UNIVERSITÉ MOHAMED PREMIER - OUJDA

Faculté Des Sciences

Département De Mathématiques et Informatique

Cours du module

Analyse Numérique

SMI S4

Prof. Mohammed BERRAJAA

Année Universitaire 2016/2017

Table des matières

Chapitre 1

Résolution de systèmes linéaires

Méthode direct

1.1 Position du problème :

Dans ce chapitre, nous considérons un système d'équations linéaires d'ordre n de la forme

$$Ax = b \qquad (1.1)$$

Ici A est une $n \times n$ matrice régulière de coefficients $a_{ij}, 1 \leq i, j \leq n$, donnés, b est un vecteur colonne à n composantes b_j, $1 \leq j \leq n$, données et x est un vecteur colonne à n composantes $x_j, 1 \leq j \leq n$, inconnues. Dans la suite, nous utiliserons les notations matricielles standards, i.e

$$\begin{pmatrix} a_{1,1} & a_{1,2} & . & . & . & a_{1,n} \\ a_{2,1} & a_{2,2} & . & . & . & a_{2,n} \\ . & . & & & & . \\ . & . & & & & . \\ . & . & & & & . \\ a_{n,1} & a_{n,2} & . & . & . & a_{n,n} \end{pmatrix}, \qquad b = \begin{bmatrix} b_1 \\ b_2 \\ . \\ . \\ . \\ b_n \end{bmatrix}, \qquad b = \begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_n \end{bmatrix}$$

Le système (4.1) peut s'écrire explicitement sous forme d'un système de n équations à n inconnues $x_1, x_2, ..., x_n$:

$$\begin{cases} a_{1,1}x_1+a_{1,2}x_2+.....+a_{1,n}x_n=b_1 \\ a_{2,1}x_1+a_{2,2}x_2+.....+a_{2,n}x_n=b_2 \\ . \\ . \\ . \\ a_{n,1}x_1+a_{n,2}x_2+.....+a_{n,n}x_n=b_n \end{cases} \quad (1.2).$$

Définition 1.1 : On dira que la matrice A est triangulaire supérieure (respectivement triangulaire inférieure) si $a_{ij}=0$ pour tout couple (i,j) tel que $1 \leq j < i \leq n$ (respectivement $1 \leq i < j \leq n$).

Définition 1.2 : Si A est une matrice triangulaire supérieure (respect. triangulaire inférieure), on dira que les systèmes (1.1) et (1.2) sont triangulaires supérieurs (resp. triangulaire inférieurs).

Supposons un instant que la matrice A soit triangulaire supérieure nous constatons alors que le déterminant de la matrice A est le produit des valeurs diagonales a_{ii} et, puisque A est supposée régulière ; nous avons $a_{ii} \neq 0, 1 \leq i \leq n$. Ainsi, quitte à diviser chaque équation de (1.2) par le terme de la diagonale, il n'est pas restrictif de supposer que $a_{ii}=1, 1 \leq i \leq n$. Dans ce cas, la matrice est une matrice triangulaire avec des valeurs 1 dans sa diagonale et, de (1.2), nous déduisons successivement les inconnues $x_n, x_{n-1}, ..., x_1$. En effet, nous avons :

$$x_n = b_n/a_{n,n}$$

et pour $i = n-1, n-2, ...3, 2, 1$:

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

Dans le cas où la matrice A est régulière mais non nécessairement triangulaire supérieure, la méthode d'élimination de Gauss aura pour but de transformer le système $Ax = b$ en un système équivalent triangulaire supérieure avec des valeurs 1 sur la diagonale.

1.2 Méthode de Gauss

1.2.1 Elimination de Gauss sur un exemple :

Soit le système linéaire c est une matrice triangulaire :

$$A = \begin{bmatrix} 4 & 8 & 12 \\ 3 & 8 & 13 \\ 2 & 9 & 18 \end{bmatrix} \quad et \quad b = \begin{bmatrix} 4 \\ 5 \\ 11 \end{bmatrix} \qquad (1.3)$$

Le système $Ax = b$ devient dans ce cas :

$$\left\{ \begin{array}{l} 4x_1+8x_2+12x_3= 4 \\ 3x_1+8x_2+13x_3= 5 \\ 2x_1+9x_2+18x_3= 11 \end{array} \right\} \qquad (1.4)$$

Première étape,

ça consiste à diviser la première équation de (1.4) par $a_{11}= 4$ (appelé pivot) pour obtenir :

$$x_1+2x_2+3x_3= 1. \qquad (1.5)$$

Ensuite nous soustrayons 3 fois (1.5) à la deuxième équation de (1.4) et 2 fois l'équation (1.5) à la troisième équation de (1.4) :

Nous obtenons un système équivalent à (1.4) qui est

$$\left\{\begin{array}{r} x_1+2x_2+3x_3=1 \\ 2x_2+4x_3=2 \\ 5x_2+12x_3=9 \end{array}\right\} \qquad (1.6)$$

Deuxième étape,

nous divisons la deuxième équation (1.6) par 2 (le deuxième pivot). Nous obtenons :

$$x_2+2x_3=1 \qquad (1.7)$$

Et par la suite :

$$\left\{\begin{array}{r} x_1+2x_2+3x_3=1 \\ x_2+2x_3=1 \\ 2x_3=4 \end{array}\right\} \qquad (1.8)$$

qui est équivalent au système (1.4).

Dernière étape,

finalement, il suffit de diviser la troisième équation par le troisième pivot, qui est ici 2, pour obtenir :

$$\left\{\begin{array}{r} x_1+2x_2+3x_3=1 \\ x_2+2x_3=1 \\ x_3=2 \end{array}\right\} \qquad (1.9)$$

De (1.9) , il est facile de déduire successivement les inconnues x_1, x_2, x_3 :

$$x_3=2 \qquad x_2=-3 \qquad x_1=1.$$

1.2.2 Algorithme d'élimination

Nous présentons maintenant un algorithme qui, effectué par un ordinateur, permet de réaliser l'élimination dont, le mécanisme a été décrit dans la section précédente. Pour réaliser cet objectif, nous appelons $A^{(i)}$ la matrice et $b^{(i)}$ le second membre obtenus avant la i*ème* étape de l'élimination. Ainsi le tableau $A^{(i)}$a la forme suivante :

$$A^{(i)} = \begin{bmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} & \cdots & a_{1i}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} & \cdots & a_{2i}^{(2)} & & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} & & a_{3i}^{(3)} & & a_{3n}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} & & a_{4i}^{(4)} & & a_{4n}^{(4)} \\ 0 & 0 & 0 & 0 & . & . & & \\ & & 0 & 0 & . & . & & \\ & & & & .a_{i-1,i-1}^{(i-1)} & . & & \\ & & & & 0 & a_{ii}^{(i)} & & a_{in}^{(i)} \\ & & & & & . & . & \\ & & & & & . & . & \\ & & & & & . & . & \\ & & & & 0 & a_{ni}^{(i)} & & a_{nn}^{(i)} \end{bmatrix}, \quad (1.10)$$

Avec $A^{(1)} = A$, *et* $Ax = b$

La i*ème* étape de l'élimination consistera à passer du tableau $A^{(i)}$ au tableau $A^{(i+1)}$ et du tableau $b^{(i+1)}$ par opération suivante :

i*ème* étape. Nous divisons la i*ème* ligne de $A^{(i)}$ par le i*ème* pivot $a_{ii}^{(i)}$ (supposé différent de zéro), puis on remplace la ligne $L_j^{(i+1)}$ par la ligne

$$L_j^{(i+1)} = L_j^{(i)} - m_{ji} * L_i^{(i)}, \qquad j = i+1, i+2, ..., n \qquad où, m_{ji} = a_{ji}^{(j)} / a_{ii}^{(i)} \qquad (1.11)$$

Nous faisons de même avec le second membre :

$$b_j^{(i+1)} = b_i^{(i)} - m_{ji} * b_i^{(i)} \qquad (1.12)$$

1.2.3 Matrice élémentaire de Gauss

Soient les matrices élémentaires de Gauss

$$M_1 = \begin{pmatrix} 1 & 0 & & & 0 \\ -m_{21} & 1 & 0 & & \\ -m_{31} & 0 & 1 & . & \\ . & . & & . & \\ . & . & & . & 0 \\ -m_{n1} & 0 & 0 & & 1 \end{pmatrix},$$

$$\downarrow k-ième\ colonne$$

$$M_k = \begin{pmatrix} 1 & & 0 & & & 0 & 0 \\ 0 & . & & & & . & . \\ 0 & & . & & & . & . \\ . & & & 0 & & . & . \\ . & & 1 & . & & . & . \\ . & & -m_{k+1,k} & .1 & & . & \\ . & & . & & . & . & . \\ . & & . & & & 0 & . \\ . & & . & & & 1 & 0 \\ 0 & & -m_{n,k} & 0 & & 0 & 1 \end{pmatrix} \overleftarrow{k-ième\ ligne} \qquad (1.13)$$

où $m_{i,k} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}$ $i = k+1, k+2, ..., n$ en supposant que $a_{kk}^{(k)} \neq 0$,

En posant $e_k = (0, ..., 1, 0, ..., 0)^T$ et $m_k = (0, 0, ..., -m_{k+1,k}, ..., -m_{n,k})^T$,on obtient $M_k = I + m_k e_k^T$ et on vérifie que M_k est inversible et que $M_k^{-1} = I - m_k e_k^T$

Remarque : Matriciellement, dans l'algorithme d'élimination la première étape est équivalente au produit matriciel $A^{(2)} = M_1 A^{(1)}$.

L'étape finale est alors donnée par :

$$A^{(n)} = U = M_{n-1} M_{n-2} \ldots M_2 M_1 A^{(1)}, A^{(1)} = A.$$

Evidemment, l'étape finale n'est accessible par ce procédé que si tous les pivots $a_{ii}^{(i)}$ sont tous non nuls.

Définition : 1.3 : A_k est la sous matrice principale d'ordre k de A si A_k est la $k \times k$ matrice de coefficient $a_{ij}, 1 \leq i, j \leq k \leq n$.

Nous avons le résultat suivant.

Théorème1 : Si toutes les sous-matrices principales A_k de la matrice de départ A sont régulières, $k = 1, 2, \ldots, n$, alors les pivots obtenus successivement dans l'élimination de Gauss sont tous non nuls. Inversement si tous les pivots obtenus au cours de l'élimination de Gauss sont non nuls, alors toutes les sous-matrices principales de A sont régulières.

$$\det A_i = a_{11}^{(1)} \times a_{22}^{(2)} \times \ldots \times a_{ii}^{(i)} \qquad (1.14)$$

Il est également facile de vérifier que les opérations faites sur la matrice A impliquent, si A_i est la sous-matrice principale d'ordre i de A :

$$\left\{\begin{array}{l} \det A_1 = \det A_1^{(1)} \\ \det A_2 = a_{11}^{(1)} \det A_2^{(2)} \\ \det A_3 = a_{11}^{(1)} a_{22}^{(2)} \det A_3^{(3)} \\ . \\ . \\ . \\ \det A_i = a_{11}^{(1)} a_{22}^{(2)} .....a_{i-1,i-1}^{(i-1)} \det A_i^{(i)} \\ . \\ . \\ . \end{array}\right\} \qquad (1.15)$$

Nous concluons de (1.14) et (1.15) que si $\det A_i \neq 0$ pour tout $i = 1, 2, ...n$ alors les valeurs $a_{11}^{(1)}, a_{22}^{(2)}, ...a_{i-1,i-1}^{(i-1)}, .., a_{nn}^{(n)}$ sont non nulles.

1.2.4 Elimination da Gauss avec changement de Pivot

Comme nous venons de le voir, l'algorithme d'élimination donné dans la section précédente ne peut être exécuté que si les pivots successifs sont non nuls, c'est à dire si toutes les sous-matrices principales de A sont régulières. Il est évident qu'il est impossible de traiter par cet algorithme le système suivant :

$$\left\{\begin{array}{c} 0x_1 + x_2 + 3x_3 = 1 \\ 5x_1 + 2x_2 + 3x_3 = 4 \\ 6x_1 + 8x_2 + x_3 = 1 \end{array}\right\} \qquad (1.16)$$

Car, dans ce cas, on ne peut pas diviser la première ligne par le premier pivot qui est nul (la première sous-matrice principale est donc singulière !). On voit immédiatement que les choses se présentent mieux si on échange la première et la troisième ligne pour obtenir

$$\left\{\begin{array}{c} 6x_1+8x_2+x_3=1 \\ 5x_1+2x_2+3x_3=4 \\ 0x_1+x_2+3x_3=1 \end{array}\right\} \qquad (1.17)$$

En effet, maintenant nous pouvons diviser la première ligne par le pivot 6.

Cette manière de faire s'appelle" pivotage partiel" ; elle consiste à échanger deux équations dont le but d'avoir le plus grand pivot possible en valeur absolue.

Le problème peut se poser même avec un pivot trop petit.

Pour éviter de diviser par des pivots trop petits pouvant conduire à des solutions absurdes.

Exemple : soit à résoudre le système

$$\left\{\begin{array}{c} 10^{-10}x_1+x_2=1 \\ x_1-x_2=0 \end{array}\right\};$$

La solution exact est $x_1=x_2=1/(1+10^{-10})\simeq 1.$

Cependant, si on suppose que les calculs sont effectués en virgule flottante, avec mantisse à 9 chiffres, la résolution du système par la méthode de Gauss donne des résultats différents selon qu'on l'applique avec ou sans pivot

(i) Si on applique la méthode de Gauss sans pivot on obtient

$$m_{21}=10^{10} \qquad et \qquad \left\{\begin{array}{c} 10^{-10}x_1+x_2=1 \\ (-1-10^{10})x_2=-10^{10} \end{array}\right\}$$

Ce qui donne $x_1\simeq 0, x_2\simeq 1$ à cause des arrondis des résultats avec neuf premiers chiffres significatifs.

(ii) Si on adopte la stratégie du pivot partiel qui consiste à mettre en première ligne celle dont le coefficient de x_1 est le plus grand en module alors on permute les lignes pour obtenir

$$\left\{\begin{array}{c} x_1 - x_2 = 0 \\ 10^{-10}x_1 + x_2 = 1 \end{array}\right\}$$

Pour lequel $m_{21} = 10^{-10}$ et qui conduit à la solution approchée $x_2 \simeq 1$ et $x_1 = x_2$.

En conclusion, on peut adopter automatiquement la stratégie du pivot partiel, c'est à dire à chaque étape k :

choisir $a_{ii}^{(i)} = \max_{k \geq i} \left| a_{ki}^{(k)} \right|$.

Matriciellement, cette opération revient à multiplier la matrice $A^{(i)}$ par une matrice de permutation P_{kl} avant d'appliquer l'élimination de Gauss. L'étape finale est données par $A^{(n)} = U = M_{n-1}P_{i-1i}M_{n-2}.....M_2P_{2i}M_1P_{1i}A$ où les M_i sont des matrices élémentaires de Gauss et les P_{kl} des matrices de permutations (elle échange les lignes k et l) pour $l \geq k$.

Si à une étape k on n'a pas besoin de pivoter, l'écriture reste valable avec $P_{kl} = I$ où I est la matrice identité.

$$P_{kl} = \begin{array}{c} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccccc} & & & & \downarrow k..... & & & & & l \downarrow & \\ 1 & 0 & & & 0 & & & & & & \\ 0 & 1 & & & 0 & & & & & & \\ 0 & 0 & 1 & & 0 & & & & & & \\ 0 & 0 & & 1 & 0 & & & & & & \\ .. & . & . & . & P_{kk} = 0 & 1 & . & P_{kl} = 1 & . & . & \leftarrow \\ & & & & 0 & 0 & & & & & \\ & & & & . & & 1 & & & & \\ & & & & P_{lk} = 1 & & & P_{ll} = 0 & & . & \\ & & & & & & & & . & & \\ 0 & & & & & & & & & 1 & \end{array} \\ \overline{k - ième\ ligne} \end{array} \qquad (1.18)$$

Remarque : $P_{kl}A$ échange les lignes k et l alors que AP_{kl} échange les colonnes k et l. on a encore : $P_{kl} = P_{kl}^{-1} = P_{kl}^T$.

Définition 1.4 : Une matrice de permutation est un produit de matrices de permutation.

1.2.5 Méthode de Gauss avec pivot total

On pourrait aussi adopter la stratégie du pivot total qui consiste, à chaque étape k, à prendre $a_{ii}^{(i)} = \max_{k>i, j>i} \left| a_{kj}^{(i)} \right|$. Ce qui reviendrait à multiplier la matrice $A^{(i)}$ par deux matrices, de permutation P et Q l'une à droite pour permuter les lignes et l'autre pour permuter les colonnes.

1.2.6 Factorisation LU

Tout va donc très bien pour ce système, mais supposons qu'on ait à résoudre 3089 systèmes avec la même matrice A mais 3089 seconds membres b différents (par exemple on peut vouloir calculer la réponse d'une structure de génie civil à 3089 changements différents). Il serait un peu dommage de recommencer les opérations ci-dessous 3089 fois, alors qu'on peut en éviter une bonne partie.

Comment faire ?

L'idée est de "factoriser" la matrice A , c'est à dire comme un produit $A = LU$, où L est triangulaire inférieure et U triangulaire supérieure.

On reformule alors le système $Ax = b$ sous forme $LUx = b$ et on résout maintenant deux systèmes faciles à résoudre car triangulaires $Ly = b$ et $Ux = y$.

La factorisation LU de la matrice A découle immédiatement de l'algorithme de Gauss.

Théorème : Décomposition LU d'une matrice Soit $A \in M_n(IR)$ une matrice inversible, il existe une matrice de permutation P telle que, pour cette matrice de permutation, il existe un et un seul couple (L,U) où L est une matrice triangulaire inférieure de termes diagonaux tous égaux à 1 et U est une matrice triangulaire supérieure, vérifiant

$$PM = LU$$

***Preuve :** L'existence de la matrice P et les matrices L,U peut s'effectuer en s'inspirant de l'algorithme "LU avec pivot partiel".

En effet, chaque étape i peut s'écrire $A^{(i)} = M_{i-1}P^{(i-1)}A^{(i-1)}$ où $A^{(1)} = A$, $P^{(i-1)}$ est la matrice de permutation qui permet le choix du pivot partiel, et M_{i-1} est une matrice élémentaire de Gauss (matrice d'élimination qui effectue les combinaisons linéaires de lignes permettant de mettre à zéro tous les coefficients de la colonne i situés en dessous de la ligne i. Pour simplifier, raisonnons sur une matrice 4×4 (le raisonnement est le même pour une matrice $n \times n$.

En appliquant l'algorithme

$$M_3P^{(3)}M_2P^{(2)}M_1P^{(1)}A = U$$

Les matrices $P^{(i+1)}$ et M_{i+1} ne permutent pas .Prenons par exemple

$$M^{(2)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & a & 1 & 0 \\ 0 & b & 0 & 1 \end{bmatrix} \qquad P^{(3)} = P_{34} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

On vérifie facilement que $M_2P^{(2)} \neq P^{(2)}M_2$,mais par contre, comme la multiplication à gauche par $P^{(i+1)}$ permute les lignes $i+1$ *et* $i+k$, pour $k \geq 1$ et que la multiplication à droite permute les colonnes $i+1$ *et* $i+k$ de M_i, La matrice $\widetilde{M}_i = P^{(i+1)}M_iP^{(i+1)}$ est encore une matrice triangulaire inférieure avec la même structure que M_i : On a juste échangé les coefficients extra diagonaux des lignes $i+1$ *et* $i+k$. On a donc

$$P^{(i+1)}M_i = \widetilde{M}_iP^{(i+1)} \qquad \text{car} \qquad P^{(i+1)}.P^{(i+1)} = I.$$

Dans l'exemple précédent, on effectue le calcul

$$P^{(3)}M_2P^{(3)} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & b & 1 & 0 \\ 0 & a & 0 & 1 \end{bmatrix} = \widetilde{M}_2$$

qui est une matrice triangulaire inférieure de coefficient tous égaux à 1, et comme $P^{(3)}P^{(3)} = I$, on donc

$$P^{(3)}M_2 = \widetilde{M}_2P^{(3)}$$

Pour revenir à notre exemple où $n = 4$ on peut donc écrire

$$M_3\widetilde{M}_2P^{(3)}\widetilde{M}_1P^{(2)}P^{(1)}A = U$$

Mais par le même raisonnement que précédemment, on $P^{(3)}\widetilde{M}_1 = \widetilde{\widetilde{M}}_1P^{(3)}$ où $\widetilde{\widetilde{M}}_1$ est encore une matrice triangulaire inférieure avec des 1 sur la diagonale. On en déduit que

$$M_3\widetilde{M}_2\widetilde{\widetilde{M}}_1P^{(3)}P^{(2)}P^{(1)}A = U,$$

soit encore $PA = LU$ où $P = P^{(3)}P^{(2)}P^{(1)}$ est bien une matrice de permutation et $L = (M_3\widetilde{M}_2\widetilde{\widetilde{M}}_1)^{-1}$ est une matrice triangulaire inférieure avec des 1 sur la diagonale.

Le raisonnement pour n=4 se généralise facilement à n arbitraire.

Dans ce cas, l'échelonnement de la matrice s'écrit

$$M_{n-1}P^{(n-1)}M_{n-2}P^{(n-2)}.....M_2P^{(2)}M_1P^{(1)}A = U$$

Et se transforme en

$$F_{n-1}F_{n-1}.....F_1P^{(n-1)}P^{(n-2)}.....P^{(1)}A = U$$

où F_i est une matrice triangulaire inférieure avec des 1 sur la diagonale.

On a ainsi démontré l'existence.

2. Unicité : Pour montrer l'unicité du couple (L,U) à P donnée, supposons qu'il existe une matrice P et des matrices L_1 et L_2 triangulaires inférieures et U_1 U_2 triangulaires supérieures telles que $PA = L_1U_1 = L_2U_2$.

Dans ce cas, on a $L_2^{-1}L_1 = U_2U_1^{-1}$, or la matrice $L_2^{-1}L_1$est une matrice triangulaire inférieure avec des 1 sur la diagonale, et la matrice $U_2U_1^{-1}$ triangulaire supérieure, on en déduit que $L_2^{-1}L_1 = U_2U_1^{-1} = I$ et donc $L_1 = L_2$ et $U_1 = U_2$.

1.3 Méthode de Choleski

On va maintenant étudier la méthode de Choleski, qui est une méthode directe adaptée au cas où la matrice A est symétrique définie positive (s.d.p). On rappelle qu'une matrice $A \in M_n(IR)$ de coefficients $(a_{ij})_{i,j=1}^n$est symétrique si $A^T = A$, où A^T désigne la

transposée de A, définie par les coefficients $(a_{ji})_{n\in IN}$ et que A est définie positive si $Ax.x > 0$ $(ou\ x^TAx > 0)$ pour tout $x \neq 0$. Dans ce cas ,$x.y$ désigne le produit scalaire de x et y de IR^n. On rappelle que si A est s.d.p elle est en particulier inversible.

1.3.1 Description de la méthode

Commençons par un exemple. On considère la matrice

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

qui est également symétrique. Calculons sa décomposition LU. Par échelonnement, on obtient

$$A = LU = \begin{bmatrix} 1 & -0 & 0 \\ -0.5 & 1 & 0 \\ 0 & -2/3 & 1 \end{bmatrix} \begin{bmatrix} 2 & -1 & 0 \\ 0 & 3/2 & -1 \\ 0 & 0 & 4/3 \end{bmatrix}$$

La structure LU ne conserve pas la symétrie de la matrice A. Pour des raisons de coût mémoire, il est important de pouvoir la conserver. Une façon de faire est de décomposer U en sa partie diagonale fois une matrice triangulaire.

On obtient

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

On a donc $U = DL^T$, comme tous les coefficients de D sont positifs, on peut écrire $D = \sqrt{D}\sqrt{D}$, où $\sqrt{D}$ est la matrice diagonale dont

les éléments diagonaux sont les racines carrées des éléments diagonaux de D, on a donc $A = L\sqrt{D}\sqrt{D}L^T = \tilde{L}\tilde{L}^T$, avec $\tilde{L} = L\sqrt{D}$.

Notons que la matrice $\tilde{L}$ est toujours triangulaire inférieure, mais ses coefficients diaginaux ne sont plus astreints à être égaux à 1. C'est la décomposition de Choleski de la matrice A.

1.3.2 Théorème : Décomposition de Choleski

Soit $A \in M_n(IR)$ $(n \geq 1)$ une matrice symétrique définie positive. Alors il existe une unique matrice $\tilde{L} \in M_n(IR)$ telle que

1. $\tilde{L}$ est triangulaire inférieure, $\tilde{L} = (l_{ij})_{i,j=1}^n$
2. $l_{ii} > 0$, pour tout $i \in \{1, 2, ..., n\}$
3. $A = \tilde{L}\tilde{L}^T$

1. Existence de la décomposition Soit $A \in M_n(IR)$ $(n \geq 1)$ une matrice symétrique définie positive. On sait déjà qu'il existe une matrice de permutation P et L triangulaire inférieure et U triangulaire supérieure telles que $PA = LU$. A l'avantage dans le cas où la matrice est s.d.p, est que la décomposition est toujours possible sans permutation. On prouve l'existence et l'unicité en construisant la décomposition, c'est à dire en construisant la matrice L.

Démonstration par récurrence sur n

1. Pour n=1, on a $A = (a_{11})$.Comme A est s.d.p $a_{11} > 0$. On a peut définir $\tilde{L} = (l_{11})$ où $l_{11} = \sqrt{a_{11}}$, et on a bien $A = \tilde{L}\tilde{L}^T$.

2. On suppose que la décomposition de Choleski s'obtient pour $A \in M_p(IR)$, pour $1 \leq p \leq n$ et démontrons que la propriété est encore vraie pour $A \in M_{n+1}(IR)$ s.d.p.

Soit donc $A \in M_{n+1}(IR)$ s.d.p ; on peut écrire A sous forme :

$$A = \begin{bmatrix} B & a \\ a^T & \alpha \end{bmatrix}$$

où $B \in M_n(IR)$, $a \in IR^n$ et $\alpha \in IR$. Montrons que B est s.d.p, c'est à dire que $By.y > 0$, pour tout $y \in IR^n - \{0\}$, et $x = \begin{pmatrix} y \\ 0 \end{pmatrix} \in IR^{n+1}$

Comme A est s.d.p, on a

$$0 < Ax.x = \begin{bmatrix} B & a \\ a^T & \alpha \end{bmatrix} \begin{pmatrix} y \\ 0 \end{pmatrix} . \begin{pmatrix} y \\ 0 \end{pmatrix} = \begin{bmatrix} Bx \\ a^T \end{bmatrix} . \begin{bmatrix} y \\ 0 \end{bmatrix} = By.y$$

Et donc B est s.d.p. Par hypothèse de récurrence, il existe une matrice $M \in M_n(IR)$, $M = (m_{ij})_{i,j=1}^n$ telle que :

1. $m_{ij} = 0 \quad si \quad j > i$ (triangulaire inférieure)
2. $m_{ii} > 0$
3. $B = MM^T$.

On va chercher L sous forme

$$\tilde{L} = \begin{bmatrix} M & 0 \\ b^T & \lambda \end{bmatrix}$$

Avec $b \in IR^n$ et $\lambda \in IR_+^*$ tels que $A = \tilde{L}\tilde{L}^T$. Pour déterminer b et λ, calculons $\tilde{L}\tilde{L}^T = A$, et on veut que les égalités suivantes soient vérifiées :

$$Mb = a \quad et \quad b^T b + \lambda^2 = \alpha$$

Comme M est inversible (en effet $\det(M) = \Pi_{i=1}^n m_{ii} > 0$), la première égalité ci-dessous donne : $b^{-1} = Ma$ et en remplaçons dans la deuxième égalité, on obtient :

$(M^{-1}a)^T(Ma) + \lambda^2 = \alpha$, et donc $a^T(M^T)^{-1}(M^{-1}a) + \lambda^2 = \alpha$ soit encore a

soit encore $a^T(MM^T)^{-1}a + \lambda^2 = \alpha$,

c'est à dire $a^T B^{-1} a + \lambda^2 = \alpha \qquad (2.1)$

Pour que (2.1) soit vérifiée, il faut que

$$\alpha - a^T B^{-1} a > 0 \qquad (2.2)$$

Montrons que cette condition est effectivement vérifiée :

Soit $z = \begin{bmatrix} B^{-1}a \\ -1 \end{bmatrix} \in IR^{n+1}$. On a $z \neq 0$ et donc $0 < Ax.x$ car A est s.d.p et donc

$$Az = \begin{bmatrix} B & a \\ a^T & \alpha \end{bmatrix} \begin{bmatrix} B^{-1}a \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ a^T B^{-1} a - \alpha \end{bmatrix}$$

On a donc $Az.z = \alpha - a^T B^{-1} a > 0$ ce qui démontre l'inégalité (2.2)

On peut choisir ainsi $\lambda = \sqrt{\alpha - a^T B^{-1} a} > 0$ de tel sorte que (2.1) soit vérifiée.

Posons :

$$\tilde{L} = \begin{bmatrix} M & 0 \\ (M^{-1}a)^{-1} & \lambda \end{bmatrix}$$

$\tilde{L}$ est bien triangulaire inférieure et vérifie $l_{ii} > 0$ et $A = \tilde{L}\tilde{L}^T$.

On a terminé ainsi la partie existence.

2. Unicité et calcul de $\tilde{L}$ Soit donc $A \in M_n(IR)$ s.d.p ; on vient de montrer qu'il existe donc $\tilde{L} \in M_n(IR)$ triangulaire inférieure telle que $l_{ii} > 0$ et $A = \tilde{L}\tilde{L}^T$. On a donc

$$a_{ij} = \sum_{k=1}^{n} l_{ik} l_{jk} \qquad \forall (i,j) \in \{1, ..., n\}^2 \qquad (2.3)$$

1. Calculons la première colonne de $\tilde{L}$; pour $j = 1$, on a

$a_{11} = l_{11} l_{11}$ $\quad (l_{1j} = 0, \forall j \geq 2)$; et donc $l_{11} = \sqrt{a_{11}}$

$a_{21} = l_{21} l_{11}$ $\quad (l_{2j} = 0, \forall j \geq 3)$; $\quad l_{21} = \frac{a_{21}}{l_{11}}$

.

.

$a_{i1} = l_{i1} l_{11}$ $\qquad l_{i1} = \frac{a_{i1}}{l_{11}} \qquad \forall i \in \{2, ..., n\}$.

2. On suppose avoir calculé les q premières colonnes de $\tilde{L}$. On calcul la colonne (q+1) en prenant j=q+1 dans (2

pour $i = q+1$ $\qquad a_{q+1q+1} = \sum_{k=1}^{q+1} l_{q+1k} l_{q+1k} \qquad ; l_{q+1k} = 0 \ pour \ k \geq q+2$

$= \sum_{k=1}^{q} l_{q+1k}^2 + l_{q+1q+1} \qquad \Longrightarrow l_{q+1q+1} = \sqrt{a_{q+1q+1} - \sum_{k=1}^{q} l_{q+1k}^2}$

Notons que $a_{q+1q+1} - \sum_{k=1}^{q} l_{q+1k}^2 > 0$ car $\tilde{L}$ existe.

On procède de la même manière pour $q+2, ..., n$ on a :

$$a_{iq+1} = \sum_{k=1}^{q+1} l_{ik} l_{q+1k} = \sum_{k=1}^{q} l_{ik} l_{q+1k} + l_{iq+1} l_{q+1q+1}$$

Et donc

$$l_{iq+1} = (a_{iq+1} - \sum_{k=1}^{q} l_{ik} l_{q+1k}) \frac{1}{l_{q+1q+1}}$$

On calcule ainsi toutes les colonnes de $\tilde{L}$. On a donc démontré que $\tilde{L}$ est unique par moyen constructif de calcul de $\tilde{L}$.

Chapitre 2

Méthodes itératives pour la résolution des systèmes linéaires

2.1 Rappels : normes, rayon spectral

Définition 1.1 : Norme matricielle-norme induite) On note $M_n(IR)$ l'espace vectoriel sur IR des matrices carrées d'ordre n.

a. On appelle norme matricielle sur $M_n(IR)$ une norme $||.||$ sur $M_n(IR)$ telle que :

$$||AB|| \leq ||A||.||B|| \qquad \forall A, B \in M_n(IR).$$

b. On considère IR^n muni de la norme $||.||$. On appelle norme matricielle induite sur $M_n(IR)$ la norme encore notée $||.||$, la norme sur $M_n(IR)$ définie par :

$$||A|| = \sup\{||Ax|| ; x \in IR^n, ||x|| = 1\} \qquad \forall A \in M_n(IR)$$

Proposition 1.2 Soit $M_n(IR)$ muni d'une norme induite $||.||$. Alors pour toute matrice $A \in M_n(IR)$, on a :

1. $||Ax|| \leq ||A||\,||X|| \qquad \forall x \in IR^n.$
2. $||A|| = \max\{||Ax|| ; x \in IR^n, ||x|| = 1\}$
3. $||A|| = \max\left\{\frac{||Ax||}{||x||} ; x \in IR^n - \{0\}\right\}$
4. $||.||$ est une norme matricielle.

Preuve 1. Soit $x \in IR^n - \{0\}$ posons $y = \frac{x}{||x||}$, alors $||y|| = 1$ donc $||Ay|| \leq ||A||$ (car $||A|| = \sup_{||x||=1} ||Ax||$) et donc $\left|\left|A\frac{x}{||x||}\right|\right| \leq ||A||$ c'est à dire $||Ax|| \leq ||A|| \times ||X|| \quad \forall x \in IR^n - \{0\}$.

si $x = 0$ Alors $Ax = 0$ et $||X|| = 0$ et $||Ax|| = 0$ et l'inégalité est encore vérifiée.

2. Soit l'application Φ définie sur IR^n dans IR : $\Phi(x) = ||Ax||$ est continue sur la sphère unité $S_1 = \{x \in IR^n, ||x|| = 1\}$ qui est un compact de IR^n. Donc Φ esr bornée et atteint ses bornes. Il existe $x_0 \in IR^n$ tel que $||A|| = ||Ax_0||$.

3. Cette égalité résulte du fait que $\frac{||Ax||}{||x||} = \left|\left|A\frac{x}{||Ax||}\right|\right|$ et $\frac{x}{||Ax||} \in S_1$ pour $x \neq 0$.

4. Soient A et $B \in M_n(IR)$ on $||A|| = \sup\{||Ax|| ; x \in IR^n, ||x|| = 1\}$, or

$$||ABx|| \leq ||A|| \times ||Bx|| \leq ||A|| \times ||B|| \times ||X|| \leq ||A|| \times ||B||,$$

on en déduit que $||.||$ est une norme matricielle.

Définition 1.3 : Rayon spectral Soit $A \in M_n(IR)$ une matrice inversible. On appelle rayon spectral de A la quantité :

$$\rho(A) = \max\left\{|\lambda| ; \lambda \in C, \lambda \text{valeur propre de } A\right\}.$$

Caractérisation de normes induites : Soit $A = (a_{ij})_{1 \leq i,j \leq n} \in M_n(IR)$

1. On munit IR^n de la norme $||.||_\infty$ et $M_n(IR)$ de la norme induite correspondante, notée aussi $||.||_\infty$

Alors :

$$||A||_\infty = \max_{i \in \{1,\ldots,n\}} \Sigma_{j=1}^n |a_{ij}|$$

2. On munit IR^n de la norme $||.||_1$ et $M_n(IR)$ de la norme induite correspondante, notée aussi $||.||_1$

Alors : $||A||_1 = \max_{j\in\{1,\dots,n\}} \Sigma_{i=1}^{n} |a_{ij}|$.

3. On munit IR^n de la norme $||.||_2$ et $M_n(IR)$ de la norme induite correspondante, notée aussi $||.||_2$.

Alors :

$||A||_2 = \left[\rho(A^T A)\right]^{\frac{1}{2}}$, en particulier si A est symétrique, $||A||_2 = \rho(A)$.

Proposition 1.5 : Approximation du rayon spectral par une norme induite Soit $A \in M_n(IR)$ et $\varepsilon > 0$. Il existe une norme spectral sur IR^n (qui dépend de A et ε) telle que la norme induite sur $M_n(IR)$, notée $||.||_{A,\varepsilon}$ vérifie :

$$||A||_{A,\varepsilon} \leq \rho(A) + \varepsilon$$

Corollaire 1.6 : convergence spectrale On munit $M_n(IR)$ d'une norme , notée $||.||$. Soit $A \in M_n(IR)$.

Alors :

$\rho(A) < 1$ si et seulement si A^k converge vers 0 quand k tend vers $+\infty$.

Preuve : Si $\rho(A) < 1$, grâce à l'approximation du rayon spectral de la proposition précédente, il existe $\varepsilon > 0$ tel que $\rho(A) < 1 - 2\varepsilon$ et d'une norme induite $||.||_{A,\varepsilon}$ tel que $||A||_{A,\varepsilon} = \mu \leq \rho(A) + \varepsilon < 1 - \varepsilon$, comme $||A||_{A,\varepsilon}$ est une norme matricielle, on a $||A^k||_{A,\varepsilon} \leq \mu^k$ converge vers 0 $qd\ k \to +\infty$. Comme $M_n(IR)$ est de dimension finie, toutes les normes sont équivalentes et donc

$$||A^k|| \to 0\ qd\ k \to +\infty.$$

Réciproquement, supposons A^k converge vers 0 qd k $tend$ $vers$ $+\infty$. et montrons que $\rho(A) < 1$.

Soit λ une valeur propre et $x \neq 0$ le vecteur propre associé. Alors $A^k x = \lambda^k x$ et si $A^k \to 0$ qd $k \to +\infty$ alors $A^k x \to 0$ et donc $\lambda^k x \to 0$, qui n'est possible que si $|\lambda| < 1$.

1.7 Proposition : convergence et rayon spectral On munit $M_n(IR)$ d'une norme, notée $||.||$. Soit $A \in M_n(IR)$. Alors $\rho(A) = \lim_{\infty} ||A^k||^{1/k}$.(admise)

1.8 Corollaire : comparaison rayon spectral et norme. On munit $M_n(IR)$ d'une norme, notée $||.||$. Soit $A \in M_n(IR)$. Alors :

$$\rho(A) \leq ||A||.$$

Par conséquent si $M \in M_n(IR)$ et $x^{(0)} \in IR^n$, pour montrer que la suite $x^{(k)} = M^k x^{(0)}$ converge vers 0 dans IR^n, il suffit de trouver une norme matricielle $||.||$ telle que $||.M|| < 1$.

Preuve Si $||.||$ est une norme matricielle, alors $||A^k|| \leq ||A||^k$ et donc par la caractérisation du rayon spectral donné dans la proposition précédente, on obtient $\rho(A) = \lim_{\infty} ||A^k||^{1/k} \leq ||A||$.

1.9 Théorème : Matrice de la forme $I + A$ 1. Soit une norme matricielle induite, I la matrice identité de $M_n(IR)$ et $A \in M_n(IR)$ telle que $||A|| < 1$. Alors la matrice $I + A$ est inversible et on a $\left|\left|(I+A)^{-1}\right|\right| \leq \frac{1}{1-||A||}$.

2. Si une matrice de la forme $I + A \in M_n(IR)$ est singulière, alors $||A|| \geq 1$ pour toute norme matricielle $||.||$.

Démonstration :

1. Si $\rho(A) < 1$, les valeurs propres de A sont toutes différentes de 1 et -1. Donc 0 n'est pas valeur propre des matrices $I + A$ et $I - A$, qui sont donc inversibles.

Supposons que $||A|| < 1$ alors on a $\rho(A) < 1$. Il est facile de vérifier que

$$\Sigma_{k=0}^{n} A^k(I-A) = I - A^{n+1}$$

Si $\rho(A) < 1$ et le corollaire 1.6 $\Longrightarrow A^k \to 0$ qd $k \to +\infty$. De plus, $I - A$ inversible.

En passant à la limite, on a donc $(I-A)^{-1} = \Sigma_{k=0}^{+\infty} A^k$. $(\bigstar)$

Réciproquement, si $\rho(A) \geq 1$ la série ne peut converger en raison du corollaire 1.6.

On a démontré plus haut que si $\rho(A) < 1$ la série de terme générale A^k est absolument convergente et qu'elle vérifie $(\bigstar)$. On en déduit que si $||A|| < 1$

$$\left|\left|(I+A)^{-1}\right|\right| \leq \Sigma_{k=0}^{+\infty} ||A^k|| \leq \Sigma_{k=0}^{+\infty} ||A||^k = \frac{1}{1-||A||}.$$

De même on a $(I-A)^{-1} = \Sigma_{k=0}^{+\infty} (-1)^k A^k$ et $\left|\left|(I+A)^{-1}\right|\right| \leq \frac{1}{1-||A||}$.

2. Si une matrice de la forme $I + A \in M_n(IR)$ est singulière, alors $\lambda = -1$ est valeur propre et donc $\rho(A) = 1 \geq 1$ et en utilisant le corollaire 1.8, on obtient que $||A|| \geq \rho(A) = 1 \geq 1$.

2.2 Méthodes itératives :

2.2.1 Définitions et propriétés

Soit $A \in M_n(IR)$ une matrice inversible et $b \in IR^n$, on cherche toujours ici à résoudre le système : *Trouver* $x \in IR^n$ tel que $Ax = b$, mais de façon itérative, c'est à dire par la construction d'une suite.

Définition 2.1.1 : Méthode itérative On appelle méthode itérative de résolution du système linéaire $Ax = b$ une méthode qui construit une suite $(x^{(k)})_{k \in IN}$ où "l'itéré" $x^{(k)}$ est calculé à partir des itérés $x^{(0)}, x^{(1)}, x^{(2)}, ..., x^{(k-1)}$ censée converger vers x solution de $Ax = b$.

Définition 2.1.2 : (méthode itérative convergente) On dit que la méthode itérative est convergente si pour tout choix initial $x^{(0)} \in IR^n$, on a, $x^{(k)} \to x \; qd \; x \to +\infty$.

Enfin, on veut que cette suite soit à calculer. Une idée est de travailler avec une matrice P inversible qui soit "proche" de A, mais plus facile à inverser que A.

On appelle matrice de pré conditionnement cette matrice. On écrit alors $A = P - (P - A) = P - N$, et on réécrit le système $Ax = b$ sous forme :

$$Px = (P - A)x + b = Nx + b$$

Cette forme suggère la construction de la suite $(x^{(k)})_{k \in IN}$ à partir d'un choix initial $x^{(0)}$ donné, par la formule suivante :

$$Px^{(k+1)} = Nx^{(k)} + b,$$

ce qui peut s'écrire également

$$x^{(k+1)} = Bx^{(k)} + c \qquad (2.1)$$

avec $B = P^{-1}N = I - P^{-1}A$ et $c = P^{-1}b$.

On introduit l'erreur d'approximation $e^{(k)}$ à l'itération définie par :

$$e^{(k)} = x^{(k)} - x, \; k \in IR^n \qquad (2.2)$$

où $x^{(k)}$ est construit par (2.1) et $x = A^{-1}b$. Il est facile de vérifier que $x^{(k)} \to x = A^{-1}b \; qd \; x \to +\infty$ si et seulement si $e^{(k)} \to 0$ qd $k \to +\infty$.

Lemme 2.1.3 : La suite $(e^{(k)})_{k \in IN}$ définie par (2.2) est également définie par

$$e^{(0)} = x^{(0)} - x \qquad et \qquad e^{(k)} = x^{(k)} - x = B^k e^{(0)} \qquad (2.3)$$

comme $c = P^{-1}b = P^{-1}Ax$, on a

$e^{(k+1)} = x^{(k+1)} - x = Bx^{(k)} - x + P^{-1}Ax = B(x^{(k)} - x), \qquad B = P^{-1}N = I - P^{-1}A$

Par récurrence sur k on a $e^{(k)} = B^k(x^{(0)} - x)$

Théorème 2.1.4 : Convergence de la suite. Soit $A, P \in M_n(IR)$ inversibles. Soit $x^{(0)}$ donné et la suite $(x^{(k)})_{k\in IN}$ la suite définie par (2.1)

1. La suite $(x^{(k)})_{k\in IN}$ définie par (2.1) converge vers $x = A^{-1}b$ si et seulement si $\rho(B) < 1$.

2. La suite $(x^{(k)})_{k\in IN}$ définie par (2.1) converge si et seulement si il existe une norme induite $||.||$ telle que $||B|| < 1$.

dém :

1. On a vu aussi que $(x^{(k)})_{k\in IN}$ définie par (2.1) converge si et seulement si $e^{(k)} \to 0 \; qd \; x \to +\infty$, on en déduit par le crollaire 1.6 que $(e^{(k)})_{k\in IN}$ converge vers 0 si et seulement si $\rho(B) < 1$

2. S'il existe une norme induite $||.||$ telle que $||B|| < 1$ et donc $\rho(B) < 1$ et donc d'aprés le corollaire 1.6 la méthode converge.

Réciproquement Si la méthode converge alors $\rho(B) < 1$, et donc il existe $\eta > 0$ tel que $\rho(B) = 1 - \eta$. Prenons $\varepsilon = \frac{\eta}{2}$ et appliquons la proposition 1.5, il existe une norme induite $||B||_{B,\varepsilon} \leq 1 - \varepsilon < 1$ d'où le résultat.

Théorème 2.1.5 : Considérons deux méthodes itératives $\tilde{x}^{(k+1)} = \tilde{T}\tilde{x}^{(k)} + \tilde{c}$ et $x^{(k+1)} = Tx^{(k)} + c$ avec $\rho(T) < \rho(\tilde{T})$ et $x^{(0)} = \tilde{x}^{(0)}$ alors $\forall \varepsilon > 0 \quad \exists k_0 > 0$ tq $k > k_0 \quad sup\frac{\tilde{e}^{(k)}}{e^{(k)}} \geq (\frac{\rho(\tilde{T})}{\rho(T)+\varepsilon})^k$.

Donc la méthode itérative de la matrice T converge plus rapidement que celle de la matrice $\tilde{T}$, en résumé, l'étude des méthodes itératives consiste à étudier les deux problèmes suivants :

1. Etant donne une méthode itérative de la matrice T, déterminer si la méthode converge, i.e si $\rho(A) < 1$ ou s'il existe une norme $||.||$ telle que $||T|| < 1$.

2. Etant donné deux méthodes itératives convergente T et $\tilde{T}$, les comparer, la méthode plus rapide est celle ayant le plus petit rayon spectral.

2.1.6 On appelle taux moyen de convergence sur k itérations le nombre $\widetilde{R} = (k,T) = -\log||T^k||^{1/k}$ **et taux asymptotique de convergence le nombre** $R(T) = \lim_{k\to+\infty} \widetilde{R}(k,T) = -\log(\rho(T))$ $R(T)$ joue le rôle de vitesse de convergence, plus $R(T)$ est grand plus rapide est la convergence.

2.3 Description des méithodes classiques

Méthode de Jacobi 3.1 : Elle consiste à choisir $P = D = diag(a_{ii})$ inversible et $N = (-a_{ij})_{i\neq j}$. Le schéma itératif est comme suit :

$$x^{(k+1)} = D^{-1}(L+U)x^{(k)} + D^{-1}b \qquad (3.1.1)$$

La matrice $B_J = D^{-1}(L+U)$ est dite matrice de Jacobi associée à la matrice A. Si $x^{(0)}$ est le vecteur initial (donné), l'algorithme de Jacobi est de la forme

$$x_i^{(k+1)} = -\frac{1}{a_{ii}}\sum_{j\neq i} a_{ij}x_j^{(k)} + \frac{b_i}{a_{ii}} \quad pour\ i = 1, 2....., n$$

Cette algorithme nécessite $a_{ii} \neq 0$ pour $i = 1, 2....., n$ c'est à dire D inversible.

Explicitement, on obtient

$a_{11}x_1^{(k+1)} = -a_{12}x_2^{(k)} - a_{12}x_2^{(k)} - ..... - a_{1n}x_n^{(k)} + b_1$

.

.

.

$a_{n1}x_1^{(k+1)} = -a_{n1}x_1^{(k)} - a_{n2}x_2^{(k)} - a_{n2}x_2^{(k)} - ..... - a_{nn-1}x_{n-1}^{(k)} + b_n$

D'après le théorème précédent, une condition suffisante pour que la méthode de JACOBI converge est $\rho(B_J) < 1\ ou\ ||B_J|| < 1$.

Théorème 3.1.2

Si A est une matrice carrée à diagonale strictement dominante en lignes alors la m éthode de Jacobi converge.

Preuve On a $\sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| < |a_{ii}|$ par définition $B_J = D^{-1}(L+U)$,

D'autre part $(b_J)_{ij} = -\frac{a_{ij}}{a_{ii}}$ pour $i \neq j$ et $(b_J)_{ii} = 0$ d'où $||B_J||_\infty = \max_i \sum_j \left|(b_J)_{ij}\right| = \max_i \left\{ \frac{1}{|a_{ii}|} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \right\}$ et on a $||B_J||_\infty < 1$.

Corollaire 3.1.3

Si A est une matrice carrée à diagonale strictement dominante en colonnes alors la méthodes de Jacobi converge.(la démonstration est identique à celle du théor ème 3.1.2 en considérant la norme $||.||_1$

3.2 Méthode Gauss-Seidel Pour cette méthode, les matrices P et N sont données par :

$P = D - L$ inversible et $N = U$, L et U proviennent de l'écriture $A = D - L - U$, le schéma itératif est :

$$(D-L)x^{(k+1)} = Ux^{(k)} + b \qquad (3.2.1)$$

ou encore

$$x^{(k+1)} = (D-L)^{-1}Ux^{(k)} + (D-L)^{-1}b \qquad (3.2.2)$$

en posant que $D - L$ est inversible (3.2.1) et (3.3.2) peuvent s'écrire sous forme

$$Dx^{(k+1)} = Lx^{(k+1)} + Ux^{(k)} + b \qquad (3.2.3)$$

et

$$x^{(k+1)} = D^{-1}Lx^{(k+1)} + D^{-1}Ux^{(k)} + D^{-1}b \qquad (3.3.3)$$

En explicitant (3.3.3) on obtient :

$a_{11}x_1^{(k+1)} = -a_{12}x_2^{(k)} - \ldots\ldots\ldots\ldots - a_{1n}x_n^{(k)} + D^{-1}b_1$

$a_{22}x_2^{(k+1)} = -a_{21}x_1^{(k+1)} - a_{23}x_3^{(k)} - \ldots\ldots\ldots\ldots - a_{2n}x_n^{(k)} + D^{-1}b_2$

.

.

$a_{ii}x_i^{(k+1)} = -a_{i1}x_1^{(k+1)} - a_{i2}x_2^{(k+1)} - \ldots - a_{ii-1}x_{i-1}^{(k+1)} - a_{i+1}x_{i+1}^{(k+1)} - \ldots - a_{in}x_n^{(k)} + D^{-1}b_i$

$a_{nn}x_n^{(k+1)} = -a_{n1}x_1^{(k+1)} - a_{n2}x_2^{(k+1)} - \ldots\ldots\ldots\ldots - a_{nn-1}x_{n-1}^{(k+1)} + D^{-1}b_n$

La matrice $B_{GS} = (D-L)^{-1}U$ est dite matrice de Gauss-Seidel associée à la matrice A.

Remarque :

$$B_{GS} = (I - D^{-1}L)^{-1}D^{-1}U.$$

Théorème 3.2.1 Si A est une matrice carrée à diagonale strictement dominante en lignes alors la méthode de Gauss-Seidel converge.

Preuve Posons $B_{GS} = (D-L)^{-1}U$ et montrons que $||B_{GS}||_\infty < 1$ où $||B_{GS}||_\infty = \sup_{x\neq 0} \frac{||B_{GS}x||_\infty}{||x||_\infty}$

Soit $y = B_{GS}x = (D-L)^{-1}Ux$ alors $(D-L)y = Ux$ ou encore $Dy = Ly + Ux$ et $y = D^{-1}Ly + D^{-1}Ux$. Considérons l'indice i_0 tq

$||y_{i_0}|| = \max_i |y_i| = ||y||_\infty = ||B_{GS}x||_\infty$

Il vient $y_{i_0} = \Sigma_{j=1}^{i_0-1}(D^{-1}L)_{i_0j}y_j + \Sigma_{j=i_0+1}^{n}(D^{-1}U)_{i_0j}x_j$

par suite $|y_{i_0}| = ||y||_\infty \leq \Sigma_{j=1}^{i_0-1}\left|\frac{a_{i_0j}}{a_{i_0i_0}}\right| . ||y||_\infty + \Sigma_{j=i_0+1}^{n}\left|\frac{a_{i_0j}}{a_{i_0i_0}}\right| . ||x||_\infty$

En regroupant les termes

$$(1 - \Sigma_{j=1}^{i_0-1}\left|\frac{a_{i_0j}}{a_{i_0i_0}}\right|)\frac{||y||_\infty}{||x||_\infty} \leq \Sigma_{j=i_0+1}^{n}\left|\frac{a_{i_0j}}{a_{i_0i_0}}\right|$$

Par hypothèse, le terme $1 - \Sigma_{j=1}^{i_0-1}\left|\frac{a_{i_0j}}{a_{i_0i_0}}\right| > 0$ d'où on tire :

$$\frac{||B_{GS}x||_\infty}{||x||_\infty} = \frac{||y||_\infty}{||x||_\infty} \leq (\Sigma_{j=i_0+1}^{n}\left|\frac{a_{i_0j}}{a_{i_0i_0}}\right|)(1 - \Sigma_{j=1}^{i_0-1}\left|\frac{a_{i_0j}}{a_{ii_0i_0}}\right|)^{-1}$$

Finalement

$$\max_{x\neq 0}\frac{||B_{GS}x||_\infty}{||x||_\infty} < 1.$$

3.3 Méthode de relaxation Si on considère les matrices P et N dépendantes d'un paramètre w, on obtient $A(w) = P(w) - N(w)$.

Prenons $P(w) = \frac{1}{w}D - L$ et $N(w) = \frac{1-w}{w}D + U$, en supposant $P(w)$ inversible, le schéma itératif qui en résulte est le suivant :

$$x^{(k+1)} = (\tfrac{1}{w}D - L)^{-1}(\tfrac{1-w}{w}D + U)x^{(k)} + (\tfrac{1}{w}D - L)^{-1}b \qquad (3.3.1)$$

L'équation 3.3.1 peut être remplaceée par :

$$x^{(k+1)} = (\tfrac{1}{w}D)^{-1}Lx^{(k+1)} + [(1-w)I + wD^{-1}U)^{-1}x^{(k)} + (wD^{-1})b \qquad (3.3.2)$$

La matrice de relaxation est donnée par :

$$B_w = (\tfrac{1}{w}D - L)^{-1}(\tfrac{1-w}{w}D + U).$$

* Si $w = 1$, on retrouve la méthode de Gauss-Seidel.

* Si $w > 1$, on parle de sur-relaxation.

* Si $w < 1$, on parle de sous-relaxation.

Ici la condition de convergence $||B_w|| < 1$ dépendra du paramètre w et par conséquent, on est amené à chercher tous les w pour lesquels il y a convergence et en suite choisir la valeur optimale w_0 de telle sorte que la vitesse de convergence soit meilleure possible.

Théorème Si A est une matrice hermitienne définie positive alors la méthode de relaxation converge si $w \in]0, 2[$.

Chapitre 3

Approximation des solutions de l'equation non linéaire $f(x) = 0$

3.1 Rappels et notations :

Définition 1 : Soit k un réel strictement positif et g une fonction définie sur un intervalle $[a,b]$ de IR à valeurs dans IR.

La fonction g est dite Lipschitzienne de rapport k (ou encore $k-Lipshitzienne$ si pour tout x et $y \in [a,b]$ on a :

$$|g(x) - g(y)| \leq k\,|x - y|\,.$$

Définition 2 : Soit g est une fonction $k-Lipschitzienne$ de rapport k sur $[a,b]$. La fonction g est dite contractante de rapport de contraction $k \in]0,1[$.

Exemple 1 : $g(x) = \sin x$ est Lipshitzienne de rapport $k = 1$.

3.1.3 Définition 3 : Soit une g fonction définie sur un intervalle $[a,b]$ de IR à valeurs dans IR, la fonction g est dite uniformément continue sur $[a,b]$, si :

$\forall \varepsilon > 0, \exists \eta > 0$ tel que $\forall x$ et $y \in [a,b]$, vérifiant $|x - y| \leq \eta$, on ait $|g(x) - g(y)| \leq \varepsilon$.

Remarque 1 : Toute fonction Lipschitzienne sur $[a,b]$ est uniformément continue sur $[a,b]$.

Théorème 1 : (Des valeurs intermédiaires) Soit f une fonction définie et continue sur un intervalle fermé borné $[a,b]$ de IR. Alors pour tout $\theta \in f([a,b])$, il existe un réel $c \in [a,b]$ tel que $\theta = f(c)$. Si de plus f est strictement monotone alors le point c est unique.

3.1.5 Théorème 2 : (TVI. cas particulier $\theta = 0$) Soit f une fonction définie et continue sur un intervalle $[a,b]$ et vérifiant $f(a) \times f(b) < 0$ alors $\exists c \in [a,b]$ tel que $f(c) = 0$. Si de plus f est trictement monotone alors c est unique.

3.15 bis :(Théorème de Rolle)

Soit f une fonction définie sur un intervalle $[a,b]$ à valeurs dans IR et si f est continue sur $[a,b]$ et dérivable $]a,b[$ et vérifie $f(b) = f(a)$ alors $\exists c \in]a,b[$ tel que $f^{(1)}(c) = 0$

3.1.6 Téorème 3 :(Des accroissemenys finis) Soit f une fonction définie sur un intervalle $[a,b]$ à valeurs dans IR et si f est continue sur $[a,b]$ et dérivable sur $]a,b[$, alors elle existe $\exists c \in]a,b[$ tel que :

$$f(b) - f(a) = (b-a) \times f^{(1)}(c)$$

où $f^{(1)}(c)$ est la dérivée de f au point c.

Théorème 4 : (Formule de Taylor) Soit f une fonction de classe C^n sur un intervalle $[a,b]$, alors il existe un réel $c \in]a,b[$ tel que

$$f(b) = f(a) + (b-a)f^{(1)}(a) + \frac{1}{2!}(b-a)f^{(2)}(a) + \ldots + \frac{1}{n!}(b-a)^n f^{(n)}(a) + \frac{1}{(n+1)!}(b-a)^{n+1} f^{(n+1)}(c).$$

Théorème 5 : (De Maclaurin) Soit f une fonction de classe C^n sur un intervalle I contenant 0, et telle que $f^{(n)}$ soit dérivable à l'interieur de l'intervalle de I. Alors $\forall\ x \in I$, il existe un réel c strictement compris entre 0 et x tel que :

$$f(x) = f(0) + xf^{(1)}(0) + \frac{1}{2!}x^2 f^{(2)}(0) + \ldots + \frac{1}{n!}x^n f^{(n)}(0) + \frac{1}{(n+1)!}x^{n+1} f^{(n+1)}(c).$$

Définition 4 : Soit θ un réel et f une fonction définie sur un intervalle I de IR à valeurs dans IR. θ est dit zéro de f si $f(\theta) = 0$.

Définition 5 : Soit θ un réel et g une fonction définie sur un intervalle $I \subset IR$. On dit que θ est un point fixe de g si $g(\theta) = \theta$.

Lemme 1 : Soit I un intervalle de IR et f une fonction définie sur I et à valeur dans IR. Alors la recherche des zéros de f est équivalente à la recherche des points fixes de la fonction g définie par $g(x) = x - f(x)$.

Lemme 2 : Soit g une fonction de classe C^1 sur $[a,b]$. S'il existe un réel $k \geq 0$ tel que : $\left|g^{(1)}(x)\right| \leq k \ \forall x \in [a,b]$ alors g est $k-Lipschitzienne$.

Preuve : Il suffit d'appliquer le théorème des accroissements finis à g sur $[x,y]$ avec $x \leq y$. Donc $\exists c \in]x,y[$ tel que

$$g(y) - g(x) = (y-x) \times g^{(1)}(c)$$

donc $|g(y) - g(x)| \leq k\,|y-x|$ puisque $\left|g^{(1)}(c)\right| \leq k$.

Définition 6 : Soit $(x_n)_{n \in IN}$ une suite admettant pour limite θ.
On appelle erreur de la n-ième étape le réel défini par :

$$e_n = x_n - \theta.$$

Définition 7 : On dit que la convergence $(x_n)_{n \in IN}$ vers θ est d'ordre p si :

$$\lim_{x \to +\infty} \frac{|e_{n+1}|}{|e_n|^p} = c$$

où c et p sont des réels positifs

* si $p = 1$ la convergence est dite linéaire
* si $p = 2$ la convergence est dite quadratique
* si $p = 3$ la convergence est dite cubique.

Définition 8 : On dira que le réel δ est une approximation du réel α avec une précision ε si $|\alpha - \delta| \leq \varepsilon$.

En particulier, on dira que le terme (x_{n_0}) d'une suite $(x_n)_{n\in IN}$ approche θ avec une précision ε si $|x_n - \theta| \leq \varepsilon$.

Exemple :

$x_n = 1/n$ tend vers zéro quand n tend vers $+\infty$.

Si on veut une précision $\varepsilon = 10^{-3}$ il suffit de prendre n_0 tel que $n_0 \geq 10^3$.

3.1.16 Théorème 6 : Soit g une fonction $k - contractante$ sur $[a,b]$ à valeurs dans $[a,b]$, et $(x_n)_{n\in IN}$ la suite récurrente définie par

$x_0 \in [a,b]$, x_0 donné et $x_{n+1} = g(x_n)$ pour tout $n \geq 0$.

Alors :

1. La suite $(x_n)_{n\in IN}$ converge vers un réel θ.
2. La fonction g admet un point fixe unique.
3. Pour tout $n \in IN^*$ on a :

$$|x_n - \theta| \leq \frac{k^n}{1-k} |x_1 - x_0|$$

Preuve : Comme $x_0 \in [a,b]$ et que g une fonction $k - contractante$ sur $[a,b]$ à $[a,b]$, on a $x_n \in [a,b]$ $\forall n \in IN$.

Le fait que g est une fonction $k - contractante$ sur $[a,b]$ implique

$$|x_{n+1} - x_n| \leq |g(x_n) - g(x_{n-1})| \leq k\,|x_n - x_{n-1}| \qquad \forall n \in IN^*.$$

Par conséquent on obtient

$$|x_{n+1} - x_n| \leq k^n |x_1 - x_0| \qquad n \geq 0 \qquad (3.1.1)$$

A l'aide de l'inégalité (3.1.1) on montre que la suite $(x_n)_{n\in IN}$ vérifie :

$$\begin{aligned}|x_{n+p}-x_n| &\leq |x_{n+p}-x_{n+p-1}|+|x_{n+p-1}-x_{n+p-2}|+\ldots\ldots\ldots\ldots+|x_{n+1}-x_n| \\ &\leq k^{n+p-1}|x_1-x_0|+k^{n+p-2}|x_1-x_0|+\ldots\ldots\ldots\ldots+k^n|x_1-x_0| \\ &\leq \tfrac{1-k^p}{1-k}k^n|x_1-x_0|.\end{aligned}$$

$$|x_{n+p}-x_n| \leq \tfrac{1}{1-k}k^n|x_1-x_0| \qquad (3.1.2)$$

L'inégalité (3.1.2) prouve que la suite est de Cauchy car $k^n \to 0 \quad qd\ k \to +\infty$

alors $\forall \varepsilon > 0, \exists n_0 > 0$ tel que pour tout $n \geq n_0$ on ait :

$$k^n \leq \tfrac{1-k}{|x_1-x_0|}\varepsilon$$

et par la suite

$$\tfrac{1}{1-k}k^n|x_1-x_0| \leq \varepsilon.$$

Donc pour tout $\varepsilon > 0, \exists n_0 > 0$ tel que pour tout $n \geq n_0$ on ait :

$$|x_{n+p}-x_n| \leq \tfrac{1}{1-k}k^n|x_1-x_0| \leq \varepsilon.$$

$(x_n)_{n\in IN}$ est de Cauchy et par conséquent elle converge vers une limite θ. Comme g est continue sur $[a,b]$, et que $x_{n+1} = g(x_n)$ et que $x_n \in [a,b]\ \forall n \in IN$ alors on a $\lim\limits_{+\infty} x_n = \theta = g(\theta)$, c'est à dire que θ est un point fixe de g.

Unicité du point fixe :

Supposons que g admet un autre point fixe $\alpha \neq \beta$ alors on a :

$|g(\alpha)-g(\beta)| = |\alpha-\beta| \leq k|\alpha-\beta|$ ou encore $(1-k)|\alpha-\beta| \leq 0$ mais comme $k < 1$, alors $\alpha = \beta$.

Enfin, en faisant tendre p vers $+\infty$ on obtient dans l'inégalité(3.1.2) :

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

on obtient :

$$|\theta - x_n| \leq \frac{1}{1-k} k^n |x_1 - x_0| \qquad \forall n \in IN^*.$$

Théorème 7 : (Condition de convergence locale) Soit g une fonction de classe C^1 au voisinage de θ. Si $g(\theta) = \theta$ et $\left|g^{(1)}(\theta)\right| < 1$, alors $\exists \varepsilon > 0$ tel que $\forall x_0 \in [\theta - \varepsilon, \theta + \varepsilon]$ la suite $(x_n)_{n \in IN} = (g(x_{n-1}))_{n \in IN}$ est définie et converge vers θ,l'unique solution de $g(x) = x$ dans $I = [\theta - \varepsilon, \theta + \varepsilon]$.

preuve : Puisque g est une fonction de classe C^1 au voisinage de θ et que $\left|g^{(1)}(\theta)\right| < 1$ on a :

$$\left|g^{(1)}(x)\right| < 1, \text{au voisinage de } \theta.$$

Par conséquent, il existe $\varepsilon > 0$ tel que :

$\forall x \in I = [\theta - \varepsilon, \theta + \varepsilon] \left|g^{(1)}(x)\right| < 1$, et puisque $g^{(1)}$ est continue sur le fermé I, on en déduit qu'$\exists\, k \in [0, 1]$ tel que

$$\forall x \in I = [\theta - \varepsilon, \theta + \varepsilon] \qquad \left|g^{(1)}(x)\right| \leq k < 1$$

Pour appliquer le théorème 6, il suffit de vérifier que : $g(I) \subset I$

Or, par application du théorème des accroissements finis on a :

$$\forall x \in I = [\theta - \varepsilon, \theta + \varepsilon] \qquad |g(x) - \theta| \leq |x - \theta|.$$

Remarque 2 : * Si $\left|g^{(1)}(\theta)\right| = 1$, la suite peut converger ou diverger.

* Si $\left|g^{(1)}(\theta)\right| > 1$, et si la suite possède une infinité de termes différents de θ, alors la suite ne peut converger.

Théorème 8 : La suite récurrente définie par $x_0 \in [a,b]$, x_0 donné et $x_{n+1} = g(x_n), \forall n \geq 0$, converge linéairement vers θ et si g est de classe C^1 sur $[a,b]$, alors

$$C = \lim_{+\infty} \frac{|e_{n+1}|}{|e_n|} = \mid g^{(1)}(\theta) \mid$$

preuve : Il suffit d'appliquer le théorème des accroissements finis dans l'intervalle d'extrémités x_n et θ,on a

$|e_{n+1}| = |x_{n+1} - \theta| = |g(x_n) - \theta| = \left|(x_n - \theta)g^{(1)}(c_n)\right|$ et de là on obtient :

$$\lim_{+\infty} \frac{|e_{n+1}|}{|e_n|} = \lim_{+\infty} \mid g^{(1)}(c_n) \mid = \mid g^{(1)}(\theta) \mid .$$

3.2 Méthode de Newton et méthode de la corde

3.2.1 Méthode de Newton (ou Newton-Raphson) :

Soit une $f : IR \to IR$ une fonction de classe C^1 et θ un zéro simple de f, c'est à dire $f(\theta) = 0$ et $f^{(1)}(\theta) \neq 0$. Supposons que l'on connaisse une valeur x_n proche de θ. Pour calculer x_{n+1} nous prenons l'intersection de l'axe Ox avec la droite de la tangente du graphe de f passant par le point $(x_n, f(x_n))$

Clairement, nous avons la relation $f(x_n)/(x_n - x_{n+1}) = f^{(1)}(x_n)$ qui donne, lorsque x_0 est choisi proche de θ, la méthode de Newton :

$$x_{n+1} = x_n - \frac{f(x_n)}{f^{(1)}(x_n)} \qquad n = 0, 1, ... \qquad (3.2.1)$$

Nous voyons ainsi que la la méthode de Newton est une méthode de point fixe pour calculer θ. En effet, il suffit de constater que si on pose :

$g(x) = x - \frac{f(x)}{f^{(1)}(x)}$ alors $f(x) = 0 \Longleftrightarrow x = g(x)$ (du moins au voisinage de θ pour lequel $f^{(1)}(x) \neq 0$) et (3.2.1) s'écrit

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

En vu d'utiliser le Théorème de la convergence locale, calculons $g^{(1)}(x)$:

si f est C^2 :

$$g^{(1)}(x) = 1 - \frac{(f^{(1)}(x))^2 - f(x)f^{(2)}(x)}{\left[f^{(1)}(x)\right]^2}$$

et par la suite , puisque $f(\theta) = 0$ et $f^{(1)}(\theta) \Longrightarrow g^{(1)}(\theta) = 0$.

Nous obtenons le résultat suivant :

Théorème 9 : Supposons f est C^2 et supposons que θ soit tel que $f(\theta) = 0$ et $f^{(1)}(\theta) \neq 0$. Alors $\exists \varepsilon > 0$ si x_0 satisfait $|\theta - x_0| \leq \varepsilon$, la suite $(x_n)_{n \in IN}$ donnée par la méthode de Newton (3.2) converge vers θ. De plus la convergence est quadratique.

Preuve :

on a $g(x) = x - \frac{f(x)}{f^{(1)}(x)}$ et $\left|g^{(1)}(\theta)\right| < 1$ alors la convergence annoncée dans ce théorème est une conséquence du théorème de la convergence locale.

A priori la covergence est linéaire

$$|g(x) - g(y)| = \left|\int_y^x g^{(1)}(t)dt\right| \leq \max_{t \in I} \left|g^{(1)}(t)\right| |x - y| < k\,|x - y|$$

si nous prenons $y = \theta$ nous tirons que

$$|g(x) - g(\theta)| \leq k\,|x - \theta| \leq |x - \theta| \leq \varepsilon.$$

Nous allons maintenant montrer que la convergence est quadratique, ceci est une conséquence du fait que $g^{(1)}(\theta) = 0$.

Nous développons f autour de x_n, nous obtenons :

$$f(x) = f(x_n) + (x - x_n)f^{(1)}(x_n) + \frac{1}{2!}(x - x_n)^2 f^{(2)}(\xi_x)$$

où $\xi_x \in$ à l'intervalle d'extrimité x et x_n. En choisisant $x = \theta$ dans l'égalité ci-dessous, en divisant par $f^{(1)}(x_n)$ et en tenant compte du fait que $f(\theta) = 0$, nous avons :

$$\frac{f(x_n)}{f^{(1)}(x_n)} + \theta - x_n + \frac{f^{(2)}(\xi_\theta)}{2f^{(1)}(x_n)}(x - x_n)^2 = 0$$

En utilisons (3.2.1) nous obtenons

$$|x_{n+1} - \theta| = \frac{1}{2}\frac{|f^{(2)}(\xi_\theta)|}{|f^{(1)}(x_n)|} \mid \theta - x_n \mid^2$$

Il suffit maintenant de poser

$$C = \frac{1}{2}\frac{\max\limits_{x\in I}|f^{(2)}(\xi_\theta)|}{\min\limits_{x\in I}|f^{(1)}(x_n)|}$$

Pour obtenir

$$|x_{n+1} - \theta| \leq C \mid x - x_n \mid^2$$

d'où la convergence quadratique.

B. Mhéthode de la corde (ou Newton modifiée) Cette méthode permet d'éviter qu'à chaque itération de (3.2.1) on ait à évaluer $f^{(1)}(x_n)$. La méthode de la corde consiste à remplacer $f^{(1)}(x_n)$ par $f^{(1)}(x_0)$ dans (3.2.1), ce qui donne :

$$x_{n+1} = x_n - \frac{f(x_n)}{f^{(1)}(x_0)} \qquad n = 0, 1.....$$

Le calcul de la suite $(x_n)_{n\in IN}$ s'effectue en prenant toujours la même pente $f^{(1)}(x_0)$, d'où l'appelation méthode de la corde. Ici encore, nous posons

$$g(x) = x - \frac{f(x)}{f^{(1)}(x_0)}$$

et constatons que $f(\theta) = 0$ si $g(\theta) = \theta$.

Ainsi on a

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

et la méthode est une méthode de point fixe.

Remarque : g dépend du point fixe de départ x_0.

Théorème 10 : Supposons f de C^2 et supposons θ soit tel que $f(\theta) = 0$ et $f^{(1)}(\theta) \neq 0$. Alors $\exists \varepsilon > 0$ tel que si $x_0 \in I = [\theta - \varepsilon, \theta + \varepsilon]$, la suite $(x_n)_{n\in IN}$ donnée par la méthode de la corde converge vers θ. La convergence est linéaire.

Preuve : f de C^2 et puisque $f^{(1)}(\theta) \neq 0$, il est facile de montrer qu'$\exists \varepsilon > 0$ et $k < 1$ tels que $x_0 \in I = [\theta - \varepsilon, \theta + \varepsilon]$ on a

$$\left|g^{(1)}(x)\right| = \left|1 - \frac{f^{(1)}(x)}{f^{(1)}(x_0)}\right| < k \qquad \forall x \in I.$$

et par la suite on a

$$|g(x) - g(y)| = \left|\int_y^x g^{(1)}(t)dt\right| \leq \max_{t\in I} \left|g^{(1)}(t)\right| |x - y| < k\,|x - y|$$

si $y = \theta$, on a $|g(x) - g(\theta)| \leq k\,|x - \theta|$, c à d $|g(x) - \theta| \leq k\,|x - \theta|$

3.3 Méthode de dichotomie :

Soit f une fonction continue sur $[a,b]$ vérifiant $f(a) \times f(b) \leq 0$.

La fonction f admet au moins un zéro dans $[a,b]$. La méthode de dichotomie consiste à approcher θ par un encadrement, en réduisant à chaque étape l'intervalle de moitié selon l'algorithme suivant

Etape I : on pose $a_0 = a$ et $b_0 = b$, on pose $c = \frac{a_0+b_0}{2}$ puis on teste si $c_0 = \theta$ c'est terminé, sinon si $f(a_0) \times f(c_0) \leq 0$ alors $\theta \in [a_0, c_0]$, on pose $a_1 = a_0$ et $b_1 = c_0$ puis $c_1 = \frac{a_1+b_1}{2}$.

Si $f(b_0) \times f(c_0) \leq 0$ alors $\theta \in [c_0, b_0]$, alors on pose $a_1 = c_0$ et $b_1 = b_0$ puis $c_1 = \frac{a_1+b_1}{2}$.

Aprés cette étape la longueur de $[a_1, b_1]$ est égale à $\frac{b_0-a_0}{2} = \frac{b-a}{2}$.

Etape II : On recommence le procédé de l'étape 1.

Etape k : A chaque étape k du procédé, soit on tombe sur $c_k = \theta$ soit on diminue la longueur de l'intervalle de moitié.

Théorème : Les a_k, b_k et c_k satisfont les propriétés suivantes :

1. $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$.
2. $b_{k+1} - a_{k+1} = \frac{b_k-a_k}{2} = \frac{b_0-a_0}{2^{k+1}}$.
3. la suite c_k converge vers θ.
4. $|c_k - \theta| \leq \frac{b-a}{2^{k+1}}$.

Preuve : 1. Pour $k \geq 0$ on $c_k = (a_k + b_k)/2$ et $[a_{k+1}, b_{k+1}] = [a_k, c_k]$ ou $[c_k, b_k]$ donc $[a_{k+1}, b_{k+1}] \subset [a_k, b_k]$.

2. On a par construction $b_{k+1} - a_{k+1} = (b_k - a_k)/2$ montre par récurrence que

$$b_k - a_k = (b - a)/2^k$$

Pour $k = 0$ la relation est vraie.

Si on suppose que la relation est vraie à l'ordre k, c'est à dire $b_k - a_k = (b - a)/2^k$.

Montrons alors que $b_{k+1} - a_{k+1} = (b - a)/2^{k+1}$.

En effet, $b_{k+1} - a_{k+1} = (b_k - a_k)/2 = \frac{1}{2}(b_k - a_k)/2 = (b - a)/2^{k+1}$.

3. Par construction $\theta \in [a_k, b_k]$ et $c_k = (a_k + b_k)/2$ est le milieu de $[a_k, b_k]$ donc :

$$|c_k - \theta| \leq (b_k - a_k)/2 \leq (b - a)/2^{k+1} \to 0 \qquad qd \qquad k \to +\infty.$$

En d'auters termes :

$$c_k \to \theta \qquad qd \qquad k \to +\infty.$$

Remarque : Le théorème précédent permet de calculer à l'avance le nombre maximal $n \in IN$ d'itérations assurant la précision ε, en effet :

Pour que c_n vérifie : $|c_k - \theta| \leq (b - a)/2^{n+1}$ à la n-ième étape, il suffit que n vérifie : $(b - a)/2^{n+1} \leq \varepsilon$, on a alors :

$$|c_k - \theta| \leq (b - a)/2^{n+1} \leq \varepsilon,$$
$$\frac{b-a}{\varepsilon} \leq 2^{n+1} \Longleftrightarrow n \geq \frac{\ln(b-a) - Ln\varepsilon}{\ln 2} - 1.$$

Exemple : $f(x) = x^3 + 4x^2 - 10$. On vérifie graphiquement que f admet une racine réellé dans $[1, 2]$ et que la méthode de dichotomie est appliquable. $f(1) \times f(2) \leq 0$.

Pour trouver une approximation de cette racine, on peut réaliser la méthode de dichotomie avec une précision égale à 10^{-10}.

On a les résultats suivants : $n\ (numérique) = 33 \qquad n(théorie) = 32, 21928$

$x = 1.3652300134 \qquad f(x) = -2.378897e - 0.11.$

3.4 Méthode de la fausse position (Fegula Falsi)

Au lieu de prendre à chaque étape c_k qui est le milieu de $[a, b]$, la méthode de fausse position prend le point d'intersection de l'axe Ox avec la droite passant par $(a_k, f(a_k))$ et $(b_k, f(b_k))$.

L'équation de cette droite est donnée par :

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

Elle coupe l'axe Ox au point : $M(c_k, 0)$ où $c_k = a_k + f(a_k)\frac{a_k - b_k}{f(a_k)-f(b_k)}$.

On suit le procédé comme dans le cas de dichotomie en testant :

Si $f(c_k) \times f(a_k) \leq 0$ alors $\theta \in [a_k, c_k]$, alors on pose $a_{k+1} = a_k$ et $b_{k+1} = c_k$.

Si $f(c_k) \times f(b_k) \leq 0$ alors $\theta \in [c_k, b_k]$, alors on pose $a_{k+1} = c_k$ et $b_{k+1} = b_k$.

Puis on cherche à nouveau la droite passant par $(a_{k+1}, f(a_{k+1}))$ et $(b_{k+1}, f(b_{k+1}))$.

Exemple : $f(x) = x^3 - 20$ et comme $f(0.75) \times f(4.5) < 0$ on peut donc appliquer la méthode de la fausse position dans l'intervalle $[0.75, 4.5]$.

La solution est $x = 2.7133$.

Chapitre 4

Problèmes d'interpolation

4.1 Position du problème :

Supposons que l'on veuille chercher un polynôme p de degré $n \geq 0$ qui, pour des valeurs $t_0, t_1, t_2, ..., t_n$ distinctes données, prennent des valeurs $p_0, p_1, p_2, ..., p_n$ respectivement, c'est à dire

$$p(t_j) = p_j, \text{ pour } 0 \leq j \leq n \qquad (4.1)$$

Une manière apparemment simple de résoudre ce problème est d'écrire

$$p(t) = a + a_1 t + a_2 t^2 + ... + a_n t^n \qquad (4.2)$$

où $a_0, a_1, a_2, ..., a_n$ sont des coefficients qui devront être déterminés. Les $(n+1)$ relations (4.1) s'écrivent alors :

$$a + a_1 t_j + a_2 t_j^2 + ... + a_n t_j^n = p_j \qquad \text{pour } 0 \leq j \leq n \qquad (4.3)$$

On obtient un système de $(n+1)$ équations à $(n+1)$ inconnues $a_0, a_1, a_2, ..., a_n$

Soit T la $(n+1) \times (n+1)$ matrice définie par :

$$T = \begin{bmatrix} 1 & t_0 & t_0^2 & . & . & . & . & . & . & t_0^n \\ 1 & t_1 & t_1^2 & . & . & . & . & . & . & t_1^n \\ . & & . & . & & & & & & . \\ . & & & . & . & & & & & . \\ . & & & & . & . & & & & . \\ . & & & & & . & . & & & . \\ . & & & & & & . & . & & . \\ . & & & & & & & . & & . \\ . & & & & & & & & . & . \\ 1 & t_n & t_n^2 & & & & & & & t_n^n \end{bmatrix}$$

C'est la matrice de Vandermonde associée aux points $t_0,\ t_1,\ t_2, ..., t_n$

Si $\vec{a} et \vec{p}$ sont $(n+1)$- vecteurs colonnes suivants :

$\vec{a} = (a_0,\ a_1,\ a_2, ..., a_n)^t \quad et \quad \vec{p} = (p_0,\ p_1,\ p_2, ...,\ p_n)^t$, nous pouvons écrire (4.3) sous forme matricielles :

$$T\vec{a} = \vec{p} \qquad (4.4)$$

Ainsi, le problème consiste à chercher le polynôme p satisfaisant (4.1) peut se réduire à résoudre le système linéaire (4.4) qui n'est pas une tâche triviale.

4.2 Interpolation de LAGRANGE

1.1 Base de Lagrange : Il est facile de résoudre le problème (4.1) lorsque toutes les valeurs p_j sont égales à zéro sauf une, qui est fixée à 1.

Soit k un entier donné entre 0 et x, et supposons que l'on ait $p_k = 1 \quad et \quad pour\ j \neq k \quad p_j = 0.$

Soit φ_k la fonction de t définie par

$$\varphi_k(t) = \frac{(t-t_0)(t-t_1).....(t-t_{k-1})(t-t_{k+1}).....(t-t_n)}{(t_k-t_0)(t_k-t_1).....(t_k-t_{k-1})(t_k-t_{k+1}).....(t_k-t_n)} \qquad (4.5)$$

Le numérateur de φ_k est un produit de n termes $(t-t_j)$, $j \neq k$ et est donc un polynôme de degré n en t.

Le dénominateur de φ_k est une constante et il est facile de vérifier que :

(i) φ_k est un polynôme de degré n

(ii) $\varphi_k(t_j) = 0$ si $j \neq k$, $0 \leq j \leq n$

(iii) $\varphi_k(t_k) = 1$

A chaque point t_k nous avons donc associé un polynôme φ_k de degré n valant 1 en t_k et *zéro* aux autres points t_j, $j \neq k$.

Les polynômes $\varphi_0, \varphi_1, \varphi_2, ..., \varphi_n$ sont linéairements indépendants. En effet $\forall t \in IR$, si $a_0, a_1, a_2, ..., a_n$ sont $n+1$ nombres réels tels que $\Sigma_{j=0}^{n} \alpha_j \varphi_j(t) = 0$ $(\forall t \in IR)$, alors pour $t = t_k$ nous obtenons :

$0 = \Sigma_{j=0}^{n} \alpha_j \varphi_j(t_k) = \alpha_k$,

et par conséquent tous les $\alpha_k = 0$ *pour* $0 \leq k \leq n$.

Notons IP_n l'espace vectoriel formé par tous les polynômes de degré $\leq n$ de dimension $n+1$ et que sa base canonique est donnée par $1, t, t^2, ..., t^n$.

Le fait que les polynômes $\varphi_0, ..., \varphi_k, ..., \varphi_n$ sont linéairements indépendants montre que ces derniers forment aussi une base de IP_n.

Définition 1.1 : Nous dirons que $(\varphi_0, ..., \varphi_k, ..., \varphi_n)$ est base de Lagrange de IP_n associée aux points $t_0, t_1, t_2, ..., t_n$.

Exemple 1.1 : Prenons $n = 2$, $t_1 = -1, t_2 = 0$ et $t_3 = 1$. La base de Lagrange de IP_n associée aux points -1, 0 et 1 est formée par les polynômes définis par : $\varphi_0(t) = \frac{(t-t_1)(t-t_2)}{(t_0-t_1)(t_0-t_2)} = 0.5t^2 - 0.5t \qquad (4.6)$

$$\varphi_0^1(t) = \frac{(t-t\ )(t-t_2)}{(t_1-t_0)(t_1-t_2)} = 1 - t \qquad (4.7)$$

$$\varphi_2(t) = \frac{(t-t_0)(t-t_1)}{(t_2-t_1)(t_2-t_1)} = 0.5t^2+0.5t \qquad (4.8)$$

Revenons au point (4.1) consistant à chercher le polynôme p de degré n qui prenne des valeurs données $p_0,\ p_1,\ p_2, ...,\ p_n$ en des points distincts $t_0,\ t_1,\ t_2, ...,\ t_n$.

Soit $(\varphi_0, ..., \varphi_k, ..., \varphi_n)$ une base de Lagrange de IP_n associée aux points $t_0,\ t_1,\ t_2, ...,\ t_n$.

Alors le polynôme p cherché est défini par ;

$$p(t) = p_0\varphi_0(t) + p_1\varphi_1(t) + ... + p_n\varphi_n(t) = \Sigma_{j=0}^{n}\, p_j\varphi_j(t) \qquad (4.9)$$

En effet, puisque p est une combinaison linéaire de $(n+1)$ polynômes $\varphi_0, ..., \varphi_k, ..., \varphi_n$ tous de degré $\mathbf{n}$, alors p est lui même de degré n, c'est à dire $p \in IP_n$.

D'autre part, si nous utilisons les propriétés des polynômes φ_k, nous avons pour $k = 0, 1, 2, ..., n$:

$$p(t_k) = \Sigma_{j=0}^{n}\, p_j\varphi_j(t_k) = p_k \qquad (4.10)$$

qui est bien la relation (1.1).

Exemple 2 : Trouver un polynôme de degré deux qui vaut en $t_0 = -1$ en $p_0 = 8$, en $t_1 = 0$ en $p_1 = 3$, en $t_2 = 3$ en $p_0 = 6$.

$$p(t) = 4t^2 - t + 3.$$

4.3 Interpolation d'une fonction continue par un polynôme

Soit une fonction $: IR \to IR$ continue donnée et soit $t_0,\ t_1,\ t_2, ...,\ t_n$ $(n+1)$ points distincts donnés.

Nous cherchons maintenant à interpoler f par un polynôme p de degré n aux points $t_0,\ t_1,\ t_2, ...,\ t_n$ c'est à dire nous cherchons un polynôme p de degré n tel que :

$$p(t_k) = f(t_k), \qquad 0 \le k \le n \qquad (4.11)$$

Si $f(t)$ est donnée, alors en posant $p_j = f(t_j)$, $0 \leq j \leq n$ et en suivant ce qui est fait dans la relation (4.3), nous obtenons

$$p(t) = p_0\varphi_0(t) + p_1\varphi_1(t) + ... + p_n\varphi_n(t) = \sum_{j=0}^{n} p_j\varphi_j(t),$$

où φ_j, $0 \leq j \leq n$ forment une base de Lagrange de IP_n associée aux points $t_0, t_1, t_2, ..., t_n$.

La solution du point (4.11) est donc définie par :

$$p(t) = \sum_{j=0}^{n} f(t_j)\varphi_j(t) \qquad \forall t \in IR \qquad (4.12)$$

Définition 1.2 : On dira que le polynôme p défini par (4.12) est l'interpolant de f de degré n aux points $t_0, t_1, t_2, ..., t_n$.

Exemple 3 : Soit $f(t) = e^t$. Trouver l'interpolants de f de degré 2 aux points $t_0 = -1$, $t_1 = 0$, $t_2 = 1$.

Soit maintenant une fonction :$[a,b] \to IR$ continue et donnée sur un intervalle $[a,b]$. Soit n un entier positif et considérons le cas où tous les points $[a,b]$, $0 \leq j \leq n$, sont équidistribués dans $[a,b]$, c'est à dire $t_j = a + jh$, $0 \leq j \leq n$, avec $h = \frac{b-a}{n}$. Soit p l'interpolant de f de degré n aux points $t_0, t_1, t_2, ..., t_n$ que nous noterons p_n pour montrer qu'il dépend bien de n choisi au départ. D'aprés (4.12), p_n est défini par

$$p_n(t) = \sum_{j=0}^{n} f(t_j)\varphi_j(t) \qquad \forall t \in IR \qquad (4.13)$$

où $(\varphi_0, ..., \varphi_k, ..., \varphi_n)$ est la base de Lagrange de IP_n associée aux points $t_0, t_1, t_2, ..., t_n$. On peut montrer le résultat suivant :

4.4 Existance et unicité de l'interpolant

1.51 Théorème 1 : Il existe un polynôme p_n unique de degré $\leq n$, interpolant f en (n+1) points, c'est à dire tel que :

$$p(t_k) = f(t_k) \qquad pour \qquad 0 \leq k \leq n.$$

Preuve : Existence :

Soit

$$L_i(t) = \frac{(t-t_0)(t-t_1).....(t-t_{i-1})(t-t_{i+1}).....(t-t_n)}{(t_i-t_0)(t_i-t_1).....(t_i-t_{k-1})(t_i-t_{i+1}).....(t_i-t_n)}$$

$p_n(t) = \Sigma_{j=0}^{n} f(t_j)L_j(t)$ *pour* $0 \leq j \leq n$, on a $L_i(t_j) = \delta_{ij}$ et par conséquent $p_n(t_i) = f(t_i)$.

Unicité :

Supposons qu'il existe deux polynômes p_n et q_n de degré $\leq n$, interpolant f aux points $t_0, t_1, t_2, ..., t_n$, en posant $d_n = p_n - q_n$, on arrive à une contradiction. En effet, d_n est un polynôme de degré $\leq n$ et par conséquent il peut avoir au plus n zéros, mais d'autre part $d_n(t_k) = 0$, *pour* $0 \leq k \leq n$ ce qui voudrait dire que d_n aurait $n+1$ zéros d'où la contradiction donc $p_n = q_n$.

1.5.2 Erreure d'interpolation.

1.5.2.1 Théorème : Soit p_n le polynôme interpolant de f aux points $a = x_0 < x_1 < ..... < x_n = b$, si f est de classe C^{n+1} sur $[a,b]$ alors :

a. $\forall x \in [a,b]$, il existe $\Theta = \Theta(x) \in [a,b]$ tel que :

$$e_n(x) = f(x) - p_n(x) = (f^{(n+1)}(\Theta)/(n+1)!)\Pi_{n+1}(x)$$

avec

$$\Pi_{n+1}(x) = \Pi_{i=0}^{n}(x - x_i)$$

b. En posant

$$M_{n+1} = \max_{x\in[a,b]} \left| f^{(n+1)}(\Theta) \right|$$

On obtient :

$$\max_{x\in[a,b]} |f(x) - p_n(x)| \leq \frac{M_{n+1}}{(n+1)!} \max_{x\in[a,b]} |\Pi_{n+1}(x)|$$

et en particulier :

$$\max_{x\in[a,b]} |f(x) - p_n(x)| \leq \frac{M_{n+1}}{(n+1)!}(b-a)^{n+1}.$$

Lemme 1 : Soit f une fonction définie sur $[a,b]$ à valeurs dans IR dérivable sur $[a,b]$, si f possède au moins $n+2$ zéros distincts sur $[a,b]$, alors $f'=f^{(1)}$ possède au moins $n+1$ zéros distincts sur $[a,b]$.

Il suffit d'appliquer le théorème de Rolle entre deux zéros consécutifs de f.

Corolaire 1 : soit f une fonction de classe C^{n+1} sur $[a,b]$. Si f possède au moins n+2 zéros distincts sur $[a,b]$ alors $f^{(n+1)}$ possède au moins 1 zéros sur $[a,b]$.

Il suffit de faire une récurrence en appliquant le Lemme 1 précédent.

Preuve du théorème : Si $x=x_i$, le résultat est évident.

Si $x\neq x_i$, posons ; $R(t)=f(t)-p_n(t)-\frac{f(x)-p_n(x)}{\Pi_{n+1}(x)}\,\Pi_{n+1}(t).$

on vérifie alors que $R\in C^{n+1}[a,b]$ et que :

$$R(x_i)=f(x_i)-p_n(x_i)-\frac{f(x)-p_n(x)}{\Pi_{n+1}(x)}\,\Pi_{n+1}(x_i)=0,\qquad i=1,2,.....,n$$

et

$$R(x)=e_n(x)-e_n(x)=0.$$

Par conséquent, R admet au moins n+2 zéros distincts sur $[a,b]$, en appliquant le corolaire précédent, on montre que $R^{(n+1)}$ possède au moins 1 zéros sur $[a,b]$, c'est à dire qu'il existe $\Theta\in[a,b]$ tel que $R^{(n+1)}(\Theta)=0$,

et donc :

$$e_n(x)=(f^{(n+1)}(\Theta)/(n+1)!)\Pi_{n+1}(x)$$

ce qui implique que

$$\max_{x\in[a,b]}|e_n(x)|\leq\frac{\max_{x\in[a,b]}|\Pi_{n+1}(x)|}{(n+1)!}M_{n+1}.$$

2 Interpolation de Newton :

2.1.1 Définition : différences divisées soit f une fonction définie sur $[a,b]$ à valeurs dans IR dérivable sur $[a,b]$ contenant (n+1) points distincts $t_0, t_1, t_2, ..., t_n$.

On définit les différences divisées d'ordre i de f aux points (x_k) comme suit :

$f[x_0] = f(x_0)$

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

.

.

$f[x_0, x_1, ..., x_i] = \frac{f[x_1,x_2,.....,x_i]-f[x_0,x_1,.....,x_{i-1}]}{x_i-x_0}$ *pour* $i \geq 2$.

Exemple : $x_0 = -1, x_1 = 1, x_2 = 1, f(x_0) = 2, f(x_1) = 1, f(x_2) = -1,$

on obtient :

$f[x_0] = 2$

$f[x_0, x_1] = -1$

$f[x_1] = 1$

$f[x_2] = -1$

$f[x_1, x_2] = -2$

$f[x_0, x_1, x_2] = -0.5.$

2.1.2 Propriétés : La valeur d'une différence divisée est indépendante de l'ordre de x_i;

$$f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0} = f[x_1, x_0] = \frac{f(x_0)}{x_0-x_1} + \frac{f(x_1)}{x_1-x_0}$$

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

$$= f[x_1, x_2, x_0] = f[x_2, x_1, x_0].$$

De façon générale :

$$f[x_0, x_1, ..., x_i] = \frac{f[x_1,x_2,.....,x_i]-f[x_0,x_1,.....,x_{i-1}]}{x_i-x_0}$$

$$= \frac{f(x_0)}{(x_0-x_1)(x_0-x_2).....(x_0-x_i)} + \frac{f(x_1)}{(x_1-x_0)(x_1-x_2).....(x_1-x_i)} + ... + \frac{f(x_i)}{(x_i-x_0)(x_i-x_2).....(x_i-x_{i-1})}.$$

2.1.3 Interpolant de Newton : On appelle interpolant de newton le polynôme p_n donné par :

$$p_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + ...$$

$$..... + f[x_0, x_1, .., x_n](x - x_0)(x - x_1)...(x - x_{n-2})(x - x_{n-1}).$$

Exemple : $x_0 = -1, f(x_0) = 2, x_1 = 0, f(x_1) = 1, x_2 = 1, f(x_2) = -1.$

$f[x_0] = 2$

$f[x_0, x_2] = -1$

$f[x_1, x_2] = -2$

$f[x_0, x_1, x_2] = -0.5$

$p_n(x) = 1 - 1.5x - 0.5x^2.$

2.1.5 Base de Newton : Soient $t_0, t_1, t_2, ..., t_n$ $(n+1)$ points deux à deux distincts d'un intervalle $[a, b]$ de IR et les polynômes N_i définis par

$N_0(x) = 1, N_1(x) = (x - x_0)$

$N_i(x) = (x - x_0)(x - x_1).....(x - x_{n-2})(x - x_{i-1}) \qquad i = 1, 2, ..., n.$

3.1.6 Les polynômes N_i ont les propriétés suivantes : 1. N_i est un polynôme degré i

2. Pour $i \geq 1$, $N_i(x)$ admet $t_0, t_1, t_2, ..., t_{i-1}$ comme racines.

3. La famille $\{N_0(x), ..., N_n(x)\}$ est une base de IP_n dite base de Newton.

Preuve : 1. est évidente d'aprés la définition de $N_i(x)$

2. est évidente d'aprés la définition de N_i

3. Il suffit de montrer que la famille $\{N_0(x), ..., N_n(x)\}$ est libre.

Soient $c_0, c_1, ..., c_n$ des constantes telles que

$$\sum_{i=0}^{n} c_i N_i(x) = 0$$

comme les x_i sont supposés deux à deux distincts, on a

$0 = \sum_{i=0}^{n} c_i N_i(x_0) = c_0$

$0 = \sum_{i=0}^{n} c_i N_i(x_1) = c_0 N_0(x_1) + c_1 N_1(x_1) = c_1(x_1 - x_0) \Rightarrow c_1 = 0$, car $x_1 \neq x_0$

.

.

$0 = \sum_{i=0}^{n} c_i N_i(x_n) = c_n N_1(x_n) = c_n(x_n - x_0)(x_n - x_1).....(x_n - x_{n-2})(x_n - x_{n-1}) \Rightarrow c_n = 0$,

car les x_i sont deux à deux distincts.

d'où le résultat.

2.1.7 Théorème : Soit f une fonction numérique définie sur un intervalle $[a, b]$.

Soit p_n un polynôme interpolant de f en (n+1) points $x_0, x_1, ..., x_n \in [a, b]$.

a. on peut exprimer $p_n(x)$ comme combinaison linéaires des $N_i(x)$ de la base de Newton :

$$p_n(x) = \sum_{i=0}^{n} D_i N_i(x).$$

on obtient le système triangulaire inférieur suivant :

$$\left\{ \begin{array}{ll} p_n(x_0) = D_0 & = f(x_0) \\ p_n(x_1) = D_0 + D_1 N_1(x_1) & = f(x_1) \\ & . \\ & . \\ p_n(x_n) = D_0 + D_1 N_1(x_n) + D_2 N_2(x_n) + ... + D_n N_n(x_n) & = f(x_n) \end{array} \right\} \quad (S_1)$$

Les D_i solutions du système (S_1) sont données par

$D_0 = f[x_0]$

$D_1 = f[x_0, x_2]$

.

.

.

$D_i = \frac{f[x_1, x_2, ..., x_i] - f[x_0, x_1, ..., x_{i-1}]}{x_i - x_0} = f[x_0, ..., x_i]$ pour $i \geq 2$.

2.1.8 Exemple : Soit la fonction f telle que

x	0.15	2.30	3.15	4.85	6.25	7.95
$f(x)$	4.79867	4.49013	4.2243	3.47313	2.66674	1.51909

Les coeffficients du polynôme interpolant de f dans la base de Newton sont :

$D_0 = 4.798670, D_1 = -0.143507, D_2 = -0.056411, D_3 = 0.001229,$

$D_4 = 0.000104,$

$D_5 = -0.000002.$

Chapitre 5

Dérivation et intégration numérique

5.1 Dérivation numérique

Dans ce paragraphe, la fonction n'est bien sûr pas connue par une formule explicite mais :

- ou bien par ses valeurs sur un ensemble discret (en supposant que les points sont assez proches pour que la notion de dérivée ait un sens).

- ou bien, le plus souvant, par algorithme de calcul ou une formule compliquée qui permet, au moins en théorie, de la calculer en tout point. On suppose bien sûr que la dérivée n'est pas accessible par un procédé analogue.

5.1.1 Dérivée première :

Supposons qu'on veuille calculer une valeur approchée de $f'(x_i)$.Une première idée, consiste à remplacer f par un polynôme d'interpolation au voisinage du point x_i et on dérive celui-ci. Les formules vont varier en fonction du nombre des points qu'on choisit pour écrire le polynôme d'interpolation (en général 2 ou 3).

Dans toute la suite, on supposera f connue ou calculable aux points $\ldots\ldots, x_{i-2}, x_{i-1}, x_i, \ldots\ldots$qu'on supposera proches. On notera $h_i = x_{i+1} - x_i$.

5.1.1 Formules à deux points : Le polynôme d'interpolation sur les deux points x_i,x_{i+1} s'écrit :

$$p(x) = f(x_i) + f[x_i, x_{i+1}](x - x_i)$$

On a donc $p'(x) = f[x_i, x_{i+1}]$, ce qui fournit la formule à droite

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \qquad (5.1)$$

On a bien sûr aussi la formule à gauche

$$f'(x_i) \simeq \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \qquad (5.2)$$

5.1.2 Formules à de 3 points : On choisit d'interpoler sur les points x_{i-1}, x_i, x_{i+1} (ce qui est normalement plus satisfaisant). Dans ce cas on a

$$p(x) = f(x_i) + f[x_i, x_{i+1}](x - x_i) + f[x_{i-1}, x_i, x_{i+1}](x - x_{i-1})(x - x_i) \qquad (5.3)$$

Donc

$$p'(x) = f[x_{i-1}, x_i] + f[x_{i-1}, x_i, x_{i+1}](x - x_{i-1})(x_i - x_{i-1})$$

ce qui fournit, après simplification la formule centrée avec :

$$h_i = x_{i+1} - x_i \qquad (5.4)$$

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_{i-1})}{2h} \qquad (5.5)$$

Remarque : La formule ci-dessus n'est autre la moyenne des deux formules décentrés dans le cas équidistants $(h_i = x_{i+1} - x_i)$.

5.1.3 Erreur : Pour le calcul théorique de l'erreur commise quand on remplace $f(x_i)$ par d'une des formules approchées ci-dessus on a :

$$\left| f'(x_i) - \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} \right| \leq M_2 \frac{h}{2}$$

$$\left| f'(x_i) - \frac{f(x_{i+1}) - f(x_{i-1})}{2h} \right| \leq M_3 \frac{h^2}{6} \qquad M_i = \max_{[a,b]} \left| f^{(i)}(x) \right|$$

5.1.2 Dérivées d'ordre supérieure :

Le principe est exactement le même, il faut simplement prendre garde que le degré du polynôme d'interpolation soit suffisant pour que sa dérivée n-ième soit non nulle !

Par exemple, pour la dérivée seconde, on choisit en général d'interpoler sur 3 points, ce qui donne

Dérivée seconde : points équidistants

$$f''(x_i) \simeq \frac{f(x_{i+1})+f(x_{i-1})+2f(x_i)}{h^2}$$

avec une erreur

$$\left| f''(x_i) - \frac{f(x_{i+1})+f(x_{i-1})+2f(x_i)}{h^2} \right| \leq M_4 \frac{h^2}{12}$$

5.2 Intégration numérique.

5.2 Généralités : Nous avons pour but de calculer numériquement des intégrales définis. Soit $f : [a, b] \to IR$ une fonction continue donnée sur $[a, b]$. Nous désirons approcher numériquement la quantité

$$\int_a^b f(x)dx \qquad (5.2.1)$$

Pour ce faire, nous commençons par partitionner $[a, b]$ en petits intervalles $[x_i, x_{i+1}]$, $i = 0, 1, 2....., N$ tels que :

$$a < x_0 < x_1 < ..... < x_N < b \qquad (5.2.2)$$

Soit

$$h = \max_{0 \leq i \leq N-1} |x_{i+1} - x_i|$$

Le réel positif caractérisant la finesse de la partition. Il est clair que, N augmente, nous pouvons nous placer les points x_i de sorte à ce que h soit petit. Lorsqu'aucune raison nous incite à choisir des intervalles de longueurs différentes, nous posons

$h = \frac{b-a}{N}$ et $x_i = a + ih, i = 0, 1, ..., N.$

Etant donné la partition (5.2.2), il est naturel d'écrire :

$$\int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{x_i}^{x_{i+1}} f(x)dx \qquad (5.2.3)$$

ce sont ainsi les intégrales

$$\int_{x_i}^{x_{i+1}} f(x)dx$$

que nous allons approcher dans la suite par des formules appelées "formules de quadrature". Mentionnons encore que souvant, pour donner des formules de quadrature sur un intervalle standart (par exemple $[-1, 1]$,on exécute un changement de variable de la forme :

$$t = 2\frac{x-x_i}{x_{i+1}-x_i} - 1 \qquad (5.2.4)$$

qui, à $x \in [x_i, x_{i+1}]$, fait correspendre $t \in [-1, 1]$ Avec ce changement de variable, nous obtenons :

$$x = x_i + (x_{i+1} - x_i)\frac{t+1}{2} \qquad (5.2.5)$$

et par la suite

$$\int_{x_i}^{x_{i+1}} f(x)dx = (x_{i+1} - x_i)\frac{1}{2} \int_{-1}^{1} g_i(t)dt \qquad (5.2.6)$$

où

$$g_i(t) = f(x_i + (x_{i+1} - x_i)\frac{t+1}{2}), t \in [-1, 1] \qquad (5.2.7)$$

Nous somme maintenant en mesure de définir la notion de la formule de quadrature pour approcher numériquement $\int_{-1}^{1} g_i(t)dt$, g_i étant une fonction continue sur $[-1, 1]$.

Définition 5.1 : Si g_i est une fonction sur $[-1,1]$, la formule de quadrature pour approcher numériquement :

$$J(g_i) \underset{déf}{=} \Sigma_{j=1}^{M} w_j g_i(t_j) \qquad (5.2.8)$$

est définie par la donnée de M points $-1 \leq t_1 < ..... < t_M \leq 1$ appelées points d'intégration et M nombres réels $w_1, ..., w_M$ appelés poids de la formule de quadrature. Ces M points et M poids devront être cherchés de façon à ce que $J(g_i)$ soit une approximation numérique de $\int_{-1}^{1} g_i(t)dt$.

Nous remarquons que la formule (5.2.8) est linéaire. En effet, si g_i et l_i sont deux fonctions continues données sur l'intervalle $[-1,1]$ et si α et $\beta \in IR$, nous vérifions facilement que :

$$J(g_i + l_i) = \alpha J(g_i) + \beta J(l_i).$$

Exemple 5.1 : Un exemple classique est la formule à 2 points $(M = 2)$:

$$t_1 = -1, t_2 = 1, w_1 = 1, w_2 = 1$$

et donc

$$J(g_i) = g_i(-1) + g_i(1) \qquad (5.2.9).$$

Nous remarquons que $J(g_i)$ correspond à l'aire du trapèze hachuré de la figure (5.2).Par conséquent, approcher $\int_{-1}^{1} g_i(t)dt$ par $J(g_i)$ correspond à approcher l'aire sous le graphe de g_i par l'aire du trapèze hachuré. Pour cette raison, la formule de quadrature (5.2.9) est appelée "formule du trapèze".

fig.

Dans les sections suivantes, nous construisons d'autres formules de quadrature que la formule du "trapèze".

Dans, l'égalité (3.7) nous approchons $\int_{-1}^{1} g_i(t)dt$ par $J(g_i)$. Ainsi la quantité $\int_{x_i}^{x_{i+1}} f(x)dx$ est approchée par :

$$\frac{(x_{i+1}-x_i)}{2} \sum_{j=1}^{M} w_j f(x_i + (x_{i+1} - x_i)\frac{t+1}{2}) \qquad (3.11)$$

et donc nous allons approcher $\int_a^b f(x)dx$ par la formule dite "formule composite" :

$$L_h(f) = \sum_{i=1}^{N-1} \frac{(x_{i+1}-x_i)}{2} \sum_{j=1}^{M} w_j f(x_i + (x_{i+1} - x_i)\frac{t+1}{2}) \qquad (3.12)$$

Exemple 3.2 : $t_1 = -1, t_2 = 1, w_1 = 1, w_2 = 1$. La formule composite (3.12) s'écrit :

$$L_h(f) = \sum_{i=1}^{N-1} \frac{(x_{i+1}-x_i)}{2} [f(x_i) + f(x_{i+1})] \qquad (3.13).$$

La formule composite (3.1) est facile à interpréter graphiquement, la quantité $L_h(f)$ correspond à l'aire hacuée de la fig. 3.2

Fig 3.2 Formule du trapèze.pour approcher $\int_a^b f(x)dx$ dans le cas $N = 4$.

En régle générale nous pouvons procéder de la manière suivante pour approcher la quantité $\int_a^b f(x)dx$ par la quantité $L_h(f)$; on définit une formule de quadrature par la donnée de M points $t_1 < ..... < t_M$ et M poids $w_1, ..., w_M$ (ces points et ces poids sont reprertoriés dans des tables numériques ou logiciels de calcules) ; on partitionne l'intervalle $[a, b]$ en intervalles $[x_i, x_{i+1}]$ les x_i satisfaisant (3.2) et on calcule $L_h(f)$ par la formule composite (3.12).

Avant de montrer comment construire des formules de quadrature, définissons une propriété de $J(g_i)$.

Définition 3.2 : On dira que la formule de quadrature :

$$J(g_i) \underset{déf}{=} \Sigma_{j=1}^{M} w_j g_i(t_j).$$

Pour calculer numériquement $\int_{-1}^{1} g_i(t)dt$ est exacte pour tout polynôme de $\deg r \geq 0$ si $J(p) = \int_{-1}^{1} p(t)dt$, pour tout polynôme de $\deg \leq r$.

Lorsque la formule de quadrature.$J(g_i)$ satisfait la propriété de la définition (3.2), il est possible d'estimer l'erreur entre la valeur exacte $\int_a^b f(x)dx$ et la valeur approchée $L_h(f)$, pour autant que f soit assez régulière.

Théorème 3.1 : Supposons que le formule de quadrature :

$$J(g_i) \underset{déf}{=} \Sigma_{j=1}^{M} w_j g_i(t_j)$$

pour calculer numériquement $\int_{-1}^{1} g_i(t)dt$ soit exacte pour des polynômes $\deg = r$. Soit f une fonction donnée sur $[a,b]$, soit $L_h(f)$ la formule composite définie par (3.12) et soit h la quantité définie par (3.3). Alors si f est assez régulière (i.e $(r+1)$ fois continûment dérivable sur $[a,b]$, il existe une constante C indépendante du choix des points x_i telle que :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^{r+1} \qquad (3.14)$$

Exemple 3.3 : Considérons l'exemple de la formule du trapèze (3.10) ainsi que la formule composite $L_h(f)$ (3.13) qui en découle. Clairement si p est un polynôme de $\deg = r = 1$,c'est à dire $p(t) = \alpha + \beta t$ α et $\beta \in IR$.Il est facile de vérifier que lorsque la formule de quadrature définie par (3.10), alors

$$J(p) = \int_{-1}^{1} p(t)dt$$

Ainsi la formule du trapèze (3.10) pour calculer numériquement $\int_{-1}^{1} g_i(t)dt$ est exacte pour tout polynôme de $\deg = r = 1$.

Si l'intervalle $[a,b]$ est divisé en en N parties égales i.e $h = \frac{b-a}{N}$, $x_i = a + ih, i = 0, 1, ..., N$ et si f est C^2 $[a,b]$, alors le théorème (3.1) fournit l'estimation d'erreur suivante :

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^2 \qquad (3.15)$$

où C ne dépend ni de N ni de h. L'estimation (3.15) indique qu'en principe, lorsqu'on utilise la formule (3.13) pour approcher numériquement $\int_a^b f(x)dx$, l'erreur est divisée par 4 chaque fois que N est multiplié par 2!

En fait, l'inégalité (3.14) montre que , lorsque la partition est finie (h petit), l'erreur obtenue en approchant $\int_a^b f(x)dx$ par $L_h(f)$ est petite. Cette erreur devient d'autant plus petite avec h et que r est grand.

Il est donc légitime de chercher des points d'intégration t_j et $w_j, j = 1, ..., M$; de sorte que la formule de quadrature $J(.)$ soit exacte pour des polynômes de $\deg = r$ aussi élevé que possible.

5.3 Poids d'une formule de quadrature.

Dans cette section, nous supposons donnés M points d'integration distincts dans $[-1, 1]$

$$-1 < t_1 < ..... < t_M < +1$$

et nous cherchons à déterminer les poids $w_j, j = 1, ..., M$, de sorte que la formule de quadrature $J(g) \underset{déf}{=} \Sigma_{j=1}^{M} w_j g(t_j)$ soit exacte pour des polynômes de $\deg = r$ aussi élevé que possible.

Pour réaliser cet objectif, considérons la base de Lagrange $\varphi_1, \varphi_2, ..., \varphi_M$ de IP_{M-1}associée aux points $t_1, ..., t_M$.

Par définition, φ_k est le polynôme de $\deg = M-1$ défini par :

$$\varphi_k(t) = \frac{(t-t_0)(t-t_1).....(t-t_{k-1})(t-t_{k+1}).....(t-t_n)}{(t_k-t_1)(t_k-t_2).....(t_k-t_{k-1})(t_k-t_{k+1}).....(t_k-t_n)} \qquad j=1,...,M \qquad (3.16)$$

Soit $g : [-1,1] \longrightarrow IR$ une fonction continue donnée. Son interpolant $\widetilde{g}$ de $\deg = M-1$ aux points $t_1, ..., t_M$ est défini par :

$$\widetilde{g}(t) = \Sigma_{j=1}^{M} g(t_j)\varphi_j(t)$$

Il semble naturel de remplacer $\int_{-1}^{1} g(t)dt$ par $\int_{-1}^{1} \widetilde{g}(t)dt$,puisque

$$\int_{-1}^{1} \widetilde{g}(t)dt = \Sigma_{j=1}^{M} g(t_j) \int_{-1}^{1} \varphi_j(t)dt,$$

nous constatons qu'il suffit de poser

$$w_j = \int_{-1}^{1} \varphi_j(t)dt$$

pour que $J(g) \underset{déf}{=} \Sigma_{j=1}^{M} w_j g(t_j)$ soit une approximation de $\int_{-1}^{1} g(t)dt$.

Théorème 3.2 : Soit $t_1 < ..... < t_M$ M points distincts de $[-1,1]$ et soit $(\varphi_1, \varphi_2, ..., \varphi_M)$ la base de Lagrange de IP_{M-1} assoviée à ces M points. Alors la formule de quadrature :

$$J(g) = \Sigma_{j=1}^{M} w_j g(t_j)$$

est exate pour les polynômes de $\deg = M-1$ si et seulement si

$$w_j = \int_{-1}^{1} \varphi_j(t)dt, \; j = 1, ..., M \qquad (3.17)$$

Preuve : i) Montrons que si la formule de quadrature $J(.)$ est exate pour les polynômes de $\deg = M-1$, alors on a les relations (3.17). Puisque

$$J(g) = \Sigma_{j=1}^{M} w_j p(t_j) = \int_{-1}^{1} p(t)dt$$

pour tout polynôme $p \in IP_{M-1}$, nous pouvons choisir $p = \varphi_k, k = 1, ..., M$ et nous obtenons

$$J(\varphi_k) = \sum_{j=1}^{M} w_j \varphi_k(t_j) = \int_{-1}^{1} \varphi_k(t)dt$$

puisque $\varphi_k(t_j) = \delta_{kj}$, nous avons bien,

$$w_k = \int_{-1}^{1} \varphi_k(t)dt$$

ii) Montrons maintenant que si les relations (3.17) sont satisfaites, alors la formule de quadrature $J(.)$ est exate pour les polynômes de $\deg = M-1$.

Soit $p \in IP_{M-1}$ que nous développons dans la base de Lagrange de IP_{M-1} associé aux points $t_1, ..., t_M, i.e$

$$p(t) = \sum_{j=1}^{M} p(t_j)\varphi_j(t)$$

Ainsi donc

$$\int_{-1}^{1} p(t)dt = \sum_{j=1}^{M} p(t_j) \int_{-1}^{1} \varphi_j(t)dt = \sum_{j=1}^{M} p(t_j)w_j = J(p).$$

Remarque 3.1 : Les relations (3.17) nous permettent donc de calculer les poids $w_k, k = 1, ..., M$, d'une formule de quadrature,étant donné les points d'intégration $t_1, ..., t_M$.De plus, $\sum_{j=1}^{M} \varphi_k(t_j)$ est le polynôme de $\deg = M-1$ qui vaut 1 aux points $t_1, ..., t_M$, et est donc la fonction identique à 1.

Par conséquent, nous obtenons, en utilisons (3.17)

$$\sum_{j=1}^{M} w_j = \int_{-1}^{1} (\sum_{j=1}^{M} \varphi_k(t_j)dt = \int_{-1}^{1} dt = 2.$$

Ce qui prouve que la somme des poids caclculés par (3.17) est toujours égale à 2.

Exemple 3.4 : $M = 2, t_1 = -1, t_2 = 1$ (formule du trapèzes) et explicitons la base de Lagrange φ_1, φ_2 associée aux points t_1, t_2:

$\varphi_1(t) = 0.5(1-t)$ *et* $\varphi_1(t) = 0.5(1+t)$

La relation (3.17) s'écrit :

$w_1 = \int_{-1}^{1} \varphi_1(t)dt = 1$ et $w_2 = \int_{-1}^{1} \varphi_2(t)dt = 1.$

Le théorème 3.2 nous assure que les formules de quadratures construites grâce à (3.17) sont exactes pour les polynômes de deg $= r$, avec r plus grand que $M-1$.

Dans la suite nous verrons qu'il se peut que ces formules de quadratures soient exactes pour les polynômes de deg $= r$, avec r plus grand que $M-1$.

3.3 Formule du rectangle : La formule du rectangle est une formule à un seul point $(M=1)$: $t_1 = 0$

La base de Lagrange de IP_0 associée à t_1 est donnée par

$\varphi_1(t) = 1 \qquad \forall t \in [-1,1]$

Ainsi (3.17) nous donne

$$w_1 = \int_{-1}^{1} \varphi_1(t)dt = 2$$

et la formule du rectangle devient

$$J(g) = 2g(0) \qquad (3.18)$$

On interprète la formule du rectangle (3.18) de la façon suivante :

Elle consiste à remplacer $\int_{-1}^{1} g(t)dt$ par l'aire du rectangle de base $[-1,1]$ et de hauteur $g(0)$ $(fig3.3)$, d'où son nom. Selon le théorème 3.2, cette formule de quadrature est exacte pour tout polynôme de degré 0, mais en fait elle est meillleure, elle est exacte pour tout polynôme $p \in IP_1$ défini par $p(t) = \alpha t + \beta, où\ \alpha, \beta \in IR$.

Il est alors facile de vérifier que $\int_{-1}^{1} p(t)dt = 2\beta = 2p(0)$.

Si nous utilisons la formule du rectangle dans la formule composite (3.12), nous obtenons

$$L_h(f) = \Sigma_{i=1}^{N-1} (x_{i+1} - x_i) f(\frac{x_{i+1}+x_i}{2}) \qquad (3.19)$$

et l'estimation (3.14) du théorème (3.1)

$$\left| \int_a^b f(x)dx - L_h(f) \right| \leq Ch^2 \qquad (3.20)$$

L'interprétation géométrique de (3.19) est la suivante :

On somme les aires des rectangles dont la base est le segment $[x_i, x_{i+1}]$ et dont la haureur est $f(\zeta_i)$, où ζ_i est le milieu de $[x_i, x_{i+1}]$.

fig 3.3 Formule du rectangle sur $[-1, 1]$

3.4 Formule de Simpson : La formule de Simpson est une formule à trois points : $M = 3, t_1 = -1, t_2 = 0, t_3 = 1$.

La base de Lagrange $\varphi_1, \varphi_2, \varphi_3$ de IP_2 associée aux 3 points t_1, t_2, t_3 s'écrit

$$\varphi_1(t) = 0.5(t^2 - t), \varphi_2(t) = (1 - t^2), \varphi_3(t) = 0.5(t^2 + t)$$

Les relations (3.17) deviennent alors :

$$w_1 = \int_{-1}^{1} \varphi_1(t)dt = \frac{1}{3}, w_2 = \int_{-1}^{1} \varphi_2(t)dt = \frac{4}{3}, w_3 = \int_{-1}^{1} \varphi_3(t)dt = \frac{1}{3}$$

La formule de Simpson s'écrit donc :

$$J(g) = \frac{1}{3}g(-1) + \frac{4}{3}g(0) + \frac{1}{3}g(1) \qquad (3.21)$$

Elle est une moyenne pondérée entre la formule du trapèze (poids $\frac{1}{3}$) et la formule du rectangle $(poids\ \frac{2}{3})$.Si nous utilisons cette formule de quadrature dans (3.12), nous obtenons :

$$L_h(f) = \Sigma_{i=1}^{N-1} \frac{(x_{i+1}-x_i)}{6} \left[f(x_i) + 4f(\frac{x_{i+1}+x_i}{2}) + f(x_{i+1}) \right] \qquad (3.22)$$

D'après le théorème 3.2, la formule de Simpson est exacte pour tout polynôme de $\deg = r = 2$. En fait, elle est exacte pour tout polynôme de $\deg = r = 3$.

En effet, si $g(t) = t^3$, alors $J(g) = 0$ et $\int_{-1}^{1} g(t)dt = 0$. L'estimation (3.14) du théorème 3.2 devient donc :

$$\left|\int_a^b f(t)dt - L_h(f)\right| \leq Ch^4.$$

La formule de Simpson donne une erreur d'ordre h^4. Cette formulle est souvent utilisée dans la pratique car $L_h(f)$ converge rapidement vers $\int_a^b f(t)dt$ lorsque $h \to 0$.