ si } \mathbf{x} \neq 0 \end{aligned}$$

Exemple 5.6 *Montrez que la matrice symétrique $\mathbf{A} = \begin{pmatrix} 2 & 4 \\ 4 & 5 \end{pmatrix}$ n'est pas définie positive.*

Calculant :

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} &= \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 2 & 4 \\ 4 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \\ &= 2x_1^2 + 8x_1x_2 + 5x_2^2 \\ &= 2(x_1 + 2x_2)^2 - 3x_2^2 \end{aligned}$$

Par exemple mettant $x_1 = -2$ et $x_2 = 1$ on aura $\mathbf{x}^T\mathbf{A}\mathbf{x} < 0$ donc $\mathbf{A}$ n'est pas définie positive.

Exemple 5.7 *Montrer qu'une matrice avec un ou plusieurs éléments diagonaux non positifs ne peut pas être une matrice SDP.*

Soit a_{ii} un élément diagonal négatif, choisissant un vecteur $\mathbf{x} = \mathbf{e_i}$ ($\mathbf{e_i}$ est un vecteur avec tous les éléments nuls à l'exception du $i^{ième}$ élément qui est 1) donc on voit bien que $\mathbf{x}\mathbf{A}\mathbf{x}^\mathbf{T} = a_{ii} < 0$ alors la matrice $\mathbf{A}$ n'est pas définie positive.

Exemple 5.8 *Pour une matrice SDP 2×2 Montrez que $a_{22} - a_{21}^2 > 0$*

Prenant un vecteur $\mathbf{x} = (a_{21} - a_{11})^T$. Alors $\mathbf{x}\mathbf{A}\mathbf{x}^\mathbf{T} = a_{11}^2a_{22} - a_{11}a_{21}^2 = a_{11}^2(a_{22} - a_{21}^2/a_{11}) > 0$ donc $a_{22} - a_{21}^2/a_{11} > 0$

Les matrices SDP apparaissent souvent dans les applications impliquant l'approximation des fonctions et la solution d'équations différentielles. Ce qui les rend important, c'est qu'il existe pour eux un algorithme spécial de factorisation et de solution, qui est environ deux fois moins coûteux que l'élimination de Gauss .Le résultat est dû au cartographe Cholesky.

Théoreme 5.2 (Théorème de la factorisation de Cholesky) *Si $\mathbf{A}$ est une matrice symétrique définie positive,il existe une et une seule matrice triangulaire inférieure à valeurs diagonales positives notée $\mathbf{L}$ telle que $\mathbf{A} = \mathbf{L}\mathbf{L}^T$.*

Pour la construction de $\mathbf{L}$,commençant par une matrice 2×2 , $\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ La matrice $\mathbf{A}$ est SDP alors $a_{ii} > 0$, il peut facilement être vérifié (par calcul direct) que

$$\mathbf{L} = \begin{pmatrix} \sqrt{a_{11}} & 0 \\ a_{21}/\sqrt{a_{11}} & \sqrt{a_{22} - a_{21}^2/a_{11}} \end{pmatrix}$$

$\mathbf{A} = \mathbf{L}\mathbf{L}^T$. Par conséquent, $\mathbf{L}$ existe, au moins pour le cas spécial de $n = 2$. Supposons maintenant que $\mathbf{L_{11}}$ existe pour $n = k - 1$ ($\mathbf{A_{11}} = \mathbf{L_{11}}\mathbf{L_{11}^T}$) ; et montrons par induction que $\mathbf{L}$ doit exister pour $n = k$. On écrit la matrice $\mathbf{A} \in \mathbb{R}^{k \times k}$ sous forme partitionnée

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{a} \\ \mathbf{a}^T & a_{kk} \end{pmatrix}$$

Avec $\mathbf{A}_{11} \in \mathbb{R}^{(k-1)\times(k-1)}$, $\mathbf{a} \in \mathbb{R}^{k-1}$ et $a_{kk} \in \mathbb{R}$. Construisons

$$\mathbf{L} = \begin{pmatrix} \mathbf{L}_{11} & 0 \\ \mathbf{l}^T & l_{kk} \end{pmatrix}$$

Avec $\mathbf{l}$ un vecteur $(k-1) \times 1$ et l_{kk} un scalaire à déterminer par identification

$$\mathbf{L}\mathbf{L}^\mathbf{T} = \begin{pmatrix} \mathbf{L_{11}}\mathbf{L_{11}^T} & \mathbf{L_{11}}\mathbf{l} \\ \mathbf{l}^T{\mathbf{L_{11}}}^T & \mathbf{l^T}\mathbf{l} + l_{kk}^2 \end{pmatrix} = \mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{a} \\ \mathbf{a}^T & a_{kk} \end{pmatrix}$$

Déjà on sait par hypothèse que ($\mathbf{A_{11}}$=$\mathbf{L_{11}}\mathbf{L_{11}^T}$) , il reste à déterminer :

$$\mathbf{L_{11}}\mathbf{l} = \mathbf{a} \tag{5.19}$$

$$\mathbf{l^T}\mathbf{l} + l_{kk}^2 = a_{kk} \tag{5.20}$$

Puisque les éléments diagonaux de $\mathbf{L_{11}}$ sont tous positifs. le système (5.19) a une solution unique $\mathbf{l}$. Ainsi l_{kk} est calculé par l'équation (5.20)

$$l_{kk} = \sqrt{a_{kk} - \mathbf{l}^T\mathbf{l}}$$

Pour que la décomposition de Cholesky soit un outil utile, il faut trouver une méthode pratique pour la calculer. L'une des façons les plus est d'écrire la décomposition $\mathbf{A} = \mathbf{L}\mathbf{L}^\mathbf{T}$ en détail et de l'étudier :

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 & \dots & 0 \\ l_{21} & l_{22} & 0 & \dots & 0 \\ l_{31} & l_{32} & l_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} & \dots & l_{n1} \\ 0 & l_{22} & l_{32} & \dots & l_{n2} \\ 0 & 0 & l_{33} & \dots & l_{n3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & l_{nn} \end{pmatrix}$$

L'élément a_{ij} est le produit de la $i^{ème}$ ligne de $\mathbf{L}$ avec la $j^{ème}$ colonne de $\mathbf{L^T}$

$$a_{i1} = l_{11}l_{i1}$$

pour $i = 1$ on a $a_{11} = l_{11}^2$ qui donne (pour matrice SDP les éléments de la diagonale sont positifs)

$$l_{11} = \sqrt{a_{11}} \tag{5.21}$$

Une fois l_{11} calculé ,on termine le calcul des éléments de la première colonne $\mathbf{L}$ qui sont les éléments de première ligne de $\mathbf{L}^T$.

$$l_{i1} = a_{i1}/l_{11}, \qquad i = 2, 3, ..., n \tag{5.22}$$

On passe maintenant à la deuxième ligne de $\mathbf{L}^T$. Puisque la deuxième ligne de $\mathbf{L}$ a seulement deux éléments non nuls

$$a_{2j} = l_{21}l_{j1} + l_{22}l_{j2}, \qquad j = 2, 3, ..., n \tag{5.23}$$

En particulier,lorsque $j = 2$ nous avons $a_{22} = l_{21}^2 + l_{22}^2$. Puisque l_{12} est déjà connu, ainsi on calcule l_{22} :

$$l_{22} = \sqrt{a_{22} - l_{21}^2} \tag{5.24}$$

Une fois l_{22} connu, la seule inconnue dans (5.23) est l_{j2}. Ainsi, nous pouvons utiliser (5.23) pour calculer le reste de la deuxième ligne de $\mathbf{L}^T$.

$$l_{j2} = (a_{2j} - l_{21}l_{j1}/l_{22}), \qquad j = 3, ..., n$$

A cette étape deux colonnes de $\mathbf{L}$ qui sont les deux lignes de $\mathbf{L}^T$ sont connus. Voyons maintenant comment calculer la $i^{ème}$ ligne de $\mathbf{L}^T$, en supposant que les $(i-1)$ premières lignes sont connus. Étant donné que seules les i premières éléments de la $i^{ème}$ ligne de $\mathbf{L}$, sont différents de zéro,

$$a_{ij} = l_{i1}l_{j1} + l_{i2}l_{j2} + ... + l_{ii}l_{ji} \tag{5.25}$$

Tous les éléments de $\mathbf{L}^T$ apparaissant dans (5.25) se trouvent dans les i premières lignes. Puisque les $i-1$ premières lignes sont connus, les seules inconnues dans (5.25) sont l_{ii} et l_{ij}. Prenant $i = j$ dans (5.25),on a

$$a_{ii} = l_{i1}^2 + l_{i2}^2 + ... + l_{ii}^2$$

qui est résolu pour l_{ii}

$$l_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2} \tag{5.26}$$

Une fois l_{ii} calculé, on peut utiliser (5.25) pour calculer les l_{ji}

$$l_{ji} = \left(a_{ij} - \sum_{k=1}^{i-1} l_{ik}l_{jk} \right) / l_{ii}, \quad j = i + 1, ..., n \tag{5.27}$$

L'évaluation de l_{ii} dans (5.26) nécessite $i-1$ multiplications, $i-1$ soustractions et une opération de racine carrée, un total de $2i-1$ opérations. Le calcul de chaque l_{ji} à partir de (5.27) nécessite également $2i-1$ opérations. Le nombre total d'opérations nécessaires à la construction de $\mathbf{L}$ est donc :

$$\sum_{i=1}^{n} \sum_{j=i}^{n} (2i - 1) = \sum_{i=1}^{n} (2i - 1)(1 + n - i) = \frac{1}{6} n(n + 1)(2n + 1) \tag{5.28}$$

Pour n large , le nombre d'opérations requises est d'environ $n^3/3$, ce qui est la moitié du nombre indiqué pour l'élimination de Gauss et la factorisation LU.

Exemple 5.9 *Résoudre par la méthode de Cholesky le système* $\mathbf{Ax} = \mathbf{b}$ *avec :*

$$\mathbf{A} = \begin{pmatrix} 4 & -2 & 4 & 2 \\ -2 & 10 & -2 & 7 \\ 4 & -2 & 8 & 4 \\ 2 & -7 & 4 & 7 \end{pmatrix}, \qquad \mathbf{b} = \begin{pmatrix} 8 \\ 2 \\ 16 \\ 6 \end{pmatrix}$$

$$l_{11} = \sqrt{4} = 2$$
$$l_{21} = -2/2 = -1$$
$$l_{31} = 4/2 = 2$$
$$l_{41} = 2/2 = 1$$
$$l_{22} = \sqrt{10-1} = 3$$
$$l_{32} = (-2-(-1)(2))/3 = 0$$
$$l_{42} = (-7-(-1)(1))/3 = -2$$
$$l_{33} = \sqrt{8-(2^2+0^2)} = 2$$
$$l_{43} = (4-((2)(1)+0(-2)))/2 = 1$$
$$l_{44} = \sqrt{7-1^2-(-2)^2-1^2} = 1$$

Alors

$$\mathbf{L} = \begin{pmatrix} 2 & 0 & 0 & 0 \\ -1 & 3 & 0 & 0 \\ 2 & 0 & 2 & 0 \\ 1 & -2 & 1 & 1 \end{pmatrix}, \qquad \mathbf{L^T} = \begin{pmatrix} 2 & -1 & 2 & 1 \\ 0 & 3 & 0 & -2 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

On résout $\mathbf{Ly} = \mathbf{b}$ par substitution avant ,puis $\mathbf{L^T x} = \mathbf{y}$ par substitution arrière on aura

$$\mathbf{y} = [4\ 2\ 4\ 2]^T \text{ puis } \mathbf{x} = [1\ 2\ 1\ 2]^T$$

5.3.4 Algorithme de Thomas pour les systèmes tridiagonaux

Dans de nombreuses applications, on rencontre des systèmes linéaires extrêmement grands qui ont une structure en bandes. Les matrices bandes se produisent souvent dans la résolution des équations différentielles ordinaires et partielles. Il est avantageux de développer des algorithmes spéciaux conçus pour de tels systèmes linéaires, car ils réduisent la quantité de stockage mémoire utilisé.

Définition 5.2 *Soit une matrice* $\mathbf{T} \in \mathbb{R}^{n\times n}$ *avec* $n \geq 3$. $\mathbf{T}$ *est dite tridiagonale si elle a des éléments non nuls uniquement sur la diagonale principale et les deux diagonales adjacentes ; c'est à dire.,*

$$t_{ij} = 0 \quad si \quad |i-j| > 1, \qquad i,j \in \{1,2,...,n\}$$

Il est facile de voir que dans le processus de factorisation **LU** d'une matrice tridiagonale $\mathbf{T}$, sans échange de lignes, la matrice triangulaire inférieure $\mathbf{L} \in \mathbb{R}^{n\times n}$ de diagonal 1 et la matrice triangulaire supérieure $\mathbf{U} \in \mathbb{R}^{n\times n}$ n'ont chacune que deux éléments dans chaque ligne. nous écrivons $\mathbf{T}$ sous la forme

$$\mathbf{T} = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & & \\ & a_3 & b_3 & c_3 & & \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ & & & & a_n & b_n \end{pmatrix} \tag{5.29}$$

la factorisation peut s'écrire $\mathbf{T} = \mathbf{LU}$ où

$$\mathbf{L} = \begin{pmatrix} 1 & & & & & \\ l_2 & 1 & & & & \\ & l_3 & 1 & & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & l_n & 1 \end{pmatrix} \tag{5.30}$$

et

$$\mathbf{U} = \begin{pmatrix} u_1 & v_1 & & & & \\ & u_2 & v_2 & & & \\ & & u_3 & v_3 & & \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots \\ & & & & & u_n \end{pmatrix} \tag{5.31}$$

avec la convention que les éléments manquants dans ces matrices sont tous égaux à zéro.Il est souvent pratique de définir $a_1 = 0$ et $c_n = 0$. La multiplication de $\mathbf{L}$ par $\mathbf{U}$ montre que $v_j = c_j$, et que les éléments l_j et u_j peuvent être calculés à partir de

$$l_j = a_j/u_{j-1}, \quad u_j = b_j - l_j c_{j-1}, \qquad j = 2, 3, \dots, n \tag{5.32}$$

en commençant par $u_1 = b_1$.
Supposons que notre objectif soit de résoudre le système d'équations linéaires $\mathbf{Tx} = \mathbf{r}$, où la matrice $\mathbf{T} \in \mathbb{R}^{n\times n}$ est tridiagonale et non singulière, et $\mathbf{r} \in \mathbb{R}^n$. Après avoir calculé les éléments des matrices $\mathbf{L}$ et $\mathbf{U}$ à l'aide de (5.32), les substitution avant et arrière sont alors très simples. Soit $\mathbf{y} = \mathbf{Ux}$, l'équation $\mathbf{Ly} = \mathbf{r}$ donne

$$y_1 = r_1 \tag{5.33}$$
$$y_j = r_j - l_j y_{j-1}, \qquad j = 2, 3, \dots, n \tag{5.34}$$

et finalement

$$x_n = y_n/u_n \tag{5.35}$$
$$x_j = (y_j - v_j x_{j+1})/u_j, \qquad j = n-1, n-2, \dots, 1 \tag{5.36}$$

La factorisation $\mathbf{LU}$ d'une matrice tridiagonale nécessite environ $3n$ opérations. Les substitution avant et arrière impliquent ensemble environ $5n$ opérations. Ainsi, l'ensemble du processus de solution nécessite environ $8n$ opérations. La quantité totale de travail est donc beaucoup moins que pour une matrice pleine dont le nombre d'opération était $2n^3/3$. La méthode que nous venons de décrire est une variante de l'algorithme de Thomas

Exemple 5.10 *Résoudre le système tridiagonal* $\mathbf{Tx} = \mathbf{r}$ *suivant*

$$\begin{pmatrix} 1 & 1 & 0 & 0 \\ 3 & 2 & 1 & 0 \\ 0 & 2 & 3 & 1 \\ 0 & 0 & -2 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 1 \\ 5 \\ 2 \\ -5 \end{pmatrix}$$

La factorisation $\mathbf{T} = \mathbf{LU}$

$$\mathbf{L} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 3 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 0 & -0.4 & 1 \end{pmatrix},$$

$$\mathbf{U} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & 5 & 1 \\ 0 & 0 & 0 & -2.6 \end{pmatrix}$$

Une fois la factorisation $\mathbf{T} = \mathbf{LU}$ réalisée ,on résout $\mathbf{Ly} = \mathbf{r}$ par substitution avant qui donne $\mathbf{y} = [1\ 2\ 6\ \ -2]^T$. Puis on résout $\mathbf{Ux} = \mathbf{y}$ par substitution arrière qui donne $\mathbf{x} = [2\ \ -1\ 1\ 1]^T$.

Chapitre 6

Méthodes Itératives pour systèmes linéaires

6.1 Introduction

Dans ce chapitre, nous revenons au problème de la résolution d'un système linéaire $\mathbf{Ax} = \mathbf{b}$ avec $\mathbf{A} \in \mathbb{R}^{n\times n}$. Ce problème peut être résolu sans difficulté par l'élimination de Gauss sur les ordinateurs d'aujourd'hui. Cependant, une fois n devient très grand (par exemple plusieurs milliers) et la matrice $\mathbf{A}$ devient très clairsemée(par exemple plus de 90% de ces éléments sont nuls),les méthodes itératives deviennent plus efficaces. Les méthodes itératives nécessitent une estimation initiale $\mathbf{x^{(0)}} \in \mathbb{R}^n$ proche de la vraie solution. Une fois $\mathbf{x^{(0)}}$ choisi, nous l'utilisons pour générer une nouvelle estimation $\mathbf{x^{(1)}}$ qui est ensuite utilisée pour générer une autre estimation $\mathbf{x^{(2)}}$ et ainsi de suite. De cette manière, nous générons une séquence d'itérations $\mathbf{x^{(k)}}$ si elle converge,elle convergera vers la vraie solution $\mathbf{x}$.

6.2 Méthodes de Jacobi et Gauss-Seidel

Le schéma itératif le plus simple est peut-être l'itération de Jacobi. Il est défini pour les matrices qui ont des éléments diagonaux non nuls. La méthode peut être motivée en réécrivant le système 3 par 3 suivant

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_2 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{cases}$$

sous la forme

$$\begin{cases} x_1 &= (b_1 - a_{12}x_2 - a_{13}x_3)/a_{11} \\ x_2 &= (b_2 - a_{11}x_1 - a_{13}x_3)/a_{22} \\ x_3 &= (b_3 - a_{11}x_2 - a_{12}x_3)/a_{33} \end{cases} \tag{6.1}$$

Supposons que $\mathbf{x^{(k)}}$ est une approximation de la solution exacte $\mathbf{x}$. Une façon naturelle de générer une nouvelle approximation $\mathbf{x^{(k+1)}}$ est de calculer

$$\begin{cases} x_1^{(k+1)} &= (b_1 - a_{12}x_2^{(k)} - a_{13}x_3^{(k)})/a_{11} \\ x_2^{(k+1)} &= (b_2 - a_{11}x_1^{(k)} - a_{13}x_3^{(k)})/a_{22} \\ x_3^{(k+1)} &= (b_3 - a_{11}x_2^{(k)} - a_{12}x_3^{(k)})/a_{33} \end{cases}$$

Ceci définit l'itération de Jacobi pour le cas $n = 3$. Pour le cas général,nous écrivons

$$x_i^{(k+1)} = \left(b_i - \sum_{\substack{i=0 \\ i \neq k}}^{n} a_{ij} x_j^{(k)} \right) / a_{ii} \qquad i = 1, 2, \ldots, n \tag{6.2}$$

On choisi une valeur initiale $\mathbf{x^{(0)}}$ et on exécute l'équation (6.2) pour $k = 0, 1, 2, \ldots$ jusqu'à convergence.

Exemple 6.1 *Résoudre le système suivant par la méthode de Jacobi en prenant* $\mathbf{x^{(0)}} = [0\ 0\ 0\ 0]^T$.

$$\begin{pmatrix} 5 & 2 & 1 & 1 \\ 2 & 6 & 2 & 1 \\ 1 & 2 & 7 & 2 \\ 1 & 1 & 2 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 29 \\ 31 \\ 26 \\ 19 \end{pmatrix}$$

$$\mathbf{x^{(0)}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{x^{(1)}} = \begin{pmatrix} 2.5155 \\ 1.5994 \\ 0.7310 \\ 0.0756 \end{pmatrix}, \quad \ldots, \mathbf{x^{(10)}} = \begin{pmatrix} 3.9017 \\ 2.8994 \\ 1.9144 \\ 0.9354 \end{pmatrix}, \ldots, \mathbf{x^{(20)}} = \begin{pmatrix} 3.9965 \\ 2.9964 \\ 1.9970 \\ 0.9977 \end{pmatrix}$$

$\mathbf{x^{(20)}}$ est une bonne approximation a la solution exacte $\mathbf{x} = [4\ 3\ 2\ 1]^T$. Notez que dans l'itération de Jacobi, on n'utilise pas les dernières informations disponibles lors du calcul de $\mathbf{x^{(k+1)}}$. Par exemple, $x_1^{(k)}$ est utilisé dans le calcul de $x_2^{(k+1)}$ pourtant la composante $x_1^{(k+1)}$ est connue. Si nous révisons l'itération de Jacobi de manière à toujours utiliser les estimations la plus récentes , nous obtenons la méthode de Gauss-Seidel. Pour le systeme (6.1) l'itération de Gauss-Seidel s'écrit

$$\begin{cases} x_1^{(k+1)} &= (b_1 - a_{12} x_2^{(k)} - a_{13} x_3^{(k)}) / a_{11} \\ x_2^{(k+1)} &= (b_2 - a_{11} x_1^{(k+1)} - a_{13} x_3^{(k)}) / a_{22} \\ x_3^{(k+1)} &= (b_3 - a_{11} x_2^{(k+1)} - a_{12} x_3^{(k+1)}) / a_{33} \end{cases} \tag{6.3}$$

Et pour le cas général

$$x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right) / a_{ii} \qquad i = 1, 2, \ldots, n \tag{6.4}$$

On choisi une valeur initiale $\mathbf{x^{(0)}}$ et on exécute l'équation (6.4) pour $k = 0, 1, 2, \ldots$ jusqu'à convergence.

Exemple 6.2 *Refaire l'exemple 6.1 par la méthode de Gauss-Seidel*

$$\mathbf{x^{(0)}} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \mathbf{x^{(1)}} = \begin{pmatrix} 5.8000 \\ 3.2333 \\ 1.9619 \\ 0.7554 \end{pmatrix}, \quad \ldots, \mathbf{x^{(4)}} = \begin{pmatrix} 3.9992 \\ 2.9962 \\ 2.0006 \\ 1.0004 \end{pmatrix}, \ldots, \mathbf{x^{(6)}} = \begin{pmatrix} 4.0003 \\ 2.9999 \\ 2.0000 \\ 1.0000 \end{pmatrix}$$

Pour les itérations de Jacobi et de Gauss-Seidel, la transition de $x_1^{(k)}$ à $x_1^{(k+1)}$ peut être décrite succinctement en termes de matrices $\mathbf{L}$,$\mathbf{D}$ et $\mathbf{U}$ définies par :

$$-\mathbf{L} = \begin{pmatrix} 0 & 0 & \dots & \dots & 0 \\ a_{21} & 0 & \dots & \dots & 0 \\ a_{31} & a_{32} & \ddots & & \vdots \\ \vdots & \vdots & & 0 & 0 \\ a_{n1} & a_{n2} & \dots & a_{n,n-1} & a_{n1} \end{pmatrix} \tag{6.5}$$

$$\mathbf{D} = diag(a_{11}, a_{22}, ..., a_{nn}) \tag{6.6}$$

$$-\mathbf{U} = \begin{pmatrix} 0 & a_{12} & \dots & \dots & a_{1n} \\ 0 & 0 & \dots & \dots & a_{2n} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & a_{n-1,n} \\ 0 & 0 & \dots & 0 & 0 \end{pmatrix} \tag{6.7}$$

Ainsi,après partition de la matrice $\mathbf{A} = \mathbf{D} - \mathbf{L} - \mathbf{U}$,l'itération de Jordan s'écrit :

$$\begin{aligned} \mathbf{D}\mathbf{x}^{(k+1)} &= (\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{b} \\ \mathbf{x}^{(k+1)} &= \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)} + \mathbf{D}^{-1}\mathbf{b} \qquad k = 0, 1, 2, \dots \end{aligned} \tag{6.8}$$

En revanche, celle de Gauss-Seidel s'écrit

$$\begin{aligned} (\mathbf{D} - \mathbf{L})\mathbf{x}^{(k+1)} &= \mathbf{U}\mathbf{x}^{(k)} + \mathbf{b} \\ \mathbf{x}^{(k+1)} &= (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}\mathbf{x}^{(k)} + (\mathbf{D} - \mathbf{L})^{-1}\mathbf{b} \qquad k = 0, 1, 2, \dots \end{aligned} \tag{6.9}$$

6.3 Étude de la convergence

Pour étudier la convergence des techniques d'itération, nous avons besoin d'analyser la formule

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)} + \mathbf{c} \qquad k = 0, 1, 2, \dots \tag{6.10}$$

des équations (6.10),(6.8) et(6.9) on a $\mathbf{T} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ pour la méthode de Jacobi et $\mathbf{T} = (\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$ pour la méthode de Gauss-Seidel

Définition 6.1 *Soit* $\mathbf{M} \in \mathbb{R}^{n\times n}$ *.Le rayon spectral de* $\mathbf{M}$*, noté* $\rho(\mathbf{M})$*, est le plus grand (en magnitude) de toutes les valeurs propres de* $\rho(\mathbf{M})$*.* $\rho(\mathbf{M}) = max|\lambda|$ *où le maximum est pris sur toutes les valeurs propres de* $\mathbf{M}$

Lemme 6.1 *Si le rayon spectral satisfait* $\rho(\mathbf{T}) < 1$*,alors* $(\mathbf{I} - \mathbf{T})^{-1}$ *existe et*

$$(\mathbf{I} - \mathbf{T})^{-1} = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots = \sum_{i=0}^{\infty} \mathbf{T}^i$$

Démonstration :Parce que $\mathbf{T}\mathbf{x} = \lambda\mathbf{x}$ est vrai précisément lorsque $(\mathbf{I} - \mathbf{T})\mathbf{x} = (1 - \lambda)\mathbf{x}$, nous avons λ comme valeur propre de $\mathbf{T}$ précisément lorsque $(1 - \lambda)$ est valeur propre de $(\mathbf{I} - \mathbf{T})$. Mais $\lambda \leq \rho(\mathbf{T}) < 1$, donc $\lambda = 1$ n'est pas une valeur propre de $\mathbf{T}$, et 0 ne peut pas être une valeur propre de $(\mathbf{I} - \mathbf{T})$. Par conséquent, $(\mathbf{I} - \mathbf{T})^{-1}$ existe. Soit

$$\mathbf{S}_m = \mathbf{I} + \mathbf{T} + \mathbf{T}^2 + \cdots + \mathbf{T}^m$$

Ainsi

$$\begin{aligned}(\mathbf{I}-\mathbf{T})\mathbf{S}_m &= (\mathbf{I}+\mathbf{T}+\mathbf{T}^2+\dots\mathbf{T}^m)-(\mathbf{T}+\mathbf{T}^2+\cdots+\mathbf{T}^{m+1})\\ &= \mathbf{I}-\mathbf{T}^{m+1}\end{aligned} \tag{6.11}$$

et, comme la matrice $\mathbf{T}$ est convergente, le théorème $(A.4)$ implique que

$$\lim_{m\to\infty}(\mathbf{I}-\mathbf{T})\mathbf{S}_m = \lim_{m\to\infty}(\mathbf{I}-\mathbf{T}^{m+1}) = \mathbf{I}$$

ainsi

$$(\mathbf{I}-\mathbf{T})^{-1} = \lim_{m\to\infty}\mathbf{S}_m = \sum_{i=0}^{\infty}\mathbf{T}^i$$

Théoreme 6.1 *Pour* $\mathbf{x}^{(0)}\in\mathbb{R}$*, la suite* $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ *définie par*

$$\mathbf{x}^{(k+1)} = \mathbf{T}\mathbf{x}^{(k)}+\mathbf{c} \qquad k=0,1,2,\dots \tag{6.12}$$

converge vers la solution unique $\mathbf{x}=\mathbf{T}\mathbf{x}+\mathbf{c}$ *si est seulement si* $\rho(\mathbf{T})<1$ *,alors*

Démonstration :Supposons que $\rho(\mathbf{T})<1$.

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \mathbf{T}\mathbf{x}^{(k)}+\mathbf{c}\\ &= \mathbf{T}(\mathbf{T}\mathbf{x}^{(k-1)}+\mathbf{c})+\mathbf{c}\\ &= \mathbf{T}^{(2)}\mathbf{x}^{(k-1)}+(\mathbf{T}+\mathbf{I})\mathbf{c}\\ &\vdots\\ &= \mathbf{T}^k\mathbf{x}^{(0)}+(\mathbf{I}+\mathbf{T}+\mathbf{T}^2+\dots\mathbf{T}^k)\mathbf{c}\end{aligned} \tag{6.13}$$

Puisque $\rho(\mathbf{T})<1$ le théorème $(A.4)$ implique que $\mathbf{T}$ est convergente et

$$\lim_{k\to\infty}\mathbf{T}^k\mathbf{x}^{(0)} = 0$$

le lemme () implique

$$\lim_{k\to\infty}\mathbf{x}^{(k+1)} = \lim_{k\to\infty}\mathbf{T}^k\mathbf{x}^{(0)}+(\sum_{i=0}^{\infty}\mathbf{T}^i)\mathbf{c} = 0+(\mathbf{I}-\mathbf{T})^{-1}\mathbf{c} = (\mathbf{I}-\mathbf{T})^{-1}\mathbf{c}$$

Ainsi $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ converge vers le vecteur $\mathbf{x}=(\mathbf{I}-\mathbf{T})^{-1}\mathbf{c}$ et $\mathbf{x}=\mathbf{T}\mathbf{x}+\mathbf{c}$
Pour prouver l'inverse, nous montrerons que pour tout $\mathbf{z}\in\mathbb{R}^n$, on a $\lim_{k\to\infty}\mathbf{T}^k\mathbf{z}=0$ (par le théorème $(A.4)$ ceci est équivalent à $\rho(\mathbf{T})<1$).
Soit $\mathbf{z}$ un vecteur quelconque et $\mathbf{x}$ la solution unique à $\mathbf{x}=\mathbf{T}\mathbf{x}+\mathbf{c}$.
Définissant $\mathbf{x}^{(0)}=\mathbf{x}-\mathbf{z}$ et pour $k\geq 0\quad \mathbf{x}^{(k+1)}=\mathbf{T}\mathbf{x}^{(k)}+\mathbf{c}$. Ainsi $\left\{\mathbf{x}^{(k)}\right\}_{k=0}^{\infty}$ converge vers $\mathbf{x}$

$$\begin{aligned}\mathbf{x}-\mathbf{x}^{(k+1)} = (\mathbf{T}\mathbf{x}+\mathbf{c})-(\mathbf{T}\mathbf{x}^{(k)}+\mathbf{c}) &= \mathbf{T}(\mathbf{x}-\mathbf{x}^{(k)})\\ &= \mathbf{T}^2(\mathbf{x}-\mathbf{x}^{(k-1)})\\ &= \vdots\\ &= \mathbf{T}^{k+1}(\mathbf{x}-\mathbf{x}^{(0)})\\ &= \mathbf{T}^{k+1}\mathbf{z}\end{aligned} \tag{6.14}$$

d'ou

$$\lim_{k\to\infty}\mathbf{T}^{k+1}\mathbf{z} = \lim_{k\to\infty}\mathbf{T}^{k+1}(\mathbf{x}-\mathbf{x}^{(0)}) = \lim_{k\to\infty}(\mathbf{x}-\mathbf{x}^{(k+1)}) = 0 \tag{6.15}$$

Mais $\mathbf{z}\in\mathbb{R}^n$ est arbitraire ,ainsi par le théorème $(A.4)$ la matrice T est convergente et $\rho(\mathbf{T})<1$

Exemple 6.3 *Vérifier la convergence pour la méthode de Jacobi et Gauss-Seidel et faite* 20 *itérations pour chaque méthode*

$$\begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 4 \\ -5 \end{pmatrix}$$

Méthode de Jacobi : $\mathbf{T} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ avec

$$\mathbf{T} = \begin{pmatrix} 0 & 0.5 & -0.5 \\ -1 & 0 & -1 \\ 0.5 & 0.5 & 0 \end{pmatrix}, \qquad \det(\lambda\mathbf{I} - \mathbf{T}) = \lambda^3 + \frac{5}{4}\lambda$$

$$\det(\lambda\mathbf{I} - \mathbf{T}) = 0 \Leftrightarrow \lambda_1 = 0, \quad \lambda_2 = -\sqrt{5}i/2, \quad \lambda_3 = \sqrt{5}i/2$$

$\rho(\mathbf{T}) = \sqrt{5}/2 > 1 \Rightarrow$ l'itération de Jacobi diverge

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}^{(1)} = \begin{pmatrix} -0.50 \\ 2.0 \\ -2.50 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 1.75 \\ 5.00 \\ -1.75 \end{pmatrix},$$

$$\mathrm{x}^{(3)} = \begin{pmatrix} 2.875 \\ 2.000 \\ 0.875 \end{pmatrix}, \ldots, \mathbf{x}^{(20)} = \begin{pmatrix} -4.58793 \\ -20.35174 \\ 4.58793 \end{pmatrix}$$

Méthode de Gauss-Seidel : $\mathbf{T} = (\mathbf{D} - \mathbf{L})^{-1}\mathbf{U}$ avec

$$\mathbf{T} = \begin{pmatrix} 0 & 0.5 & -0.5 \\ 0 & -0.5 & -0.5 \\ 0 & 0 & -0.5 \end{pmatrix}, \qquad \det(\lambda\mathbf{I} - \mathbf{T}) = \lambda(\lambda + \frac{1}{2})^2$$

$$\det(\lambda\mathbf{I} - \mathbf{T}) = 0 \Leftrightarrow \lambda_1 = 0, \quad \lambda_{2,3} = \pm\sqrt{1/2},$$

$\rho(\mathbf{T}) = \sqrt{1/2} < 1 \Rightarrow$ l'itération de Gauss-Seidel converge

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}^{(1)} = \begin{pmatrix} 0.50 \\ 2.5 \\ -1.5 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} 1.50 \\ 2.00 \\ -0.75 \end{pmatrix},$$

$$\mathrm{x}^{(3)} = \begin{pmatrix} 0.875 \\ 1.875 \\ -1.125 \end{pmatrix}, \ldots, \mathbf{x}^{(20)} = \begin{pmatrix} 0.999984 \\ 2.000017 \\ -0.999999 \end{pmatrix}$$

La solution exacte est $\mathbf{x} = [1 \ 2 \ -1]$

Exemple 6.4 *Vérifier la convergence pour la méthode de Jacobi et Gauss-Seidel et faite quelques itérations pour chaque méthode*

$$\begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 7 \\ 2 \\ 5 \end{pmatrix}$$

Méthode de Jacobi : $\mathbf{T} = \mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})$ avec

$$\mathbf{T} = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}, \qquad \det(\lambda\mathbf{I} - \mathbf{T}) = \lambda^3 = 0,$$

$$\det(\lambda\mathbf{I}-\mathbf{T})=0\Leftrightarrow\lambda_{1,2,3}=0$$

$\rho(\mathbf{T})=0<1\Rightarrow$ l'itération de Jacobi converge

$$\mathbf{x}^{(0)}=\begin{pmatrix}0\\0\\0\end{pmatrix},\mathbf{x}^{(1)}=\begin{pmatrix}1\\2\\-1\end{pmatrix},\quad\mathbf{x}^{(2)}=\begin{pmatrix}1\\2\\-1\end{pmatrix},$$

Méthode de Gauss-Seidel : $\mathbf{T}=(\mathbf{D}-\mathbf{L})^{-1}\mathbf{U}$ avec

$$\mathbf{T}=\begin{pmatrix}0&-2&2\\0&2-3&\\0&0&2\end{pmatrix},\qquad\det(\lambda\mathbf{I}-\mathbf{T})=\lambda(\lambda-2)^2$$

$$\det(\lambda\mathbf{I}-\mathbf{T})=0\Leftrightarrow\lambda_1=0,\quad\lambda_{2,3}=2$$

$\rho(\mathbf{T})=2>1\Rightarrow$ l'itération de Gauss-Seidel diverge

$$\mathbf{x}^{(0)}=\begin{pmatrix}0\\0\\0\end{pmatrix},\mathbf{x}^{(1)}=\begin{pmatrix}7\\-5\\1\end{pmatrix},\quad\mathbf{x}^{(2)}=\begin{pmatrix}19\\-18\\3\end{pmatrix},$$

$$\mathrm{x}^{(3)}=\begin{pmatrix}49\\-50\\7\end{pmatrix},\ldots,\mathbf{x}^{(4)}=\begin{pmatrix}121\\-126\\15\end{pmatrix},\ldots..$$

La solution exacte est $\mathbf{x}=[1\ 2\ -1]$

Théoreme 6.2 — *Si $\mathbf{A}\in\mathbb{R}$ est a diagonale dominante , alors les méthodes de Jacobi et Gauss-Seidel convergent, et Gauss-Seidel converge plus rapidement si $\rho(\mathbf{T_G})<\rho(\mathbf{T_J})$ ou $\mathbf{T_J}$ est la matrice d'itération pour méthode de Jacobi, et $\mathbf{T_G}$ est la matrice d'itération pour Gauss-Seidel*

— *Si $\mathbf{A}$ est une matrice SDP, alors Jacobi et Gauss-Seidel convergeront.*

Exemple 6.5 *Pour l'exercice (6.1) vérifier que la matrice $\mathbf{A}$ est a diagonale dominante, comparer $\rho(\mathbf{T_J})$ et $\rho(\mathbf{T_G})$ .Conclusion*

$\rho(\mathbf{T_J})=0.7157$ et $\rho(\mathbf{T_G})=0.192208$

6.4 Méthode de relaxation

Une méthode de relaxation est une méthode itérative de résolution de systèmes linéaires basée sur l'itération de Gauss-Seidel et implique un paramètre $0\leq\omega\leq2$ dans le but d'accélérer le taux de convergence de la méthode itérative vers la solution du système linéaire. En multipliant le système linéaire $\mathbf{Ax}=\mathbf{b}$ par ω et en replaçant $\mathbf{A}$ par $\mathbf{D}-\mathbf{U}-\mathbf{L}$ donne

$$(\omega\mathbf{D}-\omega\mathbf{U}-\omega\mathbf{L})\mathbf{x}=\omega\mathbf{b}\Rightarrow((1-(1-\omega))\mathbf{D}-\omega\mathbf{U}-\omega\mathbf{L})\mathbf{x}$$

d'où

$$(\mathbf{D}-\omega\mathbf{L})\mathbf{x}=((1-\omega)\mathbf{D}+\omega\mathbf{U})\,\mathbf{x}+\omega\mathbf{b}$$

La méthode itérative est obtenue par

$$(\mathbf{D}-\omega\mathbf{L})\mathbf{x}^{(k+1)}=((1-\omega)\mathbf{D}+\omega\mathbf{U})\,\mathbf{x}^{(k)}+\omega\mathbf{b}\tag{6.16}$$

Quand $\omega<1$,l'itération (6.16) est une sous-relaxation .Quand $\omega=1$ la méthode est identique à l'itération de Gauss-Seidel, et quand $\omega>1$ on a une sur-relaxation. En écrivant $\mathbf{S}_\omega=\mathbf{D}-\omega\mathbf{L}$, $\mathbf{T}_\omega=(1-\omega)\mathbf{D}+\omega\mathbf{U}$,$\mathbf{B}_\omega=\mathbf{S}_\omega^{-1}\mathbf{T}_\omega$ et $\mathbf{c}_\omega=\mathbf{S}_\omega^{-1}\omega\mathbf{b}$, l'itération de la méthode de relaxation est :

$$\mathbf{x}^{(k+1)}=\mathbf{B}_\omega\mathbf{x}^{(k)}+\mathbf{c}_\omega\tag{6.17}$$

Exemple 6.6 *Résoudre le système suivant avec une tolérance* $||\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}||_\infty \leq \epsilon = 10^{-3}$ *;on prend* $\omega = 1.25$

$$\begin{pmatrix} 2 & -1 & 1 \\ -2 & 5 & -1 \\ 1 & -2 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} -1 \\ 1 \\ 3 \end{pmatrix}$$

$$\mathbf{x}^{(0)} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \mathbf{x}^{(1)} = \begin{pmatrix} -1.1914 \\ -0.0566 \\ 1.0010 \end{pmatrix}, \quad \mathbf{x}^{(2)} = \begin{pmatrix} -0.9882 \\ 0.0203 \\ 1.0088 \end{pmatrix}, \quad \mathbf{x}^{(3)} = \begin{pmatrix} -0.9957 \\ -0.0008 \\ 0.9960 \end{pmatrix},$$

$$\mathbf{x}^{(4)} = \begin{pmatrix} -1.0008 \\ -0.0002 \\ 1.0000 \end{pmatrix}, \mathbf{x}^{(5)} = \begin{pmatrix} -0.9990 \\ -0.0003 \\ 1.0005 \end{pmatrix}, \mathbf{x}^{(6)} = \begin{pmatrix} -0.9999 \\ 0.0001 \\ 1.0000 \end{pmatrix}, \quad \mathbf{x}^{(7)} = \begin{pmatrix} -0.9999 \\ 0.0001 \\ 1.0000 \end{pmatrix}$$

Annexe A

A

$$\sum_{k=1}^{n} k = \frac{n(n+1)}{2} \tag{A.1}$$

$$\sum_{k=1}^{n} k^2 = \frac{n(n+1)(2n+1)}{6} \tag{A.2}$$

Théoreme A.1 *Théorème des valeurs intermédiaires :*
Soit une fonction f à valeurs réelles définie et continue sur l'intervalle $[a,b]$.Si $f(a).f(b) \leq 0$, alors il existe c dans $[a,b]$ tel que $f(c) = 0$.

Théoreme A.2 *Théorème des accroissements finis :*
Étant donnés des réels a et b tels que $a < b$ ainsi qu'une fonction f continue sur $[a,b]$, dérivable sur $]a,b[$, il existe un réel $c \in]a,b[$ tel que :

$$f'(c) = \frac{f(b) - f(a)}{b - a}$$

L'expansion de Taylor pour une fonction à une seule variable :

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \frac{h^2}{2!}f^{(2)}(x) + ... + \frac{h^n}{n!}f^{(n)}(x) + ... \\ &= f(x) + hf'(x) + \frac{h^2}{2!}f^{(2)}(x) + ... + \frac{h^n}{n!}f^{(n)}(x) + \mathcal{R}_n(x) \\ \mathcal{R}_n(x) &= \frac{h^{(n+1)}}{(n+1)!}f^{(n+1)}(\xi(x)) \end{aligned} \tag{A.3}$$

L'expansion de Taylor pour une fonction à deux variables :

$$\begin{aligned} f(x+\Delta x, y+\Delta y) &= f(x,y) + \left(\frac{\partial f(x,y)}{\partial x}\Delta x + \frac{\partial f(x,y)}{\partial y}\Delta y\right) \\ &+ \frac{1}{2!}\left(\frac{\partial^2 f}{\partial x^2}(x,y)\Delta x^2 + 2\frac{\partial^2 f}{\partial x \partial y}(x,y)\Delta x \Delta y + \frac{\partial^2 f}{\partial y^2}(x,y)\Delta y^2\right) \\ &\vdots \\ &+ \frac{1}{n!}\sum_{l=0}^{n}\binom{n}{l}\frac{\partial^n f}{\partial x^l \partial y^{n-l}}(x,y)\Delta x^l \Delta y^{n-l} \\ &+ \frac{1}{(n+1)!}\sum_{l=0}^{n}\binom{n+1}{l}\frac{\partial^{n+1} f}{\partial x^l \partial y^{n+1-l}}(\xi,\psi)\Delta x^l \Delta y^{n+1-l} \end{aligned} \tag{A.4}$$

Théoreme A.3 *Théorème de la valeur moyenne pour les intégrales :*

Soit f continue et g intégrable sur $[a,b]$ et supposons que $g(x) \geq 0$ pour $x \in [a,b]$. Alors il existe un point ξ dans $[a,b]$ tel que :

$$\int_a^b g(x)f(x)dx = f(\xi)\int_a^b g(x)dx \tag{A.5}$$

Théoreme A.4 *les affirmations suivantes sont équivalentes.*

1. $\mathbf{A}$ *est une matrice convergente.*
2. $\lim_{n\to\infty} ||\mathbf{A}^n|| = 0$, *pour toutes les normes naturelles.*
3. $\rho(\mathbf{T}) < 1$
4. $\lim_{n\to\infty} \mathbf{A}^n\mathbf{x} = 0$, *pour chaque x.*

Bibliographie

[1] Richard L. Burden,and J. Douglas Faires *Numerical Analysis*, Cengage Learning 20 Channel Center StreetBoston, MA 02210 USA 10th édition 2016.

[2] Steven C. Chapra,and Raymond P. Canale *Numerical methods for engineers*, McGraw-Hill Education 2 Penn Plaza, New York, NY 10121. 7th edition 2015.

[3] Epperson, James F.,*An introduction to numerical methods and analysis*, John Wiley & Sons, Inc. Hoboken, New Jersey, USA 2nd edition 2013.

[4] G. Miller,*Numerical Analysis for Engineers and Scientists*, University Printing House, Cambridge CB2 8BS, United Kingdom, 1st edition 2014.

[5] Eugeniy E. Mikhailov,*Programming with MATLAB for Scientists*, Taylor Francis Group 6000 Broken Sound Parkway NW, Published in 2017 .

[6] André Fortin,*Analyse numérique pour ingénieurs*,Presses internationales Polytechnique Quatrième édition ,2011.