



Ministère de l'Enseignement Supérieur et de la Recherche
Scientifique
Université Dr. Tahar Moulay de Saida
Faculté de Technologie

Département de génie civil et hydraulique

Méthodes Numériques

Cours et exercices corrigés

Présenté par :

Dr. SELLAH HAMID

Maître de conférences « B » en génie civil

Octobre 2017

Avant-propos

Ce polycopié se trouve être un support pédagogique des cours et exercices pratiques, destiné aux étudiants de 2ème Année des Licences LMD assurées au département de génie civil et hydraulique. Ce polycopié regroupe un certain nombre de méthodes étudiées dans les différents chapitres du cours de méthodes numériques.

Table des matières

Introduction.....	1
Chapitre 01 : Rappels et définitions	
1.1.1 Erreurs de calcul	2
1.1.2 Mesures de l'erreur.....	2
1.1.3 Les systèmes usuels.....	4
1.1.3.1 Le système binaire ($B = 2$).....	4
1.1.3.2 Le système octal ($B = 8$).....	4
1.1.3.3 Système hexadécimal ($B = 16$).....	3
1.1.4 Notation scientifique des nombres.....	3
1.2 Quelques rappels sur les matrices	5
1.2.1. Définition	5
1.2.2. Opérations sur les matrices	6
1.2.3. Transposition (A^T ou A').....	7
1.2.4. Rang d'une matrice $r(A)$	8
1.2.5. Déterminant($\det(A)$ ou $ A $).....	8
1.2.6. Matrice adjointe.....	9
1.2.7. Matrice définie positive.....	9
Chapitre02 : Résolution d'équations non linéaires.....	10
2.1 Méthode de dichotomie.....	10
2.1.1. Initialisation.....	11
2.2 Théorème du point fixe.....	11
2.2.1. Proposition	12
2.2.2. Théorème	12
2.2.3. Démonstration.....	12
2.3 Méthode de la sécante.....	12
2.4. Méthode de Newton.....	15
2.4.1. Théorème	15
2.4.2. Démonstration	15
2.4.3 Interprétation graphique	16
2.4.4 Rapidité de convergence de la méthode de Newton.....	16
Travaux dirigés I :.....	17
La Solution de fiche TD	18
Chapitre03 : Interpolation et approximation polynômiale.....	21
3.1 Introduction.....	21
3.1.1 Démonstration	21
3.2 Polynômes d'interpolation de Lagrange.....	22
3.3 Interpolation de Newton et différences divisées.....	23
3.3.1. Démonstration.....	24
3.4 Erreur d'interpolation.....	25
3.4.1. Proposition	25
3.4.2. Démonstration	25
3.4.3. Théorème	25
Fiche TD II.....	26
La solution de fiche TD II.....	26

Chapitre 04: Intégration numérique.....	29
4-1 Introduction :.....	29
4.2 Intégration de type interpolation.....	30
4.2.1 Généralités.....	30
4.2.1. Définition	30
4.2.2. Proposition.....	30
4.2.3. Théorème	30
4.2.2 Quadratures simples.....	31
4.2.4 Quadratures simples.....	31
4.3 Formules de Newton-Cotes.....	31
4.3.1. Théorème.....	31
4.3.2. Proposition	31
4.3.3. Conséquence	31
4.4. Méthode des trapèzes.....	31
4.5. Méthode de Simpson	32
4.5.1. Problème.....	32
4.5.2. Deuxième méthode pour trouver la formule.....	33
4.5.3 Formule pour la méthode de Simpson	33
4.5.1 Formule pour la méthode de Simpson.....	34
4.6 Formules de Gauss.....	35
4.6.1. Gauss-Hermite.....	37
4.6.2. Gauss-Laguerre.....	37
4.6.3. Gauss-Jacobi.....	38
4.6.4. Gauss-Tchebychev	38
4.7. Méthode de Romberg (1955).....	38
4.71 L'algorithme de Romberg	38
Fiche TD III	40
La solution de fiche TD III.....	41
 Chapitre 05 : Résolution de systèmes linéaires.....	 45
5.1. Méthode Gauss.....	45
5.1.1. Problème	45
5.2. Méthode de Croûte.....	46
5.3. Factorisation	47
5.5. Méthode d'élimination de Gauss Jordan.....	48
5.6.1. Théorème	48
5.6.2. Démonstration	48
5.6.3. Proposition	49
5.6.4. Démonstration.....	49
5.7.1. Théorème.....	50
5.7.2. Démonstration	50
5.7.3 Factorisation LDMT Cholesky.....	52
FICHE TD IV.....	54
La solution FICHE TD iv.....	54
 Chapitre 06 : Méthodes itératives.....	 60
6.1.1. Définition	60
6.1.2. Convergence des méthodes itératives	60
6.1.2.1. Théorème	60

6.2. Méthode de Jacobi.....	60
6.3. Méthode de Gauss-Seidel.....	61
Fiche TD IV.....	63
La solution de fiche TD IV.....	64

Chapitre 07 : Équations et systèmes d'équations différentielles.....	68
7.1 Introduction	66
7.1.1. Existence et unicité des solutions.....	66
7.2. Méthodes d'Euler.....	66
7.3. Méthodes de Runge-Kutta.....	68
Fiche TD V.....	70
La solution	71

Liste de figure

Figure 1 : Méthode de dichotomie	10
Figure 2 : Méthode de point fixe	12
Figure 3 : Méthode de sécante.....	13
Figure 4 : Interprétation graphique de méthode de Newton	16

Liste des équations

1. 1Les mesures de l'erreur	2
1.2 L'erreur relative	2
1.4 Matrice nulle	5
1.5 Matrice diagonale	5
1.7 Matrice identité	5
1.8 Matrice triangulaire supérieure	5
1.9 matrice triangulaire inférieure	6
1.10 La multiplication par un scalaire k	6
1.11 L'addition de deux matrices	6
1.12 Multiplication de deux matrices	6
1.13 La transposée d'une matrice	7
1.14 Matrice inverse	8
1.15 Le déterminant	8
1.16 Matrice adjointe	9
2.1 Le milieu méthode dichotomie	10
2.2 L'erreur de méthode dichotomie	10
2.3 La convergence des suites	11
2.4 Théorème du point fixe	11
2.5 L'erreur de point fixe	12
2.6 Algorithme de méthode sécante	13
2.7 La convergence de la méthode de la sécante	15
2.8 Méthode de newton	15
3.1 Polynômes de degré inférieur	21
3.2Polynômes d'interpolation de Lagrange	22
3.3 Formule d'interpolation de newton	23
4.1 Formule de Riemann.	29
4.2 Formule l'intégrale du polynôme de lagrange	30
4.3 L'erreur d'intégration	30
4.4 Formule de quadrature à 2 points	30

4.5 Formule générale de newton-cotes	31
4.6 Formule de précisement	33
4.7 Formule des trois niveaux	34
4.8 Formule pour la méthode de Simpson	34
4.9 L'erreur de méthode de Simpson	35
4.10 Formules de gauss	35
4.11 Méthode de gauss-Hermite	37
4.12 Méthode de gauss-Laguerre	37
4.13 Méthode de gauss-Jacobi	37
4.14 Méthode de gauss-Tchebychev	38
4.15 Méthode de romberg	38
5.1 Méthode de gauss	44
5.2 Méthode de croûte	45
5.3 Matrice triangulaire (inférieure)	47
5.4 Matrice Triangulaire (supérieure)	47
5.5 Factorisation de cholesky	51
5.6 Formule cholesky	52
6.1 Méthodes itératives	60
6.2 Méthode de Jacobi	61
6.3 Méthode de gauss-Seidel	61
7.1 Équation différentielle	66
7.2 Méthodes d'Euler	67
7.3 L'erreur de méthodes d'Euler	67
7.4 Méthodes de Runge-kutta d'ordre2	67
7.5 L'algorithme de Runge-kutta	67
7.6 Méthode de Runge-kutta d'ordre5	69
7.7 L'erreur de Runge-kutta	69
7.7 Méthode de Runge-kutta d'ordre6	70

Introduction générale

L'objectif de ce polycopié est de fournir à l'étudiant les bases des méthodes de résolution numériques appelées aussi analyse numérique. Par souci de synthèse, nous avons volontairement omis la plupart des démonstrations. Des théorèmes donnant les conditions de convergence ne sont pas toujours spécifiés. Le lecteur devra donc se référer à la littérature pour faire varier la validité de ses choix. Nous donnons autant que possible les algorithmes de calcul des différentes méthodes décrites, dans leurs grandes lignes sinon dans le détail. Ce polycopié est organisé comme suit :

La première partie :

Le chapitre 1 donne quelques rappels d'analyse et d'algèbre matricielle. Il traite la représentation générale des nombres puis le cas de la représentation sur calculateur et les problèmes de précision qui peuvent surgir.

Les chapitres 2 traitent respectivement la résolution des équations non linéaires.

Le chapitre 3 développe les méthodes d'interpolation polynômiale.

Le chapitre 4 donne les différentes méthodes d'intégration numériques

Les chapitres 5 et 6 sont consacrés aux problèmes d'optimisation linéaire et méthode itératif

Chapitres 7 sont respectivement consacrés aux équations différentielles ordinaires et aux dérivées partielles.

Chapitre I : Rappels et définitions

1.1. Erreurs de calcul :

Résoudre un problème signifie trouver une solution x qui dépend de données a et de formules. Ces formules sont issues en général d'une modélisation d'un problème scientifique (en physique, chimie, biologie, économie, etc.). Nous ne discutons pas ici de l'erreur de modélisation qui est l'écart entre le phénomène réel et le phénomène simulé. Dans la suite, les variables x et a appartiennent à des espaces vectoriels normés et il existe une fonction F telle que le problème s'écrive $F(x, a)=0$. En général, il n'est possible de résoudre qu'un problème approché. Ce problème discret, formulé en dimension finie, doit être résolu par un algorithme, qui peut être direct ou itératif. Enfin, l'algorithme est exécuté sur un ordinateur avec une précision finie. Dans le résultat d'un calcul scientifique, il apparaît ainsi plusieurs sources d'erreur :[1]

- l'erreur due à l'approximation des données,
- l'erreur d'approximation ou de discrétisation,
- l'erreur due à l'algorithme s'il est itératif,
- l'erreur d'arrondi due à la précision finie.

Les différentes sources d'erreur interfèrent avec les erreurs d'arrondi lors d'un calcul. Par exemple, la résolution d'un problème très sensible aux variations sur les données donne lieu à des calculs avec de grandes erreurs d'arrondi.

1. 1.1 Mesures de l'erreur

Tout calcul devrait s'accompagner d'une estimation des erreurs d'approximation commises. Pour mesurer celles-ci, nous définissons les erreurs absolues et relatives, ainsi que la notion de chiffres significatifs.

Définition :soit $x \in \mathbb{R}$ et $\tilde{x} \in \mathbb{R}$ une approximation de x .l'erreur absolue sur $\tilde{x}$ est $|x - \tilde{x}|$ si $x \neq 0$, l'erreur relative est

$$|x - \tilde{x}|/|x|. \dots\dots\dots (1,1)$$

1. Définition : Le nombre de chiffres significatifs de x est le nombre de chiffres `a partir du premier chiffre non nul.

Par exemple, le nombre de chiffres significatifs de $x = 0.0034560$ est 5. L'erreur relative sur $\tilde{x} = 0.00346$ vaut environ $1.16 \cdot 10^{-3}$ et le nombre de chiffres significatifs exacts est ici 2.

Dans le cas de vecteurs, on définit les erreurs sur les normes ou les erreurs composantes par composante.[2]

2. Définition : Soit $x \in \mathbb{R}^n$, $\tilde{x} \in \mathbb{R}^n$, une approximation de x et $\|\cdot\|$ une norme définie sur $\mathbb{R}^n$.

L'erreur absolue sur la norme est $\|x - \tilde{x}\|$.

Si $x \neq 0$, l'erreur relative sur la norme est $\|x - \tilde{x}\|/\|x\| \dots\dots\dots(1.1)$

Si $x_i \neq 0, i = 1, \dots, n$, l'erreur relative composante par composante est $\max_{1 \leq i \leq n} |x_i - \tilde{x}_i|/|x_i|$

1.1.2 Les systèmes usuels

1.1.2.1 Le système binaire ($B = 2$)

Le système binaire est le système de la base 2. Il ne comprend que le chiffre 0 et 1. Le passage de la base décimale à la base 2 s'effectue par division successive par 2.

Exemple 6 $(67)_{10} = 1000011_2$

Les poids sont les puissances de 2.

Soit a et b deux chiffres binaires. Ils peuvent donc prendre la valeur 0 ou 1. Nous avons 4 cas possibles pour chaque opération.

Le produit vaut 1 si les deux chiffres valent 1 simultanément. Elle est notée $P = a.b$

La somme vaut 1 si un seulement des deux chiffres vaut 1. On l'appelle dilemme ou encore « OU Exclusif ». Elle est notée : $S = a \oplus b$. La retenue est identique au produit : $R = a.b$ [3]

1.1.2.2 Le système octal ($B = 8$)

Le système octal est le système de la base 8. Il utilise donc les chiffres compris entre 0 et 7. Le passage du système binaire au système octal est immédiat. Il suffit de diviser le nombre binaire en tranches de trois chiffres à partir de la droite ($2^3 = 8$).

Exemple 7 Le nombre écrit en binaire

1|011|100|111|010

1|3|4|7|2

La conversion inverse s'effectue selon le même procédé. C'est très rapide. L'intérêt de ce système réside justement dans cette rapidité de conversion.

1.1.2.3 Système hexadécimal ($B = 16$)

Le système hexadécimal est un système de numération de base 16, donc à 16 chiffres de 0 à 15 notés : 0 1 2 3 4 5 6 7 8 9 A B C D E F [3]

Exemple 8 $(31)_{10} = 16 + 15 = 1F_{16}$ (H pour hexadécimal)

1.1.3 Notation scientifique des nombres

La notation scientifique d'un nombre utilise un signe, une mantisse

$$r \left(\frac{1}{10} \leq r \leq 1 \right) \text{ et un exposant } \pm r \times 10^n \dots\dots\dots(1.3)$$

Exemple

$$32.213 = 0.32213 \times 10^2$$

Le nombre de chiffres alloués à la mantisse étant infini, il y aura forcément des erreurs d'arrondi ou de troncature. Certaines précautions sont alors nécessaires quand on effectue des opérations

d'addition et de soustraction sur les nombres.

_ Il faut additionner les nombres dans l'ordre croissant. A titre d'exemple supposant que l'on alloue 5 chiffres significatifs à la mantisse alors les deux opérations suivantes donnent des résultats très différents :

$$1000000 + \underbrace{1 + 1 + \dots + 1}_{1 \text{ million de fois}} = (0.1 \times 10^7 + 0.0000 \times 10^7) + 1 + 1 + \dots 1 = 1000000$$

$$\underbrace{1 + 1 + \dots + 1}_{1 \text{ million de fois}} + 1000000 = 2000000$$

On voit que la première manière d'additionner donne un résultat faux.

_ Il faut éviter de soustraire des nombres très peu différents. Prenons le cas de la fonction

$x - \sin x$ pour x proche de zéro :

$$x = \frac{1}{15} \text{ rad}$$

$$x = 0.6666666667 \times 10^{-1}$$

$$\sin x = 0.6661729492 \times 10^{-1}$$

$$x - \sin x = 0.0004937175 \times 10^{-1}$$

$$= 0.4937175000 \times 10^{-4}$$

Plusieurs artifices existent pour éviter ces soustractions :

- pour la fonction $\sqrt{x^2 + 1} - 1$ au voisinage de 0, on utilisera l'égalité

$$\sqrt{x^2 + 1} - 1 = \frac{x^2}{\sqrt{x^2 + 1} + 1}$$

_ Pour la fonction $x - \sin x$, on utilisera le développement en séries de Taylor

Idem la fonction $e^x - e^{-2x}$ au voisinage de 0 (ajouter les puissances identique

Pour la fonction $\cos^2 x - \sin^2 x$ pour $x \approx \frac{\pi}{4}$, on utilisera l'égalité

$$\cos^2 x - \sin^2 x = \cos 2x$$

Pour la fonction $\ln x - 1$ pour $x \approx e$, on utilisera la fonction $\ln \frac{x}{e}$

1.2 Quelques rappels sur les matrices

1.2. Définition

Une matrice $m \times n$ est un tableau de nombres à m lignes et n colonnes. Les nombres qui composent la matrice sont appelés les éléments de la matrice (ou aussi les coefficients). Une matrice à m lignes et n colonnes est dite matrice d'ordre (m, n) ou de dimension $m \times n$. L'ensemble des matrices à m lignes et n colonnes à coefficients réels se note $M_{m,n}(\mathbb{R})$. [4]

Matrice particulières

$$\text{Matrice de dimension } m \times n ; (a_{ij}) = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \dots\dots\dots (1.4)$$

Matrice nulle

$$\text{Tous les éléments } a_{ij} = 0 \quad \text{Exemple } A = \begin{pmatrix} 0 & 10 \\ 2 & 0 \end{pmatrix}$$

Matrice carrée d'ordre n

Nombre de lignes = nombre de colonnes = n

$$\text{Matrice diagonale} \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ 0 & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \dots\dots\dots (1.5) \quad \text{Exemple } A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 7 \end{pmatrix}$$

$$\text{Matrice identité d'ordre } n : I_n = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \dots\dots\dots (1.7) \quad \text{Exemple } A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$\text{Matrice triangulaire supérieure} \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{nn} \end{pmatrix} \dots\dots\dots (1.8)$$

$$\text{Exemple } A = \begin{pmatrix} -1 & 5 & 2 & 8 \\ 0 & 2 & -5 & 4 \\ 0 & 0 & 1 & 2 \\ 0 & 0 & 0 & 3 \end{pmatrix}$$

$$\text{Matrice triangulaire inférieure} \begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix} \dots\dots\dots (1.9)$$

$$\text{Exemple } A = \begin{pmatrix} -1 & 0 & 0 & 0 \\ 5 & 2 & 0 & 0 \\ 2 & -5 & 1 & 0 \\ 8 & 4 & 2 & 3 \end{pmatrix}$$

1.2.2. Opérations sur les matrices :

$$\text{Multiplication par un scalaire } K : kA = \begin{pmatrix} ka_{11} & ka_{12} & \cdots & ka_{1n} \\ ka_{21} & ka_{22} & \cdots & ka_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ ka_{m1} & ka_{m2} & \cdots & ka_{mn} \end{pmatrix} \dots\dots\dots (1.10)$$

$$\text{Exemple } A \cdot 5 = 5 \cdot \begin{pmatrix} -1 & 2 & 9 \\ 5 & 2 & 7 \\ 8 & 4 & 3 \end{pmatrix} = \begin{pmatrix} -1 \times 5 & 2 \times 5 & 9 \times 5 \\ 5 \times 5 & 2 \times 5 & 7 \times 5 \\ 8 \times 5 & 4 \times 5 & 3 \times 5 \end{pmatrix} = \begin{pmatrix} -5 & 10 & 9 \\ 25 & 10 & 35 \\ 40 & 20 & 15 \end{pmatrix}$$

Addition de deux matrices de même dimension (m x n) A = (a_{ij}) et B = (b_{ij})

$$A + B = \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & a_{m2} + b_{m2} & \cdots & a_{mn} + b_{mn} \end{pmatrix} \dots\dots\dots (1.11)$$

$$\text{Exemple } A = \begin{pmatrix} -1 & 2 & 9 \\ 5 & 2 & 7 \\ 8 & 4 & 3 \end{pmatrix} \text{ et } B = \begin{pmatrix} 3 & 1 & -3 \\ 2 & 2 & 4 \\ 6 & 4 & 7 \end{pmatrix} \rightarrow A + B = \begin{pmatrix} -1+3 & 2+1 & 9-3 \\ 5+2 & 2+2 & 7+4 \\ 8+6 & 4+4 & 3+7 \end{pmatrix}$$

$$A + B = \begin{pmatrix} 2 & 3 & 6 \\ 7 & 4 & 11 \\ 14 & 8 & 10 \end{pmatrix}$$

Multiplication de deux matrices A et B de dimensions respectives m x n et n x p : AB =

$$\begin{pmatrix} a_{11} & 0 & \cdots & 0 \\ a_{21} & a_{22} & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & a_{mi} & a_{mn} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1p} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ b_{n1} & b_{n2} & a_{ni} & b_{nj} & \cdots & b_{np} \end{pmatrix} \dots\dots\dots (1.12)$$

$$\begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1p} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2p} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{ip} \\ \vdots & \vdots & \cdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mj} & \cdots & a_{mp} \end{pmatrix} = C \text{ (dimension } m \times p \text{)}$$

Avec

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + a_{i3}b_{3j} + \cdots + a_{in}b_{nj} = \sum_{k=1}^n a_{ik}b_{kj} \text{ (} i = 1, 2, m; j = 1, 2, \dots, p \text{)}$$

$$\text{Exemple } A = \begin{pmatrix} -1 & 2 & 9 \\ 5 & 2 & 7 \\ 8 & 4 & 3 \end{pmatrix} \text{ et } B = \begin{pmatrix} 3 & 1 & -3 \\ 2 & 2 & 4 \\ 6 & 4 & 7 \end{pmatrix}$$

1.2.3. Transposition (AT ou A')

La transposée d'une matrice S'obtient en remplaçant les lignes de la matrice par ses colonnes. Si la matrice A est de dimension m x n, la transposée AT, sera de dimension n x m

$$A^T = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}^T = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{1n} & a_{2n} & \cdots & a_{nm} \end{pmatrix} \dots\dots\dots(1.13)$$

Propriétés : soit A et B deux matrice et k un scalaire [5]

$$(A+B)^T = A^T + B^T$$

$$(A^T)^T = A$$

$$(kA)^T = kA^T$$

$$(A B)^T = B^T A^T$$

Pour toute matrice A, le produit $A^T A$ est une matrice carrée symétrique et les éléments de sa diagonale principale sont non négatifs

1.2.4. Rang d'une matrice r(A)

Le rang d'une matrice A de dimension m et n correspond au nombre de lignes non nulles de sa forme échelonnée réduite. On dit que A est de « plein rang » si $r(A) = m$

Remarque : Le rang d'une matrice donne le nombre maximum de ses lignes linéairement indépendantes ainsi que le nb max de ses colonnes linéairement indépendantes.

Propriétés :

1. Si B peut être obtenue de A par applications successives d'opérations élémentaires

sur ses lignes, alors $r(A) = r(B)$

2. $r(A^T) = r(A)$

3. Si le produit matriciel AB est défini, alors $r(AB) \leq \min\{r(A); r(B)\}$

Matrice inverse

Soit A une matrice carrée nxn .l'inverse de A (notée A^{-1}) , si elle existe , est la matrice qui satisfait $AA^{-1} = A^{-1}A = I_n \dots\dots\dots(1.14)$

Si l'inverse existe, on peut l'obtenir de la façon suivante : Considérer la matrice augmentée

$$(A : I) = \begin{pmatrix} a_{11} & a_{21} & \cdots & a_{1n} & : & 1 & 0 & \cdots & 0 \\ a_{12} & a_{22} & \cdots & a_{2n} & : & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & : & \cdots & \cdots & \ddots & \cdots \\ a_{1n} & a_{2n} & \cdots & a_{nn} & : & 0 & 0 & \cdots & 1 \end{pmatrix}$$

Effectuer des opérations élémentaires sur les lignes de matrice augmentée jusqu'à ce qu'elle devienne $(A : I)$. la matrice B alors l'inverse de A i.e. $B = A^{-1}$ [3]

Propriétés

Si A est inversible, alors A^{-1} est aussi inversible et $(A^{-1})^{-1} = A$ (1.15)

Si A est inversible, alors $(A^{-1})^T = (A^T)^{-1}$ (1.16)

Si A et B sont deux matrice carres inversibles de même dimension, alors leur produit

AB est aussi inversible et $(AB)^{-1} = B^{-1} A^{-1}$

1.2.4.Determinant (det(A) ou |A|)

Soit A une matrice carree nxn matrice 2x2 : $\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{21}a_{12}$ (1.17)

Ordre supérieur : le déterminant est égala la somme des produits obtenus en multipliant les éléments d'une ligne quelque (ou d'une colonne) par leur cofacteurs respectifs cofacteur $= A_{ij} = (-1)^{i+j} |M_{ij}|$ ou M_{ij} (mineur) est la sous matrice carrée $(n-1) \times (n-1)$ obtenue en supprimant la ieme ligne et la j^{eme} colonne de A [6]

ainsi $|A| = a_{11}A_{11} + a_{22}A_{12} + a_{nn}A_{1n}$

Propriétés :

1. Si A possède une ligne (ou colonne) de « 0 », alors $|A| = 0$
2. Si A possède 2 lignes (colonnes) identiques, alors $|A| = 0$
3. Si A est triangulaire, alors $|A| =$ produit de ses éléments diagonaux. En particulier, $|I_n| = 1$.
4. Si B est obtenue de A en multipliant une seule de ses lignes (colonnes) par un scalaire k, alors $|B| = k|A|$.
5. Si B est obtenu en permutant 2 lignes (ou colonnes) de A, alors $|B| = -|A|$
6. Si B est obtenu de 2 en additionnant le multiple d'une ligne (colonne) à une autre, alors $|B| = |A|$
7. $|A^T| = |A|$
8. Si A et B sont 2 matrices carrées de même dimension, alors $|AB| = |A||B|$

9. A est inversible si $|A| \neq 0$. On dit que la matrice est non singulière.

1.2.5. Matrice adjointe

Soit A une matrice carrée d'ordre n. La matrice adjointe de A (notée $\text{adj } A$) est définie comme la

transposée de la matrice des cofacteurs de A i.e. $A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}$ ou $A_{ij} = (-1)^{i+j} |M_{ij}|$

si A est une matrice carrée telle que $|A| \neq 0$, alors A inversible

et $A^{-1} = \frac{1}{|A|} \text{adj } (A)$ (1.18)

1.2.6 matrice définie positive

une matrice n x n symétrique A est dite « définie positive » si pour tout vecteur X (n x 1), le produit $X^T A X > 0$.

elle est « semi-définie positive » si $X^T A X \geq 0$ pour tout X

une matrice n x n symétrique A est dite « définie négative » si pour tout vecteur X (n x 1), le produit $X^T A X < 0$.

Elle est « semi-définie négative » si $X^T A X \leq 0$ pour tout X.

Chapitre 02 : Résolution d'équations non linéaires

On considère une fonction réelle f définie sur un intervalle $[a, b]$, avec $a < b$, et continue sur cet intervalle et on suppose que f admet une unique racine sur $I =]a, b[$ c'est-à-dire qu'il existe un unique $\alpha \in I$ tel que $f(\alpha) = 0$. [7]

2.1 Methode de dichotomie

On considère un intervalle $[a, b]$ et une fonction f continue de $[a, b]$ dans $\mathbb{R}$. On suppose que $f(a)f(b) < 0$ et que l'équation $f(x) = 0$ admet une unique solution α sur l'intervalle $[a, b]$.

La méthode de dichotomie consiste à construire une suite (x_n) qui converge vers α de la manière suivante :

2.1.1. Initialisation : on prend pour x_0 le milieu de $[a, b]$. La racine se trouve alors dans l'un des deux intervalles $]a, x_0[$ ou $]x_0, b[$ ou bien elle est égale à x_0 .

- si $f(a)f(x_0) < 0$, alors $\alpha \in]a, x_0[$. on

Pose $a_1 = a, b_1 = x_0$

- si $f(a)f(x_0) = 0$, alors $\alpha = x_0$

- si $f(a)f(x_0) > 0$, alors $\alpha \in]x_0, b[$. on

pose $a_1 = x_0, b_1 = b$

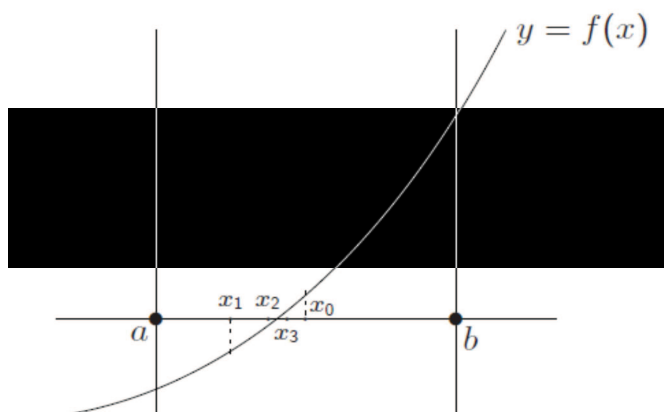


Figure 1 :methode de dichotomie

On prend alors pour x_1 le milieu de $[a_1, b_1]$

On construit ainsi une suite $x_0 = \frac{a+b}{2}, \dots, x_n = \frac{a_n + b_n}{2}$ (2.1)

$$x_1 = \frac{a_1 + b_1}{2}, \dots, x_n = \frac{a_n + b_n}{2}$$

telle que $|\alpha - x_n| \leq \frac{(a-b)}{2^{n+1}}, \dots, x_n$ (2.2)

Etant donne une précision ε , cette méthode permet d'approcher α en un nombre previsible d'itérations.

Les principes de construction suivants consistent à transformer l'équation $f(x) = 0$ en une équation équivalente $g(x) = x$

On peut poser par exemple $g(x) = x + f(x)$, mais on prendra plus généralement $g(x) = x + u(x)f(x)$ où u est une fonction non nulle sur l'intervalle

I. Il reste à choisir u pour que la suite définie par $x_0 \in I$ et la relation de récurrence

$$x_{n+1} = x_n + u(x_n)f(x_n) \dots \dots \dots (2.3)$$

Soit bien définie et converge vers la racine α de f géométriquement, on a remplacé la recherche de l'intersection du graphe de la fonction f avec l'axe Ox , par la recherche de l'intersection de la droite d'équation $y = x$ avec la courbe d'équation $y = g(x)$.

Le choix d'une méthode est conditionné par les réponses aux questions suivantes :

- 1) la suite (x_n) converge-t-elle ?
- 2) si la suite converge, sa limite est-elle α ?
- 3) si on veut la solution à ε près, comment arrêter les itérations dès que cette condition est remplie ?
- 4) comme dans tout calcul, on désire obtenir rapidement le résultat approché, il faudra donc estimer la manière dont évolue l'erreur $e = x_n - \alpha$ au cours des itérations.

Les deux premières questions sont purement mathématiques. Les deux dernières sont numériques, car on ne peut effectuer qu'un nombre fini d'itérations pour le calcul.

La continuité des fonctions considérées permet de répondre à la question 2 : si la suite converge, elle converge vers une racine de l'équation ; si, de plus, $x_n \in I$, pour tout n , alors par unicité de la racine dans I , la suite converge vers α .

2.2 Theoreme du point fixe

On dit que l'application $g : [a, b] \rightarrow \mathbb{R}$ est strictement contractante si

$$\exists L \in]0,1[, \forall x \in [a, b] , \forall y \in [a, b] , |g(x) - g(y)| \leq L|x - y| \dots \dots \dots (2.4)$$

2.2.1. Proposition – Soit g une application de classe C^1 de l'intervalle $[a, b]$ de $\mathbb{R}$ dans $\mathbb{R}$. On suppose que g vérifie[8]

$$\max\{|g'(x)|; x \in [a, b]\} \leq L < 1$$

alors l'application g est strictement contractante dans l'intervalle $[a, b]$.

2.2.2. Théorème

– Si g est une application définie sur l'intervalle $[a, b]$ à valeurs dans $[a, b]$,

alors la suite (x_n) définie par $x_0 \in [a, b]$ et la relation de récurrence $x_{n+1} = g(x_n)$ converge vers l'unique solution α de l'équation $x = g(x)$ avec $\alpha \in [a, b]$.

2.2.3. Démonstration : la suite (x_n) est bien définie $\text{carg}([a, b]) \subset [a, b]$. Montrons, par l'absurde, que l'équation $x = g(x)$ a au plus une solution. Supposons qu'il y ait deux solutions α_1 et α_2 alors [9]

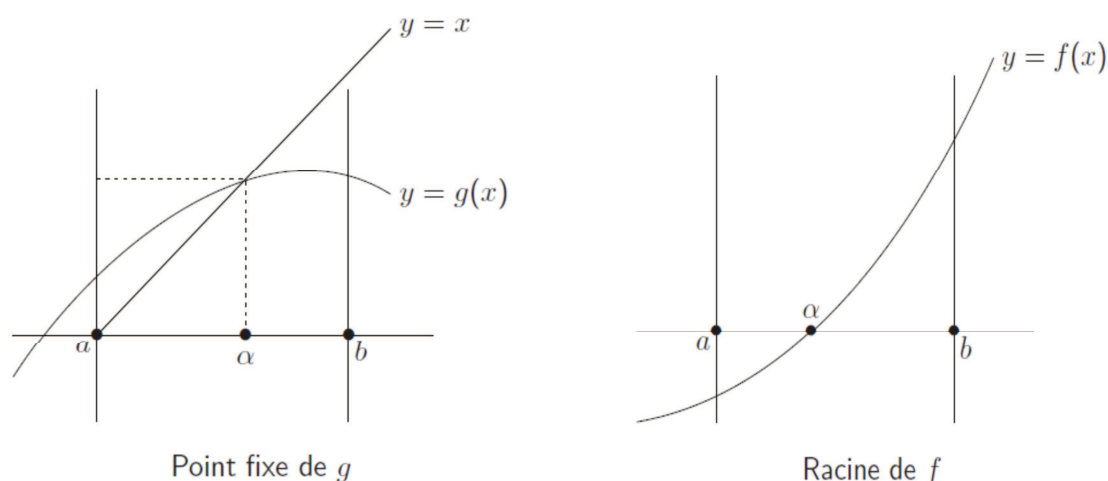


Figure 2 : méthode de point fixe

$$|\alpha_1 - \alpha_2| = |g(\alpha_1) - g(\alpha_2)| \leq L|\alpha_1 - \alpha_2| \text{ Or } L < 1$$

Donc nécessairement $\alpha_1 = \alpha_2$.

Montrons que la suite (x_n) est convergente. On a

$$|x_{n+1} - x_n| = |g(x_n) - g(x_{n-1})| \leq L|x_n - x_{n-1}|.$$

$$\text{et par récurrence } |x_{n+1} - x_n| \leq L^n |x_1 - x_0|$$

On en déduit que

$$|x_{n+p} - x_n| \leq L^n |x_1 - x_0| \frac{1 - L^n}{1 - L} \leq |x_1 - x_0| \frac{L^n}{1 - L}$$

Cette suite vérifie le critère de Cauchy ; elle est donc convergente vers $\alpha \in [a, b]$, or l'application g est continue donc la limite α vérifie $g(\alpha) = \alpha$.

On a aussi une évaluation de l'erreur en faisant tendre p vers l'infini, on obtient

$$|\alpha - x_n| \leq |x_1 - x_0| \frac{L^n}{1 - L} \dots \dots \dots (2.5)$$

On constate que, pour n fixe, l'erreur est d'autant plus petite que L est proche de 0.

2.3 Methode de la secante

Cette méthode est également appelée méthode de Lagrange, méthode des parties proportionnelles.

On considère un intervalle $[a, b]$ et une fonction f de classe C^2 de $[a, b]$ dans $\mathbb{R}$. On suppose que $f(a)f(b) < 0$ et que f' ne s'annule pas sur $[a, b]$, alors l'équation $f(x) = 0$ admet une unique solution sur l'intervalle $[a, b]$. [10]

La méthode de la sécante consiste à construire une suite (x_n) qui converge vers α de la manière suivante : soit Δ_0 la droite passant par $(a, f(a))$ et $(b, f(b))$, elle coupe l'axe O_x en un point d'abscisse $x_0 \in]a, b[$. On approche donc la fonction f par un polynôme P de degré 1 et on résout $P(x) = 0$.

Ensuite, suivant la position de α par rapport à x_0 , on considère la droite passant par $(a, f(a))$ et $(x_0, f(x_0))$ si $f(x_0)f(a) < 0$ ou celle passant par $(x_0, f(x_0))$ et $(b, f(b))$ si $f(x_0)f(b) < 0$.

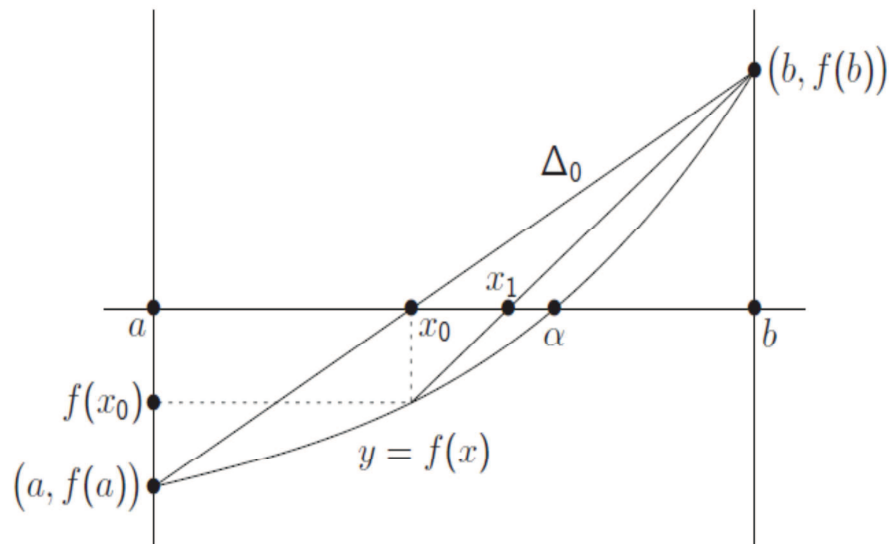


Figure 3 :méthode de secante

On appelle x_1 l'abscisse du point d'intersection de cette droite avec l'axe O_x .

On réitère ensuite le procédé.

Plaçons-nous dans le cas où $f' > 0$ est dérivable et f est convexe (i.e. $f'' \geq 0$), alors la suite (x_n) est définie par

$$\begin{cases} x_0 = a \\ x_{n+1} = \frac{bf(x_n) - x_nf(b)}{f(x_n) - f(b)} \end{cases} \dots \dots \dots (2.6)$$

En effet, si f est convexe, on remplace l'intervalle $[x_n, b]$ par l'intervalle $[x_{n+1}, b]$. L'équation d'une droite passant par

$$(c, f(c)) \text{ et } (b, f(b)) \text{ avec } c \neq b \text{ est } y - f(b) = \frac{f(b) - f(c)}{b - c} (x - b)$$

On cherche son intersection avec l'axe O_x donc on prend $y = 0$ et on obtient la formule donnée plus haut.

Remarque - si f est convexe, alors sa représentation graphique est au dessus des tangentes et en dessous des cordes.

$$\text{on pose } g(x) = \frac{bf(x_n) - x_nf(b)}{f(x_n) - f(b)} = x - \frac{b-x}{f(b) - f(x)}f'(x) \text{ d'où } x_{n+1} = g(x_n)$$

Montrons que cette suite (x_n) est définie. Pour cela, il suffit de montrer que $(g[a, b]) \subset [a, b]$. Vérifions d'abord que g est de classe C^1 sur $[a, b]$. Elle l'est de manière évidente sur $]a, b[$.

$$\text{L'application } g \text{ est continue en } b \text{ et } g(b) = b - \frac{f(b)}{f'(b)}.$$

On a, pour tout

$$x \in [a, b] \quad g'(x) = 1 - f'(x) \frac{b-x}{f(b)-f(x)} + f(x) \frac{f(b)-f(x)-(b-x)f'(x)}{(f(b)-f(x))^2} = f(b) \frac{f(b)-f(x)-(b-x)f'(x)}{(f(b)-f(x))^2}$$

on en déduit que g' est continue en b et $g'(b) = f(b)f''(b)/(2f'(b)^2)$.

De plus, l'application f est convexe, donc $f(b) - f(x) - (b-x)f'(x) \geq 0$.

On a également $f(b) > 0$

Car f croissante et $f(a)f(b) < 0$. la fonction g est donc croissante sur $[a, b]$. On a alors $g([a, b]) \subset [g(a), g(b)]$. De plus, $g(a) = a - f(a) \frac{(b-a)}{f(b)-f(a)}$, or $f(a) < 0$ et $\frac{(b-a)}{f(b)-f(a)} \geq 0$

Par croissance de f ; donc $g(a) \geq a$.

de même $g(b) = b - \frac{f(b)}{f'(b)}$, or $f'(b) > 0$ et $f(b) > 0$ donc $g(b) \leq b$.

La croissance de g montre que la suite (x_n) est croissante car $x_1 = g(a) \geq a = x_0$, or elle est dans l'intervalle $[a, b]$, donc elle converge vers $I \in [a, b]$ tel que, par continuité de g ,

$g(I) = I$. De plus, $x_0 \leq \alpha$ donc une récurrence immédiate et la croissance de g montrent que, pour tout entier n , $x_n \leq \alpha$. On en déduit que $I \leq \alpha$, c'est-à-dire que $I \in [a, b]$.

Soit $x \in]a, b[$ tel que $f(x) = 0$, alors il est immédiat que $g(x) = x$. Réciproquement, soit

$$x \in]a, b[\text{ tel que } g(x) = x, \text{ alors } f(x) \frac{(b-x)}{f(b)-f(x)} = 0, \text{ or } x \neq b \text{ et } f(x) \neq f(b)$$

Car f est strictement croissante et $x \in]a, b[$. on en déduit que $xf(x) = 0$. Il y a unicité de la racine de f , donc $I = \alpha$ et la suite (x_n) converge vers α [11]

Rapidité de convergence de la méthode de la sécante

On a $x_n - \alpha = g'(\varepsilon)(x_{n-1} - \alpha)$. On en déduit, par continuité de g' en α , que

$$\lim_{n \rightarrow +\infty} \frac{|x_n - \alpha|}{|x_{n-1} - \alpha|} = |g'(\alpha)| = g'(\alpha) \dots \dots \dots (2.7)$$

2.4. Méthode de Newton

On cherche les points fixes de la fonction $g(x) = x + u(x)f(x)$, et on a vu que

$$\lim_{n \rightarrow +\infty} \frac{|x_n - \alpha|}{|x_{n-1} - \alpha|} = |g'(\alpha)| = g'(\alpha)$$

Pour obtenir une convergence plus rapide, on peut chercher u tel que $g'(\alpha) = 0$. On a, si les fonctions ont les régularités nécessaires,

$$g'(x) = 1 + u'(x)f(x) + u(x)f'(x) \text{ et on en déduit que } g'(\alpha) = 0$$

si $u(\alpha) = -1/f'(\alpha)$. Si la fonction f' ne s'annule pas, on peut donc choisir

$u(x) = -1/f'(x)$. On obtient alors la méthode de Newton.

On considère une fonction réelle définie sur un intervalle $I = [a, b]$ de classe C^2 telle que $f(a)f(b) < 0$;

on suppose que les fonctions f' et f'' ne s'annulent pas et gardent chacune un signe constant sur I .

$$\text{On pose } g(x) = x - \frac{f(x)}{f'(x)} \dots \dots \dots (2.8)$$

Si f' et f'' est positive (respectivement négative) sur $[a, b]$, on pose $x_0 = b$ (respectivement a).

On définit alors la suite (x_n) par la donnée de x_0 et la relation de récurrence

$$x_{n+1} = g(x_n).$$

2.4.1. Théorème : La suite (x_n) converge vers α l'unique racine de f sur $[a, b]$.

2.4.2. Démonstration : pour simplifier la rédaction, on va supposer que f' est strictement positive et f'' est strictement négative sur I . Les autres cas se traitent de manière similaire. Ces hypothèses assurent l'existence et l'unicité de $\alpha \in I$ tel que $f(\alpha) = 0$. [12]

On va montrer que la suite (x_n) est croissante et majorée par α . On a $x_0 = a$, donc $x_0 \leq \alpha$.

Supposons que $x_n \leq \alpha$, alors comme $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ et que la fonction f'

est positive, donc f est croissante, on peut déduire immédiatement que $x_{n+1} \geq x_n$ et la suite (x_n) est croissante. De plus, $x_{n+1} - \alpha = g(x_n) - g(\alpha) = g'(\varepsilon)(x_n - \alpha)$ avec $\varepsilon \in]x_n, \alpha[$

$$\text{Or } g'(x) = \frac{f(x)f''(x)}{(f'(x))^2} \text{ donc } g'(\varepsilon) > 0 \text{ et } x_{n+1} \leq \alpha.$$

La suite (x_n) est croissante et majorée ; elle est donc convergente . Notons sa limite l . la continuité de g permet d'écrire que $l = g(l)$ et donc $f(l) = 0$ d'où $l = \alpha$

2.4.3 Interprétation graphique

La méthode de Newton consiste à remplacer la courbe par sa tangente en une de ses deux extrémités. Le point x_1 est l'intersection de cette tangente avec l'axe O_X .

Pour faire le dessin, on va se placer dans le cas étudié pour la méthode de la sécante,

$f' > 0$ et $f'' < 0$. On prend alors $x_0 = b$.

Traçons la tangente à la courbe représentative de f passant par $(b, f(b))$. L'équation de cette tangente est $y = f'(b)(x - b) + f(b)$. Son intersection avec l'axe O_X a une ordonnée nulle et son abscisse vaut α . On trace ensuite la tangente à la courbe

au point $(x_1, f(x_1))$. Le réel x_2 est l'abscisse de l'intersection de cette deuxième tangente avec l'axe O_X et on répète ce procédé.

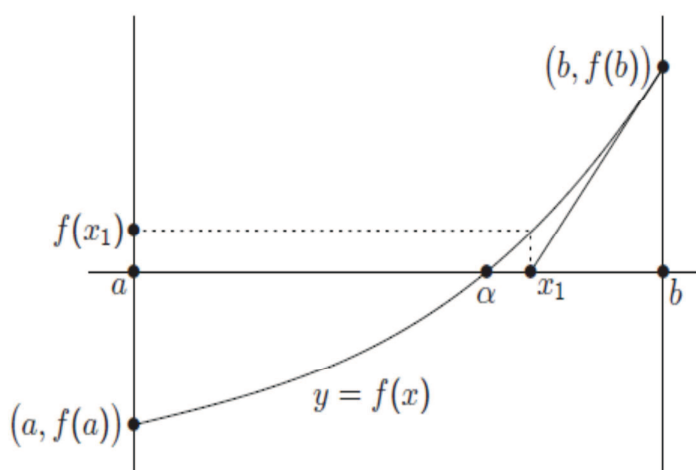


Figure 4 : Interprétation graphique de méthode de Newton

Remarque - Si on prenait la tangente à la courbe en $(a, f(a))$, son intersection avec l'axe O_X ne serait pas, sur ce dessin, dans l'intervalle $[a, b]$.

2.4.4 Rapidité de convergence de la méthode de Newton

On suppose que la fonction f est de classe $\mathcal{C}^3$ sur $I = [a, b]$, que l'on a $f(a)f(b) < 0$

et que les fonctions f' et f'' sont toutes deux strictement positives sur $[a, b]$. Ceci nous

garantit l'existence et l'unicité d'une racine simple α de f sur $[a, b]$. On a donc $f(\alpha) = 0$ et

$f'(\alpha) \neq 0$.

On pose $g(x) = x - \frac{f(x)}{f'(x)}$ et on définit la suite (x_n) par $x_0 = b$ et la relation de récurrence $x_{n+1} = g(x_n)$.

Travaux dirigés I

Exercice 01 : Utiliser la méthode de la bisection pour calculer une approximation de la solution de l'équation : $1 - xe^x = 0$ dans l'intervalle $[0,1]$ avec une précision $\varepsilon = 10^{-3}$

Exercice 02 : On considère la fonction $F(x) = x^2 + x - 1 = 0$.

Déduire les fonctions $g(x)$ 1 ; $g(x)$ 2.

Calcule la divergence de $g(x)$ 1 = $(1-x)0.5$ avec $x_0=3.25$

On pose $g(x)$ 2 : Montrer que la méthode du point fixe converge vers la solution $\alpha=1$ dans l'intervalle $I = [-1.5,0]$.

Exercice 03 : On considère la fonction $F(x) = x^2e^x - 1 = 0$

1- Trouver un intervalle $I = [a,b]$ de longueur 1 contenant la racine α de $F(x)$

2- Après avoir vérifié l'applicabilité de la méthode de Dichotomie sur cet intervalle, calculer les cinq premiers itérés.

3- Calculer le nombre d'itération qu'il faut pour avoir la solution avec une précision $\varepsilon = 10^{-4}$

Exercice 04 : On veut résoudre l'équation $x^3 - x - 1$ par la méthode du point fixe dans l'intervalle $[1,2]$

1. montrer que la fonction $x = g(x) = \sqrt[3]{x+1}$ vérifie les conditions de convergence

2. calculer la solution approximative avec la précision $\varepsilon = 10^{-2}$ et $x_0 = 1.5$

Exercice 05 : Ecrire un processus itératif $(x_n)_{n \in \mathbb{N}}$ du point fixe définie par la fonction d'itération suivante : $g(x) = x + \frac{1}{4}(e^{-x} - x^2)$

Montrer que la méthode du point fixe converge vers la solution α dans l'intervalle $I = [0,1]$

b- Déterminer le nombre d'itérations nécessaires pour calculer une solution approché avec une précision $\varepsilon = 10^{-10}$

Calculer les cinq premières itérations.

Exercice 06 : Soit l'équation suivante : $f(x) = x - 0.8 - 0.2\sin x = 0$

En utilisant la méthode de Newton- Raphson résoudre l'équation dans l'intervalle $\left[\frac{\pi}{4}, \frac{\pi}{2}\right]$ avec précision $\varepsilon = 10^{-5}$ et $x_0 = \frac{\pi}{4}$

Exercice 07 : Soit $f(x) = xe^x - 2x, x \in \mathbb{R}$

Montre que fonction f admet un minimum unique α dans l'intervalle $[0,1]$

Montrer que l'algorithme de newton permet de déterminer ce minimum α . écrire la suite de newton x_n

Calculer x_5 sachant que $x_0=1$

La Solution de la fiche de TD

Exercice 1 : En utilisant la méthode de la bisection, on va calculer la solution approchée de l'équation :

$$f(x) = 1 - xe^x = 0 \text{ ou } x \in [a, b] \text{ et } \varepsilon = 10^{-3}$$

$$\begin{cases} f \text{ est continue sur } [a, b] \\ f(0) = 1, f(1) = -1.718; f(0).f(1) < 0 \end{cases} \rightarrow \exists C \in [0,1]; f(C) = 0$$

$$\text{de plus, } f'(x) = -e^x(1+x) < 0 \forall x \in [1,2] \rightarrow f^\downarrow.$$

f est monotone la solution c est unique

on prendra 4 chiffres après la virgule et le nombre d'itérations n est

$$n \geq \frac{\ln \frac{b-a}{\varepsilon}}{\ln 2} = \frac{\ln 10^3}{\ln 2} = 9.97 \rightarrow n \text{ 10 itérations}$$

n	a	b	$c = \frac{a+b}{2}$	$b-c$	$b(x)$
1	0.000	1.000	0.5000	0.5000	+0.1756
2	0.5000	1.0000	0.7500	0.2500	-0.5877
3	0.5000	0.7500	0.6250	0.1250	-0.1676
4	0.5000	0.6250	0.5625	0.0625	+0.0128
5	0.5625	0.6250	0.5937	0.0313	-0.0750
6	0.5625	0.5937	0.5781	0.0156	-0.0305
7	0.5625	0.5781	0.5703	0.0078	-0.0087
8	0.5625	0.5703	0.5664	0.0039	+0.0020
9	0.5664	0.5703	0.5683	0.0020	-0.0032
10	0.5664	0.5683	0.5673	0.0010	+0.0004

$$b - c = 0.5683 - 0.5673 = 0.0010 \leq \varepsilon = 0.0010 \text{ Donc on arrête les calculs et la}$$

solution est $c \approx 0.5673$ et on peut écrire $c = 0.5673 \pm 0$.

Exercice 03 : $f(x) = x - 0.8 - 0.2\sin x = 0, x \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right], \varepsilon = 10^{-5}$ et $x_0 = \frac{\pi}{4}$

Etude de la convergence :

La fonction f est définie sur l'intervalle $\left[\frac{\pi}{4}, \frac{\pi}{2}\right]$ tel que

$$\begin{cases} f\left(\frac{\pi}{4}\right) = -0.16 \\ f\left(\frac{\pi}{2}\right) = 0.57 \end{cases} \rightarrow f\left(\frac{\pi}{4}\right) f\left(\frac{\pi}{2}\right) < 0$$

b) $f'(x) = 1 - 0.2\cos x$

$$f'(x) = 0 \rightarrow 1 - 0.2\cos x = 0 \rightarrow \cos x = 5 \text{ impossible de résoudre puisque } \cos x \in [-1, 1]$$

Donc

$$f'(x) \neq 0 \forall x \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right]$$

$$f''(x) = 0.2 \sin x \neq 0 \forall x \in \left[\frac{\pi}{4}, \frac{\pi}{2}\right]$$

La fonction f vérifie les 3 conditions, alors la méthode de Newton-Raphson est convergence vers une solution unique

La formule récursive est :

$$x_k = x_{k-1} - \frac{f(x_{k-1})}{f'(x_{k-1})} = x_{k-1} - \frac{x_{k-1} - 0.8 - 0.2\sin(x_{k-1})}{1 - 0.2\cos(x_{k-1})}$$

k	x_k	$ x_k - x_{k-1} $
1	0.967120	0.181722
2	0.964335	0.002785
3	0.964334	0.000001

$$|x_k - x_{k-1}| = |0.964334 - 0.964335| = 0.000001 \leq \varepsilon = 10^{-5}$$

La solution approchée est : $c \cong 0.964334$

Exercice 06 : $f(x) = x^3 - x - 1 = 0, x \in [1, 2]$

1. vérification des conditions de convergence

$$g(x) = \sqrt[3]{x+1}$$

On a deux conditions vérifié

a. $g[1, 2] \subset [1, 2]$

La fonction $g(x)$ est continue sur l'intervalle $[1, 2]$

$$g'(x) = \frac{1}{3\sqrt[3]{(x+1)^2}} > 0 \forall x \in [1, 2] \text{ donc } g \uparrow$$

$$\begin{aligned} g(1) &= 1.26 \in [1, 2] \\ g(2) &= 1.44 \in [1, 2] \end{aligned} \rightarrow \begin{aligned} \forall x \in [1, 2], 1.26 \leq g(x) \leq 1.44 \\ g[1, 2] &\subset [1, 2] \end{aligned}$$

b. $|g'(x)| \leq k \leq 1 \quad \forall x \in [1, 2].$

$$g'(x) = \frac{1}{3\sqrt[3]{(x+1)^2}} \rightarrow |g'(x)| = \frac{1}{3\sqrt[3]{(x+1)^2}}$$

$$\forall x_1 \in [1,2], \forall x_2 \in [1,2] : x_1 < x_2 \rightarrow \frac{1}{3\sqrt[3]{(x_1+1)^2}} > \frac{1}{3\sqrt[3]{(x_2+1)^2}}$$

$$\rightarrow |g'(x_1)| > |g'(x_2)| \rightarrow |g'(x)| \downarrow$$

$$\text{Donc } |g'(x)| \leq |g'(1)| = k = 0.21 < 1$$

Les conditions a et b sont vérifiées donc la méthode du point converge

La suite est définie par la formule récursive suivante

$$x_k = g(x_{k-1}) = \sqrt[3]{x_{k-1} + 1}$$

Calcul de la solution approximative

n	x_k	$ x_k - x_{k-1} $
0	1.500	—
1	1.357	0.143
2	1.331	0.019
3	1.326	0.005

La solution approchée est $c \cong 1.326$

Chapitre 3 : Interpolation et approximation polynômiale

3.1 Introduction

Soit n un entier naturel puis $x_0, x_1, \dots, x_n$, $(n+1)$ réels distincts et $f: \mathbb{R} \rightarrow \mathbb{R}$ une fonction.

Alors il existe un unique polynôme P de degré au plus n qui vérifie

$$P(x_i) = f(x_i) \quad , \forall i \in \{0, \dots, n\} \quad [13]$$

3.1.1 Démonstration. On cherche une solution P dans $\mathbb{R}_n[X]$, l'espace vectoriel des polynômes de degré inférieur ou égal à n , c'est à dire,

$$P(X) = \sum_{k=0}^n a_k X^k \quad \dots\dots\dots (3.1)$$

Les équations qui doivent être satisfaites sont donc

$$\sum_{k=0}^n a_k x_i^k = f(x_i) \quad \forall i \in 0, \dots, n.$$

Soit F l'application linéaire :

$$F: \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$$
$$(a_0, a_1, \dots, a_n) \rightarrow \left(\sum_{k=0}^n a_k (x_0)^k, \sum_{k=0}^n a_k (x_1)^k, \dots, \sum_{k=0}^n a_k (x_n)^k \right)$$

D'après le théorème du rang

$$\dim(\ker(F)) + \dim(\beta(A)) = n + 1$$

C'est-à-dire

$$F \text{ injective} \leftrightarrow F \text{ surjective} \leftrightarrow F \text{ bijective}$$

On va montrer que F est injectif

Soit $(b_0, b_1, \dots, b_n)$, vérifiant, pour tout $i \in 0, 1, \dots, n$,

$$\sum_{k=0}^n b_k (x_i)^k = 0$$

Alors le polynôme $Q(X) = \sum_{k=0}^n b_k X^k$

Possède $n+1$ racines distinctes. D'après le Théorème fondamental de l'algèbre, on a, $Q(X) = 0$ [14]

L'application F est donc injective. On peut donner une autre preuve. L'équation (4.1) est équivalente à :

$$\begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

$$A = \begin{pmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \dots & x_n^n \end{pmatrix}$$

3.2 Polynômes d'interpolation de Lagrange

Soit n un entier naturel puis $x_0, x_1, \dots, x_n$, $(n+1)$ complexes deux à deux distincts.

On cherche $(n+1)$ polynômes $L_0, \dots, L_n$ tous de degré au plus n vérifiant les égalités :
 $\forall (i, j) \in \llbracket 0, n \rrbracket^2, L_i(x_j) = \delta_{i,j}$ [15]

où $\delta_{i,j}$ est le symbole de KRONECHER défini par $\delta_{i,j} = \begin{cases} 0 & \text{si } i \neq j \\ 1 & \text{si } i = j \end{cases}$

Soit i un entier naturel élément de $\llbracket 0, n \rrbracket$ donné. Le polynôme L_i est de degré au plus n et admet les n complexes deux à deux distincts x_j , $j \neq i$ pour racines. Par suite, nécessairement, il existe une constante C telle que

$$L_i = C \prod_{j \neq i} (X - x_j)$$

L'égalité $L_i(x_i) = 1$ fournit alors $C = \frac{1}{\prod_{j \neq i} (x_i - x_j)}$ et donc nécessairement

$$L_i = \prod_{j \neq i} \left(\frac{X - x_j}{x_i - x_j} \right) \dots \dots \dots (3.2)$$

Réciproquement, si pour tout $i \in \llbracket 0, n \rrbracket$, $L_i = \prod_{j \neq i} \left(\frac{X - x_j}{x_i - x_j} \right)$ alors L_i est bien défini car les x_i sont deux à deux distincts, de degré n exactement et enfin les polynômes L_i vérifient clairement les égalités de dualité $\forall (i, j) \in \llbracket 0, n \rrbracket^2, L_i(x_j) = \delta_{i,j}$

Soient n un entier naturel puis $x_0, x_1, \dots, x_n$, $(n+1)$ complexes deux à deux distincts données. Il existe une et une seule famille, notée $(L_i)_{0 \leq i \leq n}$, de $(n+1)$ polynômes de degré au plus n vérifiant : $\forall (i, j) \in \llbracket 0, n \rrbracket^2, L_i(x_j) = \delta_{i,j}$ [16]

De plus : $\forall (i, j) \in \llbracket 0, n \rrbracket^2, L_i = \prod_{j \neq i} \left(\frac{X - x_j}{x_i - x_j} \right)$

La famille $(L_k)_{0 \leq k \leq n}$ est une base de $C_n[X]$

Les L_k sont tous dans $C_n[X]$. Montrons que la famille $(L_k)_{0 \leq k \leq n}$ est une famille libre de $C_n[X]$.

Soit $(\lambda_i, \lambda_m) \in C^{n+1}$.

$$\sum_{i=0}^n \lambda_i L_i = 0 \Rightarrow \forall j \in \llbracket 0, n \rrbracket, \sum_{i=0}^n \lambda_i L_i(x_j) = 0 \Rightarrow \forall j \in \llbracket 0, n \rrbracket, \\ \sum_{i=0}^n \lambda_i \delta_{i,j} = 0 \Rightarrow \forall j \in \llbracket 0, n \rrbracket, \lambda_j = 0$$

Donc la famille $(L_k)_{0 \leq k \leq n}$ est une famille libre de $C_n[X]$. Comme $\text{card}(L_k)_{0 \leq k \leq n} = n + 1 = \dim(C_n[X]) < +\infty$, la famille $(L_k)_{0 \leq k \leq n}$ est une de $C_n[X]$

Coordonnées d'un polynôme de degré au plus n dans la Base $(L_k)_{0 \leq k \leq n}$

En conservant les notations précédentes, on se donne de plus $n + 1$ uplet quelconque $(y_0, \dots, y_n)$ de nombres complexes. On cherche les polynômes P de degré au plus n vérifiant $\forall j \in \llbracket 0, n \rrbracket, P(x_j) = y_j$

soit $P \in C_n[X]$. notons $(\lambda_0, \dots, \lambda_n)$ les coordonnées de P dans la base $(L_i)_{0 \leq i \leq n}$ de $C_n[X]$ e. on a danc

$P = \sum_{i=0}^n \lambda_i L_i$. maintenant, pour j element de $\llbracket 0, n \rrbracket$.

$$P(x_j) = \sum_{i=0}^n \lambda_i L_i(x_j) = \sum_{i=0}^n \lambda_i \delta_{i,j} = \lambda_j$$

Ainsi,

$\forall j \in \llbracket 0, n \rrbracket, P(x_j) = y_j \Leftrightarrow \forall j \in \llbracket 0, n \rrbracket, \lambda_j = y_j$.

On a montre que

$$\forall P \in C_n[X], P = \sum_{i=0}^n P(x_i) L_i$$

3.3 Interpolation de Newton et différences divisées

On veut maintenant déterminer une nouvelle forme du polynôme d'interpolation

P qui ne nécessite pas de recalculer toutes les fonctions de base lors de l'ajout

D'un nouveau point. Si l'on a qu'un point d'interpolation x_0 alors,

$$P(x) = f(x_0)$$

Si l'on ajoute un point x_1 alors

$$P(x) = f(x_0) + a_1(x - x_0)$$

$$P(x_1) = f(x_1)$$

implique
$$a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

Si l'on a $n + 1$ points $x_0, x_1, \dots, x_n$ on cherche

$$P(x) = a_0 + a_1(x - x_0) + \dots + a_n \prod_{k=0}^{n-1} (x - x_k)$$

Remarquer que si l'on a déterminé $a_0, a_1, \dots, a_{i-1}$ alors de l'égalité

$$P(x_i) = f(x_i) = a_0 + a_1(x_i - x_0) + \dots + a_i(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})$$

On déduit que a_i est déterminé par la formule

$$a_i = \frac{f(x_i) - a_0 - a_1(x_i - x_0) - \dots - a_{i-1}(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-2})}{(x_i - x_0)(x_i - x_1) \dots (x_i - x_{i-1})}$$

Ce qui montre que a_i ne dépend que des points $x_0, \dots, x_i$, on pose alors pour tout $k \in \{1, \dots, n\}$

$$a_k = f[x_0, \dots, x_k]$$

Avec cette notation, on a

$$P(x) = f(x_0) + f[x_0, x_1](x - x_0) + \dots + f[x_0, \dots, x_n] \prod_{k=0}^{n-1} (x - x_k)$$

Lemme 2 : Soient $n \in \mathbb{N}$, et soient $x_0, x_1, \dots, x_n$ $n+1$ réels distincts. Pour toute permutation σ sur $\{0, \dots, n\}$ on a

$$f[x_0, \dots, x_n] = f[x_{\sigma(0)}, \dots, x_{\sigma(n)}]$$

3.3.1. Démonstration. On pose pour tout $j \in \{0, \dots, n\}$

$$y_j = x_{\sigma(j)}$$

On considère la décomposition de P dans les deux bases :

$$f, 1, (x - x_0), \dots, \prod_{k=0}^{n-1} (x - x_k) \text{ et } 1, x - y_0, \dots, \prod_{k=0}^{n-1} (x - y_k)$$

$$(a_k)_{0 \leq k \leq n}, (b_k)_{0 \leq k \leq n},$$

Les coefficients de P dans ces deux bases.

Alors, le coefficient du terme x^n est a_n dans la première décomposition et b_n dans la seconde. Par unicité du polynôme d'interpolation, on en déduit que $a_n = b_n$.

Lemme 2 on a la relation de récurrence

$$f[x_0, \dots, x_n] = \frac{f[x_1, \dots, x_n] - f[x_0, \dots, x_{n-1}]}{x_n - x_0}$$

On pose $y_j = x_{n-j}$

$$\text{On a } P(x) = a_0 + a_1(x - x_0) + \dots + a_n \prod_{k=0}^{n-1} (x - x_k)$$

$$= b_0 + b_1(x - y_0) + \dots + b_n \prod_{k=0}^{n-1} (x - y_k)$$

On sait déjà que $a_n = b_n$. on identifie les termes de degré $n - 1$. On trouve

$$a_{n-1} - a_n \sum_{k=0}^{n-1} x_k = b_{n-1} - b_n \sum_{k=0}^{n-1} y_k = b_{n-1} - a_n \sum_{k=0}^{n-1} x_{n-k} = b_{n-1} - a_n \sum_{k=1}^n x_k.$$

On a donc

$$a_{n-1} - b_{n-1} = a_n(x_0 - x_n)$$

or $a_{n-1} = f[x_0, \dots, x_{n-1}]$ et $b_{n-1} = f[x_1, \dots, x_n]$ d'où le resultat

3.4 Erreur d'interpolation

Soit P_n le polynôme d'interpolation de f aux points $x_0, x_1, \dots, x_n$. On pose :

$$e_n(x) = f(x) - P_n(x)$$

3.4.1. Proposition. Pour tout réel $x \in \mathbb{R}$, on a

$$e_n(x) = \begin{cases} x & \text{si } x \in \{x_0, \dots, x_n\} \\ f[x_0, \dots, x_n, x] \prod_{j=0}^n (x - x_j) & \text{sinon} \end{cases}$$

3.4.2. Démonstration .on suppose que $x \notin \{x_0, \dots, x_n\}$.soit le P_{n+1} le polynome d'interpolation de f aux points $\{x_0, \dots, x_n, x\}$. Alors

$$e_n(x) = f(x) - P_n(x) = P_{n+1}(x) - P_n(x) = f[x_0, \dots, x_n, x] \prod_{j=0}^n (x - x_j)$$

3.4.3. Théorème. Soient f une fonction définie sur $\mathbb{R}$ de classe C^k , et $x_0, \dots, x_k$ des réels distincts on pose $a = \min\{x_0, \dots, x_k\}$ et $b = \max\{x_0, \dots, x_k\}$.alors il existe $\varepsilon \in]a, b[$ tel que $f[x_0, \dots, x_k] = \frac{f^{(k)}(\varepsilon)}{k!}$ [17]

dans le cas ou $k = 1$, c'est une application du théorème des accroissemnts finis . supposons $k > 1$ Quelconque. soit P_k le polunome d'interpolation de f aux point $x_0, \dots, x_k$. alors la fonction

$$e_x(x) = f(x) - P_x(x)$$

s'annule au moins en $k + 1$ points distincts : $x_0, \dots, x_k$. Donc, d'après le théorème de Rolle, la fonction $(e_k)'$ s'annule au moins en k points dans $]a, b[$. La fonction $(e_k)''$ s'annule au moins en $k - 1$ points dans $]a, b[$. De même, la fonction $(e_k)^{(k)}$ s'annule en au moins un point dans $]a, b[$. Donc :

$$\exists \varepsilon \in]a, b[\text{ tel que } (e_k)^{(k)}(\varepsilon) = f^{(k)} - P_k^{(k)}(\varepsilon) = 0$$

$$\text{Or, } P_k^{(k)}(\varepsilon) = k! f[x_0, \dots, x_k]$$

Donc $\exists \varepsilon \in]a, b[$ tel $f[x_0, \dots, x_k] = \frac{f^{(k)}(\varepsilon)}{k!} \dots \dots \dots (3.5)$

Travaux dirigés II

Exercice 1 : On suppose que $f(x) = \sqrt[3]{x}$ et que $(x_0, y_0) = (0, 0)$, $(x_1, y_1) = (1, 1)$ et $(x_2, y_2) = (8, 2)$

Déterminer le polynôme $P_2(x)$ d'interpolation polynomiale qui passe par les points $(x_i, y_i)_{i=0,2}$

Exercice 2 : déterminer l'erreur d'interpolation dans le cas de l'interpolation parabolique

On approche la fonction $f(x)$ par la parabole passant par les points $(x_0, y_0) = (0, 1)$, $(x_1, y_1) = (1, 2)$ et $(x_2, y_2) = (2, 5)$

Exercice 3 : Soient $a_0, \dots, a_n$ des réels deux à deux distincts, pour tout $i = 0, \dots, n$ on pose

$$L_i(X) = \prod_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{X - a_j}{a_i - a_j}$$

Les L_i sont appelés polynômes interpolateurs de Lagrange. calculez $L_i(a_j)$

Soient $b_0, \dots, b_n$ de réels fixes. Montrez que $P(X) = \sum_{i=0}^n b_i L_i(X)$ est l'unique polynôme de degré inférieur ou égal à n vérifiant $P(a_j) = b_j$ pour tout

$j = 0, \dots, n$

Trouver le polynôme P de degré inférieur ou égal à 3 tel que

$$P(0) = 1 \text{ et } P(1) = 0 \text{ et } P(-1) = -2 \text{ et } P(2) = 4$$

on a

$$L_i(a_i) = \prod_{\substack{1 \leq j \leq n \\ j \neq i}} \frac{a_i - a_j}{a_i - a_j} = 1 \quad \text{et} \quad L_i(a_j) = 0 \text{ si } j \neq i$$

La solution de fiche TD II

Exercice 1

D'après la méthode de Lagrange.

$$P_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x)$$

$$= y_0 \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} + y_1 \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} + y_2 \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}$$

$$= 0 \frac{(x-1)(x-2)}{(0-1)(0-2)} + 1 \frac{(x-0)(x-8)}{(1-0)(1-8)} + 2 \frac{(x-0)(x-1)}{(8-1)(8-2)}$$

$$P_2(x) = 1 \frac{x(x-8)}{-7} + 2 \frac{x(x-1)}{56} = -\frac{3}{28}x^2 + \frac{31}{28}x$$

$$\text{On a bien } P_2(0) = 0, P_2(1)=1 \text{ et } P_2(8) = -\frac{3}{28}8^2 + \frac{31}{28}8 = 2$$

Calculer $P_2(x)$ et $f(x) = \sqrt[3]{x}$ pour $x = 0.5, 0.95, 1, 1.5, \dots$ et 3 on a

x	f(x)	$P_2(x) = -\frac{3}{28}x^2 + \frac{31}{28}x$
0.5	0.7937	0.52679
0.95	0.98305	0.95509
1	1	1
1.5	1.1447	1.4196
3	1.4422	$\frac{33}{14} = 2.3571$

Exercice 2

Le polynôme d'interpolation $P_2(x)$ de degré 2 tel que $P_2(x) = f(x_i)$, $i = 0, 1$ et 2

Avec $y_i = f(x_i)$ $i = 0, 1$ et 2, $(x_0, y_0) = (0, 1)$, $(x_1, y_1) = (1, 2)$ et $(x_2, y_2) = (2, 5)$

D'après la méthode de Lagrange

$$P_2(x) = y_0 L_0(x) + y_1 L_1(x) + y_2 L_2(x)$$

$$= 1 \frac{(x-1)(x-2)}{(-1)(-2)} + 2 \frac{(x)(x-2)}{(1)(-1)} + 5 \frac{(x)(x-1)}{(2)(1)}$$

$$= x^2 + 1$$

D'après le théorème précédent,

$$\begin{aligned} f(x) - P_2(x) &= \frac{f^{(3)}(C)}{3!} (x-x_0)(x-x_1)(x-x_2) \\ &= \frac{f^{(3)}(C)}{3!} x(x-1)(x-2) \end{aligned}$$

$$\text{si } |f^{(3)}(x)| \leq M \quad \text{alors}$$

$$\forall x \in [0, 2], |f(x) - P_2(x)|$$

$$|f(x) - P_2(x)| \leq \frac{M}{6} |x(x-1)(x-2)|$$

$$\leq \frac{M}{6} |x(x-1)(x-2)|$$

$$\leq 6.4 * 10^{-2} * M$$

le maximum de $u(x) = x(x-1)(x-2)$

$$\text{est atteint } x^* = \frac{3-\sqrt{3}}{3} ; d'où u(x^*) = \frac{1}{6} \frac{3-\sqrt{3}}{3} \left(\frac{3-\sqrt{3}}{3} - 1 \right) \left(\frac{3-\sqrt{3}}{3} - 2 \right) = 0.0$$

$$6415 \sim 6.4 * 10^{-2}$$

Exercice 3

puisque le produit contient un facteur qui est nul : $(a_j - a_j)$ Puisque les L_i sont tous de degré n , le polynôme P est de degré inférieur ou égal à n , et $P(a_j) = \sum_{i=0}^n b_i L_i(a_j) = b_j$

Il reste à montrer qu'un tel polynôme est unique. Supposons que Q convienne aussi, alors $P - Q$ est de degré inférieur ou égal à n et s'annule en $n+1$ points (les a_i), donc il est identiquement nul, i.e. $P = Q$

Pour l'application on utilise les polynômes interpolateurs de Lagrange avec $a_0 = 0, b_0 = 1 ; a_1 = 1, b_1 = 0 ; a_2 = -1, b_2 = -2 ; a_3 = 2, b_3 = 4$. on sait qu'un tel polynome $P(x)$ est unique et s'écrit

$$P(X) = 1.L_0(X) + 0.L_1(X) - 2.L_2(X) + 4.L_3(X)$$

$$\text{Ou } L_0(x) = \frac{(x-1)(x+1)(x-2)}{(0-1)(0+1)(0-2)} = \frac{1}{2}(x^3 - 2x^2 - x + 2)$$

$$L_1(x) = \frac{(x-0)(x+1)(x-2)}{(1-0)(1+1)(1-2)} = \frac{-1}{2}(x^3 - 2x^2 - 2x)$$

$$L_2(x) = \frac{(x-0)(x-1)(x-2)}{(-1-0)(-1-1)(-1-2)} = \frac{-1}{6}(x^3 - 3x^2 + 2x)$$

$$L_3(x) = \frac{(x-0)(x-1)(x+1)}{(2-0)(2-1)(2+1)} = \frac{1}{6}(x^3 - x)$$

$$P(x) = \frac{3}{2}x^3 - 2x^2 - \frac{1}{2}x + 1$$

Chapitre 04 : intégration numérique

4-1 Introduction :

Il s'agit d'une manière générale de déterminer, le mieux possible, pour l'essentiel, Une valeur approchée de l'intégrale au sens de Riemann.

$$I = \int_a^b f(x)dx \dots\dots\dots(4.1)$$

Ou a et b sont des réels.

Les problèmes de quadrature (intégration) numérique se rencontrent lorsque la fonction f est continue mais n'a pas de primitive explicite connue, ou lorsque la fonction f n'est donnée que par un nombre fini de couples (x_i, y_i) , $1 \leq i \leq n$.

Une idée importante consiste à utiliser les méthodes d'interpolation polynomiale, puisque les primitives des fonctions polynômes sont faciles à calculer.

On a fréquemment recours à ces techniques de quadrature numérique pour obtenir des solutions approchées d'équations différentielles (E.D.O. et E.D.P.).[16]

4.2 Intégration de type interpolation

[a, b] un intervalle, on note $(x_i)_i$ une subdivision ordonnée : $a = x_0 < x_1 < \dots < x_n = b$

4.2.1 Généralités

Définition 1: Une formule d'intégration à n + 1 points pour calculer une valeur approchée de I est une relation de la forme $J_n(f) = \sum_{k=0}^{n} A_{nk} f(x_k)$.

REMARQUE : Les coefficients A_{nk} ne dépendent bien sûr pas de f, ils servent pour tous les calculs d'intégrales relatifs à cet intervalle et à ces points de subdivision.

Polynôme d'interpolation de Lagrange Lorsque f est connue au moins en n+1 points (distincts) de subdivision et est de classe C^{n+1} , on écrit facilement le polynôme d'interpolation de Lagrange P_n de f et l'erreur,

$$f(x) = P_n(x) + e(x) = \sum_{i=0}^{n} A_i(x) f(x_i) + \frac{1}{(n+1)!} L(x) f^{(n+1)}(C_x)$$

$$\text{avec } L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x - x_j}{x_i - x_j} \text{ et } L(x) = \prod_{0 \leq i \leq n} (x - x_i) \text{ ou } C_x \in [a, b]$$

Ainsi l'intégrale du polynôme de Lagrange donne :

$$\int_a^b P_n(x) dx = \sum_{i=0}^n \left(\int_a^b L_i(x) dx \right) f(x_i) = J_n(f) \dots\dots\dots(4.2)$$

4.2.1. Définition La formule $J_n(f)$ est une formule de quadrature de type interpolation

pour (1) On mesure la précision des quadratures numériques à l'aide des notions :

- L'erreur d'intégration est le réel $E(f) = I - J_n(f)$
- Une formule de quadrature est dite exacte sur F s.e.v. de $C^\circ[a, b]$ si $\forall g \in F \ E(g) = 0$
- Une formule de quadrature a un degré de précision $p \in \mathbb{N}$ si elle est exacte sur l'espace vectoriel des polynômes IP_p (de degré $\leq P$), avec $E(x^{p+1}) \neq 0$

Ces définitions sont reliées entre elles !

4.2.2. Proposition . L'erreur d'intégration, pour une formule de type interpolation peut se calculer en

$$E(f) = \frac{f^{(n+1)}(\theta)}{(n+1)!} \int_a^b L(x) dx \dots\dots\dots (4.3)$$

pour un certain θ , lorsque le polynôme L reste de signe constant sur $[a, b]$.

4.2.3. Théorème. Une formule de quadrature à $n + 1$ points est exacte sur IP_n si et seulement si elle est de type interpolation

4.2.4 Quadratures simples

On retrouve les formules classiques (FAIRE DES FIGURES!):

Méthode des rectangles : on utilise la valeur de f en un seul point, voici deux exemples

$$I = (b - a)f(a) + \frac{(b - a)^2}{2} f'(\theta) = (b - a)f(b) + \frac{(b - a)^2}{2} f'(\theta)$$

Méthode des trapèzes : formule de quadrature à 2 points

$$I = (b - a) + \frac{f(a)+f(b)}{2} - \frac{(b-a)^3}{12} f^{(2)}(\hat{\theta}) \dots\dots\dots (4.4)$$

Ces formules (illustrer par quelques figures) s'obtiennent, par exemple, par la méthode des coefficients indéterminés (pour les coefficients An_k), et en utilisant la Proposition ci-dessus ou un développement de Taylor judicieux (pour l'erreur)...en supposant la régularité adéquate des fonctions f . [17]

4.3 Formules de Newton-Cotes

Ce sont des formules de quadrature de type interpolation avec subdivision régulière.

- Si les deux extrémités de l'intervalle sont des points d'interpolation il s'agit de Newton-Cotes fermé (méthodes des trapèzes, de Simpson...)
- Si les deux bornes de l'intervalle d'intégration ne sont pas des points d'interpolation il s'agit de Newton-Cotes ouvert (méthode de Poncelet...)

La régularité de la subdivision permet d'obtenir des formules qui sont très générales.

4.3.1. Théorème. Pour une subdivision régulière en n parties de l'intervalle $[a, b]$, la formule générale de Newton-Cotes fermé pour $\int_a^b f(x)dx$ est

$$I_n(f) = (b-a) \sum_{j=0}^{j=n} B_{n,j} f(a+jh) \text{ avec } h = \frac{b-a}{n}$$

Ou les coefficients sont

$$B_{n,j} = \frac{(-1)^{n-j}}{j!(n-j)!n} \int_0^n \prod_{k \neq j} (t-k) dt = B_{n,n-j} \dots\dots\dots (4.5)$$

4.3.2. Proposition Dans les formules de Newton-Côtes fermé de pas $h = (b-a)/n$, le degré de précision est $n+3$ si n est pair $n+2$ si n est impair.

4.3.3. Conséquence : si on a le choix une formule avec un nombre impair de points (c'est à dire une formule "centrée") est préférable.

4.4. Méthode des trapèzes

Soit f une fonction continue sur $[a, b]$, dérivable sur $]a, b[$ et $a = x_0$

$a = x_0 < x_1 < \dots < x_{n-1} < x_n = b$ Une subdivision régulière de l'intervalle

$[a, b]$. On note h le pas de cette subdivision. Dans la méthode des trapèzes,

La fonction f est remplacée sur chaque intervalle $[x_i, x_{i+1}]$ par la droite

Joignant les points soit $(x_i, f(x_i))$ et $(x_{i+1}, f(x_{i+1}))$, soit

$$h(x) = \frac{(x-x_i)f(x_{i+1}) - (x-x_{i+1})f(x_i)}{x_{i+1} - x_i} \quad x \in [x_i, x_{i+1}]$$

La méthode s'écrit
$$\int_a^b f(x) dx \approx \sum_{i=0}^{n-1} (x_{i+1} - x_i) \frac{f(x_i) + f(x_{i+1}))}{2}$$

Lorsque la subdivision se réduit a sa plus simple expression $x_0 = a, x_1 = b$ on a

$$\int_a^b f(x) dx \approx \frac{1}{2} (b-a) (f(a) + f(b))$$

La méthode des trapèzes est une méthode d'ordre 1. L'erreur dans la méthode des trapèzes est donnée par l'expression[18]

$$\left| \int_a^b f(x) dx - S \right| \leq \frac{1}{2} \frac{(b-a)^3}{n^2} \sup_{x \in [a,b]} |f''(x)|$$

La somme S s'exprime par

$$S = \frac{h}{2} (f(a) + f(b) + \sum_{i=1}^{n-1} f(x_i))$$

Pour améliorer la précision, on considère parfois la formule des trapèzes

Corrigée suivante

$$\int_a^b f(x) dx \approx \frac{h}{2} (f(a) + f(b) + \sum_{i=1}^{n-1} f(x_i)) - \frac{h^2}{12} (f'(b) - f'(a))$$

4.5. Méthode de Simpson

On sait que la méthode de Simpson consiste à remplacer, sur chaque intervalle $[x_i, x_{i+1}]$, la fonction f par un polynôme de degré 2 qui prend les mêmes valeurs que f en x_i, x_{i+1} et $\frac{x_i + x_{i+1}}{2}$

4.5.1.Problème: retrouver les coefficients !

1° Première méthode pour retrouver la formule

Idee Pour une raison un peu mystérieuse, la méthode de Simpson s'obtient à partir de la méthode des trapèzes par la méthode dite d'accélération de la convergence de Romberg.

Supposons que l'inégalité de la méthode des trapèzes soit un équivalent (on ne sait pas le démontrer, d'ailleurs c'est faux pour certaines fonctions f , mais ce n'est pas grave, car on indique seulement un procédé heuristique) (peut-être que $C = 1/12$, mais on s'en fiche) :

$$T_n \sim \frac{C}{n^2}$$

Pour faire intervenir $\frac{x_i^{(n)} + x_{i+1}^{(n)}}{2} = x_i^{(2n)}$, on veut faire jouer T_{2n} . plus précisément, on veut faire une combinaison linéaire de T_n et T_{2n} de la forme $S_n = \alpha T_n + \beta T_{2n}$ telle que :

$$\lim_{n \rightarrow +\infty} S_n = \int_a^b f(x) dx \text{ (c'est le minimum !)}$$

$$S_n = 0 \left(\frac{1}{n^2} \right) \text{ (pour pouvoir dire qu'on a accéléré la convergence)}$$

La première condition impose $\alpha + \beta = 1$. pour la deuxième, constatons que :

$$T_{2n} \sim \frac{C}{4n^2} \dots \dots \dots (4.6)$$

Et donc $\alpha T_n + \beta T_{2n} = \alpha \frac{c}{n^2} + \beta \frac{c}{4n^2} + 0 \left(\frac{1}{n^2} \right) = C \frac{\alpha + \frac{\beta}{4}}{n^2} + 0 \left(\frac{1}{n^2} \right)$

On est amené à prendre $\alpha + \frac{\beta}{4} = 0$. Ainsi, on est conduit à choisir

$$\begin{cases} \alpha + \beta = 1 \\ \alpha + \frac{\beta}{4} = 0 \end{cases} \text{ ou } \begin{cases} \alpha = -\frac{1}{3} \\ \beta = \frac{4}{3} \end{cases} \text{ ce qui donne } S_n = \frac{4T_{2n} - T_n}{3}$$

Il se trouve que « c » est la méthode de Simpson.

Remarque Vu qu'on a supprimé le terme en $1/n^2$, on s'attend à ce que la méthode donne un terme d'erreur en $1/n^3$. En réalité, on verra que l'erreur est en $1/n^4$.

4.5.2. Deuxième méthode pour trouver la formule

Lemme (quasiment inutile) Etant donné trois réels distincts a, b, c et trois réels quelconque y_a, y_b, y_c il existe un unique polynôme P de degré au plus 2 tel que

$$P(a) = y_a, P(b) = y_b, P(c) = y_c$$

On note $R_2[X]$ l'espace des polynômes de degré au plus 2. Pour $x \in R$, l'évaluation $ev_x : P \rightarrow P(x)$ est une forme linéaire sur cet espace. Il en résulte que l'application.

$$\begin{aligned} R_2[X] &\rightarrow R^3 \\ P &\rightarrow (P(a), P\left(\frac{a+b}{2}\right), P(b)) \end{aligned}$$

est linéaire. Or, elle est injective, car un polynôme de degré au plus 2 ayant trois racines distinctes est nécessairement nul. Par égalité des dimensions à la source et au but, il en résulte qu'elle est bijective. Le lemme en résulte.

Proposition (Formule des trois niveaux) Etant donné deux réels $a < b$ et $c = \frac{a+b}{2}$, il existe un unique triplet $(\alpha, \gamma, \beta) \in R^3$ tel que pour tout polynôme P de degré au plus 2,

$$\frac{1}{b-a} \int_a^b P(x) dx = \alpha P(a) + \gamma P\left(\frac{a+b}{2}\right) + \beta P(b) \dots\dots\dots (4.7)$$

de plus, on a : $\alpha = \dots, \beta = \dots, \gamma = \dots$ (Dur à se rappeler!)

Idée On comprend la formule de la proposition comme une expression de la forme linéaire

$$I: R_2[X] \rightarrow R, P \rightarrow \frac{1}{b-a} \int_a^b P$$

comme combinaison linéaire de (ev_a, ev_c, ev_b) :

On reprend les notations du lemme. D'après la preuve du lemme, la famille (ev_a, ev_c, ev_b) est une famille libre, donc c'est une base du dual de $R_2[X]$. D'où l'existence et l'unicité de α, γ, β coefficients de la forme linéaire $P \rightarrow \frac{1}{b-a} \int_a^b P$ dans cette base.

Pour calculer α, γ, β on teste sur la base $1, X, X^2$ de $R_2[X]$:

$$\begin{aligned} P = 1 \quad 1 &= \frac{1}{b-a} \int_a^b 1 \, dx = \alpha + \gamma + \beta \\ P = X \quad \frac{a+b}{2} &= \frac{b^2 - a^2}{2(b-a)} = \frac{1}{b-a} \int_a^b x \, dx = \alpha a + \gamma \frac{a+b}{2} + \beta b \\ P = X^2 \quad \frac{b^3 - a^3}{3(b-a)} &= \frac{1}{b-a} \int_a^b x^2 \, dx = \alpha a^2 + \gamma \left(\frac{a+b}{2}\right)^2 + \beta b^2 \end{aligned}$$

On récrit la dernière égalité $\frac{a^2+ab+b^2}{3} = \left(\alpha + \frac{\gamma}{4}\right) a^2 + \frac{\gamma}{2} ab + \left(\frac{\gamma}{4} + \beta\right) b^2$

En identifiant coefficient par coefficient (ce qui n'est pas justifié mais fonctionne), on trouve une solution, dont on vérifie qu'elle fonctionne ; c'est la seule d'après le début de la preuve :

$$\gamma = \frac{2}{3}, \quad \alpha = \beta = \frac{1}{6}$$

4.5.3 Formule pour la méthode de Simpson

Maintenant, on l'applique sur $[x_i^{(n)}, x_{i+1}^{(n)}]$ et on fait la somme sur i : le coefficient associé à

$\varepsilon = \frac{x_i^{(n)} + x_{i+1}^{(n)}}{2}$ est $\gamma = \frac{2}{3}$, car en apparaît que dans un seul intervalle.

Le coefficient associé à $x_i^{(n)}$ est $\alpha + \beta = \frac{1}{3}$ si $1 \leq i \leq n-1$, car $x_i^{(n)}$ apparaît une fois comme extrémité gauche, et une fois comme extrémité droite d'un intervalle ; et c'est $\alpha = \beta = \frac{1}{6}$ si $i = 0$ ou $i = n$. En d'autres termes, on pose :

$$S_n = \frac{b-a}{n} \left[\frac{1}{6} f(x_0^{(n)}) + \frac{1}{3} \sum_{i=1}^{n-1} f(x_i^{(n)}) + \frac{2}{3} \sum_{i=0}^{n-1} f\left(\frac{x_i^{(n)} + x_{i+1}^{(n)}}{2}\right) + \frac{1}{6} f(x_n^{(n)}) \right] \dots \dots \dots (4.8)$$

Remarque Il se trouve que la formule des trois niveaux est en fait valable pour les polynômes de degré 3 également. C'est ce qui explique que l'erreur est en $1/n^4$ et non en $1/n^3$.

4 majorations de l'erreur

Proposition si f est C^4 sur $[a, b]$ et $u_4 = \sup_{[a,b]} |f^{(4)}|$, on a

$$\left| \int_a^b f - S_n \right| \leq \frac{u_4(b-a)^5}{2880n^4} \dots \dots \dots (4.9)$$

(La valeur exacte 2 880 n'est pas indispensable. Il est plus intéressant de savoir expliquer comment retrouver les coefficients à partir d'une idée simple.)

Idée de la preuve

Pour $n = 1$, on part de l'expression : $\int_a^b (x-a)^2 (b-x)^2 f^{(4)}(x) dx$,

Et quatre intégrations par parties montrent que c'est (à peu près) l'erreur

$$\int_a^b f - S_1,$$

4.6 Formules de Gauss

Dans les méthodes précédentes, nous avons vu qu'en changeant les coefficients de pondération des valeurs de la fonction μ à intégrer aux abscisses régulièrement espacées, on pouvait grandement améliorer la convergence de la méthode. Les méthodes de Gauss ajoutent aux méthodes précédentes de pouvoir utiliser des abscisses non régulièrement espacées.

Soit $W(x)$ une fonction strictement positive sur l'intervalle $[a; b]$, appelée fonction de poids, on choisit une suite de points x_i telle que l'intégrale soit approchée par une somme discrète de la forme

$$\int_a^b W(x)f(x)dx = \sum_{i=1}^N w_i f_i \dots\dots\dots(4.10)$$

Quand les abscisses sont régulièrement espacées, les coefficients inconnus sont les poids w_i ; cela implique que pour N points d'intégration, on peut obtenir une évaluation exacte de l'intégrale jusqu'à un polynôme de degré $N-1$ [19]

Si N est pair (trapèzes) , si N est impair (Simpson). Les méthodes de Gauss utilisent le fait si les abscisses et les poids sont des inconnues à déterminer ; la formule d'intégration de Gauss à N points devient exacte jusqu'à des polynômes de degré $2N - 1$, ce qui augmente la précision de l'évaluation sans qu'il soit

Nécessaire d'augmenter le nombre de points à calculer.

Pour déterminer ces $2N$ paramètres, on s'appuie sur la construction de polynômes orthogonaux. Le principe de cette construction remonte à Gauss et Jacobi et a été largement développé par Christoffel. Soit un intervalle $(a; b)$, on introduit le produit scalaire de deux fonctions f et g avec la fonction de poids

$$W \text{ par : } \langle f/g \rangle \equiv \int_a^b W(x)f(x)dx$$

Les fonctions sont dites orthogonales si leur produit scalaire est nul. La fonction est normalisée quand $\langle \frac{f}{g} \rangle = 1$. Un ensemble orthonormé de fonctions est un ensemble de fonctions toutes normalisées et orthogonales deux à deux.

On peut construire de manière systématique une suite de polynômes telle que le coefficient du monôme de degré le plus élevé soit égal μ à un et telle qu'ils soient tous orthogonaux. [20]

La relation de récurrence est la suivante.

$$\begin{aligned} p_{-1}(x) &= 0 \\ p_0(x) &= 1 \\ p_{i+1}(x) &= (x - a_i)p_i(x) - b_i p_{i-1}(x) \end{aligned}$$

avec les coefficients a_i et b_i déterminés de la manière suivante

$$\begin{aligned} a_i &= \frac{\langle x p_i / p_i \rangle}{\langle p_i / p_{i-1} \rangle} \\ b_i &= \frac{\langle x p_i / p_i \rangle}{\langle p_{i-1} / p_{i-1} \rangle} \end{aligned}$$

Si l'on divise chaque polynôme par p_i , on obtient alors des polynômes normalisés.

Une propriété fondamentale de ces polynômes ainsi construits est la suivante : Le polynôme p_i a exactement i racines distinctes placées sur l'intervalle $[a; b]$. Chaque racine du polynôme p_i se trouve entre deux racines consécutives du polynôme p_{i+1} .

Les N racines du polynôme p_N sont choisis comme abscisses dans l'évaluation de l'intégrale de Gauss et les poids w_i sont calculés à partir de la formule.

$$\begin{aligned} w_i &= \frac{\langle p_{N-1} \left| \frac{1}{x-x_i} \right| p_{N-1} \rangle}{p_{N-1}(x_i) p'_N(x_i)} \\ &= \frac{d_N}{d_{N-1}} \frac{\langle p_{N-1} / p_{N-1} \rangle}{p_{N-1}(x_i) p'_N(x_i)} \end{aligned}$$

Où le symbole prime désigne la dérivée du polynôme et d_N le coefficient du monôme le plus élevé du polynôme p_N .

Avec ce choix, on peut montrer en utilisant la relation de récurrence, Eq. (1.18),

$$\langle \frac{p_i}{p_i} \rangle = 0$$

A titre d'exemples voici les fonctions de poids classiques que l'on utilise pour calculer une intégrale par la méthode de Gauss

Gauss-Legendre La fonction de poids est $W = 1$ sur l'intervalle $[-1; 1]$. La relation de récurrence pour les polynômes de Legendre est

$$(i + 1)P_{i+1} = (2i + 1)xP_i - iP_{i-1}$$

4.6.1. Gauss-Hermite

La fonction de poids est $W = \exp(-x^2)$ sur la droite réelle. La relation de récurrence pour les polynômes d'Hermite est

$$H_{i+1} = 2xT_i - 2iT_{i-1} \dots\dots\dots(4.11)$$

4.6.2. Gauss-Laguerre

La fonction de poids est $W = x^\alpha e^{-x}$ sur l'intervalle $[0, +\infty[$. La relation de récurrence pour les polynômes de Laguerre est

$$(i+1)L_{i+1}^\alpha = (-x + 2i + \alpha + 1)L_i^\alpha - (i + \alpha)L_{i-1}^\alpha \dots\dots\dots(4.12)$$

4.6.3. Gauss-Jacobi

La fonction de poids est $W = (1-x)^\alpha(1+x)^\beta$ sur l'intervalle $]-1; 1[$. La relation de récurrence pour les polynômes de Jacobi est

$$c_i P_{i+1}^{(\alpha, \beta)} = (d_i + e_i x) P_i^{(\alpha, \beta)} - f_i P_{i-1}^{(\alpha, \beta)}$$

ou les coefficients c_i , d_i , e_i et f_i sont donnés par les relations

$$\begin{aligned} c_i &= 2(i+1)(i+\alpha+\beta+1)(2i+\alpha+\beta) \\ d_i &= (2i+\alpha+\beta+1)(\alpha^2 - \beta^2) \\ e_i &= (2i+\alpha+\beta)(2i+\alpha+\beta+1)(2i+\alpha+\beta+2) \dots\dots\dots(4.13) \\ f_i &= 2(i+\alpha)(i+\beta)(2i+\alpha+\beta+2) \end{aligned}$$

On peut utiliser des fonctions de poids qui sont intégrables sur l'intervalle, sans être nécessairement bornées.

4.6.4. Gauss-Tchebychev

La fonction de poids est $W = (1-x^2)^{-1/2}$ sur l'intervalle $[-1; 1]$. La relation de récurrence pour les polynômes de Tchebychev.

$$T_{i+1} = 2xT_i - 2iT_{i-1} \dots\dots\dots(4.14)$$

4.7. Méthode de Romberg (1955)

Désignons la somme des aires des n trapèzes de la méthode des trapèzes par $S_0(n)$. Lorsque $n \rightarrow \infty$, la suite $S_0(n)$ converge en général assez lentement vers la valeur exacte de l'intégrale. La méthode de Romberg permet d'accélérer considérablement la convergence.

L'idée est la suivante :

Selon la formule des trapèzes améliorée, $S_0(n)$ est liée à la valeur exacte de l'intégrale par une relation de la forme :

$$S_0(n) = \int_a^b f(x) dx + Ch^2 + O(h^4) \quad , \quad h = \frac{b-a}{n}$$

où la constante C ne dépend pas de h .

Si on double le nombre d'intervalles on a :

$$S_0(2n) = \int_a^b f(x)dx + C \frac{h^2}{4} + O(h^4) \quad , \quad h = \frac{b-a}{n}$$

Par élimination de l'inconnue C on obtient :

$$\int_a^b f(x)dx = \frac{4S_0(2n) - S_0(n)}{3} + O(h^4) \quad \dots\dots\dots(4.15)$$

On est donc passé d'une précision en h_2 (formule des trapèzes) à une précision en h_4 .

Introduisons une nouvelle suite :

$$S_1(2n) = \frac{4S_0(2n) - S_0(n)}{3} \quad , \quad n = 1, 2, 4, 8, \dots$$

Cette suite converge plus vite vers la valeur de l'intégrale que la suite $S_0(n)$ (si $C \neq 0$).

On voit aisément que $S_1(2n)$ est égale à l'approximation fournie par la méthode de Simpson pour une subdivision de $[a, b]$ en $2n$ intervalles égaux.

On a, en particulier :

$$S_1(2) = \frac{b-a}{6} [f(a) + 4f(\frac{a+b}{2}) + f(b)]$$

Il en résulte donc que :

$$S_1(n) = \int_a^b f(x)dx + C^* h^4 + O(h^6) \quad , \quad h = \frac{b-a}{n}$$

Où C^* est une constante indépendante de h .

En procédant comme ci-dessus, c'est-à-dire en doublant le nombre de sous-intervalles et en éliminant l'inconnue C^* on trouve :

$$\int_a^b f(x)dx = \frac{16S_1(2n) - S_1(n)}{15} + O(h^6)$$

On est ainsi passé d'une précision en h_4 à une précision en h_6 , la suite

$$S_2(2n) = \frac{16S_1(2n) - S_1(n)}{15} \quad , \quad n = 2, 4, 8, \dots$$

Converge donc en général plus vite vers la valeur exacte de l'intégrale.

Cette démarche se généralise (pour une fonction suffisamment dérivable) et donne

l'algorithme de Romberg :

Pour $n = 1, 2, 4, 8, \dots$ calculer :

$$S_0(n) = \frac{h}{2}[f_0 + 2f_1 + 2f_2 + \dots + 2f_{n-1} + f_n], h = \frac{b-a}{n}$$

Pour $n = 2m-1, 2m, 2m+1, \dots$; $m = 1, 2, 3, \dots$

calculer :

$$S_m(2n) = \frac{4^m S_{m-1}(2n) - S_{m-1}(n)}{4^m - 1}$$

Remarque : pour obtenir une précision élevée en utilisant peu de termes de la suite $S_0(n)$ il suffit de prendre des valeurs de n plus grandes en respectant toutefois leur progression géométrique, par exemple en choisissant $n = 100, 200, 400, 800, 1600, 3200$.

Travaux dirigés III

Exercice 1 : Déterminer par la méthode des trapèzes puis par celle de Simpson $\int_0^{\frac{\pi}{2}} f(x) dx$ sur la base du tableau suivant :

x	0	$\frac{\pi}{8}$	$\frac{\pi}{4}$	$\frac{3\pi}{8}$	$\frac{\pi}{2}$
f(x)	0	0.382683	0.707107	0.923880	1

Ces points d'appui sont ceux donnant $\sin x$, comparer alors les résultats obtenus avec la valeur exacte.

Exercice 2 : Calculer à l'aide de la méthode des trapèzes l'intégrale

$$I = \int_0^{\pi} \sin x^2 dx \text{ avec le nombre de points d'appui } n = 5 \text{ puis } n = 10.$$

Exercice 3 : Trouver le nombre n de subdivisions nécessaires de l'intervalle d'intégration $[-\pi, \pi]$, pour évaluer $0.5 \cdot 10^{-3}$ près, grâce à la méthode de Simpson, l'intégrale $I = \int_{-\pi}^{\pi} \cos x dx$

Exercice 4 : méthode des trapèzes $I = \int_2^1 \frac{1}{x} dx$

On considère l'intégrale

1. Calculer la valeur exacte de I .
2. Évaluer numériquement cette intégrale par la méthode des trapèzes avec $n = 3$ sous-intervalles.
3. Pourquoi la valeur numérique obtenue à la question précédente est-elle supérieure à $\ln(2)$? Est-ce vrai quelque soit n ? Justifier la réponse. (On pourra s'aider par un dessin.)
4. Quel nombre de sous-intervalles n faut-il choisir pour avoir une erreur inférieure à 10^{-4} ? On rappelle que l'erreur de quadrature associée s'écrit, si $f \in C^2([a; b])$,

$$|E_n| = \left| \frac{(b-a)^4}{12n^2} f''(\varepsilon) \right|, \varepsilon \in]a; b[.$$

Exercice 5 Soit f une fonction $C^\infty(\mathbb{R}, \mathbb{R})$ On se donne les points $\{x_i\}_{i=0}^{2n}$ de subdivision de l'intervalle $[a, b]$: $x_i = a + ih$ avec $h = \frac{b-a}{2n}$

Le but de l'exercice est de trouver une formule de quadrature à $2n + 1$ basée sur la formule de Simpson pour approcher $\int_a^b f(x)dx \dots \dots \dots (2)$

On propose dans un premier temps (question 1 à 5) de construire la formule de quadrature à 3 points de Simpson $\int_{-1}^1 g(x)dx \approx \alpha g(-1) + \beta g(0) + \alpha g(1) \dots \dots (3)$

où les réels α et β sont à déterminer.

1. Sous quelle condition (portant sur α et β) la formule de quadrature (3) est exacte pour une fonction g constante ?
2. Sous quelle condition (portant sur α et β) la formule de quadrature (3) est exacte pour une fonction g polynomiale de degré au plus 2 ?
3. En déduire le choix de α et β rendant la formule de quadrature (3) exacte pour une fonction g polynomiale de degré au plus 2.
4. La formule de quadrature est-elle exacte pour tout polynôme de degré 3 ? La formule de quadrature est-elle exacte pour tout polynôme de degré 4 ?

La solution de fiche TD III

Exercice 1

$$I = \int_0^{\frac{\pi}{2}} f(x)dx$$

Soit T l'approximation de I par la méthode des trapèzes, le pas h donné par $h = \frac{x_n - x_0}{n} = \frac{\pi}{8}$

$$\begin{aligned} T &= \frac{h}{2} (f(x_0) + f(x_4) + 2 \sum_{i=1}^3 f(x_i)) \\ &= \frac{\pi}{16} (0 + 1 + 2(0.382683 + 0.707107 + 0.92388)) \\ &= 0.987116 \end{aligned}$$

Soit S l'approximation de I par la méthode des simpson. celle ci s'écrit

$$\begin{aligned}
S &= \frac{h}{3} (y_0 + y_4 + 4(y_1 + y_3) + 2y_2) \\
&= \frac{\pi}{8} \frac{1}{3} [(0 + 1 + 4(0.38 \dots + 0.92 \dots) + 2 \times 0.707 \dots)] \\
&= 1.000135
\end{aligned}$$

Les points d'appui donnés dans cet exercice correspondent à la fonction $f(x) = \sin x$ et $I = \int_0^{\pi/2} \sin x \, dx = 1$. On constate donc que l'approximation de I donnée par la méthode de Simpson est meilleure que celle par les trapèzes, puisque $|S - I| = 0.000135$ et $|T - I| = 0.012884$.

Exercice 2 : soit $I = \int_0^{\pi} \sin x^2 \, dx$

$n = 5$ donc le pas d'intégration est $h = \frac{\pi}{5}$. Calculons I par la méthode des trapèzes.

$$\begin{aligned}
I &= \frac{h}{2} (f(x_0) + f(x_n) + 2 \sum_{i=1}^{n-1} f(x_i)) \\
&= \frac{\pi}{10} (0 + 1 + 2 (\sin(\pi)^2 + \sin 0 + 2 \sin\left(\frac{\pi}{5}\right)^2 + \sin\left(\frac{2\pi}{5}\right)^2 + \sin\left(\frac{3\pi}{5}\right)^2 + \sin\left(\frac{4\pi}{5}\right)^2)) \\
&= 0.504431
\end{aligned}$$

2. $n = 10$ donc le pas d'intégration est $h = \frac{\pi}{10}$.

$$\begin{aligned}
&= \frac{\pi}{10} (0 + 1 + 2 (\sin(\pi)^2 + \sin 0 + 2 \sin\left(\frac{\pi}{10}\right)^2 + \sin\left(\frac{2\pi}{10}\right)^2 + \sin\left(\frac{3\pi}{10}\right)^2 + \sin\left(\frac{4\pi}{10}\right)^2 + \sin\left(\frac{5\pi}{10}\right)^2 + \sin\left(\frac{6\pi}{10}\right)^2 + \sin\left(\frac{7\pi}{10}\right)^2 + \sin\left(\frac{8\pi}{10}\right)^2 + \sin\left(\frac{9\pi}{10}\right)^2)) \\
&= 0.0722338
\end{aligned}$$

alors que la valeur « exacte » est approximativement 0, 772651. Avec ce pas plus petit l'approximation numérique est meilleure.

Exercice 3 : soit $I = \int_{-\pi}^{\pi} \cos x \, dx$

Le pas d'intégration est $h = \frac{b-a}{n} = \frac{2\pi}{n}$. D'autre part l'erreur théorique sur la méthode de Simpson est donnée par

$$E(h) = \frac{-(b-a)}{180} h^4 f^{(4)}(\xi) = \frac{-2\pi}{180} \left(\frac{2\pi}{n}\right)^4 \cos(\xi)$$

Où $\xi \in [a, b]$, par conséquent, $|E(h)| = \frac{2\pi}{180} \left(\frac{2\pi}{n}\right)^4$

Ainsi pour que

$|E(h)| \leq 0.5 \cdot 10^{-3}$ il suffit que n vérifie

$\frac{\pi}{90} \frac{16\pi^4}{n^4} \leq 0.5 \cdot 10^{-3}$, donc $n^4 \geq \frac{1}{0.5 \cdot 10^{-3}} \frac{\pi}{90} 16\pi^4$. Ainsi n vérifie $n \geq 18.6$ on prendra par exemple

$n = 20$, car pour la méthode de Simpson, le nombre de subdivisions de l'intervalle $[a, b]$ doit toujours être pair

Exercice 4 : On considère l'intégrale $I = \int_1^2 \frac{1}{x} dx$

1. La valeur exacte est $I = [\ln(x)]_{x=1}^{x=2} = \ln(2)$

2. La méthode des trapèzes composite à n points pour calculer l'intégrale d'une f sur l'intervalle $[a, b]$ s'écrit $\int_a^b f(t) dt \cong h(\frac{1}{2}f(a) + \sum_{i=1}^{n-1} f(a + ih) + \frac{1}{2}f(b))$

Avec $h = \frac{b-a}{n}$

On a $f(x) = \frac{1}{x}$, $a = 1$, $b = 2$, $n = 3$ d'où $h = \frac{1}{3}$ et on obtient

$$I \cong \frac{1}{3} \left(\frac{1}{2}f(1) + f\left(1 + \frac{1}{3}\right) + f\left(1 + \frac{2}{3}\right) + \frac{1}{2}f(2) \right) = \frac{1}{3} \left(\frac{1}{2} + \frac{3}{4} + \frac{3}{5} + \frac{1}{4} \right) = \frac{21}{30} = 0.7$$

3. La valeur numérique obtenue à la question précédente est supérieure à $\ln(2)$ car la fonction $f(x) = \frac{1}{x}$ est convexe. On peut se convaincre à l'aide d'un dessin que les trapèzes sont au-dessus de la courbe $y = \frac{1}{x}$ l'aire sous les trapèzes sera donc supérieure à l'aire sous la courbe. Pour bien visualiser la construction considérons $n = 1$:

Cela reste vrai quelque soit le pas h choisi car la fonction est convexe ce qui signifie qu'une corde définie par deux points de la courbe $y = \frac{1}{x}$ sera toujours au-dessus de la courbe et par le raisonnement précédant l'aire sous les trapèzes sera supérieure à l'aire

4. L'erreur est majorée par

$$|E| \leq \frac{(b-a)^4}{12n^2} \sup_{\epsilon \in]1,2[} f''(\epsilon) \quad \text{on a que } |E| \leq \frac{1}{12n^2} \max_{\epsilon \in]1,2[} \frac{2}{\epsilon^3} = \frac{1}{6n^2}$$

Pour que $|E| < 10^{-4}$ il faut que $\frac{1}{6n^2} < 10^{-4}$

i.e. $n > \frac{10^2}{\sqrt{6}} \approx 40.8$. A partir de 41 sous intervalles, l'erreur de quadrature

est inférieure à 10^{-4}

Exercice 5 : 1. Soit $g(x) = c$ alors on a

$$\int_{-1}^1 g(x) dx = \alpha g(-1) + \beta g(0) + \alpha g(1)$$

$$\nabla$$

$$\frac{c[x]_{-1}^1}{2c} \alpha c + \beta c + \alpha c$$

$$(2\alpha + \beta)c$$

d'où la relation $2 = (2\alpha + \beta)$

Soit $g(x) = c + dx + ex^2$ alors on a

$$\int_{-1}^1 g(x) dx = \alpha g(-1) + \beta g(0) + \alpha g(1)$$

$$\nabla$$

$$\left[cx + d \frac{x^2}{2} + e \frac{x^3}{3} \right]_{-1}^1 = \alpha(c - d + e) + \beta c + \alpha(c + d + e)$$

$$2c + \frac{2}{3}e = (2\alpha + \beta)c + \alpha e$$

D'où les relations $\begin{cases} 2 = (2\alpha + \beta) \\ \frac{2}{3} = \alpha \end{cases}$

On dispose de 2 équations et 2 inconnues .si on résout le système linéaire

$$\begin{cases} 2 = (2\alpha + \beta) \\ \frac{2}{3} = \alpha \end{cases}$$

On obtient l'unique solution $\alpha = \frac{1}{3}$, $\beta = \frac{4}{3}$

On a retrouvé la formule de Simpson :

$$\int_{-1}^1 g(x) dx \approx \frac{1}{3} [g(-1) + 4g(0) + g(1)]$$

On sait (cf. cours) que cette formule est exacte pour les polynômes de degré au plus 3 comme on va vérifier ci-dessous.

Soit $g(x) = c + dx + ex^2 + fx^3$ alors on a

$$\int_{-1}^1 g(x) dx = \left[cx + d \frac{x^2}{2} + e \frac{x^3}{3} + f \frac{x^4}{4} \right]_{-1}^1 = 2c + \frac{2}{3}e$$

$$\frac{1}{3} [g(-1) + 4g(0) + g(1)] = \frac{1}{3} [(c - d + e - f) + 4(c) + (c + d + e + f)] = 2c + \frac{2}{3}e$$

donc la formule de quadrature est exacte pour tout polynôme de degré 3.

On vérifie qu'elle n'est pas exacte pour les polynômes de degré 4 : soit $g(x) = x^4$, alors

$$\int_{-1}^1 g(x) dx = \left[\frac{x^5}{5} \right]_{-1}^1 = \frac{2}{5}$$

CHAPITRE 5 Résolution de systèmes linéaires

5.1. Méthode Gauss

L'idée de base consiste à transformer (1) en un problème que l'on sait résoudre.

(1) Si la matrice $A = D$ avec D une matrice diagonale, alors on sait résoudre (1).

(2) Si la matrice $A = U$ (ou L) avec U (ou T) triangulaire supérieure (ou inférieure) alors on sait résoudre (1). [15]

5.1.1. Problème : Comment transformer une matrice en une matrice triangulaire inférieure ou supérieure

Cas $n = 3$:

$AX = b$ s'écrit :

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \end{cases} \quad (A \ b) = \begin{pmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ a_{21} & a_{22} & a_{23} & b_2 \\ a_{31} & a_{32} & a_{33} & b_3 \end{pmatrix} \dots\dots\dots (5.1)$$

on suppose que $a_{11} \neq 0$, par élimination, on obtient

$$(A_1 \ b_1) = \begin{pmatrix} a_{11} & a_{12} & a_{13} & b_1 \\ 0 & \hat{a}_{22} & \hat{a}_{23} & \hat{b}_2 \\ 0 & \hat{a}_{32} & \hat{a}_{33} & \hat{b}_3 \end{pmatrix} \quad (1) \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \rightarrow (l_1) \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 = b_2 \rightarrow (l_2) \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 = b_3 \rightarrow (l_3) \end{cases}$$

On note par (li) la i ème équation du système précédent.

On suppose que $a_{11} \neq 0$ On pose :

$$(\hat{l}_2) = a_{11}(l_2) - a_{21}(\hat{l}_1) \text{ et } (\hat{l}_3) = a_{11}(l_3) - a_{31}(\hat{l}_1) \quad (2) \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \rightarrow (l_1) \\ \hat{a}_{22}x_2 + \hat{a}_{23}x_3 = \hat{b}_2 \rightarrow (\hat{l}_2) \\ \hat{a}_{32}x_2 + \hat{a}_{33}x_3 = \hat{b}_3 \rightarrow (\hat{l}_3) \end{cases}$$

On suppose que $\hat{a}_{22} \neq 0$ On pose : $(\hat{\hat{l}}_3) = \hat{a}_{22}(\hat{l}_3) - \hat{a}_{32}(\hat{l}_2)$ alors (2) s'écrit

$$(2) \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 = b_1 \rightarrow (l_1) \\ \hat{a}_{22}x_2 + \hat{a}_{23}x_3 = \hat{b}_2 \rightarrow (\hat{l}_2) \\ \hat{\hat{a}}_{33}x_3 = \hat{\hat{b}}_3 \rightarrow (\hat{\hat{l}}_3) \end{cases}$$

Remarque :

(1) Les termes diagonaux sont appelés les pivots

(2) Si un pivot a_{ii} est nul on change de ligne (on permute) de i à n (pivotage partiel)

(3) Cette méthode se généralise assez facilement bien qu'il faut être prudent avec le choix du pivot. En pratique, il faut éviter de prendre des pivots "trop" petits.

5.2. Méthode de Croûte

Pour calculer explicitement les matrices L et U, on a intérêt à procéder par substitution : c'est la méthode de Crout. Ecrivons le produit LU :

$$LU = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ m_{21} & 1 & 0 & \dots & 0 \\ m_{i1} & m_{i2} & 1 & 0 & \dots \\ \vdots & \vdots & \vdots & 1 & \vdots \\ m_{n1} & m_{n2} & \vdots & m_{nn-1} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & 0 & \dots & \dots & 0 \\ 0 & u_{22} & 0 & \dots & 0 \\ \vdots & 0 & u_{33} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & u_{nn} \end{pmatrix} \dots\dots\dots (5.2)$$

Ecrivons l'égalité des coefficients ligne par ligne

Ligne 1

Pour $j = 1, \dots, n$, $a_{1j} = u_{1j}$, ce qui permet de calculer

$$j = 1, \dots, n, a_{1j} = u_{1j}$$

Ligne 2

Colonne 1 $a_{21} = l_{21} u_{11}$, et puisque u_{11} est maintenant connu, on en déduit $l_{21} = \frac{a_{21}}{u_{11}}$

Colonne j, pour $j \geq 2$, $a_{2j} = l_{21} u_{1j} + u_{2j}$, et donc

$$j = 2, \dots, n, u_{2j} = a_{2j} - l_{21} u_{1j}$$

Ligne i : supposons que nous avons été capables de calculer

$$\begin{array}{cccc} u_{11} & u_{12} & \dots & u_{1n} \\ l_{21} & u_{22} & \dots & u_{2n} \\ \vdots & & \dots & \\ \vdots & & \dots & \\ l_{i-11} & \dots & l_{i-1,i-2} & u_{i-1n} \end{array}$$

Colonne 1 : $a_{i1} = l_{i1} u_{11}$, on en déduit $l_{i1} : l_{i1} = \frac{a_{i1}}{u_{11}}$

Colonne $j < i$: $a_{ij} = l_{i1} u_{1j} + l_{i2} u_{2j} + \dots + l_{ij} u_{jj}$, d'où

$$j = 1, \dots, j, l_{ij} = \frac{a_{ij} - l_{i1} u_{1j} - \dots - l_{ij-1} u_{j-1j}}{u_{jj}}$$

Colonne $j \geq i$: $a_{ij} = l_{i1} u_{1j} + l_{i2} u_{2j} + \dots + l_{ij} u_{jj}$, d'où

$$j = i, \dots, n : u_{ij} = a_{ij} - l_{i1} u_{1j} - \dots - l_{ii-1} u_{i-1j}$$

Remarquons qu'à la ligne i nous utilisons les valeurs de A la ligne i et les valeurs de L et U calculées précédemment. D'un point de vue informatique, on mettra L et U à la place de A ligne par ligne.

Calculons le nombre d'opérations nécessaires à la décomposition LU.

5.3. Factorisation

L'étape de triangularisation de la méthode précédente revient à factoriser la matrice A en un produit de deux matrices triangulaires l'une supérieure l'autre inférieure. En effet, les différentes étapes de la triangularisation peuvent s'écrire

sous forme matricielle. Si on pose $A = A^{(1)}$, la matrice $A^{(2)}$ du second système est obtenue en

multipliant à gauche $A^{(1)}$ par $M^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{a_{21}}{a_{11}} & 1 & 0 \\ -\frac{a_{31}}{a_{11}} & 0 & 1 \end{pmatrix}$ on a donc

$$A^{(2)} = M^{(1)}A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} \end{pmatrix}$$

de même la matrice $A^{(2)}$ du troisième système vérifie $A^{(3)} = M^{(2)}A^{(2)}$

$$\text{Ou } M^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{a_{32}^{(2)}}{a_{22}^{(2)}} & 1 \end{pmatrix} \text{ on a donc } A^{(3)} = M^{(1)}M^{(2)}A^{(1)} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & 0 & a_{33}^{(3)} \end{pmatrix}$$

Ainsi $A^{(1)} = (M^{(2)}M^{(1)})^{-1}A^{(3)} = (M^{(1)})^{-1}(M^{(2)})^{-1}A^{(3)}$. les matrices $(M^{(1)})$ et $(M^{(2)})$ sont inversibles car leurs déterminant sont tous égaux à 1 et leurs inverses sont faciles à calculer. en effet, on peut vérifier qu'on les obtient en transformant[20]

en leur opposé les termes figurant sous la diagonale, les autres termes restant inchangés. On trouve

$$\text{alors } L = (M^{(1)})^{-1}(M^{(2)})^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(2)}}{a_{22}^{(2)}} & 1 \end{pmatrix} \dots\dots\dots(5.3)$$

si on pose $U = A^{(3)}$, qui est triangulaire supérieure, A peut alors s'écrire comme le produit des matrices

$$L = \begin{pmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{pmatrix}, U = \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{pmatrix}$$

$$A = LU = \begin{pmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{pmatrix} * \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{pmatrix} \dots\dots\dots(5.4)$$

$$\begin{cases} U_{11} = a_{11} \\ U_{12} = a_{12} \\ U_{13} = a_{13} \end{cases} \begin{cases} L_{21}U_{11} = a_{21} \\ L_{21}U_{12} + U_{22} = a_{22} \\ L_{21}U_{13} + U_{23} = a_{23} \end{cases} \begin{cases} L_{31}U_{11} = a_{31} \\ L_{31}U_{12} + L_{32}U_{22} = a_{32} \\ L_{31}U_{13} + L_{32}U_{23} + U_{33} = a_{33} \end{cases}$$

Finalement, on obtient la solution X du système initial $AX = B$

En résolvant successivement les deux systèmes suivants :
$$\begin{cases} LY = B \\ UX = Y \end{cases}$$

Ceci se généralise à une matrice d'ordre n . le lecteur pourra le traiter en exercice .complément écrit sur brouillon.

Remarque 5 Lorsque n est “grand”, le nombre d’opérations de l’algorithme de résolution du système triangulaire est équivalent à n^2 , car il requiert pour chaque x_i à: $(n-i)$ multiplications, $(n-i)$ soustractions et une division, soit au total $\sum_{i=1}^n (2(n-i) + 1)$ opérations. D’autre part, le nombre d’opérations de l’algorithme de triangularisation est de l’ordre de n^3 . Donc, la résolution complétée nécessite au total un nombre d’opérations équivalent à n^3 . Pour $n = 15$, sur la même configuration de machine que celle citée en début du chapitre, il faudra de l’ordre d’une seconde pour la résolution totale du système.

5.5. Méthode d'élimination de GaussJordan

La méthode d'élimination de GaussJordan 10 est une variante de la méthode d'élimination de Gauss ramenant toute matrice sous forme échelonnée réduite. Dans le cas d'une matrice A inversible, cette méthode revient à chercher une matrice M telle que la matrice MA soit non pas triangulaire supérieure mais diagonale. Pour cela, on procède comme pour l'élimination de Gauss, mais en annulant à chaque étape tous les éléments de la colonne considérée situés au-dessous et au-dessus de la diagonale.[21]

Si elle est bien moins efficace 11 que la méthode d'élimination de Gauss pour la résolution de systèmes linéaires, la méthode d'élimination de GaussJordan est utile pour le calcul de l'inverse d'une matrice A carrée d'ordre n . Il s'agit de résoudre simultanément les n systèmes linéaires

$$Ax_j = e_j, \quad 1 \leq j \leq n,$$

en appliquant à chaque second membre e_j , les transformations nécessaires à l'élimination de Gauss_Jordan. D'un point de vue pratique, on a coutume d'« augmenter » la matrice A à inverser avec la matrice identité d'ordre n (les n seconds membres « élémentaires ») et d'appliquer la méthode de Gauss-Jordan à la matrice écrite par blocs $(A \ I_n)$. Au terme du processus d'élimination, le premier bloc contient la matrice identité et, si aucun échange de lignes n'a eu lieu, le second l'inverse de A .

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{pmatrix}$$

$$A^{(2)} = \begin{pmatrix} 1 & a_{12}^{(2)} & a_{13}^{(2)} & a_{14}^{(2)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & a_{34}^{(2)} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & a_{13}^{(3)} & a_{14}^{(3)} \\ 0 & 1 & a_{23}^{(3)} & a_{24}^{(3)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \end{pmatrix}$$

$$A^{(3)}$$

$$A^{(4)} = \begin{pmatrix} 1 & 0 & 0 & a_{14}^{(4)} \\ 0 & 1 & 0 & a_{24}^{(4)} \\ 0 & 0 & 1 & a_{34}^{(4)} \end{pmatrix}$$

Exemple :

Resoudre le systment par methodeGauss_Jordan $A = \begin{pmatrix} 2 & 1 & -4 \\ 3 & 3 & -5 \\ 4 & 5 & -2 \end{pmatrix}$ et $b = \begin{pmatrix} 8 \\ 14 \\ 16 \end{pmatrix}$

$$A^{(1)} = \begin{pmatrix} 1 & 1/2 & -2 & 4 \\ 0 & 3/2 & 1 & 2 \\ 0 & 3 & 6 & 0 \end{pmatrix} \begin{array}{l} \text{ligne 1}/2 \\ \text{ligne 2} - 3 * \text{ligne 1} \\ \text{ligne 3} - 4 * \text{ligne 1} \end{array}$$

$$A^{(2)} = \begin{pmatrix} 1 & 0 & -7/3 & 10/3 \\ 0 & 1 & 2/3 & 4/3 \\ 0 & 0 & 4 & -4 \end{pmatrix} \begin{array}{l} \text{ligne 2} / \frac{3}{2} \\ \text{ligne 1} - \frac{1}{2} * \text{ligne 2} \\ \text{ligne 3} - 3 * \text{ligne 2} \end{array}$$

$$A^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & -1 \end{pmatrix} \begin{array}{l} \text{ligne 3} / 4 \\ \text{ligne 1} + \frac{7}{3} * \text{ligne 2} \\ \text{ligne 2} - \frac{2}{3} * \text{ligne 3} \end{array}$$

On a directement les racines dans la 4e colonne.

$$\begin{cases} x_1 = 1 \\ x_2 = 2 \\ x_3 = -1 \end{cases}$$

5.6. Factorisation LDMT

Cette méthode considère une décomposition sous la forme d'un produit d'une matrice triangulaire inférieure, d'une matrice diagonale et d'une matrice triangulaire supérieure. Une fois obtenue la factorisation de la matrice A (d'un coût identique à celui de la factorisation LU), la résolution du système linéaire (2.1) fait intervenir la résolution d'un système triangulaire inférieur (par une méthode de descente), puis celle (triviale) d'un système diagonal et en « n » la résolution d'un système triangulaire supérieur (par une méthode de remontée), ce qui représente un coût de $n^2 + n$ opérations[18].

5.6.1. Théorème: Soit A une matrice d'ordre n. La factorisation LU de A, avec $l_{ii} = 1$ pour $i = 1, \dots, n$ existe et est unique si toutes les sous-matrices principales

$$A_k = \begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}, 1 \leq k \leq n,$$

extraites de A sont inversibles.

5.6.2. Démonstration. Il est possible de montrer l'existence de la factorisation LU de manière constructive, en utilisant

le procédé d'élimination de Gauss. En supposant que les n sous-matrices principales extraites de A sont inversibles, on va ici prouver en même temps l'existence et l'unicité par un raisonnement par récurrence. [13]

Pour $k = 1$, on a $A_1 = a_{11} \neq 0$,

et il suffit de poser $L_1 = 1$ et $U_1 = a_{11}$. Montrons à présent que s'il existe une unique factorisation de la sous matrice

A_{k-1} , $2 \leq k \leq n$, de la forme $A_{k-1} = L_{k-1}U_{k-1}$, avec $l_{k-1,ii} = 1, i = 1, \dots, k-1$, alors il existe une unique factorisation de ce type pour A_k . Pour cela, décomposons A_k en blocs

$$A_k = \begin{pmatrix} A_{k-1} & b \\ l^T & d \end{pmatrix},$$

avec b et c des vecteurs de R_k et d'un nombre réel, et cherchons une factorisation de A_k de la forme

$$\begin{pmatrix} A_{k-1} & b \\ l^T & d \end{pmatrix} = \begin{pmatrix} L_{k-1} & 0 \\ l^T & 1 \end{pmatrix} \begin{pmatrix} U_{k-1} & u \\ 0^T & u \end{pmatrix},$$

avec l et u des vecteurs de R_k et u un nombre réel. En effectuant le produit de matrices et en identifiant par blocs avec A_k , on obtient

$$L_{k-1}U_{k-1} = A_{k-1}, \quad \text{avec} \quad l_{k-1}u = b, \quad l^T U_{k-1} = c^T \quad \text{et} \quad l^T u + u = d$$

Si la première de ces égalités n'apporte aucune nouvelle information, les trois suivantes permettent de déterminer les vecteurs l et u et le scalaire u. En effet, on a par hypothèse $0 \neq \det(A_{k-1}) = \det(L_{k-1})\det(U_{k-1})$, les matrices (L_{k-1}) et (U_{k-1}) sont donc inversibles. Par conséquent, les vecteurs l et u existent et sont uniques et $u = d - l^T u$ Ceci achève la preuve par récurrence.

Dans cette preuve, on utilise de manière fondamentale le fait les termes diagonaux de la matrice L sont tous égaux à 1. On aurait tout aussi bien pu choisir d'imposer d'autres valeurs (non nulles) ou encore décider de fixer les valeurs des éléments diagonaux de la matrice U. Ceci implique que plusieurs factorisations LU existent, chacune pouvant être déduite d'une autre par multiplication par une matrice diagonale convenable.[13]

La factorisation LU est particulièrement avantageuse lorsque l'on doit résoudre plusieurs systèmes linéaires ayant tous A pour matrice, mais des seconds membres différents. En effet, il suffit de conserver les matrices L et U obtenues à l'issue de la factorisation pour ramener ensuite la résolution de chaque système linéaire $Ax = b$ à celle de deux systèmes triangulaires,

$$L_y = b, \text{ puis } U_x = y,$$

5.6.3. Proposition : Sous les conditions du théorème, il existe une unique matrice triangulaire inférieure L , une unique matrice diagonale D et une unique matrice triangulaire supérieure M^T , les éléments diagonaux de L et M étant tous égaux à 1, telles que $A = LD M^T$:

5.6.4. Démonstration. Les hypothèses du théorème 2.2 étant satisfaites, on sait qu'il existe une unique factorisation LU de la matrice A . En choisissant les éléments diagonaux de la matrice D égaux à u_{ii} , $1 \leq i \leq n$, (tous non nuls puisque la matrice U est inversible), on a $A = LU = LDD^{-1}U$

Il suffit alors de poser MD^T pour obtenir l'existence de la factorisation. Son unicité est une conséquence de l'unicité de la factorisation LU.

L'intérêt de cette dernière factorisation prend son sens lorsque la matrice A est symétrique, puisque dans ce cas $M = L$.

La factorisation résultante $A = LDL^T$ peut alors être calculée avec un coût environ deux fois moindre. La factorisation LD M^T permet également de démontrer le résultat suivant.

5.7. Factorisation de Cholesky

Une matrice symétrique définie positive vérifie les conditions de la proposition et admet par conséquent une factorisation LDL T , dont la matrice diagonale D est de plus à termes strictement positifs. Cette observation conduit à une factorisation ne faisant intervenir qu'une seule matrice triangulaire inférieure, appelée factorisation de Cholesky. Plus précisément, on a le résultat suivant.

5.7.1. Théorème . Soit A une matrice symétrique définie positive d'ordre n . Alors, il existe une unique matrice triangulaire inférieure B dont les éléments diagonaux sont strictement positifs telle que

$$A = BB^T :$$

5.7.2. Démonstration. En notant, les n sous-matrices principales extraites de A on peut écrire, pour tout vecteur w de $\mathbb{R}^k$,

$$w^T A_k w = v^T A v, \text{ avec } v_i = w_i, 1 \leq i \leq k, \text{ et } v_i = 0, k+1 \leq i \leq n,$$

ce qui montre que les sous-matrices A_k sont symétriques définies positives. La matrice A vérifie donc les conditions du théorème et admet une unique factorisation LU. Les éléments diagonaux de la matrice U sont de plus strictement positifs, car on a

$$\prod_{i=1}^k u_{ii} = \det(A_k) > 0, 1 \leq k \leq n.$$

En introduisant la matrice diagonale Δ définie par $\Delta_{ii} = \sqrt{u_{ii}}$, $1 \leq i \leq n$, la factorisation se réécrit $A = L\Delta\Delta^{-1}U$,

En posant $B = L\Delta$ et $C = \Delta^{-1}U$, la symétrie de A entraîne que $BC = C^T B^T$, d'où

$C(B^T)^{-1} = B^{-1}C^T = I_n$ (une matrice étant triangulaire supérieure, l'autre triangulaire inférieure et toutes deux à coefficients diagonaux égaux à 1) et donc $C = B^T$. On a donc montré l'existence d'au moins une factorisation de Cholesky. Pour montrer l'unicité de cette décomposition, on suppose qu'il existe deux matrices triangulaires inférieures B_1 et B_2 telles que

$$A = B_1 B_1^T = B_2 B_2^T$$

d'où $B_2^{-1}B_1 = B_2^T(B_1^T)^{-1}$. il existe donc une matrice diagonale D que $B_2^{-1}B_1 = D$ et, par conséquent $B_1 = B_2 D$. finalement, on a

$$A = B_2 B_2^T = B_2 D D^T B_2^T$$

et donc $D^2 = I_n$. Les coefficients diagonaux d'une matrice de factorisation de Cholesky étant par hypothèse positifs, on a nécessairement $D = I_n$ et donc $B_1 = B_2$

Pour la mise en œuvre de cette factorisation, on procède de la manière suivante. On pose $B = (b_{ij})$ $1 \leq i, j \leq n$ avec $b_{ij} = 0$ si $i < j$ et l'on déduit alors de l'égalité

$A = BB^T$ que

$$a_{ij} = \sum_{k=1}^n b_{ik} b_{jk} = \sum_{k=1}^{\min(i,j)} b_{ik} b_{jk}, \quad 1 \leq i, j \leq n. \dots \dots (5.5)$$

La matrice A étant symétrique, il suffit que les relations ci-dessus soient vérifiées pour $j \leq i$

(par exemple), et l'on va donc construire les colonnes de B à partir des colonnes de A . On fixe donc j à 1 et on fait varier i de 1 à n :

$$\begin{aligned} a_{11} &= (b_{11})^2, \text{ d'où } b_{11} = \sqrt{a_{11}} \\ a_{21} &= b_{11} b_{21}, \text{ d'où } b_{21} = \frac{a_{21}}{b_{11}} \\ &\vdots \\ a_{n1} &= b_{11} b_{n1}, \text{ d'où } b_{n1} = \frac{a_{n1}}{b_{11}} \end{aligned}$$

pour déterminer la première colonne de B . La $j^{\text{ième}}$ colonne de B est obtenue en utilisant les relations

$$\begin{aligned}
a_{ji} &= (b_{j1})^2 + (b_{j2})^2 + \dots + (b_{jj})^2, \text{ d'ou } b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} (b_{jk})^2} \\
a_{j+1j} &= b_{j1}b_{j+11} + b_{j2}b_{j+12} + \dots + b_{jj}b_{j+1j}, \text{ d'ou } b_{j+1j} = \frac{a_{j+1j} - \sum_{k=1}^{j-1} b_{jk}b_{j+1k}}{b_{jj}} \dots\dots\dots (5.6) \\
&\vdots \\
a_{nj} &= b_{j1}b_{n1} + b_{j2}b_{n2} + \dots + b_{jj}b_{nj}, \text{ d'ou } b_{nj} = \frac{a_{nj} - \sum_{k=1}^{j-1} b_{jk}b_{j+1k}}{b_{jj}}
\end{aligned}$$

Remarque

1. La méthode de Cholesky est moins générale que la factorisation LU puisqu'elle n'est applicable qu'aux matrices SDP.
2. On montre que le nombre d'opérations nécessaires à la résolution d'un système linéaire de taille n est de l'ordre de $n^3/6$, soit environ 2 fois moins que pour la méthode générale LU
3. Un autre avantage: la méthode de Cholesky est très robuste numériquement, peu sensible aux erreurs de troncature. Le pivot âgé est en pratique beaucoup moins nécessaire pour qu'il ne l'est pour l'élimination de Gauss.
4. Même divisé par 2, le nombre d'opérations reste proportionnel au cube de la taille du système à résoudre. Cela devient très problématique lorsque n est de l'ordre de 106, ..., 109 comme on le rencontre en calcul intensif ... Il faut trouver autre chose !

Travaux dirigés IV

Exercice 1: Effectuer une élimination de Gauss sur les système linéaires suivants.

$$\begin{aligned}
(1) \quad & \begin{cases} 2x_1 + 4x_2 + 4x_3 = 2 \\ x_1 + 3x_2 + x_3 = 1 \\ x_1 + 5x_2 + 6x_3 = -6 \end{cases} \\
(2) \quad & \begin{cases} x_1 - 6x_3 + 2x_4 = 6 \\ 8x_1 - 2x_3 - 2x_4 = -2 \\ 2x_1 + 9x_2 + x_3 + 3x_4 = -8 \\ 2x_1 + x_2 - 3x_3 + 10x_4 = -4 \end{cases}
\end{aligned}$$

Exercice 2 : On revient sur la première matrice donnée dans l'exercice 2 :

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

Effectuer une factorisation LU de cette matrice ou L est une matrice triangulaire inférieure ayant des 1 sur sa diagonale et U est une matrice triangulaire supérieure

$$|2| = 2 \neq 0, \begin{vmatrix} 2 & 4 \\ 1 & 3 \end{vmatrix} = 2 \neq 0, \begin{vmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{vmatrix} = 14 \neq 0,$$

Exercice 3 : Décompose à l'aide de la méthode de cholosky la matrice a symétrique en un produit de deux matrice triangulaire en suite résoudre le système

$$A = \begin{pmatrix} 64 & 40 & 24 \\ 40 & 29 & 17 \\ 24 & 17 & 19 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

La solution FICHE TD IV

Exercice 1: Effectuer une élimination de Gauss sur les système linéaires suivants.

$$(1) \begin{cases} 2x_1 + 4x_2 + 4x_3 = 2 \\ x_1 + 3x_2 + x_3 = 1 \\ x_1 + 5x_2 + 6x_3 = -6 \end{cases}$$

$$(2) \begin{cases} x_1 - 6x_3 + 2x_4 = 6 \\ 8x_1 - 2x_3 - 2x_4 = -2 \\ 2x_1 + 9x_2 + x_3 + 3x_4 = -8 \\ 2x_1 + x_2 - 3x_3 + 10x_4 = -4 \end{cases}$$

Nous écrivons le premier système sous la forme du tableau

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 1 & 3 & 1 & 1 \\ 1 & 5 & 6 & -6 \end{pmatrix} \begin{matrix} L_1 \\ L_2 \\ L_3 \end{matrix}$$

On effectue l'élimination de Gauss. On a successivement

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 3 & 4 & -7 \end{pmatrix} \begin{matrix} L_2 \leftarrow L_2 - 0.5L_1 \\ L_3 \leftarrow L_3 - 0.5L_1 \end{matrix}$$

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 7 & -7 \end{pmatrix} L_3 \leftarrow L_3 - 3L_2$$

On obtient alors le système triangulaire suivant

$$\begin{cases} 2x_1 + 4x_2 + 4x_3 = 2 \\ x_2 - x_3 = 0 \\ 7x_3 = -7 \end{cases}$$

En utilisant l'algorithme de remontée (1.2) on a successivement

$$x_3 = -1, x_2 = -1, x_1 = 5$$

L'élimination de Gauss ci-dessus est dite sans permutation.

On peut par exemple à l'étape (2.1) ci-dessus, remplacer le pivot 1 par le coefficient 3

de x_2 de la dernière ligne, parce que $3 > 1$ donne plus de stabilité numérique. Dans ce

cas on dit que l'on fait une élimination de Gauss avec pivot partiel. Dans ce contexte on obtient

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 7 & -7 \end{pmatrix} L_3 \leftrightarrow L_2$$

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 3 & 4 & -7 \\ 0 & 0 & -\frac{7}{3} & \frac{7}{3} \end{pmatrix} L_3 \leftarrow L_3 - \frac{1}{3} \times L_2$$

D'où on obtient le système triangulaire supérieur suivant

$$\begin{cases} 2x_1 + 4x_2 + 4x_3 = 2 \\ 3x_2 + 4x_3 = -7 \\ -\frac{7}{3}x_3 = -\frac{7}{3} \end{cases}$$

En appliquant l'algorithme de remonté à ce système on obtient

$$x_3 = -1, x_2 = -1, x_1 = 5$$

On peut enfin par exemple à l'étape (2.1) ci-dessus, remplacer le pivot 1 par le coefficient le plus grand en module dans la sous-matrice.

$$\begin{pmatrix} 1 & -1 \\ 3 & 4 \end{pmatrix}$$

Ceci rend la méthode plus stable numériquement. Ici on trouve 4 comme nouveau pivot.

Dans ce cas on dit que l'on fait une élimination de Gauss avec pivot partiel. Dans ce contexte on obtient

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 3 & 4 & -7 \\ 0 & 1 & -1 & 0 \end{pmatrix} L_3 \leftrightarrow L_3$$

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 4 & 3 & -7 \\ 0 & -1 & 1 & 0 \end{pmatrix} c_3 \leftrightarrow c_3$$

$$\begin{pmatrix} 2 & 4 & 4 & 2 \\ 0 & 4 & 3 & -7 \\ 0 & 0 & \frac{7}{4} & -\frac{7}{4} \end{pmatrix} L_3 \leftarrow L_3 + \frac{1}{4}L_2$$

Cette dernière transformation donne le système linéaire suivant

$$\begin{cases} 2x_1 + 4x_3 + 4x_2 = 2 \\ +4x_3 + 3x_2 = -7 \\ \frac{7}{4}x_2 = -\frac{7}{4} \end{cases}$$

Par application de l'algorithme de remontée au système triangulaire ci-dessus on obtient $x_3 = -1, x_2 = -1, x_1 = 5$

On met le deuxième exemple sous forme de la matrice suivant
$$\begin{pmatrix} 1 & 0 & 6 & 2 & 6 \\ 8 & 0 & -2 & -2 & -2 \\ 2 & 9 & 1 & 3 & -8 \\ 2 & 1 & -3 & 10 & -4 \end{pmatrix} \begin{matrix} L_1 \\ L_2 \\ L_3 \\ L_4 \end{matrix}$$

puis on effectue
$$\begin{pmatrix} 1 & 0 & 6 & 2 & 6 \\ 0 & 0 & -50 & -18 & -50 \\ 0 & 9 & -11 & -1 & -20 \\ 0 & 1 & -15 & 6 & -16 \end{pmatrix} \begin{matrix} L_2 \leftarrow L_2 - 8L_1 \\ L_3 \leftarrow L_3 - 2L_1 \\ L_4 \leftarrow L_4 - 2L_1 \end{matrix}$$

La matrice obtenue après la 1^{ère} étape d'élimination (2.2) a pour pivot 0. Pour continuer la méthode de Gauss, on peut soit utiliser la stratégie de pivot partiel ou soit celle de pivot total.

Pivot partiel : on prend comme pivot le plus grand élément de la colonne $\begin{pmatrix} 0 \\ 9 \\ 1 \end{pmatrix}$

Cela revient à échanger la 2^{ème} et la 3^{ème} lignes. On obtient

$$\begin{pmatrix} 1 & 0 & 6 & 2 & 6 \\ 0 & 9 & -11 & -1 & -20 \\ 0 & 0 & -50 & -18 & -50 \\ 0 & 1 & -15 & 6 & -16 \end{pmatrix} L_2 \leftrightarrow L_3$$

On continue l'élimination

$$\begin{pmatrix} 1 & 0 & 6 & 2 & 6 \\ 0 & 9 & -11 & -1 & -20 \\ 0 & 0 & -50 & -18 & -50 \\ 0 & 0 & -\frac{124}{9} & \frac{55}{9} & -\frac{124}{9} \end{pmatrix} L_4 \leftrightarrow L_4 - \frac{1}{9}L_2$$

$$\begin{pmatrix} 1 & 0 & 6 & 2 & 6 \\ 0 & 9 & -11 & -1 & -20 \\ 0 & 0 & -50 & -18 & -50 \\ 0 & 0 & 0 & \frac{2491}{225} & 0 \end{pmatrix} L_4 \leftrightarrow L_4 - \frac{1}{50} \frac{124}{9} L_3$$

On déduit le système triangulaire supérieur suivant

$$\begin{cases} x_1 + 6x_3 + 2x_4 = 6 \\ 9x_2 - 11x_3 - x_4 = -20 \\ -50x_3 - 18x_4 = -50 \\ + \frac{2491}{225}x_4 = 0 \end{cases}$$

D'où par la formule de remontée on trouve

$$x_4 = 0, \quad x_3 = 1, x_2 = -1, x_1 = 0$$

• Pivot total : On part de l'étape (2.2) de l'élimination de Gauss. Le plus grand élément en module

de la sous-matrice
$$\begin{pmatrix} 0 & -50 & -18 \\ 9 & -11 & -1 \\ 1 & -15 & 6 \end{pmatrix}$$

est -50, qui se trouve à la 2^{ème} ligne et à la 3^{ème} colonne de la matrice de départ. On

Positionne -50 en pivot, en échangeant la 2ième colonne et la 3ième colonne :

$$\begin{pmatrix} 1 & 0 & 0 & 2 & 6 \\ 0 & -50 & 0 & -18 & -50 \\ 0 & -11 & 9 & -1 & -20 \\ 0 & -15 & 1 & 6 & -16 \end{pmatrix} c_2 \leftrightarrow c_3$$

On continue l'élimination

$$\begin{pmatrix} 1 & 6 & 0 & 2 & 6 \\ 0 & -50 & 0 & -18 & -50 \\ 0 & 0 & 9 & \frac{74}{25} & -9 \\ 0 & 0 & 1 & \frac{285}{225} & -1 \end{pmatrix} \begin{matrix} L_3 \leftrightarrow L_3 - \frac{11}{50}L_2 \\ L_4 \leftrightarrow L_4 - \frac{15}{50}L_2 \end{matrix}$$

$$\begin{pmatrix} 1 & 6 & 0 & 2 & 6 \\ 0 & -50 & 0 & -18 & -50 \\ 0 & 0 & 9 & \frac{74}{25} & -9 \\ 0 & 0 & 0 & \frac{2491}{225} & 0 \end{pmatrix} L_4 \leftrightarrow L_4 - \frac{1}{9}L_3$$

Ce qui conduit au système linéaire suivant

$$\begin{cases} x_1 + 6x_3 + 4x_4 = 6 \\ -50x_3 - 18x_4 = -50 \\ 9x_2 + \frac{74}{25}x_4 = -9 \\ + \frac{2491}{225}x_4 = 0 \end{cases}$$

dont la solution est

$$x_4 = 0, \quad x_2 = -1, x_3 = 1, x_1 = 0$$

Exercice 2 : On revient sur la première matrice donnée dans l'exercice 2 :

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

Effectuer une factorisation LU de cette matrice où L est une matrice triangulaire

Inférieure ayant des 1 sur sa diagonale et U est une matrice triangulaire supérieure

$$|2| = 2 \neq 0, \begin{vmatrix} 2 & 4 \\ 1 & 3 \end{vmatrix} = 2 \neq 0, \begin{vmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{vmatrix} = 14 \neq 0,$$

Décomposition LU

Les mineurs principaux de la matrice proposée sont

Donc on peut la factoriser sous la forme LU où L est une matrice triangulaire inférieure ayant des 1 sur sa diagonale et U est une matrice triangulaire supérieure.

Identification directe Comme la matrice proposée est d'ordre 3, on peut écrire complètement le produit LU et faire des identifications. On écrit :

$$\begin{pmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ 0 & U_{22} & U_{23} \\ 0 & 0 & U_{33} \end{pmatrix}$$

D4ou on obtient

$$\left\{ \begin{array}{l} U_{11} = 2 \\ U_{12} = 4 \\ U_{13} = 4 \\ L_{21}U_{11} = 1 \\ L_{31}U_{11} = 1 \\ L_{21}U_{12} + U_{22} = 3 \\ L_{21}U_{13} + U_{23} = 1 \\ L_{31}U_{12} + L_{32}U_{22} = 5 \\ L_{31}U_{13} + L_{32}U_{23} + U_{33} = 6 \end{array} \right. \leftrightarrow \left\{ \begin{array}{l} U_{11} = 2 \\ U_{12} = 4 \\ U_{13} = 4 \\ L_{21} = \frac{1}{2} \\ L_{31} = \frac{1}{2} \\ U_{22} = 1 \\ U_{23} = -1 \\ L_{32} = 3 \\ U_{33} = 7 \end{array} \right.$$

On trouve $\begin{pmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & 4 \\ 0 & 1 & -1 \\ 0 & 0 & 7 \end{pmatrix}$

Cas général on écrit $A = A^{(1)} = L^{(1)}U^{(1)}$ ou

$$A = \begin{pmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & 0 & 1 \end{pmatrix} \begin{pmatrix} U_{11} & U_{12} & U_{13} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} \end{pmatrix}$$

Par identification on a $\left\{ \begin{array}{l} a_{1j} = a_{1j}^{(1)} = U_{1j} \quad , j = 1, \dots, 3 \\ a_{i1} = a_{i1}^{(1)} = L_{i1}U_{11} \quad , i = 2, 3 \\ a_{ij} = a_{ij}^{(1)} = L_{i1}U_{1j} + a_{ij}^{(2)} \quad , i, j = 2, 3 \end{array} \right.$

$$\left\{ \begin{array}{l} U_{1j} = a_{1j} \quad , j = 1, 2, 3 \\ L_{i1} = \frac{a_{i1}}{U_{11}} \quad , i = 2, 3 \\ a_{ij}^{(2)} = a_{ij}^{(1)} - L_{i1}U_{1j} \quad , i, j = 2, 3 \end{array} \right.$$

$$\Leftrightarrow \begin{cases} U_{11} = 2 \\ U_{12} = 4 \\ U_{13} = 4 \\ L_{21} = \frac{1}{2} \\ L_{31} = \frac{1}{2} \\ a_{22}^{(2)} = 3 - \frac{1 \times 4}{2} = 1 \\ a_{23}^{(2)} = 1 - \frac{1 \times 4}{2} = -1 \\ a_{32}^{(2)} = 5 - \frac{1 \times 4}{2} = 3 \\ a_{33}^{(2)} = 6 - \frac{1 \times 4}{2} = 4 \end{cases}$$

on pose $A^{(2)} = \begin{pmatrix} 1 & -1 \\ 3 & 4 \end{pmatrix}$

On décompose

$$A^{(2)} = L^{(1)}U^{(1)} \text{ avec } A^{(2)} = \begin{pmatrix} 1 & -1 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ L_{32} & 1 \end{pmatrix} \begin{pmatrix} U_{22} & U_{23} \\ 0 & U_{33} \end{pmatrix}$$

$$\text{Donc } \begin{cases} a_{22}^{(2)} = U_{22} \\ a_{23}^{(2)} = U_{23} \\ a_{32}^{(2)} = L_{32} \\ a_{33}^{(2)} = L_{32}U_{23} + U_{33} \end{cases} \Leftrightarrow \begin{cases} U_{22} = a_{22}^{(2)} \\ U_{23} = a_{23}^{(2)} \\ L_{32} = a_{32}^{(2)} \\ U_{33} = a_{33}^{(2)} - L_{32}U_{23} \end{cases}$$

$$\begin{cases} U_{22} = 1 \\ U_{23} = -1 \\ L_{32} = 3 \\ U_{33} = 7 \end{cases}$$

D'où

$$A^{(2)} = \begin{pmatrix} 1 & -1 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 0 & 7 \end{pmatrix}$$

Il vient $A = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 7 \end{pmatrix} \begin{pmatrix} 2 & 4 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

Ou encore $A = \left[\begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix} \right] \left[\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 0 & 7 \end{pmatrix} \begin{pmatrix} 2 & 4 & 4 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right]$

Ou bien encore

$$\begin{pmatrix} 2 & 4 & 4 \\ 1 & 3 & 1 \\ 1 & 5 & 6 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1/2 & 1 & 0 \\ 1/2 & 3 & 1 \end{pmatrix} \begin{pmatrix} 2 & 4 & 4 \\ 0 & 1 & -1 \\ 0 & 0 & 7 \end{pmatrix}$$

D'où la décomposition LU.

Exercice 3 : A est une matrice définie positive

$$A = T^t T \quad T \text{ est matrice triangulaire supérieure } T = \begin{pmatrix} T_{11} & T_{12} & T_{13} \\ 0 & T_{22} & T_{23} \\ 0 & 0 & T_{33} \end{pmatrix}$$

$$\text{Algorithme : } t_{11} = \sqrt{a_{11}} \quad , t_{1j} = \frac{a_{j1}}{t_{11}} \quad 1 \leq j \leq n$$

$$t_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} (t_{ki})^2} \quad 1 \leq i \leq n$$

$$t_{ij} = (a_{ij} - \sum_{k=1}^{i-1} (t_{ki} t_{kj})) / t_{ii}$$

$$t_{11} = \sqrt{64} = 8, \quad t_{12} = \frac{40}{8} = 5, \quad t_{13} = \frac{24}{8} = 3$$

$$t_{22} = \sqrt{29 - 25} = 2, \quad t_{23} = \frac{17 - 5 \cdot 3}{2} = 2, \quad t_{33} = \sqrt{19 - (3^2 + 1)} = 3$$

$$T = \begin{pmatrix} 8 & 5 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix}$$

2)

$$\begin{cases} Ax = b \\ A = T^t T \end{cases} \rightarrow T^t T x = b \quad \text{on pose } Tx = Z \dots \dots [1] T^t Z = b \dots \dots [2] [2] =$$

$$\begin{pmatrix} 8 & 0 & 0 \\ 5 & 2 & 0 \\ 3 & 1 & 3 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \begin{cases} 8Z_1 = 1 \\ 5Z_1 + 2Z_2 = 1 \\ 3Z_1 + Z_2 + 3Z_3 = 1 \end{cases} \leftrightarrow \begin{cases} Z_1 = \frac{1}{8} \\ Z_2 = \frac{3}{16} \\ Z_3 = \frac{7}{48} \end{cases}$$

$$[1] = \begin{pmatrix} 8 & 5 & 3 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1/8 \\ 3/16 \\ 7/48 \end{pmatrix} \begin{cases} 8x_1 + 5x_2 + 3x_3 = 1/8 \\ 2x_2 + x_3 = 3/16 \\ 3x_3 = 7/48 \end{cases} \leftrightarrow \begin{cases} x_1 = -\frac{5}{96} \\ x_2 = \frac{5}{72} \\ x_3 = \frac{7}{144} \end{cases}$$

Chapitre 06 : Méthodes itératives

6.1.1. Définition : On va voir un type de méthodes itératives de résolution du système linéaire

$Ax = b$ sous la forme

$$(*) \begin{cases} x^{(0)} \text{ vecteur arbitraire} \\ x^{(k+1)} = Bx^{(k)} + c, k \geq 0 \end{cases} Ax = b \leftrightarrow x = Bx + c, \dots\dots\dots (6.1)$$

La matrice B et le vecteur c sont en fonction de A et b.

Définition la méthode itérative (*) est convergente si $\lim_{k \rightarrow +\infty} x^{(k)} = x, \forall x^{(0)}$ remarque

Si on pose $e^{(k)} = x^{(k)} - x$, pour $k = 0, 1, 2, \dots$

comme $x = Bx + c$ et $x^{(k+1)} = Bx^{(k)} + c$, on a

$$e^{(k)} = Bx^{(k+1)} - Bx = Be^{(k-1)} = \dots = B^k e^{(0)}$$

$$\lim_{k \rightarrow +\infty} x^{(k)} = x \leftrightarrow \lim_{k \rightarrow +\infty} e^{(k)} = 0 \leftrightarrow \lim_{k \rightarrow +\infty} B^k e^{(0)} = 0$$

la méthode itérative (*) est convergente

$$\lim_{k \rightarrow +\infty} B^k v = 0 \quad \forall v$$

Ce qui équivaut à $\lim_{k \rightarrow +\infty} \|B^k v\| = 0, \forall v$ pour toute norme vectorielle

6.1.2. Convergence des méthodes itératives :

6.1.2.1. Théorème Les propositions suivantes sont équivalentes :

- (1) la méthode itérative (*) est convergente ;
- (2) $\rho(B) < 1$;
- (3) $(\|B\|) < 1$ Pour au moins une norme matricielle subordonnée $\|\cdot\|$.

6.2. Méthode de Jacobi

Ces méthodes sont des cas particuliers de la méthode suivante :

$A = M - N$ avec M inversible et assez simple. On aurait alors :
 $A = M - N$ avec M inversible et assez simple . on aurait alors

$$Ax = b \leftrightarrow Mx = Nx + b \leftrightarrow x = M^{-1}Nx + M^{-1}b$$

$$\leftrightarrow x = Bx + c, \text{ avec } B = M^{-1}N \text{ et } c = M^{-1}b$$

Méthode de Jacobi : en posant $A = D - (E + F)$

$$M = D \text{ est la diagonale de } A \text{ et } N = E + F.$$

$$Ax = b \leftrightarrow Dx = (E + F)x + b$$

on suppose que D est inversible, c'est-à-dire $a_{ii} \neq 0, 1 \leq i \leq N$.

$$J = D^{-1}(E + F) = I_N - D^{-1}A$$

est appelée la matrice de Jacobi

$$\begin{cases} x^{(0)} \text{ donné} \\ Dx^{(k+1)} = (E + F)x^{(k)} + b, k \geq 0 \end{cases} \text{ A chaque étape, } x_N^{(k+1)} \dots \dots \dots (6.2)$$

on calcule les N composantes $x_1^{(k+1)}, \dots$,

$$\text{du vecteur } x_1^{(k+1)} \begin{cases} a_{1,1}x_1^{(k+1)} = -a_{1,2}x_2^{(k)} - \dots - a_{1,N}x_N^{(k)} + b_1 \\ a_{2,2}x_2^{(k+1)} = -a_{2,1}x_1^{(k)} - a_{2,3}x_3^{(k)} - \dots - a_{2,N}x_N^{(k)} + b_2 \\ \vdots \\ a_{N,N}x_N^{(k+1)} = -a_{N,1}x_1^{(k)} - \dots - a_{N,N-1}x_{N-1}^{(k)} + b_N \end{cases}$$

6.3. Méthode de Gauss-Seidel.

M est la partie triangulaire inférieure de A : $M = D - E$ et $N = F$.

On pourrait améliorer la méthode précédente en utilisant les quantités déjà calculées, on calcule successivement les N composantes [17]

$$x_1^{(k+1)}, x_N^{(k+1)} \text{ du vecteur } x^{(k+1)}$$

$$\begin{cases} a_{1,1}x_1^{(k+1)} = -a_{1,2}x_2^{(k)} - \dots - a_{1,N}x_N^{(k)} + b_1 \\ a_{2,2}x_2^{(k+1)} = -a_{2,1}x_1^{(k+1)} - a_{2,3}x_3^{(k)} - \dots - a_{2,N}x_N^{(k)} + b_2 \dots \dots \dots (6.3) \\ \vdots \\ a_{N,N}x_N^{(k+1)} = -a_{N,1}x_1^{(k+1)} - \dots - a_{N,N-1}x_{N-1}^{(k+1)} + b_N \end{cases}$$

Qui revient à écrire

$$Dx^{(k+1)} = Ex^{(k+1)} + Fx^{(k)} + b$$

En encore

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

$$J_1 = (D - E)^{-1}F \text{ est la matrice de Gauss - Seidel}$$

Elle est inversible si $a_{ii} \neq 0, 1 \leq i \leq N$.

Travaux dirigés IV

Exercice 1 : soit le système linéaire suivant :

$$\begin{cases} -x_1 + 5x_2 - x_3 = 3 \\ 9x_1 - 2x_2 + x_3 = 8 \\ x_1 - 2x_2 + 9x_3 = 9 \end{cases}$$

- (a) Faire 2 itérations de la méthode de Jacobi en partant du point $\vec{x}_0 = (0 \ 0 \ 0)^T$.
- (b) Faire 1 itération de la méthode de Gauss-Seidel en partant du point $\vec{x}_0 = (0 \ 0 \ 0)^T$.
- (c) Réordonner les équations de façon à assurer la convergence des deux méthodes.
- (d) Soit le système linéaire $A\vec{x} = \vec{b}$ suivant:

$$\begin{pmatrix} 3 & 1 \\ 2 & 5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4 \\ 7 \end{pmatrix}$$

Écrire la matrice d'itérations T_j de la méthode de Jacobi et en déduire son rayonspectral. Est-ce que la méthode Jacobi converge pour ce système linéaire?

Exercice 2 soit le système d'équations algébriques suivant

$$\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$$

- (a) Calculer la solution exacte $\vec{x} = (x_1 \ x_2)^T$ du système linéaire.
- (b) Faire quelques itérations de la méthode de Gauss-Seidel en partant de l'approximation initiale $\vec{x}^0 = (0 \ 0)^T$ pour vérifier que l'algorithme diverge.
- (c) Proposer une simple modification du système qui conduise à la convergence de la Méthode de Gauss-Seidel en partant de l'approximation initiale $\vec{x}^0 = (0 \ 0)^T$. Vérifier en itérant jusqu'à ce que $\|\vec{x}^K - \vec{x}\|_\infty < 10^{-1}$
- (d) Montrer en utilisant deux résultats différents du cours que la méthode de Jacobi converge aussi pour le système modifié obtenu en (c). (Ne pas faire d'itérations de la méthode de Jacobi).

Exercice 3. Soit le système d'équations algébriques suivant

$$\begin{pmatrix} 1 & a & 0 \\ a & 1 & a \\ 0 & a & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Où a est un paramètre réel tel que $a \neq \mp \frac{1}{\sqrt{2}}$

- (a) Pour $a = -\frac{1}{2}$, faire une itération de la méthode de Gauss-Seidel en partant de l'approximation initiale $\vec{x}^0 = (0 \ 0)^T$
- (b) Pour quelles valeurs de a , la convergence de la méthode de Gauss-Seidel est-elle assurée?

(c) Pour $a = -\frac{1}{2}$, donner la matrice d'itérations TJ de la méthode de Jacobi. Sachant que la matrice TJ possède les valeurs propres $\lambda_1 = 0, \lambda_2 = -\frac{1}{\sqrt{2}}$ et $\lambda_3 = \frac{1}{\sqrt{2}}$, est-ce que la méthode Gauss-Seidel converge pour ce système linéaire?

Exercice 4 . Soit le système d'équations algébriques suivant

$$\begin{cases} 3x_2 - x_3 + 8x_4 = -11 \\ 2x_1 - x_2 + 10x_3 = -11 \\ 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 = 25 \end{cases}$$

(a) Pourquoi ne peut-on pas utiliser directement les méthodes de Jacobi et de Gauss-Seidel avec ce système linéaire?

(b) Réordonner les équations de telle sorte que la matrice associée soit à diagonale strictement dominante.

(c) Calculer une itération des méthodes de Jacobi et de Gauss-Seidel, en partant du point $\vec{x}^0 = (0 \ 0 \ 0 \ 0)^T$.

Exercice 05 : Soit la matrice A et le membre de droite $\vec{b}$

$$\begin{bmatrix} 9 & -3 & 3 \\ -3 & 10 & -4 \\ 3 & -4 & 18 \end{bmatrix} \vec{b} = \begin{pmatrix} 3 \\ -4 \\ 18 \end{pmatrix}$$

(a) Calculer la factorisation LU la plus adéquate de A Justifier votre choix ;

(b) Utiliser votre factorisation pour calculer la solution du système $A\vec{x} = \vec{b}$;

(c) La méthode de Gauss-Seidel appliquée au système $A\vec{x} = \vec{b}$ sera-t-elle convergente si le point de départ $\vec{x}^0$ est bien choisi ?

(d) Effectuer deux itérations de la méthode de Gauss-Seidel à partir de $\vec{x}^0 = (1 \ 1 \ 0)^T$

La solution de fiche TD IV:

Exercice 1 (a) En partant de l'approximation initiale $\vec{x}^0$, on obtient avec la méthode de Jacobi les itérations $\vec{x}^1 = (-3 \ -4 \ 1)^T$ et $\vec{x}^2 = (-24 \ -17 \ \frac{4}{9})^T$

(b) En partant de l'approximation initiale $\vec{x}^0$, on obtient avec la méthode de Gauss-Seidel l'itération $\vec{x}^1 = (-3 \ -\frac{35}{2} \ -\frac{23}{9})^T$

$$(c) \begin{cases} 9x_1 - 2x_2 + x_3 = 8 \\ -x_1 - 5x_2 - x_3 = -3 \\ x_1 - 2x_2 + 9x_3 = 9 \end{cases}$$

On obtient une matrice à diagonale strictement dominante.

(d) On sait que

$$T_J = -D^{-1}(T_I + T_S) = I - D^{-1}A = T_J \begin{pmatrix} 0 & -\frac{1}{3} \\ -\frac{2}{5} & 0 \end{pmatrix}$$

Son rayon spectral est $\rho(T_J) = \sqrt{\frac{2}{15}} < 1$ et donc la méthode de Jacobi converge.

Exercice 2 : (a) La solution est $\bar{x} = (1 \ 1)^T$:

(b) En partant de l'approximation initiale $\bar{x}^0 = (0 \ 0)^T$, on obtient avec la méthode

de Gauss-Seidel les itérations suivantes: $\bar{x}^1 = (3 \ -3)^T$, $\bar{x}^2 = (9 \ -15)^T$ et

$\bar{x}^3 = (33 \ 63)^T$ et ce qui confirme que la méthode de Gauss-Seidel diverge.

(c) En partant de l'approximation initiale $\bar{x}^0 = (0 \ 0)^T$, on obtient avec la méthode

de Gauss-Seidel les 3 itérations suivantes: $\bar{x}^1 = (\frac{3}{2} \ \frac{3}{4})^T$, $\bar{x}^2 = (\frac{9}{8} \ \frac{15}{16})^T$ et $\bar{x}^3 = (\frac{33}{32} \ \frac{63}{64})^T$

avec $\|\bar{x}^{(3)} - \bar{x}\|_{\infty} = 0.031 < 0.1$

(d) La méthode de Jacobi converge car la matrice A est à diagonale strictement dominante. En plus le rayon spectral de la matrice d'itérations $\rho(T_J) = \frac{1}{2} < 1$

Exercice 3 : (a) en partant de l'intégration initiale $\bar{x}^0 = [0 \ 0 \ 0]^T$, on obtient avec la méthode de

gauss-seidel l'itération suivante $\bar{x}^1 = [1 \ \frac{3}{2} \ \frac{7}{4}]^T$

(b) $-\frac{1}{2} < a < \frac{1}{2}$, la convergence de la méthode de Gauss-Seidel est assurée car la matrice A est à diagonale strictement dominante.

(c) La matrice d'itération de Jacobi est

$$T_J = -D^{-1}(T_I + T_S) = \begin{pmatrix} 9 & -3 & 3 \\ -3 & 10 & -4 \\ 3 & -4 & 18 \end{pmatrix}$$

En se servant des valeurs propres de la matrice T_J , on a $\rho(T_J) = \frac{1}{\sqrt{2}} < 1$. La méthode

de Jacobi converge, ce qui entraîne la convergence de la méthode de Gauss-Seidel car

elle est plus rapide.

Exercice 4 : (a) On ne peut pas appliquer directement les méthodes de Jacobi et de Gauss-Seidel puisque l'un des coefficients diagonaux est nul ($A_{11} = 0$).

(b) On obtient le système d'équations algébriques suivant:

$$\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ 2x_1 - x_2 + 10x_3 = -11 \\ 3x_2 - x_3 + 8x_4 = -11 \end{cases}$$

(c) en partant $\vec{x}^0 = [0 \ 0 \ 0 \ 0]^T$, la méthode de Jacobi donne l'itération $\vec{x}^1 = \left[\frac{3}{5} \ \frac{25}{11} \ \frac{-11}{10} \ \frac{-11}{8} \right]^T$ la méthode de Jacobi donne l'itération $\vec{x}^2 = \left[\frac{3}{5} \ \frac{128}{55} \ \frac{-543}{550} \ \frac{-726}{349} \right]^T$

Exercice 5 :

(a) La matrice A est symétrique et définie positive, la décomposition de Cholesky est la plus adéquate.

$$A = LL^T \begin{bmatrix} 3 & 0 & 0 \\ -1 & 3 & 0 \\ 1 & -1 & 4 \end{bmatrix} \begin{bmatrix} 3 & -1 & 1 \\ 0 & 3 & -1 \\ 0 & 0 & 4 \end{bmatrix}$$

b) on résout $L\vec{y} = [3 \ -4 \ 18]^T$ pour obtenir $\vec{y} = [1 \ -1 \ 4]^T$ et ensuite $L^T\vec{x} = \vec{y}$ pour calculer la solution $\vec{x} = [0 \ 0 \ 1]^T$

(c) La convergence de la méthode de Gauss-Seidel est assurée qu'elle que soit la solution de départ utilisée car la matrice A est à diagonale strictement dominante.

(d) en partant $\vec{x}^0 = [1 \ 1 \ 0]^T$, on trouve $\vec{x}^1 = \left[\frac{2}{3} \ \frac{-1}{5} \ \frac{38}{40} \right]^T$ et ensuite $\vec{x}^2 = \left[-\frac{2}{135} \ \frac{-1}{15} \ \frac{80}{81} \right]^T$

Chapitre 7 : Équations et systèmes d'équations différentielles

7.1 Introduction :

On rappelle, dans les premiers paragraphes, les résultats fondamentaux des équations différentielles. Les paragraphes suivants sont consacrés aux méthodes numériques usuelles. Dans le traitement numérique des équations différentielles, on distingue les méthodes à pas séparés (ou à un seul pas) qui permettent de calculer y_{n+1} à partir de la seule connaissance de y_n et les méthodes à pas liés (ou à pas multiples) qui nécessitent la connaissance de $y_n, y_{n-1}, \dots, y_{n-p}$ pour calculer y_{n+1} . Les méthodes numériques de résolution[15]

(dites à différences finies) sont fondées sur le développement de Taylor.

7.1.1. Existence et unicité des solutions

Soit U un ouvert de $\mathbb{R} \times \mathbb{R}^n$ et $f: U \rightarrow \mathbb{R}^n$ une application continue. On considère une équation différentielle de la forme $\dot{x} = f(t, x)$ munie d'une condition initiale $x(t_0) = x_0 \in \mathbb{R}^n$ donnée, avec la notation $\dot{x} = \frac{dx}{dt}$. Une solution de cette équation est (par définition) une application différentiable $y = [a, b] \rightarrow \mathbb{R}^n$ vérifiant les conditions

$$\begin{cases} \dot{x} = f(t, x) \\ x(t_0) = x_0 \end{cases} \quad \forall t \in [a, b] \dots \dots \dots (7.1)$$

Ce problème admet, sous certaines conditions, une solution. On démontre en effet que si la fonction f est une fonction continûment différentielle alors il existe deux valeurs maximales (éventuellement infinies) t_1 et t_2

pour lesquelles la solution de l'équation différentielle existe et est unique pour tout t dans l'intervalle $[t_0 - t_1, t_0 + t_2]$. On démontre aussi que si f est une fonction continue dans $[a, b] \times \mathbb{R}^n$ et vérifie la condition de Lipchitz de rapport $k > 0$ suivante

$$\forall t \in [a, b], \forall x_1, x_2 \in \mathbb{R}^n, \|f(t, x_2) - f(t, x_1)\| \leq k \|x_2 - x_1\| [17]$$

Alors le problème avec condition initiale (t_0, x_0) admet une solution unique dans $[a, b]$, donnée par $x(t) = x_0 + \int_{t_0}^t f(s, x(s)) ds$

7.2. Méthodes d'Euler

La méthode de Leonhard Euler (1707-1783) est une méthode à pas séparé du premier ordre. Elle consiste à remplacer l'opérateur de dérivation d/dx par le schéma discret $(y_{i+1} - y_i)/h$. La résolution du problème

$$\begin{cases} y'(x) = f(x, y) \\ y(x_0) = y_0 \end{cases}$$

Conduit au schéma $\begin{cases} y_{i+1} = x_i + h \\ y_{i+1} = y_i + hf(x_i, y_i) \end{cases} \dots \dots \dots (7.2)$

En pratique, la méthode d'Euler n'est pas utilisée, car elle n'offre pas une précision suffisante. Cette méthode est convergente et du premier ordre, car l'erreur de consistance vaut

$$|y(x_i) - y_i| = \frac{1}{2}h^2 f'(c, y_i) \text{ avec } c \in [x_{i-1}, x_i] \dots \dots \dots (7.3)$$

Mais la méthode explicite est souvent instable. par exemple, si la fonction f est linéaire

$$f(x, y) = -ay \text{ avec } a > 0, \text{ le schéma d'Euler}$$

$$y_{i+1} = y_i - ah y_i = (1 - ah)y_i$$

est instable dès que $h > 2/a$, car dans ce cas, y_i tend vers l'infini lorsque i tend vers l'infini. En revanche, le schéma rétrograde

$$y_{i+1} = y_i - h y'_{i+1} = \frac{y_i}{1 + ah}$$

conduit à une méthode implicite qui est universellement stable puisque y_{i+1} tend vers zéro, quand le pas h tend vers l'infini.

Dans la méthode des trapèzes, la fonction f est remplacée par une fonction affine par morceaux. Le schéma de discrétisation

$$y_{i+1} = y_i - \frac{h}{2}(y'_{i+1} + y'_i)$$

conduit à une méthode d'ordre 2

7.3. Méthodes de Runge-Kutta

Carl Runge (1856-1927) et Martin Kutta (1867-1944) ont proposé en 1895 de résoudre le problème de Cauchy.

$$\begin{cases} y'(x) = f(x, y) \\ y(x_0) = y_0 \end{cases}$$

Conduit au schéma $\begin{cases} x_{i+1} = x_i + h_i \\ y_{i+1} = y_i + h\phi(x_i, y_i, h_i) \end{cases} \dots \dots \dots (7.4)$

Où la fonction d'incrément ϕ est une approximation de $f(x, y)$ sur l'intervalle $[x_i, x_{i+1}]$

Supposons donnés un entier r , une matrice A dont les éléments triangulaires supérieurs sont nuls et un vecteur $b = (b_1, \dots, b_r)$ l'algorithme de Runge-Kutta est le suivant

$$\begin{cases} y_{i+1} = y_i + h(b_1 k_1 + \dots + b_r k_r) \\ x_{i+1} = x_i + h \\ k_j = f(x_i + c_j h, y_i + h(a_{j1} k_1 + \dots + a_{jr} k_r)) \end{cases} \dots \dots \dots (7.5)$$

Le vecteur b vérifie $b_1 + \dots + b_r = 1$. Les coefficients c_i sont les sommes des éléments d'une ligne de la matrice A . Les éléments supérieurs de A sont nuls $a_{ij} = 0$ si $j \geq i$. Dans ces méthodes, le pas h peut facilement varier (h_i). Une méthode de Runge-Kutta est entièrement déterminée par la donnée de l'entier r , du vecteur b et de la matrice A .

Méthodes d'ordre 1. Pour $b = 1$ et $a_{11} = 0$, l'algorithme $y_{i+1} = y_i + h.f(x_i, y_i)$ se réduit à la méthode d'Euler. Pour $b = 1$ et $a_{11} = 1$, l'algorithme de Runge-Kutta $y_{i+1} = y_i + h.f(x_{i+1}, y_{i+1})$ conduit à la méthode d'Euler rétrograde (méthode implicite).

Méthodes d'ordre 2. Pour déterminer toutes les méthodes d'ordre 2, cherchons une fonction ϕ de la forme $\phi = b_1 k_1 + b_2 k_2$

Où les coefficients k_1 et k_2 sont données par
$$\begin{cases} k_1 = f(x_i, y_i) \\ k_2 = f(x_i + ch, y_i + ahk_1) \end{cases}$$

Développons y_{i+1} au voisinage du point (x_i, y_i)

$$y_{i+1} = y(x_i) + hf(x_i, y(x_i)) + \frac{h^2}{2} \left(\frac{\partial f}{\partial x}(x_i, y(x_i)) + f(x_i, y(x_i)) \frac{\partial f}{\partial y}(x_i, y(x_i)) \right)$$

De même, développons sk_2 au voisinage du point (x_i, y_i)

$$y_{i+1} = y_i + hb_1 k_1 + hb_2 \left(f(x_i, y_i) + hc \frac{\partial f}{\partial x}(x_i, y_i) + ahf(x_i, y_i) \frac{\partial f}{\partial x}(x_i, y_i) \right) + O(h^2)$$

En identifiant les deux expressions, il vient

$$\begin{cases} y(x_i) = y_i \\ f(x_i, y_i) = (b_1 + b_2)f(x_i, y_i) \\ \frac{1}{2}f'_x + \frac{1}{2}f'_y f = b_2(cf'_x + af'_y f) \end{cases}$$

On en déduit que $b_1 + b_2 = 1$ et $b_2 c = b_2 a = 1/2$. Soit en posant

$$b_2 = \emptyset \text{ et } b_1 = 1 - \emptyset$$

$c = a = \frac{1}{2}$, on retrouve les trois cas standards. La méthode d'Euler s'obtient pour $\emptyset = 0$

$$y_{i+1} = y_i + hf(x_i, y_i)$$

La méthode de Heun (ou d'Euler-Cauchy) est obtenue pour $\emptyset = 1/2$

$$\begin{cases} y_{i+1} = y_i + h(k_1 + k_2)/2 \\ k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h, y_i + hk_1) \end{cases}$$

La méthode de Runge-Kutta (proprement dite) est obtenue pour $\emptyset = 1$

$$\begin{cases} y_{i+1} = y_i + hk_2 \\ k_1 = f(x_i, y_i) \\ k_2 = f(x_i + h/2, y_i + hk_1) \end{cases}$$

Méthodes d'ordre 3. L'algorithme de Runge-Kutta classique correspond au

cas $b = (1/6, 2/3, 1/6)$ et à la matrice

$$A = \begin{pmatrix} 0 & 0 & 0 \\ 1/2 & 0 & 0 \\ -1 & 2 & 0 \end{pmatrix}$$

L'algorithme effectue à chaque pas le calcul de trois facteurs k_i

$$\begin{cases} y_{i+1} = y_i + h(k_1 + 4k_2 + k_3)/6 \\ k_1 = f(x_i, y_i) \\ k_2 = f\left(x_i + \frac{h}{2}, y_i + \frac{hk_1}{2}\right) \\ k_3 = f(x_i + h, y_i - hk_1 + 2hk_2) \end{cases}$$

Pour améliorer l'efficacité du calcul, on utilise des méthodes à pas variable, c'est-à-dire des méthodes dans lesquelles le pas varie à chaque itération.

Une des méthodes classiques consiste à employer deux méthodes de Runge-Kutta emboîtées. La première méthode d'ordre r sert à calculer la solution approchée, tandis que la seconde méthode d'ordre r' sert à estimer l'erreur de consistance pour contrôler le pas. On dit que la méthode est d'ordre (r', r) Proposée en 1957, la méthode de Merson est la première méthode [17]

de Runge-Kutta emboîtée. Elle consiste à calculer.

$$\begin{cases} k_1 = f(x_i, y_i) \\ k_2 = f\left(x_i + \frac{h}{3}, y_i + \frac{hk_1}{3}\right) \\ k_3 = f(x_i + h/3, y_i - hk_1/6 + hk_2/6) \\ k_4 = f\left(x_i + \frac{h}{2}, y_i + hk_1/8 + 3hk_2/8\right) \dots\dots\dots(7.6) \\ k_5 = f\left(x_i + h, y_i + \frac{hk_1}{2} - 3hk_3/2 + 2hk_1\right) \\ y_{i+1} = y_i + h(k_1 + 4k_2 + k_3)/6 \\ y_{i+1}^* = y_i + h(k_1 - 3k_3 + 4k_4)/2 \end{cases}$$

L'erreur $\Delta_i = |y_{i+1} - y_{i+1}^*| \dots\dots\dots(7.7)$ est évaluée à chaque pas.

Si ε désigne la tolérance acceptée, l'algorithme

de Merson divise le pas par facteur 2 quand $\Delta_i > \varepsilon$, multiplie le pas par deux quand $\Delta_i \leq \frac{\varepsilon}{64}$, et conserve le pas actuel dans les autres cas. La méthode de Merson est d'ordre 5 pour le calcul de la solution et d'ordre 4 pour le contrôle du pas. Dans la méthode proposée par Fehlberg en 1969,

on calcule

$$\left\{ \begin{array}{l} k_1 = hf(x_i, y_i) \\ k_2 = hf\left(x_i + \frac{h}{5}, y_i + \frac{k_1}{5}\right) \\ k_3 = hf\left(x_i + \frac{3h}{10}, y_i + \frac{3k_1}{40} + \frac{9k_2}{40}\right) \\ k_4 = hf\left(x_i + \frac{3h}{5}, y_i + \frac{3h}{10} - \frac{9k_2}{10} + \frac{6k_3}{5}\right) \\ k_5 = hf\left(x_i + h, y_i + \frac{11k_1}{54} + \frac{5k_2}{2} + \frac{70k_3}{27} + \frac{35k_4}{27}\right) \dots\dots\dots(7,7) \\ k_6 = hf\left(x_i + \frac{7h}{8}, y_i + \frac{1631k_1}{55296} + \frac{175k_2}{512} + \frac{575k_3}{13824} + \frac{44275k_4}{110592} + \frac{253k_5}{4096}\right) \\ y_{i+1} = y_i + h\left(\frac{37}{378}k_1 + \frac{250}{621}k_3 + \frac{125}{594}k_4 + \frac{512}{1771}k_6\right) \\ y_{i+1}^* = y_i + h\left(\frac{2825}{27648}k_1 + \frac{18575}{48348}k_3 + \frac{13525}{55296}k_4 + \frac{1}{4}k_6\right) \end{array} \right.$$

Travaux dirigés IV

Exercice 1. Faire trois itérations avec $h = 0,1$ des méthodes d'Euler explicite, d'Euler modifiée, du point milieu et de Runge-Kutta d'ordre 4 pour les équations différentielles suivantes :

- a) $\begin{cases} y'(t) = t \sin(y(t)) \\ y(0) = 2 \end{cases}$
- b) $\begin{cases} y'(t) = t^2 + (y(t))^2 + 1 \\ y(1) = 0 \end{cases}$
- b) $\begin{cases} y'(t) = y(t)e^t \\ y(0) = 2 \end{cases}$

Exercice 2 : L'équation différentielle :

$$y'(t) = y(t) + e^{2t} \quad (y(0) = 2)$$

possède la solution analytique $y(t) = e^t + e^{2t}$

- a) En prenant $h = 0,1$, faire 3 itérations de la méthode d'Euler modifiée et calculer l'erreur commise sur y_3 en comparant les résultats avec la solution analytique $y(0; 3)$.
- b) En prenant $h = 0,05$, faire 6 itérations de la méthode d'Euler modifiée et calculer l'erreur commise sur y_6 en comparant les résultats avec la solution analytique $y(0; 3)$.
- c) Faire le rapport des erreurs commises en a) et en b) et commenter le résultat en fonction de l'erreur de troncature locale liée à la méthode utilisée.
- d) Utiliser l'extrapolation de Richardson pour obtenir une meilleure approximation de $y(0; 3)$.

Exercice 3 : L'équation différentielle : $y'(t) = 2y(t) \quad (y(0) = 5)$

a) Vérifier que la solution analytique $y(t) = 5e^{2t}$

b) En posant $h = \frac{1}{N}$, montrer que les approximations fournies par la méthode d'Euler

explicite peuvent s'écrire comme $y_n = 5(1 + 2h)^n$, pour $n = 0, \dots, N$.

c) Vérifier numériquement que l'erreur $e(h)$ se comporte suivant la relation $e(h) \approx K$ où K est une constante.

La solution FICHE TD IV

Exercice 1 :

a) on $y'(t) = t \sin(y(t))$, $y(0) = 2$ et $h = 0.1$ a donc que $t_0 = 0$, que $y_0 = 2$ et que $f(t_n, y_n) = t_n \sin(y_n)$

$$\text{Euler } \begin{cases} y_{n+1} = y_n + hf(t_n, y_n) \\ t_{n+1} = t_n + h \end{cases}$$

$$\begin{cases} y_0 = 2 \\ y_1 = 2 + 0.1 \times \sin 2 = 2 \\ y_2 = 2 + 0.1 \times 0.1 \sin 2 = 2.0090929 \\ y_3 = 2.0090929 + 0.1 \times 0.2 \times \sin 2.0090929 = 2.02720249 \end{cases}$$

$$\text{Euler modifiée } \begin{cases} \hat{y} = y_n + hf(t_n, y_n) \\ y_{n+1} = y_n + \frac{h}{2} [f(t_n, y_n) + f(t_{n+1}, \hat{y})] \end{cases}$$

$$f(t_0, y_0) = f(0, 2) = 0$$

$$\hat{y} = y_0 + h \times (t_0, y_0) = 2 + h \times 0 = 2$$

$$y_1 = y_0 + \frac{h}{2} [f(t_0, y_0) + f(t_0 + h, \hat{y})] = 2 + 0.05[0 + 0.1 \times \sin(2)] = 2.00454648$$

$$f(t_1, y_1) = 0.1 \times \sin(2.004546487) = 0.907396$$

$$\hat{y} = 2.004546487 + 0.1 \times 0.907396 = 2.01362044$$

$$f(t_2, \hat{y}) = 0.2 \times \sin(2.01362044) = 0.18070903$$

$$y_2 = 2.004546487 + 0.05[0.907396 + 0.18070903] = 2.018118919$$

$$\text{de meme, on a que } y_3 = 2.040539939$$

$$\text{Runge kutta d'ordre 4} \left\{ \begin{array}{l} k_1 = hf(t_n, y_n) \\ k_2 = hf\left(t_n + \frac{h}{2}, y_n + \frac{k_1}{2}\right) \\ k_3 = hf\left(t_n + \frac{h}{2}, y_n + \frac{k_2}{2}\right) \\ k_4 = hf(t_n + h, y_n + k_3) \\ y_{n+1} = y_i + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4) \\ t_{n+1} = t_n + h \end{array} \right.$$

on a que $h = 0,1, t_0 = 0, y_0 = 2$ et que $f(t_n, y_n) = t_n \sin(y_n)$

premiere itération

$$\begin{aligned} k_1 &= hf(t_0, y_0) = 0.1 \times 0 \times \sin 2 = 0 \\ k_2 &= hf\left(0 + 0.05, 2 + \frac{0}{2}\right) = 0.1f(0.05, 2) = 0.1 \times 0.05 \times \sin 2 = 0.004546487 \\ k_3 &= 0.1f\left(0.05, 2 + \frac{0.004546487}{2}\right) = 0.1 f(0.05, 2.0022773244) \\ &= 0.1 \times 0.05 \times \sin(2.002273244) = 0.004541745 \\ k_4 &= 0.1f(0.1, 2.004541745) = 0.00907398 \\ y_1 &= 2 + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4) = 2.004541741 \end{aligned}$$

de meme , on trouve que

Deuxieme itération

$$k_1 = 0.009074 \quad k_2 = 0.013582 \quad k_3 = 0.013568 \quad k_4 = 0.018032$$

$$y_2 = 2.01810947$$

Troisième itération

$$k_1 = 0.018032 \quad k_2 = 0.022442 \quad k_3 = 0.022418 \quad k_4 = 0.026751$$

$$y_2 = 2.04052645$$

b) on a $y'(t) = t^2 + (y(t))^2 + 1$ et $y(1) = 0$ et $h = 0.1$. donc , on également que $t_0 = 1, y_0 = 0$ et que $f(t_n, y_n) = t_n^2 + (y_n)^2 + 1$

$$\text{Euler} \left\{ \begin{array}{l} y_1 = y_0 + hf(t_0, y_0) = 0 + 0.1 \times (1^2 + 0 + 1) \\ y_1 = 0.2 \quad t_1 = 1.1 \\ y_2 = y_1 + hf(t_1, y_1) = 0.2 + 0.1 \times (1.1^2 + 0.2^2 + 1) \\ y_2 = 0.425 \quad t_2 = 1.2 \\ y_3 = y_2 + hf(t_2, y_2) = 0.425 + 0.1 \times (1.2^2 + 0.425^2 + 1) \\ y_3 = 0.6870625 \quad t_3 = 1.3 \end{array} \right.$$

Euler modifiée

Première itération

$$\hat{y} = y_0 + hf(t_0, y_0)$$

$$y_1 = y_0 + \frac{h}{2} [f(t_0, y_0) + f(t_0 + h, \hat{y})]$$

$$y_1 = 0 + 0.05 \times (f(1,0) + f(1.1, 0.2)) = 0.2125$$

$$t_1 = 1.1$$

Deuxième itération

$$\hat{y} = y_1 + hf(t_1, y_1) = 0.2125 + 0.1 \times f(1.1, 0.2125) = 0.4380156$$

$$y_2 = y_1 + \frac{h}{2} [f(t_1, y_1) + f(t_1 + h, \hat{y})]$$

$$y_2 = 0.2125 + 0.05 \times (f(1.1, 0.2125) + f(1.2, 0.4380156)) = 0.45685059$$

$$t_2 = 1.2$$

Troisième itération

$$\hat{y} = y_2 + hf(t_2, y_2) = 0.45685069 + 0.1 \times f(1.2, 0.45685069) = 0.7217219$$

$$y_3 = y_2 + \frac{h}{2} [f(t_2, y_2) + f(t_2 + h, \hat{y})]$$

$$y_3 = 0.45685069 + 0.05 \times (f(1.2, 0.45685069) + f(1.3, 0.7217219)) = 0.74983045$$

$$t_3 = 1.3$$

Runge-Kutta

O(h4)

:

première itération

$$k_1 = 0.2$$

$$k_2 = 0.211250$$

$$k_3 = 0.211366$$

$$k_4 = 0.225468$$

$$y_2 = 0.2117831$$

Deuxième itération

$$k_1 = 0.225485$$

$$k_2 = 0.242782$$

$$k_3 = 0.243351$$

$$k_4 = 0.264715$$

$$y_2 = 0.45552718$$

Troisième itération

$$k_1 = 0.264751$$

$$k_2 = 0.290813$$

$$k_3 = 0.292362$$

$$k_4 = 0.420788$$

$$y_2 = 0.748199$$

c) $\{ y'(t) = y(t)e^t \text{ et } y(0) = 2 \text{ et } h = 0.1 \}$. donc , on égalment que $t_0 = 0, y_0 =$

2 et que $f(t_n, y_n) = y_n e^{t_n}$

$$\text{Euler} \begin{cases} y_1 = 2.2 \\ y_2 = 2.4431376 \\ y_3 = 2.741543 \end{cases} \quad \text{Euler modifiée} \begin{cases} \hat{y} = 2.2 \rightarrow y_1 = 2.2215688 \\ \hat{y} = 2.4670901 \rightarrow y_2 = 2.494994 \\ \hat{y} = 2.7997344 \rightarrow y_3 = 2.836326 \end{cases}$$

Runge-Kutta

O(h4)

:

première itération

$$k_1 = 0.2$$

$$k_2 = 0.220767$$

$$k_3 = 0.211859$$

$$k_4 = 0.245553$$

$$y_2 = 2.2218007$$

Deuxieme itération

$$k_1 = 0.245547 \quad k_2 = 0.272401 \quad k_3 = 0.273961 \quad k_4 = 0.0304833$$

$$y_2 = 2.495651$$

Troisième itération

$$k_1 = 0.304820 \quad k_2 = 0.340018 \quad k_3 = 0.342278 \quad k_4 = 0.383080$$

$$y_2 = 2.8377328$$

Exercice 2 : Commey(t) = $e^t + e^{2t}$, alors $y(0,3) = 3.17197608$. Ainssi , on a que $f(t_n, y_n) = y_n + e^{2t_n}$ que $y(0) = 2$ et donc que $y_0 = 2$ et que $t_0 = 0$

On fait 3 itérations avec $h = 0.1$ et avec la méthode d'Euler modifiée.

$$\begin{cases} \hat{y} = 2.3 \rightarrow y_1 = 2.2215688 \\ \hat{y} = 2.68081743 \rightarrow y_2 = 2.712075889 \\ \hat{y} = 3.132465948 \rightarrow y_3 = 3.1700001557 \end{cases}$$

L'erreur est alors donnée par :

$$|y(0,3) - y_3| = |3.171977608 - 3.1700001557| = 0.00197745$$

On fait 6 itérations avec $h = 0.05$ et avec la méthode d'Euler modifiée.

$$\begin{cases} \hat{y} = 2.15 \rightarrow y_1 = 2.1563793 \\ \hat{y} = 2.3194568 \rightarrow y_2 = 2.3264395 \\ \hat{y} = 2.5038316 \rightarrow y_3 = 2.5114778 \\ \hat{y} = 2.7045447 \rightarrow y_4 = 2.7129205 \\ \hat{y} = 2.9231577 \rightarrow y_5 = 2.9323361 \\ \hat{y} = 3.1613890 \rightarrow y_6 = 3.1714502 \end{cases}$$

L'erreur est alors donnée par :

$$|y(0,3) - y_6| = |3.171977608 - 3.1714502| = 0.00052739$$

Le ratio des erreurs est $\frac{0.00197745}{0.00052739} = 3.75 \approx 4$, ce qui confirme que la méthode

d'Euler modifiée est d'ordre 2.

L'extrapolation de Richardson d'ordre n est donnée par l'équation suivante :

$$Q_{\text{exa}} \cong \frac{2^n + Q_{\text{app}}\left(\frac{h}{2}\right) - Q_{\text{app}}(h)}{2^n - 1}$$

$$\text{On a donc } \frac{4 \times 3.171450217 - 3.170001557}{3} = 3.171933572 \text{ (zrreur} = 0.44 \times 10^{-4})$$

Exercice 3 :

si $y(t) = 5e^{2t}$, alors a que $y'(t) = 10e^{2t} = 2y(t)$. De plus, on a $y(0) = 5e^0 = 5$

Donc, $y(t) = 5e^{2t}$ est bien une solution analytique de l'équation différentielle.

Puisque $y'(t) = 2y(t)$, on a que $f(t_n, y_n) = 2y_n$. on a également que $y_0 = 5$. l'algorithme de la méthode d'Euler explicite nous dit que.

$$y_{n+1} = y_n + hf(t_n, y_n) = y_n + h(2y_n) = (1 + 2h)y_n$$

pour tout n . Par conséquent, on obtient par récursivité que

$$\begin{aligned} y_{n+1} &= (1 + 2h)((1 + 2h)y_{n-1}) \\ &= (1 + 2h)^2((1 + 2h)y_{n-2}) \\ &= (1 + 2h)^{n+1}y_0 = 5(1 + 2h)^{n+1} \end{aligned}$$

On peut chercher la constante en calculant e_n/h_n .

n	h_n	e_n	e_n/h_n
2	0.5	16.9462	33.9
10	0.1	5.9866	59.9
20	0.05	3.3077	66.7
100	0.01	0.7220	72.2
200	0.005	0.3651	73.0

et donc $K \approx 73$.

Bibliographie

- [1] F. Jedrzejewski, 2005. « Introduction aux méthodes numériques Deuxième édition ». Springer-Verlag France, Paris
- [2] J. Erhel, N. Nassif, B. Philippe, 2004. « Calcul matriciel et systèmes linéaires ». DEA Informatique et modélisation Beyrouth, Liban
- [3] Q. Louveaux, « Introduction aux Méthodes Numériques ». Université de Liège Faculté des Sciences Appliquées
- [4] ES. Term, 2013. « Les matrices ». Spécialité Mathématiques
- [5] G. Legendre, 2010. « Méthodes numériques Introduction à l'analyse numérique et au calcul scientifique »
- [6] M.A. Aziz Alaoui et C. Bertelle, 2002. « Méthodes numériques appliquées cours, exercices corrigés et mise en œuvre en java ». Faculté des Sciences et Techniques Le Havre Cedex – France
- [7] J.B. Clément, 2008. « analyse numérique ». Institut Galilée, Villetaneuse

- [8] P. Viot, 2006 . « Méthode d'analyse numérique ». Institut Universitaire Professionnalisé d'Évry
- [9] S. Mammar, 1999. « Méthodes numérique ». Génie Électrique et Informatique Industrielle 3ème Année
- [10] J. Fabbri, 2008. « Cours d'Analyse Numérique -version 3.0 ». EPU Université de Tours DI3ème année Département Informatique
- [11] D. Pastre, 2004. « Méthodes numériques ». licence de mathématiques et licence MASS Université René Descartes
- [12] S. El Bernoussi, 2009. « Cours Méthodes Numériques et Programmation ». Université de Nice Sophia-Antipolis Licence L3 Mathématiques
- [13] D. N'dri et S. Dufour, 2016. « Recueil d'exercices Exercices pour les cours de calcul scientifique pour ingénieurs MTH2210C Édition du 3 ». École Polytechnique de Montréal.
- [14] J. P. Grivet, 2009. « Méthodes numériques appliquées pour le scientifique et l'ingénieur Réalisation et mise en pages ». Centre technique Grenoble Sciences .
- [15] B. MEZIANI, 2010. « Analyse Numérique ». Université A. Mira de Béjaia Année 2009/2010 Faculté de la Technologie 2ème Année LMD/ST MATH VI
- [16] Université de Savoie L3 MASS Méthodes numériques TD 3. http://www.lama.univ-smb.fr/~perrin/methodes_num/td3_cor.pdf Consulté le 15/05/2017
- [17] U. A. B. Tlemcen 2015/2016 Département de Mathématiques. « https://www2.mat.ulaval.ca/fileadmin/Cours/MAT-2910/A-11/Solution_7.pdf » Consulté le 13/04/2017
- [18] S. Melliani et L. S. Chadli, 2010. « Module : Analyse numérique par Applications linéaires, matrices, déterminants ». Université Sultan Moulay Slimane Systèmes d'équations linéaires Corrections d'Arnaud Bodin
- [19] <http://exo7.emath.fr/ficpdf/fic00160.pdf>. Université de Nice Sophia-Antipolis Consulté le 22/06/2017
- [20] T. Takahashi, 2009. « Licence L3 Mathématiques » <http://math.unice.fr/~jabin/CTD3-6.pdf> Consulté le 14/06/2017.
- [21] M. Pierre et A. Henrot, 2014. « Analyse Numérique Cours de Polycopié rédigé ». Ecole des Mines de Nancy - Campus Artem .