

Emmanuel Kossi ATCHONOUUGLO

Analyses Numériques

Document de base

Table des matières

I INTRODUCTION A L'ANALYSE NUMERIQUE	4
I.1 Qu'est-ce que l'analyse numérique?	4
I.1.1 Définition	4
I.1.2 Objectifs de l'analyse numérique	4
I.1.3 Problèmes liés à l'analyse numérique	4
I.2 Les outils de l'analyse numérique	4
I.2.1 L'algorithme	4
I.2.2 L'organigramme	5
II RESOLUTION DES EQUATIONS NON LINEAIRES	6
II.1 Résolution d'une équation quelconque	6
II.1.1 Séparation des racines	6
II.1.1.1 Méthode graphique	6
II.1.1.2 Méthode de balayage	7
II.1.2 Amélioration d'une racine séparée	7
II.1.2.1 Critères d'arrêt	7
II.1.2.2 Méthode de dichotomie	7
II.1.2.3 Méthode de Newton-Raphson	7
II.1.2.4 Méthode de la sécante ou de la corde	8
II.1.2.5 Méthode d'interpolation linéaire	8
III RESOLUTION DES SYSTEMES D'EQUATIONS LINEAIRES	10
III.1 Rappels sur les matrices	10
III.1.1 Notations et définitions	10
III.1.2 Opérations sur les matrices	10
III.2 Systèmes linéaires	11
III.3 Résolution de systèmes linéaires non homogènes	12
III.3.1 Algorithmes de résolution directe	12
III.3.1.1 Système à matrice diagonale	12
III.3.1.2 Système à matrice triangulaire	12
III.3.2 Transformation des systèmes linéaires	13
III.3.2.1 Méthode de Gauss sans stratégie de pivot	13
III.3.2.2 Méthode de Gauss-Jordan	15
III.3.2.3 Méthode de Crout ($A = LR$)	16

III.3.2.4 Méthode de Cholesky ($A = LL'$)	16
III.3.3 Méthode de résolution indirecte ou méthode itérative	17
III.3.3.1 Méthode de Jacobi	18
III.3.3.2 Méthode de Gauss-Seidel	19
III.4 Exercices	20
IV INTERPOLATION LINEAIRE	22
IV.1 Introduction	22
IV.2 Interpolation polynomiale	22
IV.2.1 Problème de l'interpolation	22
IV.3 Approximation polynomiale	22
IV.3.1 Méthode d'interpolation de Lagrange	22
IV.3.2 Polynôme d'interpolation de Newton : Méthode des différences divisées	23
IV.4 Approximation au sens des moindres carrés	24
IV.4.1 Approximation polynomiale	24
IV.4.2 Approximation quelconque	25
V EQUATIONS DIFFÉRENTIELLES	27
V.1 Introduction	27
V.2 Dérivation numérique	28
V.3 Méthodes d'Euler	29
V.3.1 Méthode d'Euler du premier ordre	29
V.3.2 Étude générale des méthodes à un pas	29
V.3.2.1 Consistance, stabilité, convergence	29
V.4 Méthodes de Runge-Kutta d'ordre 2	30
VI Intégration numérique	32
VI.1 Introduction	32
VI.2 Exemples	32
VI.3 Evaluation de l'erreur. Noyau de Peano	34

Chapitre I

INTRODUCTION A L'ANALYSE NUMERIQUE

I.1 Qu'est-ce que l'analyse numérique ?

I.1.1 Définition

L'analyse numérique consiste à élaborer des processus qui permettent de résoudre des problèmes concrets par la seule voie du calcul numérique, donc à l'aide d'un ordinateur. En général, on procède comme suit :

- ☞ Étude de la solution du problème direct ;
- ☞ Création et étude de la ou des méthode(s) de résolution ;
- ☞ Expérimentation de la ou des méthode(s) afin de tester leur rapidité.

Remarque I.1.1 *A la place d'une étude théorique complète d'un algorithme, il est pratique et plus rapide de l'expérimenter.*

I.1.2 Objectifs de l'analyse numérique

Les différents objectifs du traitement numérique sont

- ① résoudre numériquement des problèmes complexes qui ne peuvent pas être traités de façon analytique
- ② limiter au maximum le temps de calcul (qui a un coût !)
- ③ réduire au maximum les erreurs, qu'elles proviennent des arrondis ou de l'algorithme utilisé.

I.1.3 Problèmes liés à l'analyse numérique

L'analyse numérique recherche le passage d'un domaine **continu, infini et théorique** à un domaine **discret, borné et concret**.

Pour les principaux problèmes, on peut retenir que :

- ① la représentation des nombres sur ordinateur n'est pas exacte ;
- ② les opérateurs arithmétiques ne sont qu'une approximation sur les ordinateurs ;
- ③ les concepts d'infiniments petits ou infiniments grands n'existent pas ;
- ④ il y a accumulation d'erreurs : les opérations de l'algorithme ne sont jamais effectuées avec exactitude, il y a une erreur d'arrondi ou de troncature.

I.2 Les outils de l'analyse numérique

I.2.1 L'algorithme

Un algorithme est un procédé de calculs qui permet de résoudre une série de problèmes d'un même type.

Le meilleur algorithme est celui qui donne la solution la plus précise avec un minimum d'opérations et en utilisant le moins de place mémoire possible.

On peut distinguer deux types d'algorithmes :

① les algorithmes génériques

- fonctionnent dans un grand nombre de cas,
- n'utilisent que peu d'hypothèses à priori,
- ne sont généralement pas les plus efficaces (temps de calcul/précision), mais couvrent un large éventail d'utilisations possibles

② les algorithmes spécifiques

- ne fonctionnent que dans des cas bien précis,
- utilisent des hypothèses à priori, qui les rendent adaptés à des problèmes précis,
- sont généralement plus efficaces, à la fois en temps de calcul et en précision, certaines hypothèses étant explicitement prises en compte dans l'algorithme utilisé.

I.2.2 L'organigramme

L'organigramme est la représentation schématique de la suite d'opérations traduisant le processus de résolution d'un problème donné.

C'est un passage entre l'algorithme et le programme, il est généralement constitué de symbols.

Généralement, l'organigramme est indépendant du langage dans lequel on souhaite traduire l'algorithme. Il existe deux types d'organigrammes :

- l'organigramme de type graphique ;
- l'organigramme de type texte.

Exercice I.2.1 1. *Qu'est ce qu'un algorithme ?*

2. *Qu'est ce qu'un organigramme ?*
3. *Quelle différence faites-vous entre un algorithme et organigramme ?*
4. *Donnez quelques difficultés liées à résolution des problèmes sur ordinateurs.*

Chapitre II

RESOLUTION DES EQUATIONS NON LINEAIRES

II.1 Résolution d'une équation quelconque

Soit f une fonction réelle continue sur un intervalle $[a, b]$.

On désire résoudre dans cet intervalle l'équation $f(x) = 0$.

- Si les études le permettent, on peut utiliser des propriétés de la fonction pour rechercher les solutions à l'aide d'une méthode particulière ;
- Si la fonction f est quelconque, on va rechercher une ou les solution(s) approchée(s) à l'aide d'une méthode plus générale.

Ce genre de problème se rencontre souvent, qu'il s'agisse de déterminer le point de fonctionnement d'une diode d'après sa caractéristique, la concentration d'une espèce chimique dans un mélange réactionnel ou la fréquence de coupure d'un filtre électrique.

On n'envisagera ici que des fonctions réelles. On peut être amené à chercher toutes les solutions de $f(x) = 0$, ou seulement quelques unes ou la plus petite.

Cas particulier important où f est un polynôme : les racines sont réelles ou deux à deux complexes conjuguées, en nombre égal au degré du polynôme.

Il est rare d'écrire une solution analytique de l'équation $f(x) = 0$; sauf pour les polynômes de degré inférieur ou égal à 4 ou pour de fonctions simples.

En conséquence, toutes les méthodes générales de recherche de racine sont des méthodes itératives ;

On doit se préoccuper de la convergence de la méthode et de la vitesse de convergence, définir un critère d'arrêt des itérations et prévoir le rôle des erreurs d'arrondi inévitables dans tout calcul numérique.

Aucune méthode connue ne fonctionne rien aveuglez : on doit toujours avoir une connaissance au moins approximative de l'emplacement de la racine.

Il est vivement recommandé de tracer le graphe de la fonction pour avoir une idée du nombre et de la position des zéros.

II.1.1 Séparation des racines

II.1.1.1 Méthode graphique

Elle consiste à tracer la courbe représentative de f et à chercher l'intersection de la courbe avec l'axe des abscisses.

Si la fonction f est plus compliquée pour son étude, on peut chercher à la décomposer comme somme de deux fonctions. Dans ce cas, on cherche les points d'intersection des courbes représentatives des deux fonctions.

Exemple II.1.1 : $f(x) = \exp(x) \sin(x) - 1$.

II.1.1.2 Méthode de balayage

Elle est basée sur le corollaire du théorème de Rolle :

Corollaire II.1.1 Si une fonction continue f sur un segment $[a, b]$ prend sur ce segment des valeurs de signes contraires, càd si $f(a) \cdot f(b) < 0$, alors ce segment contient au moins une racine de l'équation $f(x) = 0$.

De plus, si la dérivée $f'(x)$ existe et $f'(x)$ est de signe constant sur $[a, b]$, alors la racine est unique.

Technique de calcul

On balise l'intervalle $[a, b]$ par des points équidistants :

$$\begin{aligned} x_0 &= a \\ x_1 &= a + h \\ x_2 &= a + 2h \\ x_3 &= a + 3h \\ &\vdots \\ x_n &= a + nh = b \end{aligned}$$

soit en général $x_i = a + ih$, $i = 0, 1, \dots, n$ avec $h = (b - a)/n$.

Si $f(x_i) \cdot f(x_{i+1}) < 0$, alors $\exists x_s \in [x_i, x_{i+1}] | f(x_s) = 0$.

Après avoir correctement utilisé cette technique, on obtient un ou des sous intervalles contenant chacun une racine et une seule : cette racine s'appelle alors *racine séparée et notée x_s* .

II.1.2 Amélioration d'une racine séparée

II.1.2.1 Critères d'arrêt

Nous optons pour deux critères d'arrêt possibles :

- ☞ erreur relative $\epsilon_r \geq \|x_n - x_{n-1}\|$
- ☞ erreur absolue $\epsilon_a \geq \|x_n - x_{n-1}\|/\|x_n\|$

II.1.2.2 Méthode de dichotomie

II.1.2.3 Méthode de Newton-Raphson

Cette méthode est attribuée à Isaac Newton (1642 - 1727) ; cependant, c'est Raphson qui publiait en 1960 la formule itérative utilisée actuellement.

Le principe de la méthode consiste, étant donné un point de départ x_0 choisi arbitrairement, à élaborer une suite $(x_i)_{0 \leq i \leq n+1}$ qui, lorsque la méthode converge, tende vers la solution x_s .

Soient M_a le point de coordonnées $(a, f(a))$ et M_b le point de coordonnées $(b, f(b))$.

On remplace l'arc $M_a M_b$ par la droite tangente à la courbe au point M_a ou M_b . On recherche l'intersection de cette tangente et de l'axe des abscisses ; cette intersection x_m est une première approximation de la racine.

Le coefficient de la pente de la droite tangente à la courbe en M_a est $f'(a)$.

Le coefficient de la pente de $(M_a x_m)$ est donnée par $\frac{f(x_m) - f(a)}{x_m - a}$ donc on doit avoir $\frac{-f(a)}{x_m - a} = f'(a)$ ou encore $x_m = a - \frac{f(a)}{f'(a)}$.

D'où généralement, on utilise l'algorithme

$$x_{n+1} = \phi(x_n), \quad \phi(x) = x - \frac{f(x)}{f'(x)}$$

Précaution à suivre

- Il est indispensable de compter les itérations et de stopper le processus itératif s'il est jugé trop long.
- Il faut arrêter le processus itératif si $f'(x_n)$ est nul ou si $f(x_n)$ ou $f'(x_n)$ est non défini.

Possibilité de non-convergence

☞ Divergence systématique : $f(x) = x^{1/3}$.

Avec $\phi(x) = -2x$, on a $x_n = (-2)^n x_0$.

☞ Bouclage du cycle itératif : $f(x) = x^{1/2}$.

Avec $\phi(x) = -x$, on a $x_n = (-1)^n x_0$.

☞ Interruption par rejet à l'infini : $f(x) = 2x^3 - 7x^2 + 8x$.

Si on choisit par exemple $x_0 = 2$, on aura

$$\phi(x) = \frac{4x^3 - 7x^2}{6x^2 - 14x + 8} \Rightarrow x_1 = 1 \quad \text{et } x_2 = \infty.$$

Avantages et inconvénients

☞ Avantages

- convergence rapide ;
- choix d'un seul point pour enclancher le processus ;

☞ Inconvénients

- nécessité de calcul de la dérivée ;
- la méthode ne peut donner qu'une seule racine ;
- la non-assurance de la convergence de la méthode.

II.1.2.4 Méthode de la sécante ou de la corde

Dans certaines situations, la dérivée de f est très compliquée à calculer ou même voir impossible à expliciter. On ne peut plus continuer par utiliser la méthode de Newton.

Dans la méthode de Newton-Raphson, on avait

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

On peut remplacer la dérivée par son expression approchée

$$f'(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

et par suite, il vient que

$$x_{n+1} = \frac{f(x_n).x_{n-1} - f(x_{n-1}).x_n}{f(x_n) - f(x_{n-1})}.$$

Contrairement à la méthode de Newton-Raphson, ici, il faut choisir deux points au départ.

II.1.2.5 Méthode d'interpolation linéaire

Soient M_a le point de coordonnées $(a, f(a))$ et M_b le point de coordonnées $(b, f(b))$.

On trace la droite $(M_a M_b)$; cette droite coupe l'axe des x au point x_m . Ce point est considéré comme une première approximation de la racine.

☞ Si $f(a).f(x_m) = 0$, alors la racine est x_m ;

☞ Si $f(a).f(x_m) < 0$, alors la racine est dans l'intervalle $[a, x_m]$;

 Si $f(a) \cdot f(x_m) > 0$, alors la racine est dans l'intervalle $[x_m, b[$;

Des équations des pentes des droites $(M_a M_b)$ et $(x_m M_b)$, on a la formule de calcul de x_m

$$x_m = \frac{f(b) \cdot a - f(a) \cdot b}{f(b) - f(a)} \quad \text{ou généralement pour la forme itérative} \quad x_{n+1} = \frac{f(x_n) \cdot x_{n-1} - f(x_{n-1}) \cdot x_n}{f(x_n) - f(x_{n-1})}.$$

Exercice II.1.1 1. Pour chacune de ces méthodes, donner l'organigramme de type graphique ;

2. Donner ensuite l'organigramme du type texte pour ces mêmes méthodes ;

Exercice II.1.2 Résolution d'une équation non linéaire

1. On cherche à résoudre sur l'intervalle $[2, 5]$ l'équation $f(x) = 0$ où f est une fonction continue sur $[2, 5]$ non linéaire.

Décrire brièvement deux approches permettant de repérer les racines éventuelles de cette équation.

2. On se propose de résoudre l'équation non linéaire $f(x) = 0$ sur l'intervalle $[2, 5]$ avec la méthode de dichotomie.

Calculer le nombre d'itérations nécessaires pour avoir une précision absolue de l'ordre de 10^{-5}

3. Donner les avantages et inconvénients de la méthode de Newton-Raphson pour la résolution d'une équation non linéaire.

Exercice II.1.3 On souhaite résoudre l'équation $x - e^{-x} = 0$, $x \in [0, +\infty[$.

1. Montrer que cette équation admet une racine unique s dans $[0, +\infty[$
2. Utiliser la méthode de point fixe pour montrer l'existence d'une solution de cette équation.

Exercice II.1.4 1. Montrer que f a une seule racine $l \in]0, +\infty[$.

2. Montrer que la méthode itérative diverge.

Exercice II.1.5 Soit l'équation

$$\ln(x) = 2 - x$$

1. Montrer que cette équation admet une solution unique α dans $[0, 2]$.

2. Étudier l'itération

$$\begin{aligned} x_0 &\quad \text{donné} \\ x_{n+1} &= 2 - \ln(x_n) \end{aligned}$$

et montrer que cette itération converge vers α .

3. Montrer que cette équation proposée est équivalente à $x = \exp(2 - x)$, et étudier l'itération

$$\begin{aligned} x_0 &\quad \text{donné} \\ x_{n+1} &= \exp(2 - x_n). \end{aligned}$$

Qu'en déduisez-vous ?

4. Écrire la méthode de Newton pour l'équation proposée et proposer un bon choix de l'initialisation x_0 .

Chapitre III

RESOLUTION DES SYSTEMES D'EQUATIONS LINEAIRES

III.1 Rappels sur les matrices

III.1.1 Notations et définitions

On note $A = (a_{ij})_{m,n}$ une matrice ayant m lignes et n colonnes et où les a_{ij} sont des nombres réels ; ce qui équivaut à

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}$$

On définit par

matrice carrée	$: A = (a_{ij})_{n,n}$
matrice colonne	$: A = (a_i)_m$
matrice ligne	$: A = (a_i)_n$
matrice diagonale	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = 0$ pour $i \neq j$
matrice triangulaire supérieure	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = 0$ pour $i < j$
matrice triangulaire inférieure	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = 0$ pour $i > j$
matrice unité notée \mathbb{I}	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = \delta_j^i$
matrice nulle	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = 0 \forall i, j$
matrice symétrique	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = a_{ji}$
matrice antisymétrique	$: A = (a_{ij})_{n,n},$ avec $a_{ij} = -a_{ji} \forall i \neq j.$

III.1.2 Opérations sur les matrices

$A = B$ si A et B sont de même ordre et $a_{ij} = b_{ij}$

$C = A + B$ si A et B sont de même ordre et $c_{ij} = a_{ij} + b_{ij}$

$D = \lambda A$ signifie que $d_{ij} = \lambda a_{ij}$

$C = A \times B$ si $A = (a_{ij})_{m,n}$ et $B = (b_{ij})_{np}$ et

$$c_{ij} = \sum_{k=1}^n a_{ik} b_{kj}.$$

La matrice transposée de A est notée A^t et si $A = (a_{ij})_{m,n}$ alors $a^t = (a_{ji})_{n,m}$

Transposée d'un produit de matrice $(AB)^t = B^t A^t$

Matrice inverse : soient deux matrices A et B telles que $AB = BA = \mathbb{I}$; B est l'inverse de la matrice A et on a : $B = A^{-1}$

Norme d'une matrice

$$\|A\| = \max_i \sum_j |a_{ij}| \quad \text{ou} \quad \|A\| = \max_j \sum_i |a_{ij}|.$$

Matrice symétrique définie positive : soit $(x_i)^t = (x_1, x_2, \dots, x_n)$ un vecteur ; une matrice symétrique A est définie positive si :

$$\sum_{i=1}^n \sum_{j=1}^m a_{ij} x_i x_j > 0.$$

III.2 Systèmes linéaires

Étant donné n variables indépendantes x_i , on qualifie de linéaire toute équation du premier ordre définie par

$$a_1 x_1 + a_2 x_2 + \dots + a_n x_n = b_n.$$

Un système linéaire est donc

$$\begin{cases} a_{11} x_1 + a_{12} x_2 + a_{13} x_3 + \dots + a_{1n} x_n = b_1 \\ a_{21} x_1 + a_{22} x_2 + a_{23} x_3 + \dots + a_{2n} x_n = b_2 \\ \vdots \\ a_{m1} x_1 + a_{m2} x_2 + a_{m3} x_3 + \dots + a_{mn} x_n = b_m \end{cases}$$

soit

$$\sum_{j=1}^n a_{ij} x_j = b_i, \quad i = 1, 2, 3, \dots, m.$$

On pose $AX = B$ où A est la matrice m lignes n colonnes, X et B sont les matrices colonnes suivantes :

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mn} \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}.$$

Remarque III.2.1 *Si le nombre d'équations est supérieur au nombre d'inconnues alors il existe des solutions dans des cas particuliers ;*

Si le nombre d'équations est inférieur au nombre d'inconnues alors il peut y avoir une ou plusieurs solutions ;

Un système est homogène si $b_i = 0$ pour tout i ; il admet la solution $x_i = 0$ pour tout i ;

Un système pour lequel $\det A = 0$ est dit singulier.

Remarque III.2.2 Les matrices que nous considérerons dans la suite sont des matrices carrées.

Propriété III.2.1 *On ne change pas la solution d'un système linéaire*

en multipliant les éléments d'une ligne (second membre compris) par un même nombre ;

en ajoutant aux éléments d'une ligne une combinaison linéaire des éléments d'une autre ligne.

Un système linéaire $AX = B$ est dit mal conditionné lorsqu'une faible variation du vecteur B ou de la matrice A entraîne une grande variation de la solution X .

III.3 Résolution de systèmes linéaires non homogènes

Soit A une matrice carrée d'ordre n . On suppose maintenant que le système $AX = B$ possède la propriété suivante :

$$\det A \neq 0$$

Suivant cette hypothèse la solution du système $AX = B$ existent et elle est unique.

III.3.1 Algorithmes de résolution directe

Ces algorithmes sont basés sur la transformation (par éliminations successives des inconnues) de la matrice A ; la matrice A devient alors soit diagonale soit triangulaire.

III.3.1.1 Système à matrice diagonale

La matrice A est telle que $a_{ij} = 0 \forall i \neq j$ et $a_{ii} \neq 0 \forall i$.

Le système s'écrit alors sous la forme $a_{ii}x_i = b_i \quad i = 1, 2, \dots, n$

La solution du système est simplement donnée par

$$x_i = \frac{b_i}{a_{ii}} \quad i = 1, 2, \dots, n.$$

III.3.1.2 Système à matrice triangulaire

Rappelons qu'on a $a_{ij} = 0 \forall i > j$ avec $a_{ii} \neq 0 \forall i$.

Le système s'écrit

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n &= b_{n-1} \\ a_{nn}x_n &= b_n \end{aligned}$$

Le système peut être résolu facilement par la méthode de *retour en arrière* :

– x_n est déduit de la dernière équation

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

– Connaissant la solution x_n , on la reporte dans l'avant dernière équation ; il vient alors

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

puis de proche en proche on obtient $x_{n-2}, x_{n-3}, \dots, x_1$.

– L'algorithme est :

$$i = n-1, n-2, \dots, 1 \begin{cases} x_n &= \frac{b_n}{a_{nn}} \\ x_i &= \frac{1}{a_{i,i}} \left(b_i - \sum_{j=i+1}^n a_{ij}x_j \right) \end{cases}.$$

III.3.2 Transformation des systèmes linéaires

III.3.2.1 Méthode de Gauss sans stratégie de pivot

Principe :

La méthode de Gauss consiste à transformer un système linéaire $AX = B$ en une suite de systèmes linéaires : (A étant une matrice carrée)

$$A^{(1)}X = B^{(1)}; A^{(2)}X = B^{(2)}; \dots A^{(n)}X = B^{(n)}$$

ayant même solution que le système initial $A^{(0)}X = B^{(0)}$ et tel que la matrice finale $A^{(n)}$ soit une matrice triangulaire supérieure.

Le système final $A^{(n)}X = B^{(n)}$ est alors résolu par la méthode de retour en arrière.

On a :

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n = b_2 \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n = b_n \end{cases} \implies \begin{cases} F_1(x) = 0 \\ F_2(x) = 0 \\ \vdots \\ F_n(x) = 0 \end{cases}$$

donc :

$$\begin{cases} F_1(x) = a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n - b_1 \\ F_2(x) = a_{21}x_1 + a_{22}x_2 + \dots + a_{2n}x_n - b_2 \\ \vdots \\ F_n(x) = a_{n1}x_1 + a_{n2}x_2 + \dots + a_{nn}x_n - b_n \end{cases}$$

On retranche à la $i^{\text{ème}}$ équation la première équation multipliée par le coefficient a_{i1}/a_{11} ; ce qui permet d'éliminer l'inconnue x_1 de toute les équations sauf de la première.

Le système devient alors :

$$\begin{cases} F_1^{(1)}(x) = a_{11}^{(1)}x_1 + a_{12}^{(1)}x_2 + \dots + a_{1n}^{(1)}x_n - b_1^{(1)} \\ F_2^{(1)}(x) = 0 + a_{22}^{(1)}x_2 + \dots + a_{2n}^{(1)}x_n - b_2^{(1)} \\ \vdots \\ F_n^{(1)}(x) = 0 + a_{n2}^{(1)}x_2 + \dots + a_{nn}^{(1)}x_n - b_n^{(1)} \end{cases}$$

avec

$$\begin{aligned} a_{22}^{(1)} &= a_{22}^{(0)} - \frac{a_{21}}{a_{11}}a_{12}, \dots, a_{22}^{(1)} = a_{2n}^{(0)} - \frac{a_{21}}{a_{11}}a_{1n} \\ &\vdots \\ a_{n2}^{(1)} &= a_{n2}^{(0)} - \frac{a_{n1}}{a_{11}}a_{12}, \dots, a_{n2}^{(1)} = a_{nn}^{(0)} - \frac{a_{n1}}{a_{11}}a_{1n} \\ b_2^{(1)} &= b_2^{(0)} - \frac{a_{21}}{a_{11}}b_1, \dots, b_n^{(1)} = b_n^{(0)} - \frac{a_{21}}{a_{11}}b_1 \end{aligned}$$

La première équation $F_1(x)$ sert à éliminer x_1 dans toutes les autres équations : c'est l'équation pivot et le coefficient a_{11} est le pivot. De façon générale l'équation k sert à éliminer l'inconnue x_k dans toutes les autres équations qui suivent. Donc une deuxième étape servira à éliminer x_2 dans les équations $3, 4, 5, \dots, n$ grâce au pivot a_{22} de l'équation pivot $F_2^{(1)}(x)$. Ainsi de proche à proche, on élimine successivement toutes les inconnues des équations d'ordre supérieur à l'inconnue. La matrice finale est alors triangulaire supérieure.

Si nous considérons la matrice augmentée du système i.e la matrice à n ligne et $n+1$ colonnes formée à partir de la matrice A et du vecteur B , nous pouvons définir la démarche suivante :

$$\begin{aligned}
 a_{ij}^{(k+1)} &= a_{ij}^{(k)} && \text{pour } \begin{cases} i = 1, \dots, k \\ j = 1, \dots, n+1 \end{cases} \\
 \text{pour } k = 1, \dots, n-1 \quad a_{ij}^{(k+1)} &= 0 && \text{pour } \begin{cases} i = k+1, \dots, n \\ j = 1, \dots, k \end{cases} \\
 a_{ij}^{(k+1)} &= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)} && \text{pour } \begin{cases} i = k+1, \dots, n \\ j = k+1, \dots, n+1 \end{cases}
 \end{aligned}$$

Exemple III.3.1 On veut résoudre par la méthode de Gauss sans stratégie de pivot le système $AX = B$ avec

$$A = \begin{pmatrix} 2 & 6 & 2 & 8 \\ 1 & 4 & 2 & 2 \\ 1 & 1 & 2 & 4 \\ 1 & 1 & 1 & 1 \end{pmatrix} \quad \text{et } B = \begin{pmatrix} 16 \\ 5 \\ 9 \\ 2 \end{pmatrix}$$

On trouvera

$$A^{(3)} \left(\begin{array}{cccc|c} 2 & 6 & 2 & 8 & 16 \\ 0 & 1 & 1 & -2 & -3 \\ 0 & 0 & 3 & -4 & -5 \\ 0 & 0 & 0 & -13/3 & -26/3 \end{array} \right)$$

d'où la solution

$$X_s = \begin{pmatrix} -1 \\ 0 \\ 1 \\ 2 \end{pmatrix}.$$

On doit remarquer que les pivots successifs a_{kk} ont été supposés non nuls dans l'algorithme précédent, alors que ceux-ci interviennent comme diviseurs. Si l'un d'entre eux est nul, le processus doit être modifié : lorsque le pivot $a_{kk} = 0$, on permute la ligne k correspondante avec l'une quelconque des lignes suivantes qui possède un pivot non nul et on reprend le processus normal.

Stratégie du pivot partiel

Il se peut que le pivot, sans être nul, soit très petit en comparaison des autres éléments de la matrice A ; on montre que c'est une cause importante d'erreur d'arrondi. Aussi pour une bonne précision des calculs, il est préférable que les pivots successifs soient les plus grands possibles en valeur absolue. On peut donc rechercher parmi les équations possibles, celle dont la valeur absolue du pivot est maximale et l'utiliser comme équation pivot.

Par suite, si l'on trouve un pivot maximal nul, c'est que tous les pivots possibles sont nuls, par conséquent, la matrice est dite singulière ($\det(A) = 0$).

Stratégie du pivot total

On peut aussi appliquer cette stratégie aux colonnes disponibles, c'est la stratégie du pivot total. Cette méthode plus générale est cependant plus délicate à appliquer car une permutation des colonnes entraîne une permutation dans l'ordre des inconnues, changement dont il faudra tenir compte lors de l'utilisation de la méthode de retour en arrière.

Remarque III.3.1 Il n'est pas nécessaire de stocker en mémoire toutes les matrices intermédiaires $A^{(k)}$. On se sert en général de deux matrices déclarées.

Pour la résolution du système triangulaire, on ne se sert pas des éléments situés en dessous de la diagonale principale de $A^{(n)}$. Qu'on les ait ou non mis à zéro n'a donc aucune importance, ce qui simplifie encore la programmation.

III.3.2.2 Méthode de Gauss-Jordan

La méthode de Gauss-Jordan constitue une variante de la méthode de Gauss : elle procède aussi par élimination des inconnues mais lorsqu'une inconnue est éliminée, elle l'est de toutes les autres équations du système, i.e aussi bien des équations précédentes que des équations suivantes par rapport au pivot utilisé. La matrice A du système linéaire est alors transformée en une matrice unité \mathcal{I} , de sorte que la solution s'obtient directement.

Considérons alors la matrice A_a carrée d'ordre n , à laquelle nous avons rajouté le second membre :

$$A_a = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} & a_{1,n+1} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} & a_{2,n+1} \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} & a_{n,n+1} \end{pmatrix}.$$

Nous supposons aussi qu'il n'y a pas d'éléments nuls sur la diagonale. La méthode de Gauss-Jordan peut se résumer en la succession d'étapes suivantes :

- ① On divise la première ligne par a_{11} ;
- ② On multiplie la première ligne par a_{k1} et on la retranche terme à terme de la $k^{textime}$ ligne, pour $k = 2, 3, \dots, n$. L'élément a_{k1} est annulé sauf pour $k = 1$;
- ③ On divise la deuxième ligne par $a_{22}^{(1)}$; l'indice supérieur représente le nombre d'opérations effectuées sur l'élément ;
- ④ On multiplie la deuxième ligne par $a_{k2}^{(1)}$ et on la retranche terme à terme de la $k^{textime}$ ligne, pour $k = 1, 3, 4, \dots, n$. Ce qui annule les termes de la deuxième colonne sauf la première ligne.
- ⑤ On répète les étapes 3. et 4. pour les lignes $j = 3, 4, \dots, n$ en faisant varier l'indice k de 1 à $j - 1$ puis de $j + 1 = n$.

Exemple III.3.2 Résoudre en utilisant la méthode de Gauss-Jordan le système linéaire suivant :

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

L'application de la méthode conduira à

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & -1 \\ 0 & 0 & 1 & 2 \end{pmatrix}$$

L'algorithme s'écrit alors

$$\begin{aligned} &\text{pour } i = j + 1, \dots, n + 1 \\ &\quad a_{ji} = \frac{a_{ji}}{a_{jj}} \\ &\text{pour } j = 1, \dots, n \\ &\quad \text{pour } k = 1, \dots, n (k \neq j) \\ &\quad\quad a_{ki} = a_{ki} - a_{ji}a_{kj} \end{aligned}$$

Remarque III.3.2 On peut introduire dans cette procédure, une stratégie de pivot qui évite la restriction du départ, à savoir aucun terme sur la diagonale n'est nul.

III.3.2.3 Méthode de Crout ($A = LR$)

Cette méthode est basée sur la transformation de la matrice A en produit de deux matrices en laissant le second membre B intact. Ce qui a l'avantage de pouvoir utiliser, avec la même décomposition, plusieurs seconds membres, sans allourdir considérablement les calculs.

On décompose la matrice A en produit de deux matrices \mathcal{L} et \mathcal{R} où :

- \mathcal{L} est triangulaire inférieure avec des 1 sur la diagonale ;
- \mathcal{R} est triangulaire supérieure.

On pose

$$\mathcal{L} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{pmatrix} \quad \mathcal{R} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\ 0 & r_{22} & r_{23} & \dots & r_{2n} \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \dots & r_{nn} \end{pmatrix}.$$

Ainsi le système initial $AX = B$ est transformé en deux systèmes triangulaires plus simple à résoudre :

$$AX = B \iff \mathcal{L}(\mathcal{R}X) = B$$

ce qui peut s'écrire sous la forme de deux égalités :

$$\begin{cases} \mathcal{L}Y = B \\ \mathcal{R}X = Y \end{cases}$$

La résolution du premier système donne un vecteur Y qui devient le second membre du deuxième système dont le vecteur solution est évidemment la solution cherchée.

Il faut donc maintenant déterminer les deux matrices \mathcal{L} et \mathcal{R} ; ceci peut se faire par identification avec la matrice A une fois que le produit est fait. On trouve comme formule :

- pour $i = 1, \dots, n$ $r_{1i} = a_{1i}$ et $l_{i1} = a_{11}/a_{11}$
- pour $i = 2, \dots, n$

$$\text{pour } j = i, \dots, n \quad r_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}r_{kj}$$

$$\text{pour } k = i+1, \dots, n \quad l_{ki} = \left(a_{ki} - \sum_{j=1}^{i-1} l_{kj}r_{ji} \right) / r_{ii}.$$

On remarque que, pour la détermination des coefficients l_{ij} , on divise par les coefficients r_{ii} , si ceux-ci sont nuls on ne peut pas décomposer la matrice A qui est alors singulière. Le système n'a donc pas de solution.

Les deux matrices \mathcal{L} et \mathcal{R} peuvent être stockées dans les cases mémoires occupées par la matrice A de départ, les termes de la diagonale de \mathcal{L} (qui sont égaux à 1) n'étant pas sauvegardés.

La méthode de Crout est une méthode plus économique en temps de calcul que les méthodes d'élimination.

III.3.2.4 Méthode de Cholesky ($A = LL^t$)

Cette méthode ne s'applique que sur les matrices carrées symétriques et définies positives. C'est-à-dire que toutes ses valeurs propres sont réelles et strictement positives.

Soit donc une matrice A symétrique et définie positive. On considère le système linéaire suivant :

$$AX = B.$$

On peut montrer que la matrice A se décompose sous la forme :

$$A = \mathcal{L}\mathcal{L}^t$$

où \mathcal{L} est une matrice triangulaire inférieure.

Le système devient alors :

$$\mathcal{L}\mathcal{L}^t X = B.$$

Comme dans le cas de la méthode précédente, nous sommes ramenés à la résolution de deux systèmes triangulaires simples

$$\begin{cases} \mathcal{L}Y = B \\ \mathcal{L}^t X = Y \end{cases}$$

Si on appelle l_{ij} les termes de la matrice \mathcal{L} , alors par identification des termes de la matrice $\mathcal{L}\mathcal{L}^t$ aux termes de la matrice A , on trouve :

$$\begin{aligned} l_{11} &= \sqrt{a_{11}} \\ l_{i1} &= a_{i1}/l_{11} \quad \text{pour } i = 2, \dots, n \\ l_{jj} &= \left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right)^{1/2} \quad \text{pour } j = 2, \dots, n \\ l_{ij} &= \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk} \right) / l_{jj} \quad \text{pour } j = 2, \dots, n; i = j+1, \dots, n \end{aligned}$$

Si la matrice A est symétrique mais non définie positive alors il existe au moins un indice j tel que

$$\left(a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2 \right) < 0.$$

Puisqu'il faut prendre la racine carrée de cette quantité pour calculer l_{ij} , il est impossible de continuer et l'algorithme s'arrête. Au total, la résolution d'un système linéaire par la méthode de Cholesky nécessite n extractions de racines carrées et de l'ordre de $n^3/3$ opérations arithmétiques.

Remarque III.3.3 Il est possible de savoir si un système $AX = B$ est bien ou mal conditionné en connaissant son conditionnement, défini par :

$$\text{cond}(A) = \|A\| \|A^{-1}\|$$

où $\|\cdot\|$ est une norme matricielle quelconque. On a toujours $\text{cond}(A) > 1$, et plus ce nombre est grand plus la résolution du système est difficile.

Définition III.3.1 On définit le conditionnement d'une matrice A symétrique définie positive comme le rapport entre la valeur maximale et la valeur minimale des ses valeurs propres :

$$\text{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

III.3.3 Méthode de résolution indirecte ou méthode itérative

L'idée des méthodes itératives est de construire une suite de vecteurs $X^{(k)}$ qui converge vers le vecteur X , solution du système $AX = B$,

$$X = \lim_{k \rightarrow \infty} X^{(k)}$$

L'intérêt des méthodes itératives, comparées aux méthodes directes, est d'être simples à programmer et de nécessiter moins de place en mémoire. En revanche le temps de calcul est souvent plus long.

Une stratégie est de considérer la relation de récurrence linéaire

$$X^{(k+1)} = GX^{(k)} + g \tag{III.1}$$

où G est la matrice d'itération de la méthode itérative (dépendant de A) et g est un vecteur (dépendant de B), tels que

$$X = GX + g.$$

Étant $X = A^{-1}B$, on obtient $g = (I - B)A^{-1}B$; la méthode itérative III.1 est donc complètement définie par la matrice G .

En définissant l'erreur au pas k comme

$$e^{(k)} = X - X^{(k)}$$

on obtient la relation de récurrence

$$e^{(k)} = Ge^{(k-1)}, \quad \text{et donc } e^{(k)} = G^{(k)}e^{(0)}, \quad k = 0, 1, \dots$$

On démontre que $\lim_{k \rightarrow \infty} e^{(k)} = 0$ pour tout $e^{(0)}$ (et donc pour tout $X^{(0)}$) si et seulement si $\rho(G) < 1$, où $\rho(G)$ est le rayon spectral de la matrice G , défini comme

$$\rho(G) = \max |\lambda_i(G)|,$$

et $\lambda_i(G)$ sont les valeurs propres de la matrice G .

Une technique générale pour construire des méthodes itératives est basée sur une décomposition (splitting) de la matrice A sous la forme $A = P - N$, où P et N sont des matrices à déterminer avec P non singulière. La matrice P est appelée matrice de *préconditionnement*. Plus précisément, $X^{(0)}$ étant donné, on peut calculer $X^{(k)}$, pour $k \geq 1$, en résolvant le système

$$PX^{(k+1)} = Nx^{(k)} + B, \quad k \geq 0 \tag{III.2}$$

Clairement, la solution exacte X satisfait $PX = NX + B$ et donc $AX = B$.

Le système III.2 peut être écrit également sous la forme III.1, avec $G = P^{-1}N$, et $g = P^{-1}B$.

III.3.3.1 Méthode de Jacobi

On remarque que si les éléments diagonaux de A sont non nuls, le système linéaire $AX = B$ est équivalent à

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1, j \neq i}^n a_{ij}x_j \right) \quad i = 1, \dots, n.$$

Pour une donnée initiale $X^{(0)}$ choisie, on calcule $X^{(k+1)}$ par

$$x_i^{(k+1)} = - \sum_{j=1; j \neq i}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \quad \text{pour } i = 1, \dots, n.$$

Et dans ce cas, on a

$$A = D - E - F$$

avec

D : matrice diagonale

$$d_{ii} = a_{ii}, \quad d_{ij} = 0 \quad \forall i \neq j$$

E : matrice strictement triangulaire inférieure de A :

$$E_{ij} = -a_{ij} \quad \forall i > j \quad \text{et} \quad E_{ii} = 0 \quad \forall i \leq j$$

F : matrice strictement triangulaire supérieure de A :

$$F_{ij} = -a_{ij} \quad \forall i < j \quad \text{et} \quad F_{ii} = 0 \quad \forall i \geq j$$

La matrice d'itération de la méthode de Jacobi est donnée par :

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

On peut montrer que cette méthode converge si la matrice A est inversible à diagonale dominante i.e :

$$|a_{ii}| \geq \sum_{j=1; j \neq i}^n \|a_{ij}\|.$$

Notons que l'algorithme de Jacobi nécessite le stockage des deux vecteurs $X^{(k)}$ et $X^{(k+1)}$.

Remarque III.3.4 *En pratique, la stricte dominance n'est pas indispensable. L'inégalité large suffit pour la plus par des matrices inversibles.*

Si A est une matrice symétrique définie positive, alors la méthode de Gauss-Seidel converge (la méthode de Jacobi pas forcément).

III.3.3.2 Méthode de Gauss-Seidel

Elle consiste à prendre :

$$A = P - N = (D - E) - F$$

avec

D : matrice diagonale

$$d_{ii} = a_{ii}, \quad d_{ij} = 0 \quad \forall i \neq j$$

E : matrice strictement triangulaire inférieure de A :

$$E_{ij} = -a_{ij} \quad \forall i > j \quad \text{et} \quad E_{ii} = 0 \quad \forall i \leq j$$

F : matrice strictement triangulaire supérieure de A :

$$F_{ij} : -a_{ij} \quad \forall i < j \quad \text{et} \quad F_{ii} = 0 \quad \forall i \geq j$$

On suppose que $D - E$ est inversible c'est-à-dire : $\forall i \ a_{ii} \neq 0$.

La matrice $G = (D - E)^{-1}F$ est appelée *matrice de Gauss-Seidel* associée à A .

Nous pouvons écrire l'algorithme de la méthode de Gauss-Seidel

la suite de vecteurs $X^{(k)}$ de composantes $(x_i^{(k)})$ est définie par :

x⁽⁰⁾ est quelconque ;

l

$$x_i^{(k+1)} = -\sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} x_j^{(k+1)} - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} x_j^{(k)} + \frac{b_i}{a_{ii}}, \quad \text{pour } i = 1, \dots, n.$$

La méthode de Gauss-Seidel converge si la matrice du système A est à diagonale dominante.

Exemple III.3.3 Utiliser les deux méthodes itératives pour résoudre le système suivant :

$$\begin{bmatrix} 1 & -1/2 & 1/2 \\ 1/2 & 1 & 0 \\ -1/2 & -1/2 & 1 \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5/2 \\ -1/2 \\ 2 \end{pmatrix}.$$

III.4 Exercices

Exercice III.4.1 On veut résoudre le système linéaire $AX = b$ où

$$a = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 2 & 5 \\ 4 & 6 & 8 \end{bmatrix} \quad \text{et} \quad b = \begin{bmatrix} 1 \\ 2 \\ 5 \end{bmatrix}$$

par la méthode d'élimination de Pivot ou la méthode de décomposition LR.

1. Vérifier que l'algorithme de Gauss sans pivoting ne peut pas être exécuté jusqu'au bout.
2. Utiliser la stratégie de pivot afin de triangulariser cette matrice puis trouver le résultat par l'algorithme de retrograde.
3. Calculer la factorisation LR de la matrice A.
4. Résoudre le système linéaire $AX = b$ en remplaçant A par LR et en utilisant les algorithmes de substitution progressive et retrograde.

Exercice III.4.2 Considérons la matrice symétrique suivante

$$A = \begin{bmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{bmatrix}$$

1. Pour quelles valeurs de a la matrice A est-elle définie positive ?
2. Ecrire la matrice J de l'itération de Jacobi.
3. Pour quelles valeurs de a la méthode de Jacobi converge-t-elle ?
4. Ecrire la matrice G de l'itération de Gauss-Seidel.
5. Calculer $\rho(G)$. Pour quelles valeurs de a cette méthode converge-t-elle plus vite que celle de Jacobi ?

Exercice III.4.3 1. Soit A une matrice définie positive. Montrer que son déterminant est positif.

2. On considère la matrice A définie par :

$$A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}.$$

Donner sa décomposition de Cholesky. En déduire le déterminant de la matrice A.

Exercice III.4.4 Soient α et β deux réels. Considérons les matrices symétriques suivantes

$$A_\alpha = \begin{pmatrix} 2 & \alpha & 0 \\ \alpha & 2 & \alpha \\ 0 & \alpha & 2 \end{pmatrix} \quad \text{et} \quad C_\beta = \begin{pmatrix} 1 & \beta & \beta \\ \beta & 1 & \beta \\ \beta & \beta & 1 \end{pmatrix}$$

1. Pour quelles valeurs de α la matrice A_α est-elle définie positive ? Donner la décomposition de Cholesky de la matrice A_α .
2. Pour quelles valeurs de β la matrice C_β est-elle définie positive ?
3. Écrire la matrice G_α de l'itération de Gauss-Seidel. Pour quelles valeurs de α la méthode converge ?
4. Écrire la matrice G_β de l'itération de Gauss-Seidel. Pour quelles valeurs de β la méthode converge ?

Exercice III.4.5 1. Considérons la matrice symétrique suivante

$$A = \begin{bmatrix} 2 & a & 0 \\ a & 4 & a \\ 0 & a & 2 \end{bmatrix}$$

- (a) Pour quelles valeurs de a la matrice A est-elle définie positive ?
- (b) Pour quelles valeurs de a la méthode de Gauss-Seidel converge-t-elle ?
- (c) Calculer le rayon spectral de la matrice A défini par

$$\rho(A) = \max |\lambda_i(A)|$$

- (d) Calculer le conditionnement de la matrice A :

$$\text{cond}(A) = \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}.$$

2. Considérons les systèmes linéaires

$$\begin{cases} 4x_1 + 3x_2 + 3x_3 = 10 \\ 3x_1 + 4x_2 + 3x_3 = 10 \\ 3x_1 + 3x_2 + 4x_3 = 10 \end{cases} \quad \begin{cases} 4x_1 + x_2 + x_3 = 6 \\ x_1 + 4x_2 + x_3 = 6 \\ x_1 + x_2 + 4x_3 = 6 \end{cases}$$

1. Rappeler une condition suffisante de convergence pour les méthodes de GAUSS-SEIDEL. Rappeler une autre condition suffisante de convergence pour la méthode de GAUSS-SEIDEL. Les des systèmes vérifient-ils ces conditions ?
2. Écrire la méthode de GAUSS-SEIDEL pour ces deux systèmes linéaires.
3. On illustrera les résultats théoriques de convergence/non-convergence de ces deux schémas en prenant comme point de départ le vecteur $(x_1, x_2, x_3) = (0, 0, 0)$ et en calculant les 3 premiers itérés avec la méthode de GAUSS-SEIDEL pour les système.

Chapitre IV

INTERPOLATION LINÉAIRE

IV.1 Introduction

Le problème de l'interpolation linéaire ou l'approximation a de nombreuses application en calcul numérique et dans les sciences expérimentales. On veut généralement représenter le phénomène observé par une fonction simple soit que l'on ne connaisse pas la loi exacte qui le régit soit que l'on veuille en rendre compte par une fonction plus simple. Cela servira à obtenir une valeur approchée du phénomène mesuré entre les points de mesure.

Dans de nombreux cas, nous cherchons une fonction g qui passe par tous les points. Généralement, on choisit une fonction simple qui soit continue et dérivable : un polynôme. C'est l'interpolation polynomiale.

Un deuxième choix est possible, nous pouvons chercher une fonction g , d'un type préalablement choisi, telle que la distance entre les points connus et les points approchés soit minimale : c'est l'approximation.

IV.2 Interpolation polynomiale

IV.2.1 Problème de l'interpolation

On considère une fonction de \mathbb{R} dans \mathbb{R} , dont on connaît les valeurs y_i qu'elle prend sur $n + 1$ abscisses x_i , $i = 0, \dots, n$.

Le problème de l'interpolation polynomiale est de déterminer un polynôme de degré inférieur ou égal à n , qui prend les valeurs y_i sur les $n + 1$ abscisses x_i . On considère le polynôme de degré inférieur ou égal à n :

$$\mathcal{P}(x) = \alpha_0 + \alpha_1 x + \alpha_2 x^2 + \dots + \alpha_i x^i + \dots + \alpha_n x^n.$$

Les $n + 1$ relations $\mathcal{P}(x_i) = y_i$ forment alors un système linéaire non homogène d'ordre $n + 1$. Il est bien évident que ce système s'écrit matriciellement sous la forme :

$$\begin{bmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \dots & \dots & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \dots & \dots & x_1 & 1 \\ \vdots & \vdots & \vdots & & & \vdots & \vdots \\ \vdots & \vdots & \vdots & & & \vdots & \vdots \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \dots & \dots & x_n & 1 \end{bmatrix} \begin{pmatrix} \alpha_n \\ \alpha_{n-1} \\ \vdots \\ \vdots \\ \alpha_0 \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ \vdots \\ y_n \end{pmatrix}$$

IV.3 Approximation polynomiale

IV.3.1 Méthode d'interpolation de Lagrange

On considère les polynômes ϕ_i , $i = 1, \dots, n$ de degré n tels que

$$\phi_i(x) = \delta_j^i, \quad i, j = 0, \dots, n$$

où $\delta_j^i = 1$ si $i = j$ et $\delta_j^i = 0$ si $i \neq j$. Ces polynômes sont définis de façon suivante :

$$\phi_i(x) = \prod_{i \neq j} \frac{(x - x_j)}{(x_i - x_j)}, \quad 0 \leq i \leq n.$$

Par suite le polynôme d'interpolation de la fonction f aux points x_i , $i = 0, \dots, n$ s'écrit :

$$\mathcal{P}_n(x) = \sum_{i=0}^n f(x_i) \phi_i(x). \quad (\text{IV.1})$$

Théorème IV.3.1 *Le problème d'interpolation $\mathcal{P}(x_i) = f(x_i)$, $i = 0, \dots, n$ admet une solution et une seule, donnée par la formule IV.1.*

Exemple IV.3.1 Trouver le polynôme d'interpolation de Lagrange pour les points suivants :

x_i	2	3	-1	4
y_i	1	-1	2	3

IV.3.2 Polynôme d'interpolation de Newton : Méthode des différences divisées

Principe

Les $n + 1$ polynômes définis par

$$\begin{aligned} N_0(x) &= 1 \\ N_1(x) &= (x - x_0) \\ N_2(x) &= (x - x_0)(x - x_1) \\ &\vdots \\ N_n(x) &= (x - x_0)(x - x_1) \dots (x - x_n) \end{aligned}$$

sont linéairement indépendants et forment une base de l'espace vectoriel des polynômes de degré inférieur ou égal à n . Ces polynômes s'appellent *polynômes de Newton*. Dans cette base, un polynôme $\mathcal{P}(x)$ s'écrit :

$$\mathcal{P}(x) = \alpha_0 N_0(x) + \alpha_1 N_1(x) + \dots + \alpha_n N_n(x).$$

Si ce polynôme est le polynôme d'interpolation de $f(x)$ i.e $\mathcal{P}(x_i) = f(x_i)$, alors il vient que les α_i sont solutions du système linéaire

$$\begin{bmatrix} 1 & 0 & 0 & \dots & & 0 \\ 1 & (x_1 - x_0) & 0 & & & \vdots \\ \vdots & \vdots & & & & \vdots \\ 1 & (x - x_0) & \dots & \dots & ((x_n - x_0) \dots (x_n - x_{n-1})) \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Les coefficients α_i peuvent être déterminés à l'aide des différences divisées de la fonction f au lieu de les déterminer par la résolution du système ci-dessus.

On appelle différence divisée d'ordre 1 les quantités :

$$f[x_0, x_1] = \frac{f(x_0) - f(x_1)}{x_0 - x_1}, \quad f[x_2, x_1] = \frac{f(x_2) - f(x_1)}{x_2 - x_1}, \dots$$

Par convention on appelle différence divisée d'ordre 0 la valeur des fonctions notée :

$$f[x_0] = f(x_0), \quad f[x_1] = f(x_1), \dots$$

Par récurrence, on définit les différences divisées d'ordre p en fonction des différences divisées d'ordre $p-1$:

$$f[x_0, x_1, \dots, x_p] = \frac{f[x_0, x_1, \dots, x_{p-1}] - f[x_1, \dots, x_p]}{x_0 - x_p}.$$

On peut montrer la commutativité des arguments, la différence divisée est donc invariante sous l'effet d'une permutation des abscisses.

On montre que

$$\begin{aligned}\alpha_0 &= f(x_0) \\ \alpha_1 &= f[x_0, x_1] \\ \alpha_p &= f[x_0, x_1, \dots, x_p], \quad p = 2, \dots, n\end{aligned}$$

Exemple IV.3.2 Retrouver le polynôme d'interpolation

$$\mathcal{P}(x) = 1 + (x - 2)(-2 + (x - 3)(-5/12 + (x + 1)41/60))$$

avec les données de l'exemple précédent.

IV.4 Approximation au sens des moindres carrés

IV.4.1 Approximation polynomiale

Soit f une fonction réelle d'une variable réelle dont on connaît les valeurs y_i qu'elle prend sur $n+1$ abscisses distinctes x_i .

Posons $\mathcal{P}(x) = a_0x^p + a_1x^{p-1} + \dots + a_p$ un polynôme de degré p avec $p < n$. Considérons le vecteur résidus $R(r_i)$ de composantes $r_i = y_i - \mathcal{P}(x_i)$ pour $i = 0, \dots, n$. On rappelle que la norme euclidienne sur \mathbb{R}^n est la norme notée ϕ_2 et définie pour tout vecteur $X = (x_i)$ par :

$$\phi_2(x) = \left(\sum_{i=0}^n |x_i|^2 \right)^{1/2}.$$

Il s'agit de minimiser la fonction $\phi_2(R)$ de \mathbb{R}^{n+1} dans \mathbb{R} , définie par :

$$\phi_2(R) = \left(\sum_{i=0}^n |r_i|^2 \right)^{1/2} = \left(\sum_{i=0}^n (y_i - \mathcal{P}(x_i))^2 \right)^{1/2}.$$

Considérons la fonction $D(a_0, a_1, \dots, a_p)$ définie par :

$$D(a_0, a_1, a_2, \dots, a_p) = \sum_{i=0}^n (y_i - \mathcal{P}(x_i))^2$$

Une condition nécessaire pour que cette fonction admette un extremum au point $P = (a_i)$ est que :

$$\frac{\partial D}{\partial a_k}(a_0, a_1, \dots, a_p) = 0 \quad k = 0, \dots, p$$

ce qui conduit aux équations

$$\sum_{i=0}^n \mathcal{P}(x_i)x_i^{p-k} = \sum_{i=0}^n y_i x_i^{p-k}, \quad k = 0, \dots, n$$

ou encore

$$a_0 \sum_{i=0}^n x_i^p x_i^{p-k} + \cdots + a_p \sum_{i=0}^n x_i^{p-k} = \sum_{i=0}^n y_i x_i^{p-k}, \quad k = 0, \dots, n.$$

Ces $p+1$ équations forment un système linéaire non homogène qui s'écrit sous la forme matricielle $PA = B$ avec

$$P = \begin{bmatrix} \sum_{i=0}^n x_i^{2p} & \sum_{i=0}^n x_i^{2p-1} & \cdots & \cdots & \sum_{i=0}^n x_i^{p+1} & \sum_{i=0}^n x_i^p \\ \sum_{i=0}^n x_i^{2p-1} & \sum_{i=0}^n x_i^{2p-2} & \cdots & \cdots & \sum_{i=0}^n x_i^p & \sum_{i=0}^n x_i^{p-1} \\ \vdots & \vdots & & & \vdots & \vdots \\ \vdots & \vdots & & & \vdots & \vdots \\ \vdots & \vdots & & & \vdots & \vdots \\ \sum_{i=0}^n x_i^{p+1} & \sum_{i=0}^n x_i^p & \cdots & \cdots & \sum_{i=0}^n x_i^2 & \sum_{i=0}^n x_i \\ \sum_{i=0}^n x_i^p & \sum_{i=0}^n x_i^{p-1} & \cdots & \cdots & \sum_{i=0}^n x_i & n+1 \end{bmatrix} \quad A = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ a_{p-1} \\ a_p \end{bmatrix} \quad B = \begin{bmatrix} \sum_{i=0}^n x_i^p y_i \\ \sum_{i=0}^n x_i^{p-1} y_i \\ \vdots \\ \vdots \\ \sum_{i=0}^n x_i y_i \\ \sum_{i=0}^n y_i \end{bmatrix}$$

On montre que P est symétrique et définie positive donc le système admet une unique solution. Puisque la matrice P est symétrique et définie positive, le système peut être résolu par la méthode de Cholesky. Toutefois, il est préférable d'utiliser la méthode de Gauss avec pivot total, car cette méthode est en général plus précise. D'autre part, le degré p du polynôme $\mathcal{P}(x)$ ne doit pas dépasser 10 car le système précédent est très souvent mal conditionné.

Exemple IV.4.1 Trouver le polynôme d'approximation de degré 2 avec les moindres carrés pour les points suivants :

IV.4.2 Approximation quelconque

Soit f une fonction réelle d'une variable réelle dont on connaît les valeurs y_i qu'elle prend sur $n+1$ abscisses distinctes x_i .

L'approximation polynomiale, bien que très employée, ne peut pas rendre compte de toutes les approximations. Nous généralisons au cas de l'approximation par une fonction quelconque g . La forme de la fonction g est choisie à priori. Des constantes sont donc à déterminer, par exemple : si

$$g(x) = a_0 x^2 \sin(a_1 x + a_2)$$

les valeurs de a_0 , a_1 et a_2 sont à déterminer.

Au point (x_i, y_i) , nous remplaçons la valeur de y_i par :

$$y_i = \sum_{k=0}^p a_k \frac{\partial g(x_i)}{\partial a_k}, \quad i = 0, \dots, n.$$

Nous obtenons un système de $n+1$ équations à $p+1$ inconnues

$$\begin{bmatrix} \frac{\partial g(x_0)}{\partial a_0} & \frac{\partial g(x_0)}{\partial a_1} & \cdots & \cdots & \frac{\partial g(x_0)}{\partial a_p} \\ \frac{\partial g(x_1)}{\partial a_0} & \frac{\partial g(x_1)}{\partial a_1} & \cdots & \cdots & \frac{\partial g(x_1)}{\partial a_p} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \frac{\partial g(x_n)}{\partial a_0} & \frac{\partial g(x_n)}{\partial a_1} & \cdots & \cdots & \frac{\partial g(x_n)}{\partial a_p} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ \vdots \\ y_p \end{bmatrix}$$

ou encore

$$GA = Y.$$

Puisque généralement les dérivées partielles sont fonction des paramètres a_k , il faut donc se donner des valeurs initiales pour a_k et recalculer de nouvelles valeurs par la résolution du système.

En général, le système $GA = Y$ n'a pas de solution numériquement puisque la matrice G est à $n + 1$ lignes et $p + 1$ colonnes et p peut être différent de n . En multipliant à gauche par la matrice transposée de G , on obtient :

$$G^t GA = G^t Y.$$

Exemple IV.4.2 On se donne à priori $g(x) = a_0\sqrt{x} + a_1$. Montrer que $a_0 = 2$ et $a_1 = 1$.

Exercice IV.4.1 On veut donner le développement limité d'ordre 3 de la fonction

$$f(x) = \exp(x^2) \sin(x), \quad x \in [-2, 2].$$

On procède la la méthode des moindres carrés pour déterminer le polynôme d'ordre 3.

1. On souhaite balayer l'intervalle $[-2, 2]$ avec un pas $h = 1/10$. Déterminer les images des points d'abscisses $x_i = -2 + ih$ dans cet intervalle
2. Calculer la matrice P et le vecteur B permettant de calculer les coefficients du polôme qui interpole ces valeurs.
3. Déterminer le polynôme de degré cherché.
4. Tracer la fonction f et le polynôme trouvé dans l'intervalle $[-2, 2]$.

Chapitre V

EQUATIONS DIFFÉRENTIELLES

V.1 Introduction

De nombreux problèmes physiques, mécaniques, astronomiques, chimiques, économiques, ..., se forment en termes d'équations différentielles. Comme il est bien souvent beaucoup trop difficile, voir impossible, d'en obtenir la solution, nous allons en effectuer l'intégration numérique. Nous étudierons les méthodes usuelles telles que celle d'Euler et celles de Runge-Kutta.

On considère une fonction continue $f : [a, b] \times \mathbb{R} \rightarrow \mathbb{R}$. Pour $y_0 \in \mathbb{R}$ donné, on cherche $y : t \in \mathbb{R}$ qui satisfait le problème suivant, appelé *problème de Cauchy* :

$$\begin{cases} y'(t) = f(t, y(t)) \text{ si } t \in [a, b] \\ y(0) = y_0 \end{cases} \quad (\text{V.1})$$

Ici f est une fonction continue par rapport à chacune de ses variables.

Théorème V.1.1 (Cauchy-Lipschitz)

Si f est continue sur $[a, b] \times \mathbb{R}$ et s'il existe une constante $L > 0$ telle que

$$|f(t, u) - f(t, v)| \leq L|u - v| \quad \forall u, v \in \mathbb{R}, \forall t \in [a, b]$$

alors le problème de Cauchy (V.1) admet une solution globale et elle est unique.

Considérons dans l'intervalle $[a, b]$ des abscisses équidistantes t_0, t_1, \dots, t_n définies par :

$$t_i = a + ih, \quad i = 0, 1, \dots, n$$

Il vient que $t_0 = a$ et $t_n = b$; et h est appelé *le pas d'intégration*.

Soit $y(t_i)$ la valeur exacte de la solution de notre problème à l'abscisse t_i . Une méthode d'intégration numérique fournira une valeur approchée y_i de $y(t_i)$ pour $i = 0, 1, \dots, n$ à partir de la condition initiale connue $y_0 = y(t_0)$. Les différentes méthodes de résolution des équations différentielles se distinguent par la manière d'obtenir ces approximations y_i .

V.2 Dérivation numérique

Soit $y : [a, b] \rightarrow \mathbb{R}$ de classe C^1 et $a = t_0, t_1, \dots, t_n = b$, $n + 1$ nœuds équirépartis dans $[a, b]$. On note $h = (b - a)/n$ la distance entre deux nœuds consécutifs. La dérivée $y'(t_i)$ est donnée par

$$\begin{aligned} y'(t_i) &= \lim_{h \rightarrow 0+} \frac{y(t_i + h) - y(t_i)}{h} \\ &= \lim_{h \rightarrow 0+} \frac{y(t_i) - y(t_i - h)}{h} \\ &= \lim_{h \rightarrow 0+} \frac{\cancel{y(t_i + h)} - \cancel{y(t_i - h)}}{2h} \end{aligned}$$

Soit maintenant $(Dy)_i$ une approximation de $y'(t_i)$. On appelle

1. différence finie progressive l'approximation

$$(Dy)_i^P = \frac{y(t_{i+1}) - y(t_i)}{h}, \quad ; i = 0, \dots, n - 1;$$

2. différence finie rétrograde l'approximation

$$(Dy)_i^R = \frac{y(t_i) - y(t_{i-1})}{h}; \quad i = 1, \dots, n;$$

3. différence finie centrée l'approximation

$$(Dy)_i^C = \frac{y(t_{i+1}) - y(t_{i-1})}{2h}; \quad i = 1, \dots, n - 1.$$

Si $y \in C^2(\mathbb{R})$, pour tout $t \in \mathbb{R}$, il existe un η entre t_i et t tel que l'on a le développement de Taylor

$$y(t) = y(t_i) + y'(t_i)(t - t_i) + \frac{y''(\eta)}{2}(t - t_i)^2$$

4. Pour $t = t_{i+1}$ on obtient pour la différence finie progressive

$$(Dy)_i^P = y'(t_i) + \frac{h}{2}y''(\eta).$$

ce qui conduit à une estimation du type

$$|y'(t_i) - (Dy)_i^P| \leq Ch,$$

où $C = \frac{1}{2} \max_{t \in [t_i, t_{i+1}]} |y''(t)|$.

5. Pour $t = t_{i-1}$ on obtient pour la différence finie rétrograde

$$(Dy)_i^R = y'(t_i) - \frac{h}{2}y''(\eta),$$

ce qui conduit à une estimation du type

$$|y'(t_i) - (Dy)_i^R| \leq Ch,$$

où $C = \frac{1}{2} \max_{t \in [t_{i-1}, t_i]} |y''(t)|$.

6. Pour $t = t_{i+1}$ et $t = t_{i-1}$ avec un développement d'ordre 2 (si $y \in C^3$)

$$y(t_{i+1}) = y(t_i) + y'(t_i)h + \frac{y''(t_i)}{2}h^2 + \frac{y'''(\eta_1)}{6}h^3,$$

$$y(t_{i-1}) = y(t_i) - y'(t_i)h + \frac{y''(t_i)}{2}h^2 - \frac{y'''(\eta_2)}{6}h^3,$$

on obtient

$$(Dy)_i^C = y'(t_i) + \frac{y'''(\eta_1) + y'''(\eta_2)}{12} h^2,$$

et donc l'estimation suivante

$$|y'(t_i) - (Dy)_i^C| \leq Ch^2$$

où $C = \frac{1}{6} \max_{t \in [t_i+1, t_{i+1}]} |y''(t)|$.

Définition V.2.1 La différence $|y'(t_i) - (Dy)_i^P|$ (et celles correspondantes aux autres différences finies) est appelée *erreur de troncature au point t_i* .

L'erreur de troncature est d'ordre 1 pour les formules progressive et rétrograde et d'ordre 2 pour la formule centrée.

V.3 Méthodes d'Euler

V.3.1 Méthode d'Euler du premier ordre

Soient $0 = t_0 < t_1 < \dots < t_n < t_{n+1} < \dots$ une suite de nombres réels équirépartis. On note $h = t_{n+1} - t_n$. On notera par

y_n une approximation de $y(t_n)$.

Dans le problème de Cauchy (V.1), pour $t = t_n$ on a

$$y'_0(t_n) = f(t_n; y(t_n)).$$

On approche la dérivée $y'_0(t_n)$ en utilisant des schémas de dérivation numérique.

Schéma d'Euler progressif :

$$\begin{cases} (y_{n+1} - y_n)/h = f(t_n; y_n) \text{ pour } n = 0, 1, 2, \dots \\ y_0 \end{cases}$$

Schéma d'Euler rétrograde :

$$\begin{cases} (y_{n+1} - y_n)/h = f(t_{n+1}; y_{n+1}) \text{ pour } n = 0, 1, 2, \dots \\ y_0 \end{cases}$$

V.3.2 Étude générale des méthodes à un pas

Les méthodes à un pas sont les méthodes de résolution numérique qui peuvent s'écrire sous la forme

$$y_{n+1} = y_n + h_n \Phi(t_n; y_n; h_n); \quad 0 \leq n < N;$$

où Φ est une fonction qu'on supposera continue.

V.3.2.1 Consistance, stabilité, convergence

Définition V.3.1 L'erreur de consistance e_n relative à une solution exacte y est l'erreur

$$e_n = y(t_{n+1}) - y_{n+1}; \quad 0 \leq n < N;$$

en supposant $y_n = y(t_n)$. On a donc

$$e_n = y(t_{n+1}) - y(t_n) - h_n \Phi(t_n; y(t_n); h_n).$$

On dit que la méthode est consistante si pour toute solution exacte y la somme des erreurs de consistance relatives à y , soit $\sum_n |e_n|$, tend vers 0 quand h_{\max} tend vers 0.¹

Une autre notion fondamentale est la notion de stabilité. dans la pratique, le calcul récurrent des points y_n est entaché d'erreurs d'arrondis ϵ_n . Pour que les calculs soient significatifs, il est indispensable que la propagation de ces erreurs reste contrôlable.

Définition V.3.2 On dit que la méthode est stable s'il existe une constante $S \geq 0$ telle que pour toutes suites (y_n) , (\tilde{y}_n) définies par

$$y_{n+1} = y_n + h_n \Phi(t_n; y_n; h_n); \quad \forall n < N; \quad \tilde{y}_{n+1} = \tilde{y}_n + h_n \Phi(t_n; \tilde{y}_n; h_n) + \epsilon_n; \quad \forall n < N;$$

on ait

$$\max_n |\tilde{y}_n - y_n| \leq S \left(|\tilde{y}_0 - y_0| + \sum_n |\epsilon_n| \right)$$

Une dernière notion importante est la suivante

Définition V.3.3 On dit que la méthode est convergente si pour toute solution exacte y , la suite (y_n) vérifie

$$\max_n |y_n - y(t_n)| \rightarrow 0$$

quand $y_0 \rightarrow y(0)$ et $h_{\max} \rightarrow 0$.

Posons $\tilde{y}_n = y(t_n)$. Par définition d'erreur de consistance on a

$$\tilde{y}_{n+1} = \tilde{y}_n + h_n \Phi(t_n; \tilde{y}_n; h_n) + e_n.$$

Si la méthode est stable, de constante S , l'erreur globale est donc

$$\max_n |y_n - y(t_n)| \leq S \left(|y_0 - y(0)| + \sum_n |e_n| \right).$$

V.4 Méthodes de Runge-Kutta d'ordre 2

Si on intègre l'équation $y'(t) = f(t, y(t))$ entre t_n et t_{n+1} on obtient :

$$y(t_{n+1}) - y(t_n) = \int_{t_n}^{t_{n+1}} f(t, y(t)) dt.$$

En utilisant la formule des trapèzes, on trouve le schéma implicite suivant, appelé schéma de Crank-Nicolson ou du trapèze :

$$y_{n+1} - y_n = \frac{h}{2} (f(t_n; y_n) + f(t_{n+1}; y_{n+1})), \quad \forall n \geq 0.$$

Ce schéma est implicite. En le modifiant afin de le rendre explicite, on identifie la méthode de Heun :

$$y_{n+1} - y_n = \frac{h}{2} (f(t_n; y_n) + f(t_{n+1}; y_n + hf(t_n; y_n))).$$

Ce deux méthodes sont d'ordre 2 par rapport à h .

Si on utilise la méthode du point milieu on trouve

$$y_{n+1} - y_n = hf \left(t_{n+\frac{1}{2}}, y_{n+\frac{1}{2}} \right).$$

Si maintenant on approche $y_{n+\frac{1}{2}}$ par

$$y_{n+\frac{1}{2}} = y_n + \frac{1}{2} f(t_n; y_n),$$

on trouve la méthode d'Euler modifiée :

$$y_{n+1} - y_n = h f \left(t_{n+\frac{1}{2}}, y_n + \frac{1}{2} f(t_n; y_n) \right).$$

 Les méthodes de Heun et d'Euler modifiées sont des cas particuliers dans la famille des méthodes de Runge-Kutta d'ordre 2. Il existe d'autres méthodes plus compliquées, comme par exemple la méthode de Runge-Kutta d'ordre 4 suivante, qui est obtenue en considérant la méthode d'intégration de Simpson.

$$y_{n+1} = y_n + \frac{h}{6} (K_1 + 2K_2 + 2K_3 + K_4),$$

où les K_i sont calculés comme suit :

$$\begin{aligned} K_1 &= f(t_n; y_n) \\ K_2 &= f\left(t_n + \frac{h}{2}; y_n + \frac{h}{2} K_1\right) \\ K_3 &= f\left(t_n + \frac{h}{2}; y_n + \frac{h}{2} K_2\right) \\ K_4 &= f(t_{n+1}, y_n + h K_3). \end{aligned}$$

Chapitre VI

Intégration numérique

VI.1 Introduction

On souhaite disposer d'un moyen d'évaluer numériquement $I(f) = \int_a^b f(t)dt$ où f est une fonction continue sur un intervalle $[a, b]$ avec $a < b$. En effet, dans de nombreuses applications, cette intégrale ne se ramène pas à des fonctions simples, et même si c'est le cas, on cherche une méthode simple et utilisable dans le cas général. Pour cela, nous chercherons une approximation de $I(f)$, notée $J(f)$ sous la forme

$$J(f) = (b - a) \sum_{i=0}^n \omega_i f(\xi_i)$$

où les points ξ_i sont dans l'intervalle $[a, b]$ et $\sum_{i=0}^n \omega_i = 1$. Les ξ_i sont appelés les points d'intégration et les coefficients ω_i les poids d'intégration.

Définition VI.1.1 . On dit qu'une méthode de quadrature est d'ordre N si la formule approchée est exacte pour tout $f \in \mathcal{P}_N$ et inexacte pour au moins un $f \in \mathcal{P}_{N+1}$.

On observera que les formules sont toujours exactes pour $f(x) = 1$ à cause de l'hypothèse $\sum_{i=0}^n \omega_i = 1$. Par linéarité, elles sont donc exactes au moins pour $f \in \mathcal{P}_0$.

VI.2 Exemples

1. Un seul point. On choisit un seul point $\xi \in [a, b]$ et on remplace f sur $[a, b]$ par le polynôme de degré 0 : $p_0(x) = f(\xi)$. On a alors

$$\int_a^b f(x)dx \approx (b - a)f(\xi).$$

Voici les choix plus courants :

- (a) $\xi = a$: méthode des rectangles à gauche (ordre 0) ;
(b) $\xi = b$: méthode des rectangles à droite (ordre 0) ;
(c) $\xi = (a + b)/2$: méthode du point milieu (ordre 1) ;
2. Interpolation linéaire. On choisit $\xi_0 = a$ et $\xi_1 = b$ et on remplace f sur $[a, b]$ par la fonction linéaire p_1 qui interpole f aux points a, b :

$$p_1(x) = \frac{(x - a)f(b) - (x - b)f(a)}{b - a}$$

On obtient la formule suivante, dite méthode des trapèzes (ordre 1) :

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

3. Méthodes de Newton-Cotes. On choisit $n + 1$ points équidistants

$$\xi_i = a + i \frac{b - a}{n}$$

Pour déterminer la formule de quadrature élémentaire, on se ramène par changement de variable à l'intervalle $[-1, 1]$, subdivisé par les points $\tau_i = -1 + i2/n$.

Le polynôme d'interpolation d'une fonction $f \in C([-1, 1])$ est donné par

$$p_n(x) = \sum_{i=0}^n f(\tau_i) \phi_i(x),$$

où ϕ_i est le polynôme de base de Lagrange $\phi_i(x) = \prod_{j \neq i} \frac{x - \tau_j}{\tau_i - \tau_j}$. On a donc

$$\int_{-1}^1 f(x) dx \approx \int_{-1}^1 p_n(x) dx = 2 \sum_{i=0}^n \omega_i f(\tau_i)$$

avec $\omega_i = 0.5 \int_{-1}^1 \phi_i(x) dx$. Par suite de la symétrie des points τ_i autour de 0, on a

$$\phi_i(-x) = \phi_i(x), \quad \tau_{n-i} = -\tau_i, \quad \phi_{n-i}(x) = \phi(-x), \quad \omega_{n-i} = \omega_i.$$

Après changement de variable, les coefficients ω_i sont inchangés, donc on obtient la formule

$$\int_a^b f(x) dx \approx (b - a) \sum_{i=0}^n \omega_i f(\xi_i).$$

Si $f \in P_n$, alors $p_n = f$, donc la méthode de Newton-Cotes de rang n est d'ordre supérieur ou égal à n . De plus, lorsque $f \in C([-1, 1])$ est un polynôme impair, on a

$$\int_{-1}^1 f(x) dx = 0 = 2 \sum_{i=0}^n \omega_i f(\tau_i)$$

Si n est pair, les formules sont donc encore exactes pour $f(x) = x^{n+1}$, et plus généralement pour $f \in \mathcal{P}_{n+1}$ par linéarité. On démontre en fait le résultat suivant que nous admettrons :

Proposition VI.2.1 Si n est pair, l'ordre de la méthode de Newton-Cotes de rang n est $n + 1$, si n est impair, l'ordre est n .

Ceci fait que, hormis le cas $n = 1$, les méthodes de Newton-Cotes ne sont utilisées que pour n pair :

(a) $n = 1$ méthode des trapèzes (ordre 1)

$$\omega_0 = \omega_1 = \frac{1}{2}.$$

(b) $n = 2$ méthode de Simpson (ordre 3)

$$\omega_0 = \omega_2 = \frac{1}{6}, \quad \omega_1 = \frac{2}{3}.$$

(c) $n = 4$ méthode de Boole-Villarceau (ordre 5)

$$\omega_0 = \omega_4 = \frac{7}{90}, \quad \omega_1 = \omega_3 = \frac{16}{45}, \quad \omega_2 = \frac{2}{15}.$$

(d) $n = 6$ méthode de Weddle-Hardy (ordre 7)

$$\omega_0 = \omega_6 = \frac{41}{840}, \quad \omega_1 = \omega_5 = \frac{9}{35}, \quad \omega_2 = \omega_4 = \frac{9}{280}, \quad \omega_3 = \frac{34}{105}.$$

Pour $n \geq 8$ il apparaît des coefficients $\omega_i < 0$, ce qui a pour effet de rendre les formules beaucoup plus sensibles aux erreurs d'arrondis. Les méthodes de Newton-Cotes ne sont donc utilisées en pratique que dans les 4 cas ci-dessus.

VI.3 Evaluation de l'erreur. Noyau de Peano

Théorème VI.3.1 . On suppose que la méthode est d'ordre $N \geq 0$. Si f est de classe C^{N+1} sur $[a, b]$, alors

$$E(f) = \int_a^b f(x)dx - (b-a) \sum_{i=0}^n \omega_i f(\xi_i) = \frac{1}{N!} \int_a^b K_N(t) f^{N+1}(t) dt,$$

où K_N est une fonction sur $[a, b]$, appelée noyau de Peano associé à la méthode, définie par

$$K_N(t) = E(x \mapsto (x-t)_+^N), \quad t \in [a, b].$$