Université Mohammed Premier
Ecole Supérieure de Technologie
Département d'Informatique

Calcul Scientifique

Cours et exercices

Filière: *Génie Civil*
Semestre : S_2
Année : 2020/2021

Prof : Tahrichi Mohamed

Table des matières

Chapitre 1

Résolution des systèmes linéaires par des méthodes directes

I Introduction

On considère une matrice inversible A et on fixe un vecteur $b \in \mathbb{R}^n$; ("le second membre"). L'objectif est d'étudier des méthodes de résolution du système linéaire :

$$Au = b \tag{1.1}$$

dont l'inconnue est un vecteur $u = \begin{pmatrix} x_1 \\ \vdots \\ x_2 \end{pmatrix} \in \mathbb{R}^n$.

On sait que sous l'hypothèse d'inversibilité de la matrice A, un tel système de n équations à n inconnues a une solution unique. L'objectif est d'étudier des méthodes de calcul de cette solution qui soient praticables pour des systèmes de grande taille (au moins $n \geq 100$). On peut se rappeler qu'il existe une expression très élégante de la solution avec des quotients de déterminants :

$$x_i = \frac{det(A_1;\ldots;A_{i-1};b;A_{i+1};\ldots;A_n)}{det(A)}.$$

Faisons remarquer que les formules de Cramer ci-dessus sont très peu utilisées (voire ne sont pas utilisées) dans la pratique car leur coût opératoire est de l'ordre de $(n+1)!$ opérations. Ainsi, en supposant qu'on dispose d'un ordinateur capable d'effectuer 10^9 opérations par seconde, alors il nous faudrait 9.61047 années pour résoudre un système linéaire ayant 50 inconnues comme le montre le tableau ci-dessous !

Taille n	nombre d'opéarations	Temps
10	$11! \simeq 3.99107$	0,04 seconde
20	$21! \simeq 5.11019$	1620 années
50	$51! \simeq 1.551066$	9,61047 années

Ajoutons à cela que même pour des systèmes linéaires de tailles "modestes", le calcul, sur des ordinateurs, de la solution par la méthode de Cramer est entaché par la propagation des erreurs d'arrondi et de troncature.
Pour toutes ses raisons, de nombreuses méthodes dites méthodes directes qui permettent de résoudre (1.1) ont été développées. Ces méthodes fournissent la solution du système en un nombre fini d'étapes. et elles sont basées sur une décomposition de la matrice A sous forme d'un produit de matrices plus "simples" à manipuler : $A = LU$, $A = QR$, A $= LDL^T$, $A = SS^T$, ...

Remarque

Un système triangulaire (T) du type ci-dessous est aisée à résoudre, par la méthode dite de substitutions successives ou de remontée.

$$(T) = \left\{ \begin{array}{lllll} a_{1,1}x_1 & +\ldots & \ldots & +a_{1,n}x_n & = b_1 \\ & \ldots & \ldots & \ldots & \ldots \\ & & \ldots & & \ldots \\ & a_{i,i}x_i & \ldots & +a_{i,n}x_n & = b_i \\ & & \ldots & & \ldots \\ & & & & \ldots \\ & & & a_{n,n}x_n & = b_n \end{array} \right.$$

Cette méthode consiste à déterminer successivement :

$$x_n = \frac{b_n}{a_{n,n}}$$

puis chaque x_i pour i variant de $n-1$ à 1, par ordre décroissant des indices, en fonction des termes déjà calculés selon la formule :

$$x_i = \frac{1}{a_{i,i}}(b_i - a_{i,i+1}x_{i+1} - \cdots - a_{i,n}x_n)$$

Remarquer que le fait que la matrice soit inversible équivaut ici à : $\forall i;\ a_{i,i} \neq 0$. Il n'est pas difficile de voir que le nombre d'opérations nécessaires pour résoudre un tel système est de l'ordre de n^2. Le coût est faible par rapport au coût de l'algorithme à mettre en oeuvre pour réduire un système quelconque à cette forme, algorithme du pivot de Gauss qui fait l'objet de la prochaine section.

II Méthode d'élimination de Gauss

L'idée de cette méthode est de se ramener à un système linéaire dont la matrice est triangulaire supérieure, on obtient ensuite la solution par simple remontée.

1 Cas d'une matrice 4×4

On cherche à résoudre le système suivant :

$$\left\{ \begin{array}{llllllll} 2x & + & y & + & z & & = & -3 \\ 4x & + & 3y & + & 3z & + & t & = & -5 \\ 8x & + & 7y & + & 9z & + & 5t & = & -7 \\ 6x & + & 7y & + & 9z & + & 8t & = & 1 \end{array} \right.$$

La méthode de Gauss, consiste à éliminer x des lignes 2, 3, 4, puis y des lignes 3, 4, puis z de la ligne 4. On obtient alors la valeur de t et on en déduit les autres valeurs en remontant.
Dans la suite on note par L_i, $2 \leq i \leq 4$ la ligne numéro i du système et par $a_{i,j}$ l'élément d'indice (i,j) dans la matrice associée.

Étape 1 :

On effectue alors pour toute ligne L_i, $2 \leq i \leq 4$:

$$L_i \leftarrow L_i - \frac{a_{i,1}}{a_{1,1}}L_1,$$

on obtient ainsi le système équivalent suivant :

$$\left\{ \begin{array}{llllllll} 2x & + & y & + & z & & = & -3 \\ & & y & + & z & + & t & = & 1 \\ & & 3y & + & 5z & + & 5t & = & 5 \\ & & 4y & + & 6z & + & 8t & = & 10 \end{array} \right.$$

Étape 2 :
On effectue pour toute ligne L_i, $3 \leq i \leq 4$:

$$L_i \leftarrow L_i - \frac{a_{i,2}}{a_{2,2}} L_2,$$

on obtient ainsi le système équivalent suivant :

$$\begin{cases} 2x + y + z = -3 \\ \quad y + z + t = 1 \\ \quad\quad 2z + 2t = 2 \\ \quad\quad 2z + 4t = 4 \end{cases}$$

Étape 3 :
Finalement on effectue

$$L_4 \leftarrow L_4 - \frac{a_{i,3}}{a_{3,3}} L_3,$$

on obtient ainsi le système équivalent suivant :

$$\begin{cases} 2x + y + z = -3 \\ \quad y + z + t = 1 \\ \quad\quad 2z + 2t = 2 \\ \quad\quad\quad 2t = 4 \end{cases}$$

Le système est donc triangulaire, en utilisant la méthode de remontée, on aboutit à la solution suivante :

$$\begin{cases} y = -1 \\ z = 0 \\ z = -1 \\ t = 2 \end{cases}$$

2 Cas général

Le système initial est $Ax = b$ avec

$$A^{(1)} = A = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & a_{1,j}^{(1)} & \dots & a_{1,n}^{(1)} \\ a_{2,1}^{(1)} & a_{2,2}^{(1)} & \dots & a_{2,j}^{(1)} & \dots & a_{2,n}^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{i,1}^{(1)} & a_{i,2}^{(1)} & \dots & a_{i,j}^{(1)} & \dots & a_{i,n}^{(1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n,1}^{(1)} & a_{n,2}^{(1)} & \dots & a_{n,j}^{(1)} & \dots & a_{n,n}^{(1)} \end{pmatrix}, \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_i \\ \vdots \\ x_n \end{pmatrix} \quad et \quad b^{(1)} = b = \begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} \\ \vdots \\ b_i^{(1)} \\ \vdots \\ b_n^{(1)} \end{pmatrix}$$

Étape 1 : A fin d'éliminer la variable x_1 des $(n-1)$ dernières lignes, on effectue

$$\begin{cases} a_{i,j}^{(2)} = a_{i,j}^{(1)} - \dfrac{a_{i,1}^{(1)}}{a_{1,1}^{(1)}} a_{1,j}^{(1)}, & i,j = 2,3,\dots,n, \\ b_i^{(2)} = b_i^{(1)} - \dfrac{a_{i,1}^{(1)}}{a_{1,1}^{(1)}} b_1^{(1)}, & i = 2,3,\dots,n, \end{cases}$$

à condition que $a_{1,1}$ soit non nul, ce que l'on supposera pour le moment.
Lorsqu'on a fait l'élimination jusqu'à la dernière ligne on obtient un système linéaire sous la forme $A^{(2)}x = b^{(2)}$ avec

$$A^{(2)} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & a_{1,j}^{(1)} & \dots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & \dots & a_{2,j}^{(2)} & \dots & a_{2,n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{i,2}^{(2)} & \dots & a_{i,j}^{(2)} & \dots & a_{i,n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & a_{n,2}^{(2)} & \dots & a_{n,j}^{(2)} & \dots & a_{n,n}^{(2)} \end{pmatrix} \quad et \quad b^{(2)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ \vdots \\ b_i^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix}$$

On dit qu'on vient de réaliser la première étape de l'élimination de Gauss, ce qui a consisté à faire apparaître des zéros sur la première colonne en dessous de la diagonale.
Étape 2 : La deuxième étape consiste à éliminer la deuxième inconnue x_2 des lignes $3,4,\dots,n$. Pour cela on commence par la troisième ligne, en supposant maintenant que $a_{2,2}^{(2)}$ est non nul. Ce qui conduit aux formules :

$$\begin{cases} a_{i,j}^{(3)} = a_{i,j}^{(2)} - \dfrac{a_{i,2}^{(2)}}{a_{2,2}^{(2)}} a_{2,j}^{(2)}, & i,j = 3,4,\dots,n, \\ b_i^{(3)} = b_i^{(2)} - \dfrac{a_{i,2}^{(2)}}{a_{2,2}^{(2)}} b_2^{(2)}, & i = 3,4,\dots,n, \end{cases}$$

Ce calcul étant fait jusqu'à la ligne n, on obtient un système sous la forme $A^{(3)}x = b^{(3)}$ avec

$$A^{(3)} = \begin{pmatrix} a_{1,1}^{(1)} & a_{1,2}^{(1)} & \dots & a_{1,j}^{(1)} & \dots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & \dots & a_{2,j}^{(2)} & \dots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & \dots & \dots & a_{3,n}^{(3)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n,j}^{(3)} & \dots & \dots & a_{n,n}^{(3)} \end{pmatrix} \quad et \quad b^{(3)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_n^{(3)} \end{pmatrix}$$

Étape k : On peut alors définir l'algorithme général en supposant qu'on a effectué $k-1$ étapes de l'élimination de Gauss, et on a un système qui s'écrit $A^{(k)}x = b^{(k)}$, avec $A^{(k)}$ et $b^{(k)}$ donnés par

$$A^{(k)} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} & \dots & a_{1,j} & \dots & a_{1,n} \\ 0 & a_{2,2}^{(2)} & a_{2,3}^{(2)} & \dots & a_{2,j}^{(2)} & \dots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & \dots & a_{3,k}^{(3)} & \dots & a_{3,n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{k,k}^{(k)} & \dots & a_{k,n}^{(k)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & a_{n,k}^{(k)} & \dots & a_{n,n}^{(k)} \end{pmatrix} \quad et \quad b^{(k)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(k)} \end{pmatrix}$$

L'étape k consiste à faire l'élimination de la variable x_k dans les lignes $k+1,k+2,\dots,n$. Ceci conduit aux formules suivantes, définies pour $i = k+1,k+2,\dots,n$,

$$\begin{cases} a_{i,j}^{(k+1)} & = a_{i,j}^{(k)} - \dfrac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} a_{k,j}^{(k)}, & i,j = k+1,k+2,\dots,n \\ b_i^{(k+1)} & = b_i^{(k)} - \dfrac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}} b_k^{(k)}, & i = k+1,k+2,\dots,n \end{cases}$$

Conclusion : Au bout de $(n-1)$ étapes, on obtient le système triangulaire suivant $A^{(n)}x = b^{(n)}$ avec

$$A^{(n)} = \begin{pmatrix} a_{1,1} & a_{1,2}^{(2)} & a_{1,3} & \dots & \dots & a_{1,n} \\ 0 & a_{2,2}^{(2)} & a_{2,3} & \dots & \dots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & a_{3,k}^{(3)} & \dots & a_{3,n}^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & a_{n-1,n-1}^{(n-1)} & a_{n-1,n}^{(n-1)} \\ 0 & 0 & 0 & \dots & 0 & a_{n,n}^{(n)} \end{pmatrix} \quad et \quad b^{(n)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_{n-1}^{(n-1)} \\ b_n^{(n)} \end{pmatrix}$$

et en utilisant la méthode de remontée, on obtient la solution de système de départ.

Définition On appelle pivots les nombres $a_{k,k}^{(k)}$.

En supposant que tous les pivots sont non nuls, l'algorithme d'élimination de Gauss se résume comme suit.

Algorithme : Méthode de Gauss sans pivotage

———————————Phase de triangularisation———————————

pour $k = 1$ *jusqu'à* $n-1$ **faire**
  pour $i = k+1$ *jusqu'à* n **faire**
    pour $j = k+1$ *jusqu'à* n **faire**
      $a_{i,j} \leftarrow a_{i,j} - \dfrac{a_{i,k}}{a_{k,k}} a_{k,j}$;
    fin pour
    $b_i \leftarrow b_i - \dfrac{a_{i,k}}{a_{k,k}} b_k$;
    $a_{i,k} \leftarrow 0$;
  fin pour
fin pour

———————————Phase de résolution du système triangulaire supérieur———————————

$x_n \leftarrow \dfrac{b_n}{a_{n,n}}$;
pour $i = n-1$ *jusqu'à* 1 **faire**
  $x_i \leftarrow \dfrac{1}{a_{i,i}}(b_i - \sum_{j=i+1}^{n} a_{i,j} x_j)$;
fin pour

3 Pivot nul : permutation de lignes

Tout ce que nous avons fait jusqu'ici suppose qu'à aucun moment on ne rencontre de pivot nul et ceci n'est pas toujours vrai même si la matrice est inversible, voit l'exemple ci-dessous.

Dans ce cas, il n'est pas possible d'appliquer la méthode d'élimination de Gauss comme vu précédemment, et comme l'ordre des équations composant un système d'équations linéaires ne joue aucun rôle, nous pouvons permuter la ligne de pivot nul avec l'une des lignes suivantes, il faut bien sûr permuter les seconds membres en parallèle. On obtiendra un nouveau pivot non nul et on pourra poursuivre. Pour le choix de la ligne avec la quelle on permutera, nous pouvons adopter l'une des stratégies suivantes :

Stratégie de pivotage partiel. Cette stratégie, appelée encore stratégie du pivot partiel, consiste à chercher dans la sous colonne k, le plus grand pivot en valeur absolue, i.e., on cherche l'indice de la ligne l tel que $a_{l,k} = \max_{k \le i \le n} | a_{i,k} |$. Ensuite, on permute les lignes l et k.

Stratégie de pivotage total. Dans cette stratégie, on cherche le maximum en valeur absolue dans toute la sous matrice $(a_{i,j})_{i,j=k,\ldots,n}$, i.e., on cherche les indices de ligne l et de colonne m vérifiant $a_{l,m} = \max_{k \le i,j \le n} | a_{i,j} |$. Ensuite, on permute les lignes l et k ainsi que les colonnes m et l. A noter que cette stratégie est appelée encore stratégie du pivot total.

Exemple Soit à résoudre le système $Ax = b$, où $b \in \mathbb{R}$ est quelconque et $A \in \mathbb{R}^3 \times \mathbb{R}^3$ la matrice

$$A = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 4 & 5 \\ 7 & 8 & 9 \end{pmatrix}$$

Après application de la première étape de la méthode de Gauss, on obtient

$$A^{(2)} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{pmatrix}$$

On remarque qu'on ne peut pas continuer, pourtant la matrice A est inversible ! Par contre il suffit de permuter les lignes 2 et 3 de la matrice A (ne pas oublier de faire la même chose avec le second membre) et obtenir

$$\widehat{A^{(2)}} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}$$

qui est déjà une matrice triangulaire ce qui termine bien l'élimination de Gauss.

III Méthode de décomposition LU

1 Définition et motivation

La méthode de décomposition LU consiste à factoriser la matrice A en un produit de deux matrices triangulaires

$$A = LU$$

où L est triangulaire inférieure (lower) et U est triangulaire supérieure (upper).
Si on dispose d'une telle décomposition de la matrice A alors le système $Ax = b$ s'écrit sous la forme

$$LUx = b.$$

La solution de ce système s'obtient alors en résolvant successivement les deux systèmes triangulaires suivants :

$$\begin{cases} Ly &= b \\ Ux &= y \end{cases}$$

Cette méthode est intéressante dans le cas où on a à résoudre plusieurs équations $Ax = b_i$ car la décomposition $A = LU$ est effectuée une fois pour toute. C'est en fait l'algorithme de l'élimination de Gauss dans le cas où on ne permute jamais. On appelle sous matrice diagonale d'ordre k de la matrice $A \in \mathcal{M}_n(\mathbb{R})$, la matrice définie par :

$$\Delta^k = \begin{pmatrix} a_{1,1} & \cdots & \cdots & a_{1,k} \\ \vdots & \ddots & \vdots & \\ a_{k,1} & \cdots & \cdots & a_{k,k} \end{pmatrix}$$

Le résultat suivant donne une condition suffisante sur la matrice A pour qu'il n'y ait pas de permutation au cours de la méthode d'élimination de Gauss.

Théorème Soit A une matrice dont toutes les sous-matrices diagonales sont inversibles. Il existe un unique couple de matrices (L, U), avec U triangulaire supérieure, et L triangulaire inférieure à diagonale unité tel que

$$A = LU$$

2 Calcul direct de la décompostion LU

Rappelons la présentation classique de l'algorithme de Doolittle. On oublie l'algorithme d'élimination de Gauss, pour chercher directement une décomposition LU de A. Pour assurer l'unicité de la décomposition, nous demandons que la diagonale de L soit unitaire ($l_{i,i} = 1,\ i = 1, \ldots, n$).
Traitons un exemple, soit la matrice

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

On cherche deux matrices

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \times & 1 & 0 \\ \times & \times & 1 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{pmatrix}$$

telle que $A = LU$.

- On identifie la première ligne de A et la première ligne de LU, cela permet d'obtenir la première ligne de U :

$$LU = \begin{pmatrix} 1 & 0 & 0 \\ \times & 1 & 0 \\ \times & \times & 1 \end{pmatrix} \begin{pmatrix} \mathbf{2} & \mathbf{1} & \mathbf{-2} \\ 0 & \times & \times \\ 0 & 0 & \times \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 \\ 4 & 5 & -3 \\ -2 & 5 & 3 \end{pmatrix}$$

- On identifie la première colonne de A avec la première colonne de LU, cela permet d'obtenir la première colonne de L :

$$LU = \begin{pmatrix} \mathbf{1} & 0 & 0 \\ \mathbf{2} & 1 & 0 \\ \mathbf{-1} & \times & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 \\ 0 & \times & \times \\ 0 & 0 & \times \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 \\ \mathbf{4} & 5 & -3 \\ \mathbf{-2} & 5 & 3 \end{pmatrix}$$

- On identifie la deuxième ligne de A avec la deuxième ligne de LU, cela permet d'obtenir la deuxième ligne de U :

$$LU = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & \times & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 \\ 0 & \mathbf{3} & \mathbf{1} \\ 0 & 0 & \times \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 \\ 4 & \mathbf{5} & \mathbf{-3} \\ -2 & 5 & 3 \end{pmatrix}$$

- On identifie la deuxième colonne de A avec la deuxième colonne de LU, cela permet d'obtenir la deuxième colonne de L :

$$LU = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & \mathbf{2} & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 \\ 0 & 3 & 1 \\ 0 & 0 & \times \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 \\ 4 & \mathbf{5} & -3 \\ -2 & \mathbf{5} & 3 \end{pmatrix}$$

- On identifie la troisième ligne de A avec la troisième ligne de LU, cela permet d'obtenir la troisième ligne de U :

$$LU = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix} \begin{pmatrix} 2 & 1 & -2 \\ 0 & 3 & 1 \\ 0 & 0 & \mathbf{-1} \end{pmatrix} = \begin{pmatrix} 2 & 1 & -2 \\ 4 & 5 & -3 \\ -2 & 5 & \mathbf{3} \end{pmatrix}$$

Une autre méthode consiste à calculer toujours explicitement une décomposition LU de la matrice A, mais cette fois-ci c'est la diagonale de U qui est formée de 1. Cette méthode conduit à l'algorithme de **Crout**.

3 Algorithme de Doolittle

En écrivant $A = LU$ et en se souvenant que les matrice L et U sont triangulaires et que les termes diagonaux de L valent 1, on obtient

$$A_i = L_i U = \sum_{k=1}^{i-1} \ell_{i,k} U_k + U_i \iff U_i = A_i - \sum_{k=1}^{i-1} \ell_{i,k} U_k$$

Nous voyons que pour calculer les éléments de la ième ligne de U, U_i, il nous faut connaître préalablement les éléments des lignes 1 à $i-1$ de U ainsi que les éléments des colonnes 1 à $i-1$ de L. On peut remarquer de plus que, pratiquement, on ne calculera que les termes $u_{i,j}$ pour j compris entre i et n, puisque les autres termes de la ligne sont connus car nuls. De manière similaire on a :

$$\widehat{A_i} = L\widehat{U_i} = \sum_{k=1}^{i-1} u_{k,i} \widehat{L_k} + u_{i,i} \widehat{L_i} \iff L_i = \frac{1}{a_{i,i}} \left(\widehat{A_i} - \sum_{k=1}^{i-1} u_{k,i} \widehat{L_k} \right)$$

Nous voyons que, pour calculer les éléments de la ième colonne de L, $\widehat{L_i}$, il nous faut connaître préalablement les éléments des lignes 1 à i de U ainsi que les éléments des colonnes 1 à $i-1$ de L. On peut remarquer de plus que, pratiquement, on ne calculera que les termes $l_{j,i}$ pour j compris entre $i+1$ et n, puisque on sait déjà que $l_{i,i} = 1$ et que les autres termes de la colonne sont nuls.
Nous allons donc calculer :

- la première ligne de U, puis la première colonne de L,
- la deuxième ligne de U, puis la deuxième colonne de L, …
- et ainsi de suite.

Algorithme de Doolittle

pour i *jusqu'à* $n-1$ **faire**
  pour $j = i$ *jusqu'à* n **faire**
    $u_{i,j} \leftarrow a_{i,j} - \sum_{k=1}^{i-1} \ell_{i,k} u_{k,j}$
  fin pour
  pour $j = i+1$ *jusqu'à* n **faire**
    $\ell_{j,i} \leftarrow \frac{1}{u_{i,i}} \left(a_{j,i} - \sum_{k=1}^{i-1} \ell_{j,k} u_{k,i} \right)$
  fin pour
  $\ell_{i,i} \leftarrow 1$;
fin pour
$u_{n,n} \leftarrow a_{n,n} - \sum_{k=1}^{n-1} \ell_{n,k} u_{k,n}$

IV Décomposition de Cholesky

1 Existence

Définition Une matrice $A \in \mathbb{M}_n(\mathbb{R})$ est dite définie positive si pour tout $x \in \mathbb{R}_n$ et $x \neq 0$, on a

$$< Ax, x > > 0.$$

Si A est une matrice définie positive alors on peut lui appliquer la factorisation LU. En effet,

$$Ax = 0 \Rightarrow < Ax, x >= 0 \Rightarrow x = 0,$$

donc A est inversible, et en prenant $x_i = 0$ pour $i > k$, on trouve que Δ^k est définie positive pour tout k.

Théorème Si A est une matrice symétrique définie positive, alors elle admet une unique factorisation sous la forme

$$A = BB^T,$$

où B est une matrice triangulaire inférieure dont tous les éléments diagonaux sont positifs.

2 Algorithme de Cholesky

On cherche une décomposition de la matrice A du système, de la forme $A = BB^T$, où B est une matrice triangulaire inférieure dont les termes diagonaux sont positifs (B^T désigne sa transposée). On désigne par B_i et $\widehat{B_i}$ la $i$ème ligne et la $i$ème colonne de B respectivement. On obtient cette décomposition par identification :

$$a_{1,1} = B_1(\widehat{B^T})_1 = b_{1,1}^2,$$

puisque $b_{1,1}$ doit être positif, alors

$$b_{1,1} = \sqrt{a_{1,1}}.$$

$$\widehat{A_1} = B(\widehat{B^T})_1 = b_{1,1}\widehat{B_1} \Rightarrow \widehat{B_1} = \frac{1}{b_{1,1}}\widehat{A_1},$$

ce qui permet de déterminer la première colonne de B.
De façon similaire on définit la deuxième colonne de B :

$$a_{2,2} = b_{2,1}^2 + b_{2,2}^2.$$

puisque $b_{2,2}$ doit être positif, on a

$$b_{2,2} = \sqrt{a_{2,2} - b_{2,1}^2}.$$

$$\widehat{A_2} = B(\widehat{B^T})_2 = b_{2,1}\widehat{B_1} + b_{2,2}\widehat{B_2} \Rightarrow \widehat{B_2} = \frac{1}{b_{2,2}}(\widehat{A_2} - b_{2,1}\widehat{B_1}).$$

ce qui termine de déterminer la deuxième colonne de B. De façon générale, lorsque l'on a déterminé les $j-1$ premières colonnes de B, on écrit :

$$a_{j,j} = B_j \widehat{B}_j^T = \sum_{k=1}^{j} b_{j,k}^2 \Rightarrow b_{j,j}^2 = a_{j,j} - \sum_{k=1}^{j-1} b_{j,k}^2.$$

Puisque $b_{j,j}$ doit être positif, on a

$$b_{j,j} = \sqrt{a_{j,j} - \sum_{k=1}^{j-1} b_{j,k}^2}.$$

$$\widehat{A}_j = B(\widehat{B}^T)_j = \sum_{k=1}^{j} b_{j,k} \widehat{B}_k \Rightarrow \widehat{B}_j = \frac{1}{b_{j,j}} (\widehat{A}_j - \sum_{k=1}^{j-1} b_{j,k} \widehat{B}_k).$$

ce qui termine de déterminer la j ème colonne de B. On déterminera successivement les colonnes $1, 2, \ldots, n-1$, puis on terminera par le calcul de

$$b_{n,n} = \sqrt{a_{n,n} - \sum_{k=1}^{n-1} b_{n,k}^2}.$$

Comme dans la factorisation de Doolittle, dans chaque colonne j on ne calcule que les termes $b_{i,j}$ pour i supérieurs ou égal à j, puisque les autres termes de cette colonne sont connus car nuls. On a donc l'algorithme :

Décomposition Cholesky

pour $j = 1$ **jusqu'à** $n-1$ **faire**

$b_{j,j} \leftarrow \sqrt{a_{j,j} - \sum_{k=1}^{j-1} b_{j,k}^2}$;

pour $i = j+1$ **jusqu'à** n **faire**

$b_{i,j} \leftarrow \frac{1}{b_{j,j}} \left(a_{i,j} - \sum_{k=1}^{j-1} b_{i,k} b_{j,k} \right)$;

fin pour

fin pour

$b_{n,n} \leftarrow \sqrt{a_{n,n} - \sum_{k=1}^{n-1} b_{n,k}^2}$;

Travaux dirigés

01

1 Résoudre par la méthode de Gauss le système linéaire suivant :

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

2 La matrice associée au système précédent admet-elle une décomposition LU ?

3 Résoudre par la méthode de Gauss avec stratégie de pivotage partiel le système linéaire suivant :

$$\begin{cases} 2x_1 & +4x_2 & -4x_3 & +x_4 & = 0 \\ 3x_1 & +6x_2 & +x_3 & -2x_4 & = -7 \\ -x_1 & +x_2 & +2x_3 & +3x_4 & = 4 \\ x_1 & +x_2 & -4x_3 & +x_4 & = 2 \end{cases}$$

4 La matrice associée au système précédent admet-elle une décomposition LU ?

02

1 Réaliser la décomposition LU de la matrice :

$$\begin{pmatrix} -1 & 1 & -3 & 0 \\ 1 & 1 & 3 & 8 \\ -2 & 2 & -5 & -1 \\ 3 & 1 & 8 & 13 \end{pmatrix}$$

2 En déduire la solution du système linéaire $Ax = b$ avec $b = (0, 2, -1, 5)^t$.

3 Sans calculer A^2, résoudre le système linéaire $A^2x = b$.

03

On pose

$$A = \begin{pmatrix} 2 & 1 & 3 \\ -4 & 1 & -4 \\ -2 & 2 & 4 \end{pmatrix}.$$

1 Donner la décomposition LU de la matrice A (i.e. $A = LU$ avec L triangulaire inférieure et U triangulaire supérieure).

2 a. Expliquer comment on peut calculer le déterminant d'une matrice A d'ordre n à partir de sa factorisation LU.

b. Appliquer cette méthode à la matrice A.

3 En déduire la solution du système linéaire $Ax = b$ où $b = (0, 1, 2)^t$.

4 Soit $B = U^tAL^t$. Sans calculs supplémentaires, donner une décomposition LU de la matrice B.

Corrigés des exercices

Exercice 01

1 Elimination de Gauss

$$\left\{\begin{array}{rrrrl} x_1 & -3x_2 & -x_3 & & = 2 \\ -x_1 & +x_2 & & +2x_4 & = 3 \\ & x_2 & -x_3 & & = 1 \\ 2x_1 & +x_2 & & -x_4 & = 0 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} x_1 & -3x_2 & -x_3 & & = 2 \\ & -2x_2 & -x_3 & +2x_4 & = 5 \\ & x_2 & -x_3 & & = 1 \\ & 7x_2 & +2x_3 & -x_4 & = -4 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} x_1 & -3x_2 & -x_3 & & = 2 \\ & -2x_2 & -x_3 & +2x_4 & = 5 \\ & & -\frac{3}{2}x_3 & +x_4 & = \frac{7}{2} \\ & & -\frac{3}{2}x_3 & +6x_4 & = \frac{27}{2} \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} x_1 & -3x_2 & -x_3 & & = 2 \\ & -2x_2 & -x_3 & +2x_4 & = 5 \\ & & -\frac{3}{2}x_3 & +x_4 & = \frac{7}{2} \\ & & & +5x_4 & = 10 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{ll} x_1 & = 1 \\ x_2 & = 0 \\ x_3 & = -1 \\ x_4 & = 2. \end{array}\right.$$

2 Oui, car on a effectuer l'élimination de Gauss sans aucun changement du pivot. Appliquer l'algorithme de *Doolittle* et vous trouverez :

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -\frac{1}{2} & 1 & 0 \\ 2 & -\frac{7}{2} & 1 & 1 \end{pmatrix} \quad \text{et} \quad U \begin{pmatrix} 1 & -3 & -1 & 0 \\ 0 & -2 & -1 & 2 \\ 0 & 0 & -\frac{3}{2} & 1 \\ 0 & 0 & 0 & 5 \end{pmatrix}.$$

3 Elimination de Gauss :

$$\left\{\begin{array}{rrrrl} 2x_1 & +4x_2 & -4x_3 & +x_4 & = 0 \\ 3x_1 & +6x_2 & +x_3 & -2x_4 & = -7 \\ -x_1 & +x_2 & +2x_3 & +3x_4 & = 4 \\ x_1 & +x_2 & -4x_3 & +x_4 & = 2 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} 3x_1 & +6x_2 & +x_3 & -2x_4 & = -7 \\ 2x_1 & +4x_2 & -4x_3 & +x_4 & = 0 \\ -x_1 & +x_2 & +2x_3 & +3x_4 & = 4 \\ x_1 & +x_2 & -4x_3 & +x_4 & = 2 \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} 3x_1 & +6x_2 & +x_3 & -2x_4 & = -7 \\ & & -\frac{14}{3}x_3 & +\frac{7}{3}x_4 & = \frac{14}{3} \\ & +3x_2 & +\frac{7}{3}x_3 & +\frac{7}{3}x_4 & = \frac{5}{3} \\ & -x_2 & -\frac{13}{3}x_3 & +\frac{5}{3}x_4 & = \frac{13}{3} \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} 3x_1 & +6x_2 & +x_3 & -2x_4 & = -7 \\ & +3x_2 & +\frac{7}{3}x_3 & +\frac{7}{3}x_4 & = \frac{5}{3} \\ & & -\frac{14}{3}x_3 & +\frac{7}{3}x_4 & = \frac{14}{3} \\ & & -\frac{32}{9}x_3 & +\frac{22}{9}x_4 & = \frac{44}{9} \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{rrrrl} 3x_1 & +6x_2 & +x_3 & -2x_4 & = -7 \\ & +3x_2 & +\frac{7}{3}x_3 & +\frac{7}{3}x_4 & = \frac{5}{3} \\ & & -\frac{14}{3}x_3 & +\frac{7}{3}x_4 & = \frac{14}{3} \\ & & & +\frac{2}{3}x_4 & = \frac{4}{3} \end{array}\right.$$

$$\Leftrightarrow \left\{\begin{array}{ll} x_1 & = 1 \\ x_2 & = -1 \\ x_3 & = 0 \\ x_4 & = 2. \end{array}\right.$$

4 La matrice associée au système (notée par A) n'admet pas la décomposition LU (car, on a trouvé un pivot nul, lors de la résolution du système par la méthode de Gauss).

Exercice 02

1 Soit

$$A = \begin{pmatrix} -1 & 1 & -3 & 0 \\ 1 & 1 & 3 & 8 \\ -2 & 2 & -5 & -1 \\ 3 & 1 & 8 & 13 \end{pmatrix}.$$

Il faut tout d'abord calculer les déterminants des sous matrices de A que vous allez trouver non nuls, vous déduisez donc que A admet une décomposition LU et puis vous appliquez l'algorithme de *Doolittle* pour trouvez que :

$$U = \begin{pmatrix} -1 & 1 & -3 & 0 \\ 0 & 2 & 0 & 8 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & -4 \end{pmatrix} \quad \text{et} \quad L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ -3 & 2 & -1 & 1 \end{pmatrix}.$$

2

$$Ax = b \quad \Leftrightarrow \quad LUx = b \quad \Leftrightarrow \quad \begin{cases} Ly = b, \\ Ux = y. \end{cases}$$

D'une part,

$$Ly = b \quad \Leftrightarrow \quad \begin{cases} y_1 & & & & = & 0 \\ -y_1 & +y_2 & & & = & 2 \\ 2y_1 & & +y_3 & & = & -1 \\ -3y_1 & +2y_2 & -y_3 & +y_4 & = & 5 \end{cases}$$

$$\Leftrightarrow \quad \begin{cases} y_1 & = & 0 \\ y_2 & = & 2 \\ y_3 & = & -1 \\ y_4 & = & 0 \end{cases}$$

D'autre part,

$$Ux = y \quad \Leftrightarrow \quad \begin{cases} -x_1 & +x_2 & -3x_3 & & = & 0 \\ & 2x_2 & & +8x_4 & = & 2 \\ & & x_3 & -x_4 & = & -1 \\ & & & -4x_4 & = & 0 \end{cases}$$

$$\Leftrightarrow \quad \begin{cases} x_1 & = & 4 \\ x_2 & = & 1 \\ x_3 & = & -1 \\ x_4 & = & 0 \end{cases}$$

3

$$A^2x = b \quad \Leftrightarrow \quad AAx = b$$

$$\Leftrightarrow \quad \begin{cases} Ay = b, & \text{(déja trouvée)} \\ Ax = y. \end{cases}$$

et

$$Ax = y \quad \Leftrightarrow \quad LUx = y \quad \Leftrightarrow \quad \begin{cases} Lz = y, \\ Ux = z. \end{cases}$$

Exercice 03

1 Après l'application de l'algorithme de *Doolittle*, en déduit que :

$$U = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 3 & 2 \\ 0 & 0 & 5 \end{pmatrix} \quad \text{et} \quad L = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -1 & 1 & 1 \end{pmatrix}.$$

2 Solution de $Ax = b$ avec $b = (0,1,2)^t$.

a. Si A est décomposée en LU (i.e. $A = LU$ avec L triangulaire inférieure et U triangulaire supérieure). Alors,

$$\begin{aligned} \det(A) &= \det(L)\det(U) \\ &= \left(\prod_{i=1}^{n} \ell_{ii}\right)\left(\prod_{i=1}^{n} u_{ii}\right), \end{aligned}$$

car L et U sont des matrices triangulaires.

b.

$$\begin{aligned} \det(A) &= \det(L)\det(U) \\ &= \Big(1 \times 1 \times 1\Big)\Big(2 \times 3 \times 5\Big) \\ &= 30. \end{aligned}$$

3

$$Ax = b \quad \Leftrightarrow \quad LUx = b \quad \Leftrightarrow \quad \begin{cases} Ly = b \\ Ux = y. \end{cases}$$

On trouve comme solution : $x = (-\frac{2}{5}, \frac{1}{5}, \frac{1}{5})^t$.

4

$$B = U^tAL^t = \underbrace{U^tL}_{L'}\underbrace{UL^t}_{U'} = L'U'.$$

Chapitre 2

Méthodes itératives pour la résolution des systèmes linéaires

I Généralités

On considère un système d'équations linéaires d'ordre n de la forme

$$\mathscr{A}x = b \tag{2.1}$$

où $\mathscr{A}$ est une matrice inversible de coefficients $(a_{ij})_{1\leq i,j\leq n}$ donnés et b un vecteur de composantes $(b_i)_{1\leq i\leq n}$ données.
Les méthodes itératives pour résoudre (2.1) consistent à construire une suite de vecteurs $x^0, x^1, \ldots, x^n$ telle que

$$\lim_{n\to+\infty} \|x - x^n\|_\infty = 0.$$

où $\|.\|_\infty$ désigne la norme infinie dans $\mathbb{R}^n$ définie par

$$\|a\|_\infty = \max_{1\leq i\leq n} |a_i| \quad \text{avec} \quad a = (a_1, a_2, \ldots, a_n).$$

Pour trouver cette suite on décompose $\mathscr{A}$ en une différence de deux matrices $\mathscr{K}$ et $\mathscr{M}$ c'est-à-dire

$$\mathscr{A} = \mathscr{K} - \mathscr{M}.$$

Alors (2.1) s'écrit

$$\mathscr{K}x = \mathscr{M}x + b$$

soit encore

$$x = \mathscr{K}^{-1}\mathscr{M}x + \mathscr{K}^{-1}b.$$

Cette dernière égalité nous suggère la **méthode itérative** suivante

- ❑ On se donne un vecteur x^0 quelconque.
- ❒ Pour $n = 0, 1, 2, \ldots$, on calcule

$$\boxed{x^{n+1} = \mathscr{K}^{-1}\mathscr{M}x^n + \mathscr{K}^{-1}b} \tag{2.2}$$

La matrice $\mathscr{K}^{-1}\mathscr{M}$ est dite *matrice de la méthode itérative*. Dans le cas pratique, pour calculer x^{n+1} a partir de x^n nous commençons par calculer le vecteur $a = \mathscr{M}x^n + b$ puis nous résolvons le système $\mathscr{K}x = a$. Ce système doit être plus facile a résoudre que le système (2.1) ce qui est le cas si $\mathscr{K}$ est diagonale ou triangulaire.

Définition On dit que la méthode itérative (2.1) est convergente si et seulement si pour tout vecteur de départ x^0 on a

$$\lim_{n\to+\infty} \|x - x^n\| = 0.$$

Définition Si $\mathscr{B}$ est une matrice de valeurs propres $\lambda_1, \lambda_2, \ldots, \lambda_n$. On appelle le ***rayon spectral*** de $\mathscr{B}$ la quantité

$$\rho(\mathscr{B}) = \max_{1\leq j\leq n} |\lambda_j| \tag{2.3}$$

où $|\lambda_j|$ est le module de λ_j.

Nous donnons sans dÃ©monstration le rÃ©sultats suivant :

Théorème La méthode itérative (2.1) converge $\Longleftrightarrow \rho(\mathscr{K}^{-1}\mathscr{M}) < 1$

II Méthode de Jacobi

Cette méthode consiste à poser

$$\mathscr{K} = \mathscr{D} \quad \text{et} \quad \mathscr{M} = \mathscr{E} + \mathscr{F}$$

où

- $\mathscr{D}$ est la matrice diagonale $(a_{ii})_{1\le i\le n}$.
- $\mathscr{E}$ la matrice formée des éléments sous-diagonaux $(-a_{ij})_{1\le j<i\le n}$.
- $\mathscr{F}$ la matrice formée des éléments sur-diagonaux $(-a_{ij})_{1\le i<j\le n}$.

On a bien $\mathscr{A} = \mathscr{K} - \mathscr{M} = \mathscr{D} - \mathscr{E} - \mathscr{F}$. D'autre part, puisque

$$\mathscr{K}^{-1} = \mathscr{D}^{-1} = diag\left(\frac{1}{a_{11}},\dots,\frac{1}{a_{nn}}\right)$$

alors d'après (2.1)

$$x_i^{n+1} = \frac{1}{a_{11}}\left(-\sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^n + b_i\right), \quad 1 \le i \le n.$$

La matrice $\mathscr{J}$ définie par $\mathscr{J} = \mathscr{K}^{-1}\mathscr{M} = \mathscr{D}^{-1}(\mathscr{E}+\mathscr{F})$ est appelée matrice de ***Jacobi.***

Définition Soit $\mathscr{A} = (a_{ij})_{1\le i,j\le n}$ une matrice d'ordre n. On dit que $\mathscr{A}$ est *à diagonale strictement dominante* si et seulement si

$$\forall i \in \{1,2,\dots,n\} \qquad |a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}|$$

Théorème Si $\mathscr{A}$ est une matrice à diagonale strictement dominante alors la méthode de **Jacobi** pour la résolution de $\mathscr{A}x = b$ converge.

Démonstration. On a besoin du lemme suivant.

Lemme Soit $\mathscr{B} = (b_{ij})_{1\le i,j\le 1}$ une matrice d'ordre n alors on a

$$\rho(\mathscr{B}) < 1 \Longleftrightarrow \|\mathscr{B}\|_\infty < 1$$

où $\|\mathscr{B}\|_\infty = \max_{1\le i\le n}\sum_{j=1}^{n}|b_{ij}|$.

La matrice $\mathscr{A}$ est à diagonale strictement dominante alors $|a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|$, donc la méthode est bien définie puisque $a_{ii} > 0$. D'autre part les coefficients de la matrice de *Jacobi* $\mathscr{J} = \mathscr{D}^{-1}(\mathscr{E}+\mathscr{F})$ sont donnés par

$$\mathscr{J}_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}}, & i \neq j \\ 0, & i = j. \end{cases}$$

D'où

$$\begin{aligned}\|\mathscr{J}\|_\infty &= \max_{1\le i\le n}\sum_{j=1}^{n}|\mathscr{J}_{ij}|\\ &= \max_{1\le i\le n}\sum_{\substack{j=1\\ j\neq i}}^{n}|\frac{a_{ij}}{a_{ii}}|\\ &= \max_{1\le i\le n}\frac{1}{|a_{ii}|}\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}| < 1.\end{aligned}$$

D'après le Lemme précédent on déduit que $\rho(\mathscr{J}) < 1$ et par conséquent la méthode de *Jacobi* converge. □

Exemple On considère la matrice $\mathscr{A}$ et le vecteur b définis par

$$\mathscr{A} = \begin{pmatrix} 10 & -1 & 0\\ -1 & 10 & -1\\ 0 & -1 & 10\end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 1\\ 1\\ 1\end{pmatrix}$$

Alors d'après les définitions des matrices $\mathscr{D}, \mathscr{E}$ et $\mathscr{F}$ on a

$$\mathscr{D} = \begin{pmatrix} 10 & 0 & 0\\ 0 & 10 & 0\\ 0 & 0 & 10\end{pmatrix}, \quad \mathscr{E} = \begin{pmatrix} 0 & 0 & 0\\ 1 & 0 & 0\\ 0 & 1 & 0\end{pmatrix}, \quad \mathscr{F} = \begin{pmatrix} 0 & 1 & 0\\ 0 & 0 & 1\\ 0 & 0 & 0\end{pmatrix}$$

Puisque les coefficients de la matrice $\mathscr{A}$ vérifient

$$\begin{aligned}|a_{11}| &= 10 > |-1| + 0,\\ |a_{22}| &= 10 > |-1| + |-1|,\\ |a_{33}| &= 10 > |-1| + 0.\end{aligned}$$

Alors on déduit que $\mathscr{A}$ est à diagonale strictement dominante donc la méthode de *Jacobi* converge. Pour le vecteur de départ $x^0 = (0,0,0)^T$ en utilisant la formule (2.1) on obtient

$$x^1 = \left(\frac{1}{10}, \frac{1}{10}, \frac{1}{10}\right)^T, \quad x^2 = (0.11, 0.12, 0.11)^T, \quad x^3 = (0.112, 0.122, 0.112)^T.$$

On observe que ces vecteurs sont proches de la solution exacte du système donnée par $x = (0.1111, 0.1111, 0.1111)^T$.

Exemple On choisit dans cet exemple

$$\mathscr{A} = \begin{pmatrix} 1 & -2 & 2\\ -1 & 1 & -1\\ -2 & -2 & 1\end{pmatrix} \quad \text{et} \quad b = \begin{pmatrix} 1\\ 0\\ 3\end{pmatrix}$$

On voit que $\mathscr{A}$ n'est pas à diagonale strictement dominante. Alors on cherche les valeurs propres de la matrice $\mathscr{K}^{-1}\mathscr{M}$ où

$$\mathscr{K}^{-1} = \begin{pmatrix} 1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 1\end{pmatrix} \quad \text{et} \quad \mathscr{M} = \begin{pmatrix} 0 & 2 & -2\\ 1 & 0 & 1\\ 2 & 2 & 0\end{pmatrix}.$$

Puisque $\mathscr{K}^{-1}\mathscr{M} = \mathscr{M}$ alors on doit résoudre l'équation

$$\det(\mathscr{M} - \lambda\mathscr{I}) = 0.$$

Plus précisément on a

$$\begin{vmatrix} -\lambda & 2 & -2\\ 1 & -\lambda & 1\\ 2 & 2 & -\lambda\end{vmatrix} = -\lambda\begin{vmatrix} -\lambda & 1\\ 2 & -\lambda\end{vmatrix} - \begin{vmatrix} 2 & -2\\ 2 & -\lambda\end{vmatrix} + 2\begin{vmatrix} 2 & -2\\ -\lambda & 1\end{vmatrix} = -\lambda^3.$$

Alors on a $\rho(\mathscr{K}^{-1}\mathscr{M}) = 0 < 1$ par suite la méthode est convergente.

III Méthode de Gauss-Seidel

Cette méthode est obtenue en choisissant

$$\boxed{\mathscr{K} = \mathscr{D} - \mathscr{E} \quad \text{et} \quad \mathscr{M} = \mathscr{F}}$$

L'égalité (2.1) peut d'écrire

$$(\mathscr{D} - \mathscr{E})x^{n+1} = \mathscr{F}x^n + b. \tag{2.4}$$

On vérifie sans difficulté que la relation (2.4) écrite composante par composante devient

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

La matrice $\mathcal{G} = \mathscr{K}^{-1}\mathscr{M} = (\mathscr{D} - \mathscr{E})^{-1}\mathscr{F}$ est appelée matrice de *Gauss-Seidel.*

Théorème Si $\mathscr{A}$ est une matrice à diagonale strictement dominante alors la méthode de **Gauss-Seidel** pour la résolution de $\mathscr{A}x = b$ converge.

Démonstration. $\mathscr{A}$ est à diagonale strictement dominante d'où $a_{ii} \neq 0\ \forall i$ ce qui rend la méthode bien définie. Soit $\mathcal{G} = (\mathscr{D} - \mathscr{E})^{-1}\mathscr{F}$ la matrice de *Gauss-Seidel.* Raisonnons par absurde et supposons qu'il existe une valeur propre λ de $\mathcal{G}$ avec $|\lambda| \geq 1$. On a d'une part l'existence d'un vecteur propre $u \neq 0$ tel que

$$\begin{aligned}(\mathscr{D} - \mathscr{E})^{-1}\mathscr{F} = \lambda u &\Longleftrightarrow \frac{1}{\lambda}\mathscr{F}u = (\mathscr{D} - \mathscr{E})u \\ &\Longleftrightarrow \left((\mathscr{D} - \mathscr{E}) - \frac{1}{\lambda}\mathscr{F}\right)u = 0 \\ &\Longleftrightarrow \det\left(\mathscr{D} - \mathscr{E} - \frac{1}{\lambda}\mathscr{F}\right) = 0\end{aligned} \tag{2.5}$$

et d'autre part les composantes de la matrice $\mathscr{D} - \mathscr{E} - \frac{1}{\lambda}\mathscr{F}$ sont

$$\left(\mathscr{D} - \mathscr{E} - \frac{1}{\lambda}\mathscr{F}\right)_{ij} = \begin{cases} a_{ij}, & i \geq j \\ \frac{a_{ij}}{\lambda}, & i < j. \end{cases}$$

et par conséquent puisque $\mathscr{A}$ est à diagonale strictement dominante et $|\lambda| \geq 1$ alors

$$\begin{aligned}|a_{ii}| > \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| &= \sum_{j<i} |a_{ij}| + \sum_{j>i} |a_{ij}| \\ &\geq \sum_{j<i} |a_{ij}| + \sum_{j>i} \left|\frac{a_{ij}}{\lambda}\right|.\end{aligned}$$

On en déduit que $\mathscr{D} - \mathscr{E} - \frac{1}{\lambda}\mathscr{F}$ est à diagonale strictement dominante donc inversible ce qui implique que $\det\left(\mathscr{D} - \mathscr{E} - \frac{1}{\lambda}\mathscr{F}\right) \neq 0$ ce qui est en contradiction avec la $3^{\text{ème}}$ ligne de (2.5). En conclusion on ne peut pas avoir une valeur propre λ telle que $|\lambda \geq 1|$ donc nécessairement $\rho(\mathscr{G}) < 1$ et par suite la méthode de *Gauss-Seidel* converge. □

Exemple On reconsidère le premier exemple donné dans la section précédente. Pour cet l'exemple la méthode de *Gauss-Seidel* converge car A est à diagonale strictement dominante et on a $\mathscr{A} = \mathscr{K} - \mathscr{M} = (\mathscr{D} - \mathscr{E}) - \mathscr{F}$ avec

$$\mathscr{K} = \begin{pmatrix} 10 & 0 & 0 \\ -1 & 10 & 0 \\ 0 & -1 & 10 \end{pmatrix} \quad \text{et} \quad \mathscr{M} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Si on choisit $x^0 = (0,0,0)^T$ en utilisant la formule (2.1) on obtient

$$x^1 = (0.1, 0.11, 0.111)^T, \quad x^2 = (0.111, 0.1222, 0.1122)^T.$$

Exemple On reconsidère le même système linéaire dans le deuxième exemple de la section précédente. Pour la méthode de *Gauss-Seidel* on a

$$\mathscr{K} = \mathscr{D} - \mathscr{E} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ -2 & -2 & 1 \end{pmatrix} \quad \text{et} \qquad \mathscr{M} = \mathscr{F} = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}.$$

Alors il est facile de vérifier que

$$(\mathscr{D} - \mathscr{E})^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 4 & 2 & 1 \end{pmatrix} \quad \text{et} \qquad (\mathscr{D} - \mathscr{E})^{-1}\mathscr{F} = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 2 & -1 \\ 0 & 8 & -6 \end{pmatrix}.$$

La matrice de *Gauss-Seidel* est définie par $\mathcal{G} = (\mathscr{D} - \mathscr{E})^{-1}\mathscr{F}$ et on a

$$\det(\mathcal{G} - \lambda\mathcal{I}) = -\lambda(\lambda^2 + 4\lambda - 4).$$

Les valeurs propres de $\mathcal{G}$ sont données par

$$\lambda_1 = 0, \quad \lambda_2 = 0.8284, \quad \lambda_3 = -4.8284.$$

D'où on déduit que $\rho(\mathcal{G}) \simeq 4.82 > 1$ donc la méthode ne converge pas.

IV Méthode de relaxation

La méthode de relaxation est un cas général de la méthode de *Gauss-Seidel*. Elle correspond aux choix

$$\mathscr{K} = \mathscr{K}_\omega = \frac{\mathscr{D}}{\omega} - \mathscr{E} \quad \text{et} \quad \mathscr{M} = \mathscr{M}_\omega = \frac{1-\omega}{\omega}\mathscr{D} + \mathscr{F}$$

où $\omega \in \mathbb{R}^*$. Pour $\omega = 1$ on trouve la méthode de *Gauss-Seidel*. L'égalité (2.1) peut s'écrire

$$\left(\frac{\mathscr{D}}{\omega} - \mathscr{E}\right)x^{n+1} = \left(\frac{1-\omega}{\omega}\mathscr{D} + \mathscr{F}\right)x^n + b \tag{2.6}$$

ce qui est équivalent à

$$x_i^{n+1} = (1-\omega)x_i^n - \omega\sum_{j>i}\frac{a_{ij}}{a_{ii}}x_j^n - \omega\sum_{j<i}\frac{a_{ij}}{a_{ii}}x_j^{n+1} + \omega\frac{b_i}{a_{ii}}, \quad 1 \le i \le n.$$

On va chercher des valeurs de ω pour lesquelles la méthode de relaxation converge. La matrice de la méthode est

$$\begin{aligned}\mathcal{G}_\omega &= \left(\frac{\mathscr{D}}{\omega} - \mathscr{E}\right)^{-1}\left(\frac{1-\omega}{\omega}\mathscr{D} + \mathscr{F}\right) \\ &= ((I) - \omega\mathscr{L})^{-1}[(1-\omega)(I) + \omega\mathscr{U}]\end{aligned}$$

où $\mathscr{L} = \mathscr{D}^{-1}\mathscr{E}$ et $\mathscr{U} = \mathscr{D}^{-1}\mathscr{F}$.

Théorème La méthode de relaxation diverge pour tout $\omega \in \mathbb{R} \setminus [0,2]$.

Démonstration. On a

$$\begin{aligned}\det(\mathcal{G}_\omega) &= \det[((I) - \omega\mathscr{L})^{-1}]\det[(1-\omega)(I) + \omega\mathscr{U}] \\ &= (1-\omega)^n = \prod_{i=1}^{n}\lambda_i \le [\rho(\mathcal{G}_\omega)]^n.\end{aligned}$$

Alors pour $\omega < 0$ ou $\omega > 1$ on aura $\rho(\mathcal{G}_\omega) \ge 1$ donc la méthode diverge. □

Définition Soit $\mathscr{A}$ une matrice symétrique. $\mathscr{A}$ est dite *définie positive* si et seulement si l'une des propriétés suivantes sont satisfaites :

❑ $\forall x \in \mathbb{R}^n,\ x \neq 0$ on a $x^T \mathscr{A} x > 0$.

❒ Toutes les valeurs propres de $\mathscr{A}$ sont réelles et > 0.

Théorème Soit $\mathscr{A}$ une matrice symétrique définie positive. Alors la méthode de relaxation pour la résolution de $\mathscr{A}x = b$ converge si $\omega \in]0, 2[$.

Démonstration. On a besoin du lemme suivant.

Lemme Soit $\mathscr{A}$ une matrice symétrique définie positive décomposée en $\mathscr{A} = \mathscr{N}_1 - \mathscr{N}_2$ où $\mathscr{N}_1$ est inversible. Supposons que la matrice $\mathscr{N}_1^T + \mathscr{N}_2$ est symétrique définie positive alors $\rho(\mathscr{N}_1^{-1}\mathscr{N}_2) < 1$. Autrement dit la méthode itérative de matrice $\mathscr{N}_1^{-1}\mathscr{N}_2$ est convergente.

On pose maintenant

$$\mathscr{N}_1 = \frac{1}{\omega}\mathscr{D} - \mathscr{E} \quad \text{et} \quad \mathscr{N}_2 = \frac{1-\omega}{\omega}\mathscr{D} + \mathscr{F}.$$

La matrice $\mathscr{D}$ est définie positive car $\mathscr{A}$ l'est. Un calcul simple donne

$$\mathscr{N}_1^T + \mathscr{N}_2 = \frac{2-\omega}{\omega}\mathscr{D} - \mathscr{E}^T + \mathscr{F} = \frac{2-\omega}{\omega}\mathscr{D}. \quad (\text{car} \mathscr{E}^T = \mathscr{F})$$

Par conséquent $\mathscr{N}_1^T + \mathscr{N}_2$ est symétrique définie positive$\Longleftrightarrow \omega \in]0, 2[$. Donc en utilisant le Lemme IV on conclut que $\rho(\mathscr{N}_1^{-1}\mathscr{N}_2) < 1$. Donc la méthode de relaxation converge. □

Exemple On reprend la matrice A donnée par

$$\mathscr{A} = \begin{pmatrix} 10 & -1 & 0 \\ -1 & 10 & -1 \\ 0 & -1 & 10 \end{pmatrix}$$

Les valeurs propres sont

$$\lambda_1 = 10 + \sqrt{2}, \quad \lambda_2 = 10, \quad \lambda_3 = 10 - \sqrt{2}.$$

Alors puisque $\mathscr{A}$ est symétrique définie positive, la méthode de relaxation converge pour $\omega \in]0, 2[$.

Vitesse de convergence d'une méthode itérative

On peut caractériser la convergence d'une méthode itérative par la vitesse avec laquelle l'erreur tend vers 0. Il est démontré qu'une méthode itérative est d'autant plus rapide à converger que le rayon spectral de sa matrice d'itération est plus petit i.e. Si nous avons deux méthodes itératives convergentes dont les matrices d'itération vérifient $\rho(\overline{B}) < \rho(B)$, alors la méthode de matrice $\overline{B}$ est plus rapide que celle de matrice B.

Il est possible, dans certains cas, de faire une analyse du rayon spectral $\rho(\mathcal{G}_\omega)$ en fonction de ω et de montrer qu'il existe un nombre ω_{opt} tel que $\rho(\mathcal{G}_{\omega_{opt}}) < \rho(\mathcal{G}_\omega)$ pour tout ω différent de ω_{opt}. Plus précisément, nous avons le résultat suivant.

Théorème Si $\mathscr{A}$ est une matrice tridiagonale définie positive, alors la méthode de Jacobi et la méthode de relaxation sont convergentes lorsque $0 < \omega < 2$. De plus , il existe un et un seul paramètre de relaxation optimal ω_{opt} égal à

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(\mathscr{J})}}, \tag{2.7}$$

où $\rho(\mathscr{J})$ est le rayon spectral de la matrice $\mathscr{J}$ de la méthode de Jacobi.

Travaux dirigés

01 Soit $\mathscr{A}$ est une matrice symétrique définie positive.

1 Montrer que la méthode de *Gauss-Seidel* pour la résolution du système $\mathscr{A}x = b$ est convergente.

2 Application : Étudier la convergence de la méthode de *Gauss-Seidel* (pour la résolution du système $\mathscr{A}x = b$) lorsque

$$\mathscr{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

02 On veut résoudre le système linéaire $\mathscr{A}x = b$ où $\mathscr{A}$ et b sont définis par

$$\mathscr{A} = \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

1 Vérifier que $\mathscr{A}$ est symétrique définie positive.

2 Effectuer deux itérations pour les méthodes de *Jacobi* et *Gauss-Seidel* en partant de $x^0 = [\frac{3}{2}, 2]^T$.

3 Quelle est la méthode qui converge plus rapidement ?

03

1 Calculer dans chacun des cas suivants le rayon spectral de la matrice de la méthode de *Jacobi* et le rayon spectral de la matrice de la méthode de *Gauss-Seidel* pour la résolution du système $\mathscr{A}x = b$

$$\mathscr{A} = \begin{bmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{bmatrix} \text{ et } \mathscr{A} = \begin{bmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{bmatrix}.$$

2 Que peut-on déduire ?.

04 On considère la matrice $\mathscr{A}$ et le vecteur b suivants

$$\mathscr{A} = \begin{bmatrix} 1 & \alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{et} \quad b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

1 Pour quelles valeurs de α la matrice $\mathscr{A}$ est définie positive ?

2 Pour quelles valeurs de α la méthode de *Jacobi* pour la résolution approchée du système $\mathscr{A}x = b$ est-elle convergente ?

3 Soit $\alpha = \frac{1}{2}$. En partant de $x^0 = [0, 0, 0]^T$. Calculer trois itérations de la méthode de *Jacobi*.

05 Soit $a \in \mathbb{R}$ et

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

1 Montrer que $\mathscr{A}$ est symétrique définie positive si et seulement si $-\frac{1}{2} < a < 1$.

2 Montrer que la méthode de *Jacobi* converge si et seulement si $-\frac{1}{2} < a < \frac{1}{2}$.

Corrigés des exercices

Exercice 02

$$\mathscr{A} = \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \qquad b = \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

1 D'une part, il est facile de voir que $\mathscr{A}$ est symétrique. D'autre part, on a

$$det(\mathscr{A} - \lambda\mathscr{I}) = \begin{vmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{vmatrix} = (2-\lambda)^2 - 1.$$

Les valeurs propres de $\mathscr{A}$ sont donc réelles et positives données par $\lambda = 1$ et $\lambda = 3$. Par conséquence $\mathscr{A}$ est bien symétrique définie positive.

2 **Schéma itératif de Jacobi :**

$$x^{(0)} = [1.5, 2]^T,$$
$$x_i^{(n+1)} = \frac{1}{a_{ii}}(b_i - \sum_{j \neq i} a_{ij} x_j^{(n)}), \quad i = 1, 2.\ n \geq 0.$$

Pour $n = 0$, on trouve

$$\begin{cases} x_1^{(1)} = \frac{1}{2}(1 + x_2^{(0)}) = \frac{1}{2}(1+2) = 1.50 \\ x_2^{(1)} = \frac{1}{2}(1 + x_1^{(0)}) = \frac{1}{2}(1+1.5) = 1.25 \end{cases}$$

et pour $n = 1$, on trouve

$$\begin{cases} x_1^{(2)} = \frac{1}{2}(1 + x_2^{(1)}) = \frac{1}{2}(1+1.25) = 1.125 \\ x_2^{(2)} = \frac{1}{2}(1 + x_1^{(1)}) = \frac{1}{2}(1+1.5) = 1.250 \end{cases}$$

Schéma itératif de Gauss-Seidel :

$$x^{(0)} = [1.5, 2]^T,$$
$$x_i^{(n+1)} = \frac{1}{a_{ii}}(b_i - \sum_{j>i} a_{ij} x_j^{(n)} - \sum_{j<i} a_{ij} x_j^{(n+1)}), \quad i = 1, 2, \ n \geq 0.$$

Pour $n = 0$, on trouve

$$\begin{cases} x_1^{(1)} = \frac{1}{2}(1 + x_2^{(0)}) = \frac{1}{2}(1+2) = 1.50 \\ x_2^{(1)} = \frac{1}{2}(1 + x_1^{(1)}) = \frac{1}{4}(3+2) = 1.25 \end{cases}$$

et pour $n = 1$, on trouve

$$\begin{cases} x_1^{(2)} = \frac{1}{2}(1 + x_2^{(1)}) = \frac{1}{2}(1+1.25) = 1.125 \\ x_2^{(2)} = \frac{1}{2}(1 + x_1^{(2)}) = \frac{1}{4}(3+1.25) = 1.0625 \end{cases}$$

c. La solution exacte du système $\mathscr{A}x = b$ est $x = (1, 1)^T$. Alors, on remarque que la suite de Gauss-Seidel converge plus rapidement que celle de Jacobi.

Exercice 03

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

1 **Le schéma de Jacobi :** La matrice de Jacobi est donnée par

$$\mathscr{J} = \mathscr{D}^{-1}(\mathscr{E} + \mathscr{F}) = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}$$

Le polynôme caractéristique de $\mathscr{J}$ est

$$\mathscr{P}_{\mathscr{J}}(\lambda) = det(\mathscr{J} - \lambda\mathscr{I}) = \cdots = -\lambda^3.$$

On en déduit que le rayon spectral $\rho(\mathscr{J}) = 0$. Par conséquence, la méthode de Jacobi converge pour cette matrice.
Le schéma de Gauss-Seidel : après calcul, on trouve que la matrice de Gauss-Seidel est

$$\mathscr{G} = (\mathscr{D} - \mathscr{E})^{-1}\mathscr{F} = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & 3 \\ 0 & 0 & 2 \end{pmatrix}$$

et que le polynôme caractéristique est

$$\mathscr{P}_{\mathscr{G}}(\lambda) = det(\mathscr{G} - \lambda\mathscr{I}) = \cdots = -\lambda(\lambda - 2)^2.$$

Donc le rayon spectrale $\rho(\mathscr{G}) = 2$. Par conséquence, la méthode de Gauss-Seidel diverge pour cette matrice.
En bref, dans le cas de la deuxième matrice

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

on trouve que le rayon spectrale associé à la matrice de jacobi est donné par $\rho(\mathscr{J}) = \frac{\sqrt{5}}{2} = 1.11 > 1$ et on déduit que la méthode de Jacobi diverge. Tandis que pour la méthode de Gauss-Seidel on trouve que $\rho(\mathscr{G}) = \frac{1}{2}$. ce qui implique que la méthode de Gauss-Seidel converge pour cette matrice.

2 On déduit que la convergence de la méthode de Gauss-Seidel n'implique pas celle de la méthode de Jacobi et vice versa. Cependant, dans le cas où on a la convergence des deux méthodes à priori la méthode de Gauss-Seidel est la plus rapide.

Exercice 04

$$\mathscr{A} = \begin{bmatrix} 1 & \alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{et} \quad b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

1 La matrice $\mathscr{A}$ est symétrique pour toute valeur de α. Calculons les valeurs propres de $\mathscr{A}$. Après un calcul simple de déterminant, on trouve

$$\mathscr{P}_{\mathscr{A}}(\lambda) = det(\mathscr{A}-\lambda\mathscr{I}) = \cdots = -(\lambda-1)(\lambda-1-\alpha)(\lambda-1+\alpha).$$

alors $\lambda_1 = 1$, $\lambda_2 = 1+\alpha$, $\lambda_3 = 1-\alpha$ sont les valeurs propres de $\mathscr{A}$. Par conséquence $\mathscr{A}$ est définie positive si et seulement si

$$\begin{cases} \lambda_2 = 1+\alpha > 0 \\ \lambda_3 = 1-\alpha > 0 \end{cases}$$

autrement dit si et seulement si $-1 < \alpha < 1$.

2 La matrice de Jacobi est donnée par

$$\mathscr{J} = \mathscr{D}^{-1}(\mathscr{E}+\mathscr{F}) = \begin{pmatrix} 0 & -\alpha & 0 \\ -\alpha & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

Les valeurs propres de cette matrice sont $\lambda_1 = 0$, $\lambda_2 = \alpha$, $\lambda_3 = -\alpha$, ce qui implique que $\rho(\mathscr{J}) = |\alpha|$. Alors, la méthode de Jacobi convergera si $|\alpha| < 1$. i.e., $-1 < \alpha < 1$.

3 à faire d'une façon similaire à celles des exercices précédents.

Exercice 05

$$\mathscr{A} = \begin{bmatrix} 1 & \alpha & 0 \\ \alpha & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{et} \quad b = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}.$$

1 $\mathscr{A}$ est bien symétrique. Déterminons ces valeurs propres :

$$\mathscr{P}_{\mathscr{A}}(\lambda) = det(\mathscr{A}-\lambda\mathscr{I}) = \cdots = ((1-a)-\lambda)^2((1+2a)-\lambda).$$

Ainsi les valeurs propres de $\mathscr{A}$ sont

$$\begin{cases} \lambda_1 = 1-a \quad \text{(valeur propre double)} \\ \lambda_2 = 1+2a \end{cases}$$

D'où $\mathscr{A}$ est symétrique définie positive si et seulement si $\frac{-1}{2} < a < 1$.

2 Déterminons la matrice de Jacobi. Il est facile de voir que

$$\mathscr{J} = \mathscr{D}^{-1}(\mathscr{E}+\mathscr{F}) = \begin{pmatrix} 0 & -a & -a \\ -a & 0 & -a \\ -a & -a & 0 \end{pmatrix}$$

Après calcul on trouve

$$\mathscr{P}_{\mathscr{J}}(\lambda) = det(\mathscr{J}-\lambda\mathscr{I}) = \cdots = (\lambda-a)^2(\lambda-2a).$$

Ainsi $\rho(\mathscr{J}) = |2a|$, par conséquence la méthode de Jacobi converge si et seulement si $|2a| < 1$, i.e., si $-\frac{1}{2} < a < \frac{1}{2}$. (à corrigé dans l'énoncé).

Chapitre 3

Résolution d'équations non-linéaires

Étant donnée une fonction $f : \mathbb{D} \to \mathbb{R}$ continue sur $\mathbb{D} \subset \mathbb{R}$, on s'intéresse dans ce chapitre au problème de recherche d'une solution (ou des solutions) de l'équation $f(x) = 0$. Lorsqu'elle existe, cette solution notée α, est dite zéro de la fonction f. Les méthodes numériques pour approcher α consistent à construire à partir d'une solution initiale x_0, une suite $x_1, x_2, x_3, \ldots, x_n, \ldots$, telle que

$$\lim_{n\to+\infty} x_n = \alpha, \quad \text{où} \quad \alpha \quad \text{satisfait} \quad f(\alpha) = 0.$$

Dans ce cas, on dit alors que la méthode est **convergente.** De plus, nous adopterons la définition suivante :

Définition Soit p un entier positif. On dit qu'une méthode *convergente* est d'ordre p s'il existe une constante c telle que

$$|\alpha - x_{n+1}| \leq c|\alpha - x_n|^p.$$

- ✧ Si $p = 1$ et $c < 1$ on parle de convergence linéaire.
- ✧ Si $p = 2$ on parle de convergence quadratique.
- ✧ Si $p = 3$ on parle de convergence cubique.

I Méthode du point fixe

Une méthode de point fixe pour résoudre numériquement $f(x) = 0$ consiste dans une première phase, à transformer le problème $f(x) = 0$ en un problème équivalent (admettant les mêmes solutions) du type

$$\boxed{x = g(x)}$$

Clairement, il existe une infinité de manières pour opérer cette transformation. Par exemple on peut poser

$$g(x) = x - \lambda f(x) \quad \text{avec} \quad \lambda \in \mathbb{R},\ \lambda \neq 0.$$

Définition On dit que α est un point fixe de g si $g(\alpha) = \alpha$ c'est-à-dire, l'image de α par g est lui-même.

Par exemple, le point 0 est un point fixe de $g(x) = \sin(x)$ car $\sin(0) = 0$.

Supposons que $\alpha \in \mathbb{R}$ est un zéro de f ou, de façon équivalente, un point fixe de g. Alors une méthode de point fixe consiste à :

- ❑ Évaluer une approximation x_0 du point fixe α de g;
- ❒ Calculer successivement $x_{n+1} = g(x_n),\ n = 0, 1, 2, \ldots$

Naturellement toute méthode de point fixe n'est pas forcement convergente. Par contre, si elle converge, c'est-à-dire si la suite (x_n) a une limite que nous notons α et si g est continue, alors cette limite α est nécessairement un point fixe de g puisque

$$\alpha = \lim_{n\to+\infty} x_{n+1} = \lim_{n\to+\infty} g(x_n) = g(\alpha). \tag{3.1}$$

Définition Soit $g : [a,b] \longrightarrow \mathbb{R}$ une fonction et k un réel vérifiant $0 < k < 1$. On dit que g est *contractante* sur $[a,b]$ si :

$$|g(x) - g(y)| \leq k|x - y|, \quad \forall x, y \in [a,b] \tag{3.2}$$

Dans ce cas le réel k est appelé **rapport de contraction**.

Théorème Soit $g : [a,b] \longrightarrow \mathbb{R}$ une fonction de classe $\mathcal{C}^1$ sur $[a,b]$. S'il existe $0 < k < 1$ tel que $\forall x \in]a,b[$ on a $|g'(x)| \leq k$ alors g est contractante de rapport de contraction k.

Démonstration. On applique le théorème des accroissement finis sur l'intervalle $[x,y]$, alors il existe $z \in]x,y[$ tel que

$$g(x) - g(y) = (x - y)g'(z).$$

Ce qui implique que

$$|g(x) - g(y)| = |x - y||g'(z)| \leq k|x - y|.$$

D'où le résultat. □

Théorème Soit $I = [a,b]$ un intervalle fermé de $\mathbb{R}$ et $g : I \longrightarrow \mathbb{R}$ une fonction donnée qui satisfait les deux propriétés suivantes :

- ❏ g est contractante sur I de rapport $k < 1$.
- ❒ $g(I) \subset I$, c'est-à-dire pour tout $x \in I$ on a $g(x) \in I$.

Alors g a un seul point fixe α dans I et, pour tout $x_0 \in I$, la suite (x_n) donnée par

$$\boxed{x_{n+1} = g(x_n), \quad n = 0, 1, 2, \ldots,} \tag{3.3}$$

converge vers α. De plus la convergence est linéaire et on a l'estimation d'erreur suivante :

$$\forall n \in \mathbb{N}^*, \quad |\alpha - x_n| \leq \frac{k^n}{1-k}|x_1 - x_0|. \tag{3.4}$$

Démonstration. Procédant par récurrence sur n. Pour $n = 1$, on a d'après (3.2)

$$\begin{aligned} |\alpha - x_1| = |g(\alpha) - g(x_0)| &\leq k|\alpha - x_0| \\ &\leq k(|\alpha - x_1| + |x_1 - x_0|) \end{aligned}$$

Par conséquent $|\alpha - x_1| \leq \frac{k}{1-k}|x_1 - x_0|$ donc l'inégalité est vrai pour $n = 1$. Supposons que le résultat est vrai pour n. Alors on a

$$\begin{aligned} |\alpha - x_{n+1}| = |g(\alpha) - g(x_n)| &\leq k|\alpha - x_n| \\ &\leq \frac{k^{n+1}}{1-k}|x_1 - x_0|. \end{aligned}$$

Finalement, on déduit que l'inégalité (3.4) est vrai pour tout $n \in \mathbb{N}^*$. D'autre part, puisque $0 < k < 1$ alors $k^n \longrightarrow 0$ quand n tend vers l'infini. D'où d'après (3.4) $\lim_{n \to +\infty} x_n = \alpha$. Puisque g est continue, alors α est un point fixe de g (voir (3.1)). La convergence est linéaire (Définition 3) car pour tout n :

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

De plus α est le seul point fixe de g dans I car si nous supposons qu'il y en a un autre dans I noté β nous avons $|\alpha - \beta| = |g(\alpha) - g(\beta)| \leq k|\alpha - \beta|$ qui implique que $\beta = \alpha$ (puisque k est plus petit que 1). Nous avons donc montré le théorème. □

Nous sommes maintenant en mesure d'énoncer un autre théorème de convergence sur les méthodes de point fixe.

Théorème Supposons $g : \mathbb{R} \longrightarrow \mathbb{R}$ une fois continûment dérivable et soit α un point fixe de g, i.e. $\alpha = g(\alpha)$. Si $|g(\alpha)| < 1$, alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\alpha - x_0| \leq \varepsilon$, alors la suite donnée par

$$x_{n+1} = g(x_n), \quad n = 0, 1, 2, \ldots,$$

converge vers α lorsque n tend vers l'infini. De plus la convergence est linéaire.

Définition Un réel v est une valeur approchée de α avec une précision ε si $|\alpha - v| \leq \varepsilon$.

Remarque

Pour avoir une valeur approchée du réel α avec précision ε il suffit de prendre la valeur de x_{n_0} où n_0 vérifie

$$\frac{k^{n_0}}{1-k}|x_1 - x_0| \leq \varepsilon \Longleftrightarrow n_0 \geq \frac{\ln(\varepsilon) + \ln(1-k) - \ln(|x_1 - x_0|)}{\ln(k)}. \tag{3.5}$$

Une fois que la fonction g vérifiant les conditions du théorème du point fixe est déterminée, on peut utiliser l'algorithme suivant pour construire les itérés de la suite (x_n). Pour que la suite donne une solution approchée de la solution exacte x^*, il est nécessaire de choisir une solution de départ $x_0 \in [a, b]$.

Algorithme point fixe

Initialisation : Choisir $x_0 \in [a, b]$.
Itération : **Pour** $n = 0, 1, 2, \ldots,$ jusqu'à convergence **faire**
Si $f(x_n) = 0$, **alors**
$\alpha \leftarrow x_n$, **stop** ;
fin du Si
$x_{n+1} \leftarrow g(x_n)$;
fin du Pour.

Exemple On cherche à résoudre l'équation $f(x) = x^3 - 4x + 1 = 0,\ x \in [0, \frac{1}{2}]$.

L'équation $f(x) = 0$ peut se récrire $g(x) = x$ avec $g(x) = \frac{1}{4}(x^3 + 1)$. On a $g'(x) = \frac{3}{4}x^2$, d'où $0 \leq g' \leq 0.1875 < 1$ donc g est contractante de rapport $k = \frac{3}{16}$. De plus $g([0, \frac{1}{2}]) = [\frac{1}{4}, \frac{9}{32}] \subset [0, \frac{1}{2}]$. Alors d'après le Théorème précédent la fonction g a un seul point fixe dans $[0, \frac{1}{2}]$ qu'on note α et la suite $x_{n+1} = \frac{1}{4}(x_n^3 + 1)$ converge vers α. Si $x_0 = 0$ alors on obtient d'après (3.4)

$$|\alpha - x_n| \leq \frac{3^n}{16^n \frac{13}{16}} \frac{1}{4} = \frac{3^n}{16^{n-1} \times 52}.$$

Si on veut obtenir une précision 10^{-5} alors :

$$n_0 \geq \frac{\ln(10^{-5}) + \ln(\frac{13}{16}) - \ln(\frac{1}{4})}{\ln(\frac{3}{16})} \simeq 6.1734$$

alors $n_0 = 7$. Par itération de g, on obtient les valeurs suivantes :

x_0	x_1	x_2	x_3	x_4	x_5	x_6
0	0.25	0.253906	0.254092	0.254101	0.254102	0.254102

II Méthode de dichotomie

Soit $f : [a, b] \longrightarrow \mathbb{R}$ une fonction continue telle que $f(a)f(b) \leq 0$ et admettant un unique zéro $\alpha \in [a, b]$. On pose $(a_0, b_0) = (a, b)$ et $x_0 = \frac{a_0 + b_0}{2}$. Si $f(x_0) = 0$ alors $\alpha = x_0$ sinon, on distingue deux cas :

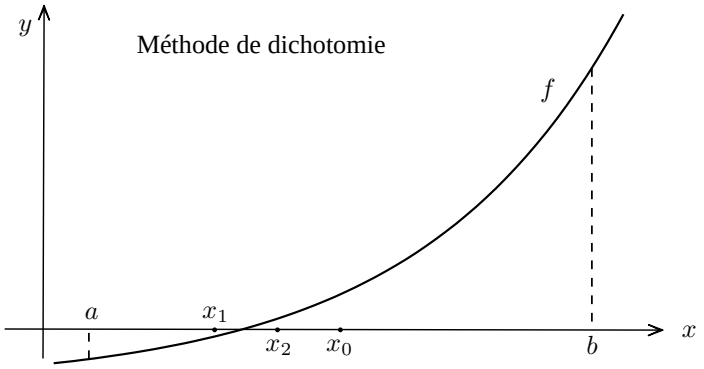
✓ Si $f(a_0)f(x_0) < 0$ alors $\alpha \in]a_0, x_0[$. On pose $(a_1, b_1) = (a_0, x_0)$.

✓ Si $f(x_0)f(b_0) < 0$ alors $\alpha \in]x_0, b_0[$. On pose $(a_1, b_1) = (x_0, b_0)$.
Ainsi $\alpha \in]a_1, b_1[$ et $b_1 - a_1 = \frac{b_0 - a_0}{2} = \frac{b-a}{2}$.
On recommence avec l'intervalle $[a_1, b_1]$ et son milieu $x_1 = \frac{a_1+b_1}{2}$.
Si $f(x_1) = 0$ alors x_1 est un zéro de f et on s'arrête. Sinon, on distingue deux cas :

✓ Si $f(a_1)f(x_1) < 0$ alors $\alpha \in]a_1, x_1[$. On pose $(a_2, b_2) = (a_1, x_1)$.

✓ Si $f(x_1)f(b_1) < 0$ alors $\alpha \in]x_1, b_1[$. On pose $(a_2, b_2) = (x_1, b_1)$.
Ainsi $\alpha \in]a_2, b_2[$ et $b_2 - a_2 = \frac{b_1 - a_1}{2} = \frac{b-a}{2^2}$.
De proche en proche on définit trois suites $(a_n)_{n\geq 0}$, $(b_n)_{n\geq 0}$ et $(x_n)_{n\geq 0}$ de la manière suivante :

$$(a_0, b_0) = (a, b), \quad x_n = \frac{a_n + b_n}{2}, \quad (a_{n+1}, b_{n+1}) = \begin{cases} (a_n, x_n), & \text{si } f(a_n)f(x_n) < 0 \\ (x_n, b_n) & \text{sinon} \end{cases}$$



Théorème La suite (x_n) converge vers l'unique zéro α de f dans $[a, b]$ et on a l'estimation d'erreur suivante :

$$|\alpha - x_n| \leq \frac{b-a}{2^{n+1}}. \tag{3.6}$$

Démonstration. Comme $\alpha \in [a_n, b_n]$ et x_n est le milieu de l'intervalle alors

$$|\alpha - x_n| \leq \frac{b_n - a_n}{2} = \frac{b-a}{2^{n+1}}.$$

D'autre part, $\frac{b-a}{2^{n+1}} \longrightarrow 0$ quand n tend vers l'infini. D'où $\lim_{n\to+\infty} x_n = \alpha$. □

Remarque

Pour avoir une valeur approchée de α avec précision ε il suffit de prendre la valeur de x_{n_0} où n_0 vérifie

$$\frac{b-a}{2^{n_0+1}} \leq \varepsilon \Longleftrightarrow n_0 \geq \frac{\ln(b-a) - \ln(\varepsilon)}{\ln(2)} - 1. \tag{3.7}$$

Algorithme dichotomie

Initialisation : $a_0 = a$ et $b_0 = b$;
Itération : **Pour** $n = 0, 1, 2, \ldots$, jusqu'à convergence **faire**
$x_n \leftarrow \frac{a_n+b_n}{2}$;
Si $f(x_n) = 0$, **alors**
$\alpha \leftarrow x_n$, **stop** ;
fin du Si
Si $f(a_n)f(x_n) < 0$ **faire**
$a_{n+1} \leftarrow a_n$; $b_{n+1} \leftarrow x_n$;
Sinon
$a_{n+1} \leftarrow x_n$; $b_{n+1} \leftarrow b_n$;
fin du Si
fin du Pour

Exemple Reprenons l'équation de l'exemple précédent $f(x) = x^3 - 4x + 1 = 0,\ x \in [0, \frac{1}{2}]$.

On a $f(0) = 1 > 0$ et $f(\frac{1}{2}) = -\frac{7}{8} < 0$ donc $\exists \alpha \in]0, \frac{1}{2}[$ tel que $f(\alpha) = 0$. De plus comme $f'(x) = 3x^2 - 4 < 0$ ce zéro est unique. En appliquant la méthode de dichotomie pour approcher α on a obtenu les résultats suivants :

n	0	1	2	3	4	5	6
$[a_n, b_n]$	$[0, \frac{1}{2}]$	$[\frac{1}{4}, \frac{1}{2}]$	$[\frac{1}{4}, \frac{3}{8}]$	$[\frac{1}{4}, \frac{5}{16}]$	$[\frac{1}{4}, \frac{9}{32}]$	$[\frac{1}{4}, \frac{17}{64}]$	$[\frac{1}{4}, \frac{33}{128}]$
x_n	$\frac{1}{4}$	0.375	0.3125	0.28125	0.265625	0.257813	0.253906

Si on veut obtenir une précision 10^{-5} alors :

$$n_0 \geq \frac{\ln(\frac{1}{2}) - \ln(10^{-5})}{\ln(2)} \simeq 14.6096$$

c'est-à-dire $n_0 = 15$.

III Méthode de Newton

On cherche à évaluer numériquement la racine α d'une équation $f(x) = 0$, en supposant qu'on dispose d'une valeur x_0 proche de cette racine.
L'idée est de remplacer la courbe représentative de f par sa tangente au point x_0 :

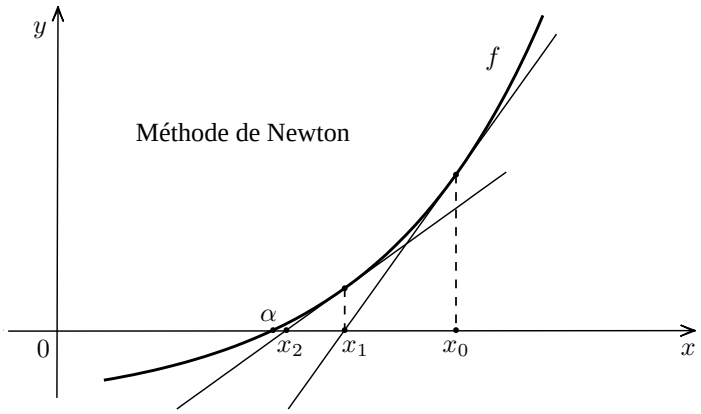
$$y = f'(x_0)(x - x_0) + f(x_0).$$

Si $f'(x_0) \neq 0$, l'abscisse x_1 du point d'intersection de cette tangente avec l'axe $y = 0$ est donnée par

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

On recommence avec le point $(x_1, f(x_1))$ la tangente à la courbe de f au point x_1 coupe l'axe des abscisses au point x_2 et on a $x_2 = x_1 - \frac{f(x_1)}{f'(x_1)}$. si $f'(x_1) \neq 0$. En itérant ce procède, on construit une suite (x_n) définie par

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n = 0, 1, 2, \ldots} \tag{3.8}$$



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

$$g(x) = x - \frac{f(x)}{f'(x)},$$

alors $f(x) = 0 \Longleftrightarrow x = g(x)$ et (3.8) est équivalent à $x_{n+1} = g(x_n)$. Supposons que f soit de classe $\mathcal{C}^2$ et que $f'(\alpha) \neq 0$. La fonction g est alors de classe $\mathcal{C}^1$ au voisinage de α et

$$g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

ce qui donne

$$g'(\alpha) = 0. \tag{3.9}$$

Nous obtenons ainsi le résultat suivant.

Théorème Supposons f deux fois continûment dérivable et supposons que α soit tel que $f(\alpha) = 0$ et $f'(\alpha) \neq 0$. Alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\alpha - x_0| \leq \varepsilon$, la suite (x_n) donnée par la méthode de Newton (3.8) converge vers α. De plus la convergence est quadratique.

Démonstration. Nous avons observé que la méthode de *Newton* est une méthode de point fixe avec $g(x) = x - \frac{f(x)}{f'(x)}$, et que $|g'(\alpha)| < 1$ en vertu de la relation (3.9). Ainsi le résultat de convergence annoncé dans ce théorème est une conséquence du Théorème I. A priori la convergence est linéaire. Nous allons maintenant démontrer que la convergence est quadratique, ceci étant une conséquence du fait que $g'(\alpha) = 0$. Si nous développons f autour du point x_n nous obtenons

$$f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{f''(\theta)}{2}(\alpha - x_n)^2$$

où θ appartient à l'intervalle d'extrémités α et x_n. En divisant par $f'(x_n)$ et en tenant compte du fait que $f(\alpha) = 0$, nous avons

$$\frac{f(x_n)}{f'(x_n)} + \alpha - x_n + \frac{f''(\theta)}{2f'(x_n)}(\alpha - x_n)^2 = 0.$$

En utilisant (3.8) nous obtenons

$$|\alpha - x_{n+1}| = \frac{|f''(\theta)|}{2|f'(x_n)|}|\alpha - x_n|^2.$$

Il suffit maintenant de poser

$$M = \frac{\max\limits_{x \in I} |f''(x)|}{2 \min\limits_{x \in I} |f'(x)|}, \quad \text{où} \quad I = [\alpha - \varepsilon, \alpha + \varepsilon]$$

pour obtenir

$$|\alpha - x_{n+1}| \leq M|\alpha - x_n|^2. \tag{3.10}$$

Cette dernière inégalité montre que la convergence est bien quadratique. □

Remarque

i) D'après (3.10) on a $M|\alpha - x_{n+1}| \leq [M|\alpha - x_n|]^2$. Par récurrence, on en déduit successivement

$$|\alpha - x_n| \leq \frac{1}{M}[M|\alpha - x_0|]^{2^n}.$$

En particulier si x_0 est choisi tel que $|\alpha - x_0| \leq \frac{1}{10M}$, on obtient

$$|\alpha - x_n| \leq \frac{1}{M}10^{-2^n}.$$

ii) Quand faut-il terminer les itérations de l'algorithme de *Newton*? Un bon critère d'arrêt est le contrôle de l'incrément : les itérations s'achèvent dès que $|x_{n+1} - x_n| < \varepsilon$ où ε est une précision fixée. En fait on peut montrer que $|\alpha - x_n| \simeq |x_{n+1} - x_n|$. L'erreur est donc plus petite que la précision ε.

Algorithme de Newton

Initialisation : Choisir x_0 une approximation initiale
Itération : **Pour** $n = 0, 1, 2, \ldots$, jusqu'à convergence **faire**
Si $f(x_n) = 0$, **alors**
$\alpha \leftarrow x_n$, **stop** ;
fin du Si
$x_{n+1} \leftarrow x_n - \frac{f(x_n)}{f'(x_n)}$;
fin du Pour

Exemple Appliquons la méthode Newton à l'exemple $f(x) = x^3 - 4x + 1 = 0$.

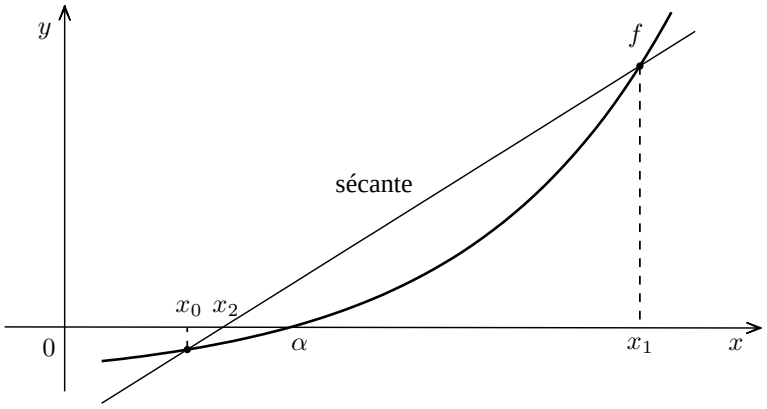
On a

$$x_{n+1} = \frac{2x_n^3 - 1}{3x_n^2 - 4}, \quad n \geq 0.$$

x_0	x_1	x_2	x_3	x_4	x_5
0	0.25	0.2540983607	0.2541016884	0.2541016884	0.2541016884

IV Méthode de la sécante

Dans certaines situations, la dérivée f' est très compliquée ou même impossible à expliciter. On ne peut alors utiliser telle quelle la méthode de *Newton*. L'idée est de remplacer f' par le taux d'accroissement de f sur un petit intervalle. Supposons qu'on dispose de deux valeurs approchées x_0, x_1 de la racine α de l'équation $f(x) = 0$ (fournies par un encadrement $x_0 < \alpha < x_1$).



Le taux d'accroissement de f sur l'intervalle $[x_0, x_1]$ est

$$\tau_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$$

et l'équation de la sécante traversant le graphe de f aux points d'abscisse x_0 et x_1 est

$$y = \tau_1(x - x_1) + f(x_1).$$

On obtient ainsi une nouvelle approximation x_2 de α en calculant l'abscisse de l'intersection de la sécante avec l'axe Ox :

$$x_2 = x_1 - \frac{f(x_1)}{\tau_1}.$$

On va bien entendu itérer ce procédé à partir des nouvelles valeurs approchées x_1 et x_2, ce qui conduit à poser

$$\boxed{\tau_n = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}, \quad x_{n+1} = x_n - \frac{f(x_n)}{\tau_n}}$$

La méthode est donc tout à fait analogue à celle de *Newton*, à ceci prés que l'on a remplacé la dérivée $f'(x_n)$ par le taux d'accroissement τ_n de f sur l'intervalle $[x_n, x_{n-1}]$. On notera que l'algorithme itératif ne peut démarrer que si on dispose deja de deux valeurs approchées x_0, x_1 de α.
D'un point de vue théorique, la convergence de la suite (x_n) est assurée par le résultat ci-dessous, qui donne

simultanément une estimation précise pour $|\alpha - x_n|$.

Théorème On suppose f est de classe $\mathcal{C}^2$ et de dérivée $f' \neq 0$ sur l'intervalle $I = [\alpha - \varepsilon, \alpha + \varepsilon]$. On introduit les quantités M_i, m_i, $i = 1, 2$, et les réels K, h tels que

$$M_i = \max_{x \in I} |f^{(i)}(x)|, \qquad m_i = \min_{x \in I} |f^{(i)}(x)|,$$
$$K = \frac{M_2}{2m_1}\left(1 + \frac{M_1}{m_1}\right), \quad h = \min\left(r, \frac{1}{K}\right).$$

Soit enfin (s_n) la suite de Fibonacci, définie par $s_{n+1} = s_n + s_{n-1}$ avec $s_0 = s_1 = 1$. Alors quel que soit le choix des points initiaux $x_0, x_1 \in [\alpha - h, \alpha + h]$ distincts, on a

$$|\alpha - x_n| \leq \frac{1}{K}[K \max(|\alpha - x_0|, |\alpha - x_1|)]^{s_n}. \tag{3.11}$$

Voici les résultats obtenus pour l'exemple des sections précédentes.

x_0	x_1	x_2	x_3	x_4	x_5	x_6
0	0.5	0.266667	0.253212	0.254104	0.254102	0.254102

Remarque

Il existe plusieurs autres méthodes pour la résolution de $f(x) = 0$. Par exemple la méthode de la corde qui consiste à remplacer $f'(x_n)$ par $f'(x_0)$ dans (3.8), ce qui donne :

$$\boxed{x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}, \quad n = 0, 1, 2, \ldots} \tag{3.12}$$

Ici encore, nous posons $g(x) = x - \frac{f(x)}{f'(x_0)}$ et constatons que $f(\alpha) = 0 \Longleftrightarrow \alpha = g(\alpha)$. Ainsi (3.12) s'écrit $x_{n+1} = g(x_n)$ et la méthode de la corde est une méthode du point fixe. On peut montrer que sous certaines conditions la suite (x_n) converge vers α et la convergence est linéaire.

Travaux dirigés

01 Soit $f : \mathbb{R} \to \mathbb{R}$ la fonction définie par

$$f(x) = x^3 + x - 1.$$

1 Montrer que l'équation $f(x) = 0$ admet une seule solution α telle que $\frac{5}{8} < \alpha < \frac{3}{4}$.

2 Soit $g :]\frac{5}{8}, \frac{3}{4}[\longrightarrow \mathbb{R}$ définie par $g(x) = \frac{1}{1+x^2}$. Montrer que $f(x) = 0 \Longleftrightarrow g(x) = x$.

3 Montrer que $\forall x \in]\frac{5}{8}, \frac{3}{4}[, \quad |g'(x)| < 0.78$.

4 Montrer que la suite (x_n) suivante converge vers α

$$\begin{cases} x_0 = \frac{3}{4}, \\ x_{n+1} = g(x_n). \end{cases}$$

5 Donner une approximation de α de précision 10^{-6} (donner le nombre d'itérations nécessaires).

6 Donner un autre choix de la fonction g de telle sorte que la suite x_n converge aussi.

02 On considère la fonction suivante

$$f(x) = x - e^{-x}.$$

1 Montrer que la fonction f admet une unique racine α dans $[0, 1]$.

2 Calculer les 3 premiers terms de la suite (x_n) obtenue en utilisant la méthode de dichotomie.

3 Donner les résultats obtenus par la méthode de la sécante avec $x_0 = 0, x_1 = 0.5$.

4 En partant de $x_0 = 0$, donner les résultats obtenus par la méthode de *Newton*.

5 Comparer les différentes méthodes sachant que la solution exacte est $\alpha = 0.56714329$.

03 On considère les fonctions définies par

$$f(x) = 3 - e^x,$$
$$g(x) = x + \lambda f(x) = x + \lambda(3 - e^x) \quad (\lambda \in \mathbb{R}^*).$$

1 Montrer que l'équation $f(x) = 0$ admet une unique solution α que l'on déterminera. Vérifier que α est aussi l'unique point fixe de g.

2 Ecrire la méthode de Newton pour la recherche de la solution de $f(x) = 0$.

3 Effectuer 4 itérations de la méthode de Newton Ã partir de $x_0 = 1$.

4 Quelle est le nombre d'itérations nécessaires pour obtenir le zéro avec 4 chiffres significatifs.

04 Soit f la fonction définie par $f(x) = e^x - e.x$, $\forall x \in \mathbb{R}$.

1 Vérifier que 1 est un zéro de f.

2 a) Ecrire la méthode de Newton pour la recherche des zéros de f.

b) Effectuer 4 itérations de la méthode de Newton Ã partir de $x_0 = 0$.

c) A l'aide de (x_2, x_3, x_4) estimer la vitesse de convergence de la méthode de Newton.

d) Montrer théoriquement le résultat de la question précédente.

e) Déduire le nombre d'itérations nécessaires pour obtenir le zéro avec 5 chiffres significatifs.

3 On considère la méthode itérative :

$$x_{n+1} = x_n - 2\frac{f(x_n)}{f'(x_n)}.$$

Effectuer 3 itérations Ã partir de $x_0 = 0$.

4 Vérifier numériquement que la convergence de cette méthode est au moins quadratique.

5 Prouver que la convergence de cette méthode est au moins quadratique.

Corrigés des exercices

Exercice 01

1 On a $f(x) = x^3 + x - 1$ est une fonction strictement croissante sur $\mathbb{R}$, car $f'(x) = 3x^2 + 1 > 0$, $\forall x \in \mathbb{R}$. De plus $f(\frac{5}{8}) = -0.13 < 0$ et $f(\frac{3}{4}) = 0.17 > 0$. Alors f admet une unique solution dans $[\frac{5}{8}, \frac{3}{4}]$.

2 On a

$$\begin{aligned} f(x) = 0 &\Longleftrightarrow x^3 + x - 1 = 0 \\ &\Longleftrightarrow x(x^2+1) = 1 \\ &\Longleftrightarrow x = \frac{1}{(x^2+1)} = g(x). \end{aligned}$$

3 Sachant que $g'(x) = -\frac{2x}{(1+x^2)^2}$, et en partant de $\frac{5}{8} < x < \frac{3}{4}$, on arrive à encadrer $g'(x)$ comme

$$-\frac{\frac{3}{2}}{(1+(\frac{5}{8})^2)^2} < g'(x) < -\frac{\frac{5}{4}}{(1+(\frac{3}{4})^2)^2}$$

On déduit que $|\, g'(x)\,| < 0.78$, $\forall x \in [\frac{5}{8}, \frac{3}{4}]$.

4 d'une part, d'après la question précédente $g(x)$ est contractante sur $[\frac{5}{8}, \frac{3}{4}]$. D'autre part, comme g est décroissante sur cet intervalle, alors

$$g\left(\left[\frac{5}{8}, \frac{3}{4}\right]\right) = \left[g\left(\frac{3}{4}\right), g\left(\frac{5}{8}\right)\right] = [0.64, 0.72] \subset \left[\frac{5}{8}, \frac{3}{4}\right].$$

par conséquence et à l'aide du théorème du point fixe, la suite définie par

$$\begin{cases} x_0 = \frac{3}{4}, \\ x_{n+1} = g(x_n). \end{cases}$$

est convergente.

5 On sait que l'erreur associée à la méthode de point fixe vérifie

$$|\, x_n - \alpha\,| \leq \frac{b-a}{2^{n+1}}.$$

avec dans ce cas $a = \frac{5}{8}$ et $b = \frac{3}{4}$. Alors pour approcher α avec une précision de 10^{-6}, il suffit de choisir n telle que

$$\frac{b-a}{2^{n+1}} \leq 10^{-6}$$

ce qui est équivalent à choisir n telle que

$$n \geq \frac{6\log(10) + \log(b-a)}{\log(2)} - 1$$

puisque n est un entier alors $n = 16$ est bien la valeur cherchée.

Exercice 02

$$f(x) = x - e^{-x} \text{ sur } [0,1].$$

1 On a $f(0) = -1 < 0$ et $f(1) = 1 - e^{-1} > 0$. De plus, $f'(x) = 1 + e^{-x} > 0$ pour tout $x \in [0,1]$, ainsi f est strictement croissante sur $[0,1]$. Donc f admet un unique zéro α.

2 En appliquant la méthode de dichotomie sur l'intervalle $[0,1]$, on obtient

n	0	1	2	3	4
x_n	0.5000	0.7500	0.6250	0.5625	0.5938

3 La suite $(x_n)_n$ obtenue par la méthode de la sécante est donnée par :

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1}) f(x_n)}{f(x_n) - f(x_{n-1})}, \quad x_0 = 0 \text{ et } x_1 = 0.5.$$

4 on obtient :

n	0	1	2	3	4
x_n	0	0.5000	0.5596	0.5670	0.5671

5 la suite $(x_n)_n$ obtenue par la méthode de Newton est donnée par :

$$\begin{aligned} x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} = \frac{x_n f'(x_n) - f(x_n)}{f'(x_n)}, \quad x_0 = 0 \\ &= \frac{(1+x_n)e^{-x_n}}{1+e^{-x_n}} = \frac{(1+x_n)}{e^{x_n}+1}, \quad \text{pour } n \geq 0. \end{aligned}$$

Alors on trouve :

n	0	1	2	3	4
x_n	0	0.5000	0.5663	0.5671	0.5671

6 D'une part, on remarque que $x_3 = 0.5670$ donnée par la méthode de la sécante est une valeur approchée de α avec une précision de 10^{-3}. Cette valeur est plus précise de $x_3 = 0.5625$ obtenue par la méthode de dichotomie. Par conséquent, la méthode de la sécante converge vers α plus rapidement que la méthode de dichotomie. D'autre part, que $x_4 = 0.5671$ donnée par la méthode de Newton est une valeur approchée de α avec une prÃ©cision de 10^{-4}, plus précise que celle donnée par la méthode de la sécante. Par conséquent, la méthode de Newton converge vers α plus rapidement que la méthode de la sécante.

Exercice 03

$$\begin{aligned} &f(x) = 3 - e^x, \\ &g(x) = x + \lambda f(x) = x + \lambda(3 - e^x) \quad (\lambda \in \mathbb{R}^*). \end{aligned}$$

1

$$f(x)=0 \quad \Leftrightarrow \quad e^x=3 \quad \Leftrightarrow \quad x=\ln(3) \Leftrightarrow \alpha=\ln(3).$$

$$g(x)=x \quad \Leftrightarrow \quad \lambda f(x)=0 \quad \Leftrightarrow \quad f(x)=0.$$

2 La méthode de Newton pour la recherche de la solution de $f(x)=0$ s'écrit :

$$\begin{aligned} x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)} &= x_n+\frac{3-e^{x_n}}{e^{x_n}} \\ &= \frac{(x_n-1)e^{x_n}+3}{e^{x_n}} \\ &= (x_n-1)+3e^{-x_n}. \end{aligned}$$

3 A partir de $x_0=1$, on calcule et on trouve : $x_1=1.1036383235$, $x_2=1.0986248980$, $x_3=1.0986122887$, $x_4=1.0986122886$.

4 Sachant que $\alpha=1.0986123$, à partir des premières valeurs de la suite $(x_n)_n$, on remarque que $x_2=1.0986248980$ est une valeur approchée de α avec une erreur de 10^{-5} et $x_3=1.0986122887$ est une valeur approchée de α avec une erreur de 10^{-8}. Donc, on a besoin de deux itérations pour obtenir le zéro avec 4 chiffres significatifs.

Exercice 04

1 On a

$$\begin{aligned} f(1)&=e-e=0; \\ f'(1)&=e-e=0; \\ f''(1)&=e\neq 0. \end{aligned}$$

alors 1 est une racine de multiplicité 2 de f.

2 a) La méthode de Newton est donnée par

$$\begin{aligned} x_{n+1}&=x_n-\frac{f(x_n)}{f'(x_n)} \\ &= x_n-\frac{e^{x_n}-ex_n}{e^{x_n}-e}. \end{aligned}$$

b) On obtient

n	0	1	2	3	4
x_n	0	0.5819	0.8055	0.9056	0.9536

c) On trouve

$$\frac{x_3-1}{x_2-1}=0.483 \text{ et } \frac{x_4-1}{x_3-1}=0.492$$

cela montre numériquement que :

$$\lim_{n\to+\infty}\frac{|x_{n+1}-1|}{|x_n-1|}=k$$

Donc la convergence est seulement linéaire.

d) Si on pose $g(x)=x-\frac{f(x)}{f'(x)}$, alors la méthode de Newton consiste à la méthode de point fixe $x_{n+1}=g(x_n)$. Pour la vitesse de convergence on a

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

avec ξ un réel proche de 1. Donc

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

D'autre part

$$g'(x)=\frac{f(x)f''(x)}{f'^2(x)} \text{ et } f'(1)=0,$$

donc, il faut étudier la $\lim_{x\to 1} g'(x)$. Pour cela, on utilise Taylor au voisinage de 1 et on trouve

$$f''(1)\neq 0 \Rightarrow \begin{cases} f(x)\sim \frac{1}{2}f''(1)(x-1)^2 \\ f'(x)\sim (x-1)f''(1) \\ f''(x)\sim f''(1) \end{cases} \Rightarrow g'(1)\sim \frac{1}{2}.$$

la convergence est donc linéaire et

$$\lim_{n\to+\infty}\frac{|x_{n+1}-1|}{|x_n-1|}=\frac{1}{2}\Rightarrow x_n-1\sim\frac{1}{2^n}.$$

e) Pour obtenir une approximation avec 5 chiffres significatifs, on doit avoir

$$|x_n-1|<10^{-5} \text{ ou } \frac{1}{2^n}<10^{-5}\Leftrightarrow n>\frac{5\log(10)}{\log(2)}.$$

on trouve $n=17$. Donc x_{17} est une approximation de 1 avec 5 chiffres significatifs.

3 ... à completer par vous même.

Chapitre 4

Interpolation Polynômiale

I Existence et unicité du polynôme d'interpolation

Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. On se donne $n+1$ points $x_0, x_1, \ldots, x_n$ dans $[a,b]$, deux à deux distincts.

Problème : Existe-t-il un polynôme p_n de degré $n \geq 0$ tel que

$$\boxed{p_n(x_i) = f(x_i), \quad \forall i = 0,1,\ldots,n?} \tag{4.1}$$

Un tel polynôme sera appelé *polynôme d'interpolation de f (ou interpolant de f)* aux points $x_0, x_1, \ldots, x_n$. Une manière apparemment simple de résoudre ce problème est d'écrire

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n \tag{4.2}$$

où $a_0, a_1, a_2, \ldots, a_n$ sont des coefficients qui devront être déterminés (clairement, si les coefficients a_j, $0 \leq j \leq n$ sont connus alors le polynôme p_n est connu). Les $(n+1)$ relations (4.1) s'écrivent alors :

$$a_0 + a_1 x_i + a_2 x_i^2 + \ldots + a_n x_i^n = f(x_i), \quad 0 \leq i \leq n. \tag{4.3}$$

Puisque les valeurs x_i et $y_i = f(x_i)$, $0 \leq j \leq n$ sont connues, les relations (4.3) forment un système de $(n+1)$ équations à $(n+1)$ inconnues $a_0, a_1, a_2, \ldots, a_n$. La forme matricielle de (4.3) est la suivante

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{4.4}$$

Le déterminant du système linéaire (4.4) est un déterminant dit de *Van der Monde* :

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix}.$$

Théorème le polynôme d'interpolation de f aux points x_i, $i = 0,\ldots,n$ existe et il est unique si et seulement si les points x_i, $i = 0,\ldots,n$ sont distincts deux à deux.

Démonstration. On peut démontrer que

$$\Delta = \prod_{0 \leq j < i \leq n} (x_i - x_j).$$

Alors si les x_i sont distincts on a $\Delta \neq 0$. Par suite le système (4.4) admet une unique solution $[a_0, a_1, a_2, \ldots, a_n]^T$. Autrement dit le polynôme p_n défini par (4.2) est unique. □

Il n'est pas recommandé de résoudre numériquement le système précèdent pour obtenir p_n. Nous verrons plus loin une méthode beaucoup plus efficace.

Exemple Pour trouver le polynôme de degré 2 qui en $x_0 = -1$ vaut $y_0 = 8$, en $x_1 = 0$ vaut $y_1 = 3$ et en $x_2 = 1$ vaut $y_2 = 6$ on résout le système suivant

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 3 \\ 6 \end{bmatrix}$$

La solution est $(a_0, a_1, a_2) = (4, -1, 3)$. Alors on obtient $p_2(x) = 4x^2 - x + 3$.

II Méthode d'interpolation de Lagrange

Définition On appelle polynômes de *Lagrange* associés aux points x_i, $i = 0, \ldots, n$ les $n+1$ polynômes définis par

$$\varphi_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{(x - x_j)}{(x_i - x_j)}, \quad 0 \leq i \leq n \tag{4.5}$$

Il est clair que φ_i est un polynôme de degré n et que

$$\varphi_i(x_j) = 0 \quad \text{si} \quad j \neq i,$$
$$\varphi_i(x_i) = 1.$$

Les polynômes $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ sont linéairement indépendants. En effet si $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_n$ sont $(n+1)$ nombres réels tels que $\sum_{i=0}^{n} \alpha_i \varphi_i(x) = 0$, $\forall x \in \mathbb{R}$, alors pour $x = x_j$ nous obtenons

$$0 = \sum_{i=0}^{n} \alpha_i \underbrace{\varphi_i(x_j)}_{\substack{0 \text{ si } i \neq j \\ 1 \text{ si } i = j}} = \alpha_j,$$

et par conséquent tous les α_j, $j = 0, 1, \ldots, n$ sont identiquement nuls. Notons maintenant $\mathbb{P}_n$ l'ensemble formé par tous les polynômes de degré inférieur ou égal à n. Il est bien connu que $\mathbb{P}_n$ est un espace vectoriel de dimension $(n+1)$ et que sa base canonique est données par $1, x, x^2, x^3, \ldots, x^n$. Le fait que $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ soient des polynômes de degré n linéairement indépendants montre que ces derniers forment aussi une base de $\mathbb{P}_n$. Ainsi nous adopterons la définition suivante.

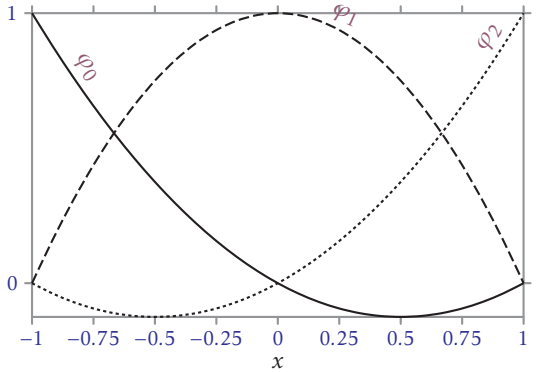
Définition Nous dirons que $\varphi_0, \varphi_1, \varphi_2, \ldots, \varphi_n$ est la base de *Lagrange* de $\mathbb{P}_n$ associé aux points $x_0, x_1, x_2, \ldots, x_n$.

Prenons $n = 2$, $x_0 = -1$, $x_1 = 0$, $x_2 = 1$. La base de *Lagrange* de $\mathbb{P}_2$ associée aux points $-1, 0$ et 1 est formée par les polynômes $\varphi_0, \varphi_1, \varphi_2$ définis par

$$\varphi_0(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} = \frac{1}{2} x(x - 1);$$

$$\varphi_1(x) = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} = -(x + 1)(x - 1);$$

$$\varphi_2(x) = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} = \frac{1}{2}(x + 1)x;$$



Revenons au problème (4.1). Le polynôme p_n cherché s'écrit dans la base de *Lagrange* de $\mathbb{P}_n$

$$p_n(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \ldots + a_n\varphi_n(x) = \sum_{i=0}^{n} a_i\varphi_i(x). \tag{4.6}$$

Si nous utilisons les propriétés des polynômes φ_i, nous avons pour $j = 0, 1, 2, \ldots, n$:

$$f(x_j) = p_n(x_j) = \sum_{i=0}^{n} a_i \underbrace{\varphi_i(x_j)}_{\substack{0 \text{ si } i \neq j \\ 1 \text{ si } i = j}} = a_j.$$

La solution du problème (4.1) est donc définie par

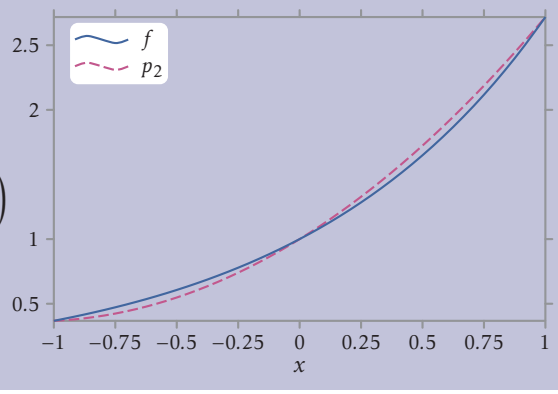
$$\boxed{p_n(x) = \sum_{i=0}^{n} f(x_i)\varphi_i(x), \quad \forall x \in [a,b]} \tag{4.7}$$

Exemple Le polynôme de degré 2 qui en $x_0 = -1$ vaut $y_0 = 8$, en $x_1 = 0$ vaut $y_1 = 3$ et en $x_2 = 1$ vaut $y_2 = 6$. s'écrit

$$\begin{aligned} p_2(x) &= 8\varphi_0(x) + 3\varphi_1(x) + 6\varphi_2(x) \\ &= 8\left(\frac{1}{2}x^2 - \frac{1}{2}x\right) + 3(1 - x^2) + 6\left(\frac{1}{2}x^2 + \frac{1}{2}x\right) \\ &= 4x^2 - x + 3 \end{aligned}$$

Soit f la fonction définie par $f(t) = e^t$. L'interpolant de f de degré 2 aux points -1,0 et 1 est donné par

$$\begin{aligned} p_2(x) &= e^{-1}\varphi_0(x) + e^0\varphi_1(x) + e\varphi_2(x) \\ &= \frac{1}{e}\left(\frac{1}{2}x^2 - \frac{1}{2}x\right) + (1 - x^2) + e\left(\frac{1}{2}x^2 + \frac{1}{2}x\right) \\ &= \left(\frac{1}{2e} - 1 + \frac{e}{2}\right)x^2 + \left(\frac{e}{2} - \frac{1}{2e}\right)x + 1. \end{aligned}$$



Remarque

L'interpolant d'un polynôme de degré $\leq n$ aux $(n+1)$ points distincts x_i, $i = 0, 1, 2, \ldots, n$ est lui même.

L'erreur d'interpolation est donnée par la formule théorique suivante.

Théorème On suppose que f est $n+1$ fois dérivable sur $[a,b]$. Alors pour tout $x \in [a,b]$, il existe un point $\theta_x \in]a,b[$ tel que

$$\boxed{f(x) - p_n(x) = \frac{f^{(n+1)}(\theta_x)}{(n+1)!}\prod_{i=0}^{n}(x - x_i)} \tag{4.8}$$

On a besoin du lemme suivant, qui découle de théorème de *Rolle*.

Lemme Soit g une fonction p fois dérivable sur $[a,b]$. On suppose qu'il existe $p+1$ points $c_0 < c_1 < \ldots < c_p$ de $[a,b]$ tels que $g(c_i) = 0$. Alors il existe $\xi \in]c_0, c_p[$ tel que $g^{(p)}(\xi) = 0$.

Démonstration. Le lemme se démontre par récurrence sur p. Pour $p = 1$, c'est le théorème de *Rolle*. Supposons le lemme démontré pour $p-1$. Le théorème de *Rolle* donne des points $\gamma_0 \in]c_0, c_1[, \ldots, \gamma_{p-1} \in]c_{p-1}, c_p[$ tels que $g'(\gamma_i) = 0$. Par hypothèse de récurrence, il existe donc $\xi \in]\gamma_0, \gamma_{p-1}[\subset]c_0, c_p[$ tel que $(g')^{(p-1)}(\xi) = g^{(p)}(\xi) = 0$. □

Démonstration du théorème On pose $\pi_n(x) = \prod_{i=0}^{n}(x - x_i)$

• Si $x = x_i$, on a $\pi_{n+1}(x_i) = 0$, tout point θ_x convient.

• Supposons maintenant x distinct des points x_i.

Soit $p_{n+1}(t)$ le polynôme d'interpolation de $f(t)$ aux points $x, x_0, \ldots, x_{n+1}$, de sorte que $p_{n+1} \in \mathbb{P}_{n+1}$. Par construction $f(x) - p_n(x) = p_{n+1}(x) - p_n(x)$. Or le polynôme $p_{n+1} - p_n$ est de degré $\leq n+1$ et s'annule aux $n+1$ points $x_0, x_1, \ldots, x_n$. On a donc

$$p_{n+1}(t) - p_n(t) = c\pi_n(t), \quad c \in \mathbb{R}.$$

Considérons la fonction

$$g(t) = f(t) - p_{n+1}(t) = f(t) - p_n(t) - c\pi_n(t).$$

Cette fonction s'annule en les $n+2$ points $x, x_0, x_1, \ldots, x_n$ donc d'après le Lemme II il existe $\theta_x \in]\min(x, x_i), \max(x, x_i)[$ tel que $g^{(n+1)}(\theta_x) = 0$. Or

$$p_n^{(n+1)} = 0, \quad \pi_n^{(n+1)} = (n+1)!$$

On a par conséquent $g^{(n+1)}(\theta_x) = f^{(n+1)}(\theta_x) - c(n+1)! = 0$, d'où

$$f(x) - p_n(x) = p_{n+1}(x) - p_n(x) = c\pi_n(x) = \frac{f^{(n+1)}(\theta_x)}{(n+1)!}\pi_n(x). \qquad \blacksquare$$

Remarque

On considère la subdivision de l'intervalle $[a, b]$ de pas constant $h = \frac{b-a}{n}$. Les points d'interpolation sont donc

$$x_i = a + ih = a + i\frac{b-a}{n}, \quad 0 \leq i \leq n.$$

Dans ce cas on peut montrer que l'erreur d'interpolation (4.8) est

$$\max_{x\in[a,b]} |f(x) - p_n(x)| \leq \frac{(b-a)^{n+1}}{(n+1)!} \max_{x\in[a,b]} |f^{(n+1)}(x)|. \tag{4.9}$$

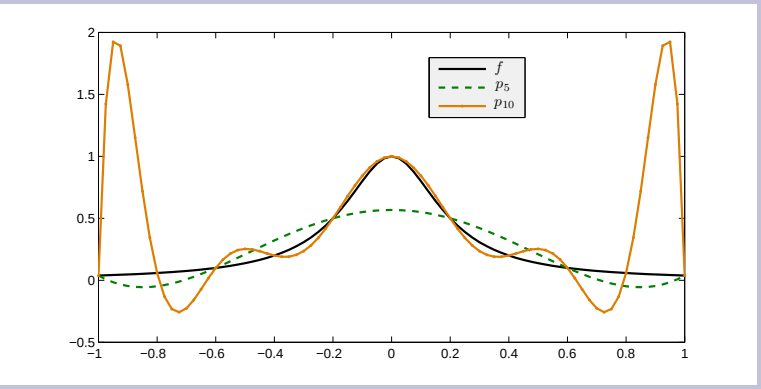
A priori nous pourrions penser que cette erreur converge vers zéro lorsque n tend vers ∞ puisque nous avons

$$\lim_{n\to+\infty} \frac{(b-a)^{(n+1)}}{(n+1)!} = 0.$$

En réalité, cette affirmation est souvent fausse car $\max_{x\in[a,b]} |f^{(n+1)}(x)|$ peut croître très rapidement avec n. Ce phénomène est illustré dans l'exemple suivant.

Exemple

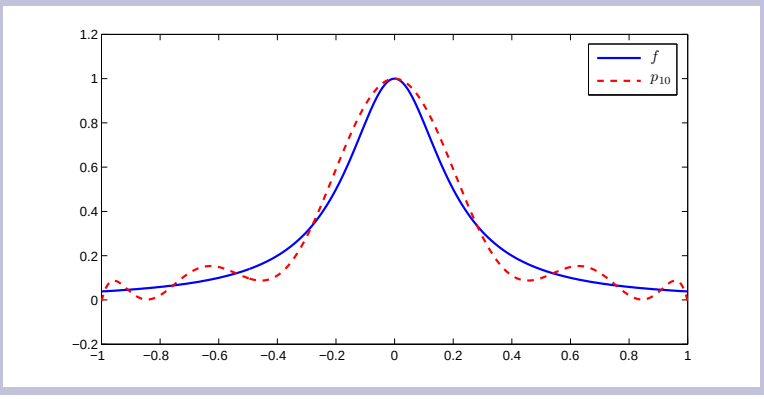
Phénomène de Runge : Soit $f(x) = \frac{1}{1+25x^2}$ que l'on considère sur l'intervalle $[-1,1]$. La fonction $f(x)$ est infiniment dérivable sur $[-1,1]$ et $|f^{(n)}(1)|$ devient très rapidement grand lorsque n tend vers l'infini. La figure suivante montre l'interpolant p_n de degré n aux points $x_i = -1 + \frac{2i}{n}$, $i = 0,1,\ldots,n$, pour $n = 5$ et $n = 10$.



Nous observons que, au voisinage des extrémités de [-1,1] l'interpolant présente de grandes oscillations (instabilités numériques). Nous concluons donc qu'il n'est pas indiqué d'interpoler une fonction par un polynôme de degré n élevé en des points $x_0, x_1, \ldots, x_n$ équidistribués.
Par contre, si nous choisissons les points dits de *Tchebychev*

$$x_i = a + \frac{(b-a)}{2}\left[1 + \cos\left(\frac{2i+1}{2n+2}\pi\right)\right], \quad i = 0,1,2,\ldots,n,$$

pour construire l'interpolant p_{10} de f, alors l'erreur $\max_{x\in[a,b]} |f(x) - p_n(x)|$ tend vers zéro lorsque n tend vers l'infini, comme le montre la figure suivante.



III Interpolation par intervalle

L'interpolation d'une fonction par des polynômes de degré élevé en des points équidistribués peut engendrer des instabilités numériques comme nous l'avons vu dans la section précédente. C'est la raison pour laquelle l'interpolation par intervalles est souvent utilisée.

Soit f une fonction continue donnée sur un intervalle $[a,b]$ et soit $(N+1)$ points $t_0 = a < t_1 < t_2 < t_3 < \ldots < x_N = b$ dans l'intervalle $[a,b]$. Pour chaque intervalle $[t_i, t_{i+1}]$, il est possible de choisir $(n-1)$ points intérieurs équirépartis notés

$$t_{i,1} < t_{i,2} < t_{i,3} < \ldots < t_{i,n-1}.$$

En posant $x_0 = t_i$, $x_j = t_{i,j}$ avec $1 \le j \le n-1$, $x_n = t_{i+1}$, nous pouvons interpoler f aux points x_j, $j = 0,1,\ldots,n$ par un polynôme de degré n comme nous l'avons fait dans la section précédente. Dans la suite nous définissons

$$h = \max_{0\le i\le N-1} |t_{i+1} - t_i|,$$

et nous construisons une fonction $f_h : [a,b] \to \mathbb{R}$ telle que f_h restreinte à chaque intervalle $[t_i, t_{i+1}]$ soit justement le polynôme d'interpolation de f de degré n aux points x_j, $j = 0, 1, \ldots, n$. On dit que f_h est l'interpolant de degré n par intervalles de la fonction f. Nous démontrons le résultat suivant :

Théorème Soit $n \in \mathbb{N}$ et soit $f : [a,b] \to \mathbb{R}$ une fonction $(n+1)$ fois continûment dérivable sur $[a,b]$. Soit f_h l'interpolant de f de degré n par intervalles. Alors il existe une constante c telle que

$$\max_{x\in[a,b]} |f(x) - f_h(x)| \le ch^{n+1}. \tag{4.10}$$

Démonstration. En utilisant l'estimation (4.9) sur l'intervalle $[t_i, t_{i+1}]$ en lieu et place de l'intervalle $[a,b]$, nous obtenons :

$$\max_{x\in[t_i,t_{i+1}]} |f(x) - f_h(x)| \le \frac{(t_{i+1}-t_i)^{n+1}}{(n+1)!} \max_{x\in[t_i,t_{i+1}]} |f^{(n+1)}(x)|.$$

Ainsi nous avons

$$\max_{x\in[t_i,t_{i+1}]} |f(x) - f_h(x)| \le c\left(\max_{0\le i\le N-1} |t_{i+1}-t_i|\right)^{n+1}, \quad i = 0, 1, 2, \ldots, N-1. \tag{4.11}$$

où la constante c est donnée par

$$c = \frac{\max_{x\in[a,b]} |f^{(n+1)}(x)|}{(n+1)!}.$$

Ainsi (4.10) est une conséquence immédiate de (4.11). □

Une interprétation du Théorème précédent est la suivante. Si on se donne un entier positif n et si on prend N points $t_1, t_2, \ldots, t_N$ avec N de plus en plus grand de façon à ce que h soit de plus en plus petit, alors $\max_{x\in[a,b]} |f(x) - f_h(x)|$ converge vers zéro lorsque h tend vers 0 (N tend vers ∞). En pratique, on prend N grand et n petit ($n = 1$ ou 2 ou 3 ou 4).

IV Méthode des différences divisées

On va décrire ici une méthode simple et efficace permettant de calculer les polynômes d'interpolation de f. Soit p_n le polynôme d'interpolation de f aux points $x_0, x_1, \ldots, x_n$.

Notation : *On désigne par $f[x_0, x_1, \ldots, x_n]$ le coefficient directeur du polynôme p_n (= coefficient de x^n dans $p_n(x)$).*

Puisque $p_n - p_{n-1}$ est un polynôme de degré $\le n$, s'annulant aux points $x_0, x_1, \ldots, x_{n-1}$, et admettant $f[x_0, x_1, \ldots, x_n]$ comme coefficient directeur. Par suite

$$p_n(x) - p_{n-1}(x) = f[x_0, x_1, \ldots, x_n](x - x_0)\ldots(x - x_{n-1}).$$

Comme $p_0(x) = f(x_0)$, on en déduit la formule fondamentale

$$\boxed{p_n(x) = f(x_0) + \sum_{k=1}^{n} f[x_0, x_1, \ldots, x_k](x - x_0)\ldots(x - x_{k-1})} \tag{4.12}$$

Pour pouvoir exploiter cette formule, il reste bien entendu à évaluer les coefficients $f[x_0, x_1, \ldots, x_k]$. On utilise pour cela une récurrence sur le nombre k, on observant que $f[x_0] = f(x_0)$.

Formule de récurrence : Pour $k \ge 1$, on a

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0}. \tag{4.13}$$

A cause de cette formule, la quantité $f[x_0, x_1, \ldots, x_k]$ est appelée *différence divisée* d'ordre k de f aux points $x_0, \ldots, x_k$.

Démonstration. Désignons par $q_{k-1} \in \mathbb{P}_{k-1}$ le polynôme d'interpolation de f aux points $x_0, x_1, \ldots, x_k$. Posons

$$r_k(x) = \frac{(x - x_0)q_{k-1}(x) - (x - x_k)p_{k-1}(x)}{x_k - x_0}.$$

Alors $r_k \in \mathbb{P}_k$, $r_k(x_0) = p_{k-1}(x_0) = f(x_0)$, $r_k(x_k) = q_{k-1}(x_k) = f(x_k)$ et pour $0 < i < k$ on a

$$r_k(x_i) = \frac{(x_i - x_0)f(x_i) - (x_i - x_k)f(x_i)}{x_k - x_0} = f(x_i).$$

Par conséquent $r_k = p_k$. Comme le coefficient directeur de q_{k-1} est $f[x_1, \ldots, x_k]$ on obtient la formule (4.13) en égalant les coefficients de x^k dans l'identité

$$p_k(x) = \frac{(x - x_0)q_{k-1}(x) - (x - x_k)p_{k-1}(x)}{x_k - x_0}.$$

□

Remarque

D'après la formule (4.13) on a pour $k = 1$, $f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}$ qui est le coefficient directeur de x dans $p_1(x)$. En effet p_1 s'écrit sous la forme de *Lagrange* de la manière suivante

$$\begin{aligned} p_1(x) &= f(x_0)\varphi_0(x) + f(x_1)\varphi_1(x) \\ &= f(x_0)\frac{(x - x_1)}{(x_0 - x_1)} + f(x_1)\frac{(x - x_0)}{(x_1 - x_0)} \\ &= \left[\frac{f(x_1) - f(x_0)}{x_1 - x_0}\right](x - x_0) + f(x_0). \end{aligned}$$

Méthode pratique : On range les valeurs $f(x_i)$ dans un tableau puis on modifie ce tableau en n étapes successives de la manière suivante

Etape 0	Etape 1	Etape 2	...	Etape n
$f(x_n)$	$f[x_{n-1}, x_n]$	$f[x_{n-2}, x_{n-1}, x_n]$	...	$f[x_0, \ldots, x_n]$
$f(x_{n-1})$	$f[x_{n-2}, x_{n-1}]$			
$f(x_{n-2})$				
$\vdots$	$\vdots$	$\vdots$		
$f(x_2)$	$f[x_1, x_2]$	$f[x_0, x_1, x_2]$		
$f(x_1)$	$f[x_0, x_1]$			
$f(x_0)$				

A l'issue de la $n^{\text{ème}}$ étape, on obtient le coefficient $f[x_0, \ldots, x_n]$ cherché et on peut alors utiliser la formule (4.12).

Exemple Cherchons le polynôme d'interpolation p_4 qui vérifie

$$p_4(-1) = 1,\ p_4(0) = 0,\ p_4(1) = -1,\ p_4(2) = 2,\ p_4(3) = 3.$$

On doit remplir le tableau suivant

Etape 0	Etape 1	Etape 2	Etape 3	Etape 4
$f(x_4)$	$f[x_3,x_4]$	$f[x_2,x_3,x_4]$	$f[x_1,x_2,x_3,x_4]$	$f[x_0,x_1,x_2,x_3,x_4]$
$f(x_3)$	$f[x_2,x_3]$	$f[x_1,x_2,x_3]$	$f[x_0,x_1,x_2,x_3]$	
$f(x_2)$	$f[x_1,x_2]$	$f[x_0,x_1,x_2]$		
$f(x_1)$	$f[x_0,x_1]$			
$f(x_0)$				

Alors on obtient

Etape 0	Etape 1	Etape 2	Etape 3	Etape 4
$f(3) = 3$	1	-1	-1	$\frac{-5}{12}$
$f(2) = 2$	3	2	$\frac{2}{3}$	
$f(1) = -1$	-1	**0**		
$f(0) = 0$	**-1**			
$f(-1) = \mathbf{1}$				

Puis on a

$$\begin{aligned} p_4(x) &= f(x_0) + \sum_{k=1}^{4} f[x_0,x_1,\dots,x_k](x-x_0)\dots(x-x_{k-1}) \\ &= f(x_0) + f[x_0,x_1](x-x_0) + f[x_0,x_1,x_2](x-x_0)(x-x_1) \\ &\quad + f[x_0,x_1,x_2,x_3](x-x_0)(x-x_1)(x-x_2) \\ &\quad + f[x_0,x_1,x_2,x_3,x_4](x-x_0)(x-x_1)(x-x_2)(x-x_3) \\ &= 1-(x+1)+\frac{2}{3}(x+1)x(x-1)-\frac{5}{12}(x+1)x(x-1)(x-2) \end{aligned}$$

Théorème Soit p_n le polynôme d'interpolation de $f : [a,b] \to \mathbb{R}$ aux points $x_0, x_1, \dots, x_n$. Pour tout $x \in [a,b]$ on a

$$f(x) - p_n(x) = f[x_0,\dots,x_n,x]\prod_{i=0}^{n}(x-x_i).$$

Démonstration. Pour $x = x_i$ la formule est évident. Supposons que $x \neq x_i$ et notons par

$$q_n(t) = p_n(t) + f[x_0,\dots,x_n,x]\prod_{i=0}^{n}(t-x_i),$$

l'interpolant de f aux points $x_0, x_1, \dots, x_n$ et x on a donc $q_n(x) = f(x)$, ce qui est la formule annoncé. □

Travaux dirigés

01

1 Écrire le polynôme P_3 d'interpolation de la fonction $f : x \to x^4$ aux 4 points $-1, 0, 1, 2$ dans la base de Lagrange.

2 Réecrire le même polynôme dans la base de Newton

3 Factoriser le polynôme $E = X^4 - P_3$.

4 Soit $f(x) = x^2 - 2$. Montrer que f est une bijection de $[0, 2]$ dans $[-2, 2]$.

5 On approche f^{-1} par p_2 son interpolant aux points $f(1), f(\frac{3}{2}), f(2)$. Quelle approximation de $\sqrt{2}$ obtient-on ?.

02 Soit P un polynôme quelconque.

1 Montrer que son polynôme d'interpolation relatif aux noeuds x_i, $0 \leq i \leq n$, est égal au reste de la division euclidienne de P par le polynôme

$$\pi_n(x) = (x - x_0)(x - x_1)...(x - x_n).$$

03 Pour deux suites de nombres $x_0; x_1; x_2; \ldots; x_r$ et $y_0; y_1; y_2; \ldots; y_r$, on définit la suite de polynômes :

$$P_{k,0} = y_k \quad \text{pour } k = 0, 1, \ldots, r$$

et pour $k = j + 1, \ldots, r$ et $j = 0, \ldots, r - 1$

$$P_{k,j+1}(x) = \frac{(x_k - x)P_{j,j}(x) - (x_j - x)P_{k,j}(x)}{x_k - x_j}.$$

1 Construire $P_{3,3}$ avec $(x_0, y_0) = (-1; -1); (x_1; y_1) = (0, 1); (x_2; y_2) = (1, 0)$ et $(x_3; y_3) = (2, 0)$.

2 Montrez par récurrence que $P_{k,j}$ avec $k \geq j$ est le polynôme d'interpolation de Lagrange pour les points $x_0, x_1, \ldots, x_{j-1}, x_k$.

3 Qu'en concluez-vous pour $P_{k,k}$?

04

1 Montrer que si la fonction elle-même est un polynôme de degré $\leq n$, alors f est identique à son polynôme d'interpolation p_n aux points $x_0, x_1, \ldots, x_n$.

2 Déduire l'écriture des monômes x^k, $0 \leq k \leq n$ dans la base de *Lagrange* $\{\varphi_i,\ j = 0, 1, \ldots, n\}$.

3 En utilisant l'expression de l'erreur d'interpolation, calculer la valeur de

$$\sum_{j=0}^{n} \varphi_j(0) x_j^k \quad 0 \leq k \leq n.$$

05 On considère la fonction f définie sur $\mathbb{N}$ par

$$f(n) = \frac{1 + (-1)^{n+1}}{2}.$$

1 Montrer que pour tout $n \geq 1$

$$f[0, 1, \ldots, n] = \frac{(-2)^{n-1}}{n!}.$$

Corrigés des exercices

Exercice 01

1 calculons les polynômes de Lagrange

$$\varphi_0 = \frac{x(x-1)(x-2)}{-1(-1-1)(-1-2)} = \frac{-x(x-1)(x-2)}{6}$$
$$\varphi_1 = \frac{(x+1)(x-1)(x-2)}{1(0-1)(0-2)} = \frac{(x+1)(x-1)(x-2)}{2}$$
$$\varphi_2 = \frac{(x+1)x(x-2)}{(1+1)1(1-2)} = \frac{-x(x+1)(x-2)}{2}$$
$$\varphi_3 = \frac{(x+1)x(x-1)}{(2+1)2(2-1)} = \frac{x(x+1)(x-1)}{6}.$$

Le polynôme d'interpolation est donné donc dans la base de Lagrange par

$$\begin{aligned} p_3(x) &= \sum_{i=0}^{3} f(x_i)\varphi_i(x) \\ &= \frac{-x(x-1)(x-2)}{6} + \frac{-x(x+1)(x-2)}{2} \\ &\quad + 16\frac{x(x+1)(x-1)}{6} \\ &= 2x^3 + x^2 - 2x. \end{aligned}$$

2 Dans la base de Newton, le polynôme p_3 est donné par

$$\begin{aligned} p_3(x) = f[-1] + f[-1,0](x+1) + f[-1,0,1](x+1)x \\ + f[-1,0,1,2](x+1)x(x-1), \end{aligned}$$

les différences divisées sont calculées par récurrence comme suit

x_i	$f[x_i]$	$f[x_i,x_{i+1}]$	$f[x_i,.,x_{i+2}]$	$f[x_i,..,x_{i+3}]$
-1	**1**			
0	0	**−1**		
1	1	1	**1**	
2	16	15	7	2

alors

$$\begin{aligned} p_3(x) &= 1 - (x+1) + (x+1)x + 2(x+1)x(x-1). \\ &= 2x^3 + x^2 - 2x. \end{aligned}$$

3 $x^4 - p_3(x)$ c'est un polyôme de degré 4. De plus $-1, 0, 1, 2$ sont quatre racines de ce polynôme alors

$$x^4 - p_3(x) = a_0 x(x+1)(x-1)(x-2)$$

et puisque ce polynôme est unitaire on déduit que $a_0 = 1$.

4 f est continue strictement croissante de $[0,2]$ vers $[-2,2]$ alors c'est une bijection.

5 Les points d'interpolation sont $x_0 = f(1)$, $x_1 = f(\frac{3}{2})$, $x_2 = f(2)$. La fonction interpolée est f^{-1}, alors le polynôme d'interpolation est donné par

$$\begin{aligned} p_2(x) &= \sum_{i=0}^{2} f^{-1}(x_i)\varphi_i(x) \\ &= \varphi_0(x) + \frac{3}{2}\varphi_1(x) + 2\varphi_2(x). \end{aligned}$$

Les fonctions de Lagrange sont donneées par :

$$\varphi_0 = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{4}{15}(x-\frac{1}{4})(x-2)$$
$$\varphi_1 = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} = \frac{-16}{35}(x+1)(x-2)$$
$$\varphi_2 = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{4}{21}(x-\frac{1}{4})(x+1)$$

On en déduit que

$$\begin{aligned} p_2(x) &= \frac{4}{15}(x-\frac{1}{4})(x-2) + \frac{3}{2}\frac{-16}{35}(x+1)(x-2) \\ &\quad + 2\frac{4}{21}(x-\frac{1}{4})(x+1) \\ &= -\frac{4}{105}x^2 + \frac{13}{35}x + \frac{148}{105} \end{aligned}$$

Puisque p_2 est un interpolant de f^{-1}, alors $p_2(x)$ est une approximation de $f^{-1}(x)$ pour tout $x \in [-2,2]$. En particulier, $p_2(0) = \frac{148}{105} = 1.40952$ est une approximation de $f^{-1}(0) = \sqrt{2}$.

Exercice 02

À l'aide la dévision euclidienne de P par π_n on trouve

$$P(x) = Q(x)\pi_n(x) + R(x), \quad deg(R) \leq n,$$

Q est dit le quotient et R le reste de la division. De plus

$$P(x_i) = R(x_i), \quad i = 0, 1, \ldots, n.$$

Par conséquence R est le polynôme d'interpolation de P aux noeuds x_i, $i = 0 \ldots, n$.

Exercice 03

1 Avec un calcul simple en utilisant la relation de récurrence on obtient

x_j	$P_{k,0}$	$P_{k,1}$	$P_{k,2}$	$P_{k,3}$
-1	-1			
0	1	$2x+1$		
1	0	$\frac{1}{2}(x-1)$	$-\frac{3}{2}x^2 + \frac{1}{2}x + 1$	
2	0	$\frac{1}{3}(x-2)$	$-\frac{5}{6}x^2 + \frac{7}{6}x + 1$	$1 - \frac{1}{6}x - \frac{3}{2}x^2 + \frac{2}{3}x^3$

2 Pour $j=0$,

$$P_{k,0}=y_k$$

est un polynôme d'interpolation (de degré 0) au point x_k pour chaque $k=0,1,\dots,n$.
Supposons que $P_{k,j}$ est un polynôme d'interpolation aux points $x_0,x_1,\dots,x_{j-1},x_k$, et montrons que $P_{k,j+1}$ est un polynôme d'interpolation aux points $x_0,x_1,\dots,x_j,x_k$.
On a

$$P_{k,j+1}(x)=\frac{(x_k-x)P_{j,j}(x)-(x_j-x)P_{k,j}(x)}{x_k-x_j}.$$

Pour $i=1,2,\dots,j-1$, on a

$$\begin{aligned}P_{k,j+1}(x_i)&=\frac{(x_k-x_i)P_{j,j}(x_i)-(x_j-x_i)P_{k,j}(x_i)}{x_k-x_j}\\&=\frac{(x_k-x_i)y_i-(x_j-x_i)y_i}{x_k-x_j}\\&=y_i.\end{aligned}$$

De plus

$$\begin{aligned}P_{k,j+1}(x_j)&=\frac{(x_k-x_j)P_{j,j}(x_j)}{x_k-x_j}\\&=y_j,\end{aligned}$$

et

$$\begin{aligned}P_{k,j+1}(x_k)&=\frac{-(x_j-x_k)P_{k,j}(x_k)}{x_k-x_j}\\&=y_k.\end{aligned}$$

On en déduit que $P_{k,j+1}$ est un polynôme d'interpolation aux points $x_0,x_1,\dots,x_j,x_k$. Ce qu'il faut démontré.

3 $P_{k,k}$ est le polynôme d'interpolation aux points $x_0,x_1,\dots,x_k$.

Exercice 04

Facile, vu dans le cours.

Exercice 05

Soit

$$f(n)=\frac{1+(-1)^{n+1}}{2},\quad n\in\mathbb{N}.$$

Montrons par récurrence que pour $n\geq 1$

$$f[0,1,\dots,n]=\frac{(-2)^{n-1}}{n!}.$$

Pour $n=1$, on a

$$f[0,1]=\frac{f(1)-f(0)}{1-0}=1,$$

d'où la propriété est vraie.
Supposons que $f[0,1,\dots,n]=\frac{(-2)^{n-1}}{n!}$ est vraie jusqu'à l'ordre n et montrons que :

$$f[0,1,\dots,n+1]=\frac{(-2)^n}{(n+1)!}.$$

Soit p le polynôme d'interpolation de f aux points $0,1,\dots,n+1$. On a

$$\begin{aligned}p(x)&=f(0)+\sum_{k=1}^{n}f[0,1,\dots,k]x(x-1)\dots(x-k+1)\\&\quad+f[0,1,\dots,n+1]x(x-1)\dots(x-n).\end{aligned}$$

Pour $x=n+1$ et en utilisant l'hypothèse de récurrence on a :

$$\begin{aligned}f(n+1)=p(n+1)&=\underbrace{f(0)}_{=0}+\sum_{k=1}^{n}\frac{(-2)^{k-1}}{k!}(n+1)n\dots(n-k+2)\\&\quad+f[0,1,\dots,n+1](n+1)!.\end{aligned}$$

Alors,

$$\begin{aligned}f[0,1,\dots,n+1](n+1)!&=f(n+1)+\frac{1}{2}\sum_{k=1}^{n}C_{n+1}^{k}(-2)^k\\&=\frac{1+(-1)^{n+2}}{2}+\frac{1}{2}\left[\sum_{k=0}^{n+1}C_{n+1}^{k}(-2)^k-1\right.\\&\quad\left.-(-2)^{n+1}\right]\\&=\frac{1}{2}\left[(-1)^{n+2}+(-1)^{n+1}-(-2)^{n+1}\right]\\&=(-2)^n.\end{aligned}$$

D'où

$$f[0,1,\dots,n+1]=\frac{(-2)^n}{(n+1)!}.$$

Chapitre 5

Intégration numérique

L'objet de ce chapitre est de décrire quelques méthodes numériques classiques permettant d'évaluer des intégrales de fonctions dont les valeurs sont connues en un nombre fini de points.

I Méthodes de quadrature élémentaires et composées

Soit $f : [a,b] \to \mathbb{R}$ une fonction continue. Nous désirons approcher numériquement la quantité $\int_a^b f(x)dx$. Pour cela, nous commenÃ§ons par partitioner l'intervalle $[a,b]$ en petits intervalles $[t_i, t_{i+1}]$, $i = 0,1,2,\ldots,N-1$, c'est-à-dire nous choisissons des points t_i tels que

$$a = t_0 < t_1 < t_2 < \ldots < t_{N-1} < t_N = b.$$

On note par h le maximum des pas des sous intervalles $h = \max_{0\le i\le N-1} |t_{i+1} - t_i|$. La formule de *Chasles* donne

$$\mathcal{I}(f) = \int_a^b f(x)dx = \sum_{i=0}^{N-1} \int_{t_i}^{t_{i+1}} f(x)dx.$$

On est donc ramené au problème d'évaluer l'intégrale de f sur un petit intervalle $[t_i, t_{i+1}]$. Ce calcul est effectué au moyen de formules approchées appelées ***méthodes de quadrature élémentaires*** du type suivant

$$\int_{t_i}^{t_{i+1}} f(x)dx \simeq (t_{i+1} - t_i) \sum_{j=0}^{n} \omega_j f(x_{i,j}) = \mathcal{I}_n(f), \tag{5.1}$$

où les $x_{i,j} \in [t_i, t_{i+1}]$ sont appelés les points d'intégration et les nombres réels ω_j sont appelés les poids de la formule de quadrature. La ***méthode de quadrature composée*** associée sera

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1} (t_{i+1} - t_i) \sum_{j=0}^{n} \omega_j f(x_{i,j}) = \mathcal{I}_h(f). \tag{5.2}$$

L'erreur entre la valeur exacte $\int_a^b f(x)dx$ et la valeur approchée est le réel $E_h(f) = \mathcal{I}(f) - \mathcal{I}_h(f)$.

Définition On dit qu'une méthode de quadrature (*élémentaire* ou *composée*) est d'ordre r si la formule est exacte pour tout $p \in \mathbb{P}_r$. Autrement dit si $E_h(p) = 0$. ($\int_{t_i}^{t_{i+1}} p(x)dx - \mathcal{I}_n(p) = 0$ pour la formule élémentaire).

Remarque

Il est clair que si $\sum_{j=0}^{n} \omega_j = 1$ alors les formules (5.1) et (5.2) sont d'ordre 0. En effet si $f \in \mathbb{P}_0$ alors $f(x) = c \in \mathbb{R}$ $\forall x$. Alors $\int_{t_i}^{t_{i+1}} f(x)dx = c(t_{i+1} - t_i)$ et $\mathcal{I}_n(f) = c(t_{i+1} - t_i)\sum_{j=0}^{n} \omega_j$.

(a) Cas le plus simple : $n = 0$.
On choisit alors un seul point $x_i = x_{i,0} \in [t_i, t_{i+1}]$ et on remplace f sur $[t_i, t_{i+1}]$ par son interpolant de degré 0 :

$p_0(x) = f(x_i)$. On a alors

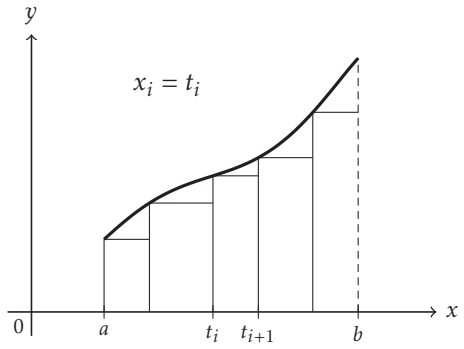
$$\int_{t_i}^{t_{i+1}} f(x)dx \simeq \int_{t_i}^{t_{i+1}} p_0(x)dx = (t_{i+1} - t_i)f(x_i),$$
$$\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1} (t_{i+1} - t_i)f(x_i).$$

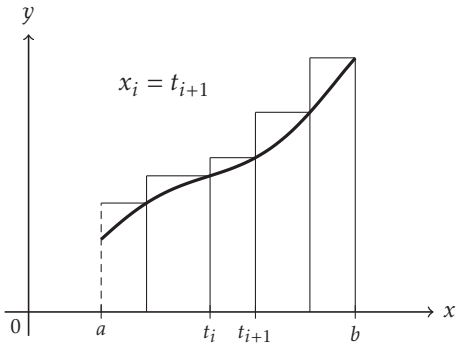
Voici les choix les plus courants :

❑ $x_i = t_i$: *méthode des rectangles à gauche* $\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1} - t_i)f(t_i)$.

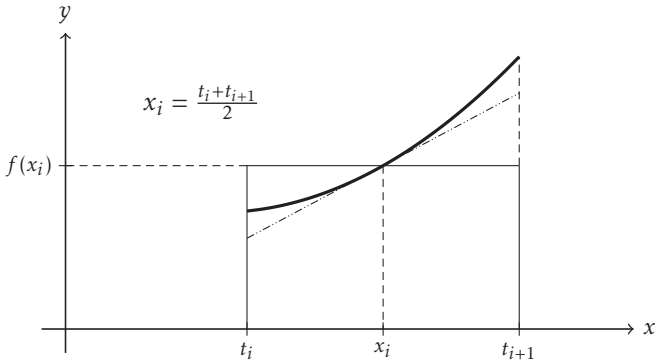
❒ $x_i = t_{i+1}$: *méthode des rectangles à droite* $\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1} - t_i)f(t_{i+1})$.

Ces méthodes sont d'ordre 0 (exactes sur $\mathbb{P}_0$).





❒ $x_i = \frac{t_i + t_{i+1}}{2}$: *méthode du point milieu* $\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1} - t_i)f(\frac{t_i+t_{i+1}}{2})$.



Cette méthode est d'ordre 1 (exacte sur $\mathbb{P}_1$). En effet pour $f(x) = x$ on a $\int_{t_i}^{t_{i+1}} f(x)dx = \frac{t_{i+1}^2 - t_i^2}{2}$ et $\mathcal{I}_n(f) = (t_{i+1} - t_i)\frac{t_{i+1}+t_i}{2}$.

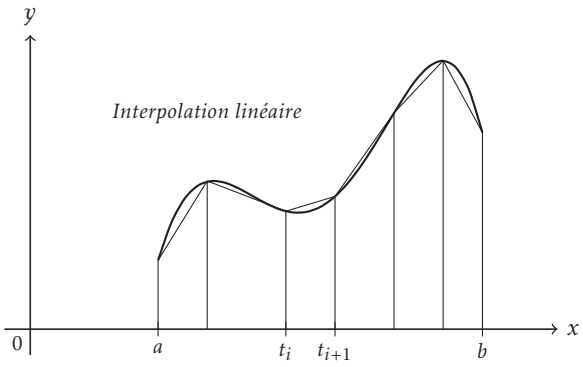
(b) Cas d'une interpolation linéaire : On choisit $n = 1$, $x_{i,0} = t_i$, $x_{i,1} = t_{i+1}$.

Si on remplace f sur $[t_i, t_{i+1}]$ par la fonction linéaire p_1 qui interpole f aux points t_i, t_{i+1}

$$p_1(x) = \frac{(x - t_i)f(t_{i+1}) - (x - t_{i+1})f(t_i)}{t_{i+1} - t_i},$$

on obtient les formules suivantes, correspondant à la *méthode dite des trapèzes* :

$$\int_{t_i}^{t_{i+1}} f(x)dx \simeq \int_{t_i}^{t_{i+1}} p_1(x)dx = (t_{i+1} - t_i)\Big(\frac{1}{2}f(t_i) + \frac{1}{2}f(t_{i+1})\Big),$$
$$\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1} - t_i)\Big(\frac{1}{2}f(t_i) + \frac{1}{2}f(t_{i+1})\Big).$$



L'ordre de cette méthode est 1 comme dans le cas précèdent. Car pour $f(x) = x$ on a $p_1(x) = x$ ce qui donne $\int_{t_i}^{t_{i+1}} f(x)d - \int_{t_i}^{t_{i+1}} p_1(x)dx = 0$.

II Formules de quadrature de Newton-Cotes :

En général dans la méthode de *Newton-Cotes* a $n+1$ points (NC_n) on prend les points équidistants

$$x_{i,j} = t_i + j\frac{t_{i+1} - t_i}{n}, \quad j = 0, 1, 2, \ldots, n$$

divisant $[t_i, t_{i+1}]$ en n sous intervalles égaux. On détermine la formule de quadrature élémentaire tout d'abord sur $[-1, 1]$, subdivisé par les points $\tau_j = -1 + j\frac{2}{n}$ et puis on se ramène par changement de variable à l'intervalle $[t_i, t_{i+1}]$. Le polynôme d'interpolation d'une fonction $g \in \mathscr{C}[-1, 1]$ aux points τ_j est donné par

$$p_n(t) = \sum_{j=0}^{n} g(\tau_j)\varphi_j(t)$$

avec $\varphi_j(t) = \prod_{\substack{k=0 \\ k\neq j}}^{n} \frac{(t - \tau_k)}{(\tau_j - \tau_k)}$. On a donc

$$\int_{-1}^{1} g(t)dt \simeq \int_{-1}^{1} p_n(t)dt = 2\sum_{j=0}^{n} \omega_j g(\tau_j) \tag{5.3}$$

avec $\omega_j = \frac{1}{2}\int_{-1}^{1} \varphi_j(x)dx$. Par suite de la symétrie des points τ_j autour de 0, on a

$$\tau_{n-j} = -\tau_j, \quad \varphi_{n-j}(x) = \varphi_j(-x), \quad \omega_{n-j} = w_j.$$

On considère le changement de variable $t = 2\frac{x - t_i}{t_{i+1} - t_i} - 1$ qui à $x \in [t_i, t_{i+1}]$ fait correspondre $t \in [-1, 1]$. Après changement de variable, les coefficients w_j restent inchangés, donc on obtient

$$\int_{t_i}^{t_{i+1}} f(x)dx = \frac{t_{i+1} - t_i}{2} \int_{-1}^{1} f\left(t_i + (t_{i+1} - t_i)\frac{t+1}{2}\right)dt.$$

Par suite si on choisi $g(t) = f(t_i + (t_{i+1} - t_i)\frac{t+1}{2})$ dans (5.3) il vient

$$\int_{t_i}^{t_{i+1}} f(x)dx \simeq (t_{i+1} - t_i)\sum_{j=0}^{n} \omega_j f\left(t_i + (t_{i+1} - t_i)\frac{\tau_j + 1}{2}\right) \tag{5.4}$$

$$\int_{a}^{b} f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1} - t_i)\sum_{j=0}^{n} \omega_j f\left(t_i + (t_{i+1} - t_i)\frac{\tau_j + 1}{2}\right). \tag{5.5}$$

Pour $n = 2$ par exemple, il vient

$$\tau_0 = -1, \quad \tau_1 = 0, \quad \tau_2 = 1, \quad \varphi_0(x) = \frac{1}{2}x(x-1), \quad \varphi_1(x) = 1 - x^2, \quad \varphi_2(x) = \varphi_0(-x),$$

$$\omega_0 = \omega_2 = \frac{1}{2}\int_{-1}^{1} x(x-1)dx = \frac{1}{6}, \qquad \omega_1 = \frac{1}{2}\int_{-1}^{1}(1-x^2)dx = \frac{2}{3}.$$

donc on obtient

$$\int_{t_i}^{t_{i+1}} f(x)dx \simeq (t_{i+1}-t_i)\left[\frac{1}{6}(f(t_i)+f(t_{i+1})) + \frac{2}{3}f\left(\frac{t_i+t_{i+1}}{2}\right)\right], \tag{5.6}$$

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1}-t_i)\left[\frac{1}{6}(f(t_i)+f(t_{i+1})) + \frac{2}{3}f\left(\frac{t_i+t_{i+1}}{2}\right)\right]. \tag{5.7}$$

Remarque

Si $f \in \mathbb{P}_n$, alors $p_n \equiv f$, donc la méthode de *Newton-Cotes* a $n+1$ points est d'ordre $\geq n$, car

$$E_h(f) = \int_a^b f(x)dx - \int_a^b p_n(x)dx = 0.$$

De plus, lorsque $f \in \mathscr{C}[-1,1]$ est une fonction impaire, on peut démontrer que

$$\int_{-1}^{1} f(x)dx = 0 = 2\sum_{j=0}^{n} w_j f(\tau_j).$$

Si n est pair, les formules sont donc encore exactes pour $f(x) = x^{n+1}$, et plus généralement pour $f \in \mathbb{P}_{n+1}$ par linéarité.

On démontre en fait le résultat suivant que nous admettrons :

Théorème

❑ Si n est pair, l'ordre de NC_n est $n+1$.
❒ Si n est impair, l'ordre de NC_n est n.

Par exemple la formule (5.5) est d'ordre 3 puisque $n = 2$ est pair. Ceci fait que, hormis le cas $n = 1$, les méthodes de *Newton-Cotes* ne sont utilisées que pour n pair :
✓ $n = 1$: méthode des *trapèzes* (ordre 1) $\omega_0 = \omega_1 = \frac{1}{2}$.
✓ $n = 2$: méthode de *Simpson* (ordre 3) $\omega_0 = \omega_2 = \frac{1}{6}$, $\omega_1 = \frac{2}{3}$.
✓ $n = 4$: méthode de *Boole-Villarceau* (ordre 5) $\omega_0 = \omega_4 = \frac{7}{90}$, $\omega_1 = \omega_3 = \frac{16}{45}$, $\omega_2 = \frac{2}{15}$.
✓ $n = 6$: méthode de *Weddle-Hardy* (ordre 7)

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

On peut montrer le résultat de convergence suivant.

Théorème

Supposons que la formule de quadrature composée (5.2) pour calculer numériquement $\int_a^b f(x)dx$ soit exacte pour des polynômes de degré $r \geq 0$. Si f est de classe $\mathscr{C}^{r+1}[a,b]$ alors il existe une constante $c > 0$ indépendante de h telle que

$$\left|\int_a^b f(x)dx - \mathcal{I}_h(f)\right| < ch^{r+1}. \tag{5.8}$$

En fait, l'inégalité (5.8) montre que, lorsque la partition est fine (h petit), l'erreur de quadratue $E_h(f)$ est petite. Cette erreur devient d'autant plus petite avec h que r est grand. Il est donc légitime de chercher une formule de quadrature exacte sur des polynômes de degré r aussi élevé que possible.

Remarque

La formule de quadrature (5.7) donne une erreur d'ordre h^4. C'est une formule souvent utilisée dans la pratique car $\mathcal{I}_h(f)$ converge très rapidement vers $\int_a^b f(x)dx$.

de sorte que la formule de quadrature (5.3) soit exacte sur des polynômes de degré r aussi grand que possible. CommenÃ§ons par la définition suivante.

Définition Le polynôme de *Legendre* de degré n est défini par

$$L_n(t) = \frac{1}{2^n n!}\frac{\partial^n}{\partial t^n}(t^2-1)^n, \quad t \in \mathbb{R}. \tag{5.9}$$

Ainsi nous avons,

$$L_0(t) = 1, \quad L_1(t) = t, \quad L_2(t) = \frac{3t^2-1}{2}, \quad \dots$$

Nous démontrons certaines propriétés des polynômes de *Legendre* L_n, $n = 0,1,2\dots$ qui nous seront utiles dans la suite.

Théorème Les polynômes de *Legendre* vérifient les propriétés suivantes :

i) La famille $\{L_0, L_1, \dots, L_n\}$ forme une base de $\mathbb{P}_n$.

ii) Si $i \neq j$ alors $\int_{-1}^{1} L_i(t)L_j(t)dt = 0$ (propriété d'orthogonalité).

iii) L_n a exactement n racines réels distincts dans l'intervalle $]-1,1[$. Ces racines sont appelés les points de *Gauss*.

Démonstration. *i*) On vérifie facilement que $L_j(t)$ est un polynôme de degré j exactement et ainsi $L_0, L_1, \dots, L_n$ sont indépendants. Ils forment donc une base de $\mathbb{P}_n$.
ii) Supposons $i > j$. On obtient alors en intégrant par parties

$$\begin{aligned}\int_{-1}^{1} L_i(t)L_j(t)dt &= \frac{1}{2^{i+j}i!j!}\int_{-1}^{1}\frac{\partial^i}{\partial t^i}(t^2-1)^i\frac{\partial^j}{\partial t^j}(t^2-1)^j dt \\ &= \frac{1}{2^{i+j}i!j!}\left\{\left[\frac{\partial^{i-1}}{\partial t^{i-1}}(t^2-1)^i\frac{\partial^j}{\partial t^j}(t^2-1)^j\right]_{-1}^{1} - \int_{-1}^{1}\frac{\partial^{i-1}}{\partial t^{i-1}}(t^2-1)^i\frac{\partial^{j+1}}{\partial t^{j+1}}(t^2-1)^j dt\right\}.\end{aligned}$$

Puisque $(t^2-1)^i$ a un zéro d'ordre i en 1 et en -1, la $(i-1)^{\text{ème}}$ dérivée de $(t^2-1)^i$ s'annule en $t = 1$ et en $t = -1$. Ainsi nous obtenons

$$\int_{-1}^{1} L_i(t)L_j(t)dt = \frac{(-1)}{2^{i+j}i!j!}\int_{-1}^{1}\frac{\partial^{i-1}}{\partial t^{i-1}}(t^2-1)^i\frac{\partial^{j+1}}{\partial t^{j+1}}(t^2-1)^j dt.$$

En intégrant par parties j fois comme ci-dessus, nous obtenons :

$$\begin{aligned}\int_{-1}^{1} L_i(t)L_j(t)dt &= \frac{(-1)^j}{2^{i+j}i!j!}\int_{-1}^{1}\frac{\partial^{i-j}}{\partial t^{i-j}}(t^2-1)^i\underbrace{\frac{\partial^{2j}}{\partial t^{2j}}(t^2-1)^j}_{(2j)!}dt \\ &= \frac{(-1)^j(2j)!}{2^{i+j}i!j!}\int_{-1}^{1}\frac{\partial^{i-j}}{\partial t^{i-j}}(t^2-1)^i dt \\ &= \frac{(-1)^j(2j)!}{2^{i+j}i!j!}\left[\frac{\partial^{i-j-1}}{\partial t^{i-j-1}}(t^2-1)^i\right]_{-1}^{1} = 0.\end{aligned}$$

□

Définition La formule de quadrature de *Gauss* a $n+1$ points est définie par

$$\int_{-1}^{1} g(t)dt \simeq 2\sum_{j=0}^{n} \omega_j g(\tau_j) = \mathcal{J}_n(g)$$

où les points d'intégration $\tau_0 < \tau_1 < \ldots < \tau_n$ sont les $n+1$ zéros du polynôme de *Legendre* L_{n+1} c'est-à-dire les $n+1$ points de *Gauss* (voir propriété iii) du Théorème III). Les poids $\omega_0, \omega_1, \ldots, \omega_n$ sont définies par

$$\omega_j = \frac{1}{2}\int_{-1}^{1} \varphi_j(t)dt, \quad j = 0, 1, \ldots, n,$$

où $\varphi_0, \varphi_1, \ldots, \varphi_n$ est la base de *Lagrange* de $\mathbb{P}_n$ associée aux $n+1$ points de *Gauss*.

Nous sommes maintenant en mesure de démontrer le résultat suivant.

Théorème La formule de Gauss à $n+1$ points ($n \geq 0$) est exacte sur les polynômes de degré r=$2n+1$.

Démonstration. Soit p un polynôme de degré $2n+1$. Clairement, nous pouvons définir pour $t \in \mathbb{R}$:

$$p_n(t) = \sum_{j=0}^{n} p(\tau_j)\varphi_j(t),$$

où $\varphi_0, \varphi_1, \ldots, \varphi_n$ est la base de *Lagrange* de $\mathbb{P}_n$ associée aux $n+1$ points de *Gauss* $\tau_0, \tau_1, \ldots, \tau_n$. Autrement dit, le polynôme p_n est donc l'interpolant de p de degré n aux points $\tau_0, \tau_1, \ldots, \tau_n$. Considérons maintenant le polynôme q défini par

$$q(t) = p(t) - p_n(t) \quad \forall t \in \mathbb{R}.$$

Le polynôme q est un polynôme de degré $2n+1$ qui s'annule en les points $\tau_0, \tau_1, \ldots, \tau_n$. Ainsi q est divisible par le polynôme v de degré $n+1$ défini par

$$v(t) = (t-\tau_0)(t-\tau_1)\ldots(t-\tau_n) \quad \forall t \in \mathbb{R},$$

c'est-à-dire qu'il existe un polynôme z de degré n tel que

$$q(t) = v(t)z(t) \quad \forall t \in \mathbb{R}.$$

Puisque L_{n+1} est un polynôme de degré $n+1$ qui s'annule aux points $\tau_0, \tau_1, \ldots, \tau_n$, alors il existe un nombre réel α tel que

$$v(t) = \alpha L_{n+1}(t) \quad \forall t \in \mathbb{R}.$$

D'autre part, puisque z est de degré n, il existe $\beta_0, \beta_1, \ldots, \beta_n \in \mathbb{R}$ (propriété i) du Théorème III) tels que

$$z(t) = \sum_{j=0}^{n} \beta_j L_j(t).$$

Ainsi donc, en utilisant la propriété ii) du Théorème III, nous avons

$$\int_{-1}^{1} q(t)dt = \int_{-1}^{1} v(t)z(t)dt = \alpha\sum_{j=0}^{n} \beta_j \int_{-1}^{1} L_{n+1}(t)L_j(t)dt = 0.$$

Par définition de q nous avons prouvé que

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

en enfin, par définition de p_n, nous obtenons

$$\int_{-1}^{1} p(t)dt = \sum_{j=0}^{n} p(\tau_j)\int_{-1}^{1} \varphi_j(t)dt = 2\sum_{j=0}^{n} \omega_j p(\tau_j).$$

Cette dernière relation est exactement ce que nous voulions montrer. □

Remarque

❑ Les poids ω_j, $j = 0, 1, \ldots, n$ d'une formule de *Gauss* à $n+1$ points sont tous positifs car φ_k^2 est un polynôme de degré $2n$, ce qui implique, en utilisant le Théorème III, que

$$0 < \int_{-1}^{1} \varphi_k^2(t)dt = 2\sum_{j=0}^{n} \omega_j \varphi_k^2(\tau_j) = 2\omega_k.$$

❒ La formule de Gauss à $n+1$ points n'est pas exacte sur l'espace $\mathbb{P}_{2n+2}$. En effet, il suffit de prendre $p(t) = \prod_{j=0}^{n}(t-\tau_j)^2$ pour obtenir $\mathcal{J}_n(p) = 0$ alors que $\int_{-1}^{1} p(t)dt > 0$.

Connaissant une formule de *Gauss* à $n+1$ points, nous pouvons calculer, pour une fonction f définie sur $[a,b]$, la quantité $\mathcal{I}_h(f)$ donnée par la formule composée (5.5). Tenant compte des Théorèmes II et III nous avons

$$\left| \int_a^b f(x)dx - \mathcal{I}_h(f) \right| < ch^{2n+2}, \tag{5.10}$$

où ici f est supposée $2n+2$ fois continÃ»ment dérivable et c est constante positive qui ne dépend pas de n.

Exemple

Formule de Gauss à un seul point On a $L_1(t) = t$ et donc le seul zéro de L_1 est donné par $\tau_0 = 0$. On a $\varphi_0(t) = 1$ et $\omega_0 = \frac{1}{2}$ alors la formule est donnée par

$$\int_{-1}^{1} g(t)dt \simeq g(0) = \mathcal{J}_0(g).$$

Nous retrouvons dans ce cas-là la formule du point milieu qui est d'ordre h^2 (voir section 1).

Exemple

Formule de Gauss à deux points Nous avons $L_2(t) = \frac{1}{2}(3t^2 - 1)$ et donc les deux zéros de L_2 sont donnés par

$$\tau_0 = -\frac{1}{\sqrt{3}} \quad \text{et} \quad \tau_1 = \frac{1}{\sqrt{3}}.$$

La base de *Lagrange* $\{\varphi_0, \varphi_1\}$ de $\mathbb{P}_2$ associée aux points τ_0, τ_1 est définie par

$$\varphi_0(t) = \frac{1-\sqrt{3}t}{2} \quad \text{et} \quad \varphi_1(t) = \frac{1+\sqrt{3}t}{2}$$

et ainsi

$$\omega_0 = \frac{1}{2}\int_{-1}^{1} \varphi_0(t)dt = \frac{1}{2} \quad \text{et} \quad \omega_1 = \frac{1}{2}\int_{-1}^{1} \varphi_1(t)dt = \frac{1}{2}.$$

La formule de *Gauss* à deux points s'écrit donc :

$$\mathcal{J}_1(g) = g\left(-\frac{1}{\sqrt{3}}\right) + g\left(\frac{1}{\sqrt{3}}\right),$$

et la formule (5.5) devient

$$\int_a^b f(x)dx \simeq \sum_{i=0}^{N-1}(t_{i+1}-t_i)\left[f\left(t_i + \frac{\sqrt{3}-1}{2\sqrt{3}}(t_{i+1}-t_i)\right) + f\left(t_i + \frac{\sqrt{3}+1}{2\sqrt{3}}(t_{i+1}-t_i)\right)\right].$$

Si f est quatre fois continuement dérivable, les Théorèmes III et II nous assurent l'existence d'une constante c telle que

$$\left| \int_a^b f(x)dx - \mathcal{I}_h(f) \right| < ch^4,$$

La formule de *Gauss* à deux points converge donc au même ordre que la formule de *Simpson*.

Travaux dirigés

01 Soit $0 < \alpha \leq 1$ un nombre réel, soit $\tau_0 = -\alpha$, $\tau_1 = 0$, $\tau_2 = \alpha$ et soit $\omega_0, \omega_1, \omega_2$ trois nombres réels. On considère la formule de quadrature définie par

$$\int_{-1}^{1} g(t)dt \simeq \sum_{j=0}^{2} \omega_j g(\tau_j) = \mathcal{I}_2(g),$$

où g est une fonction continue sur $[-1,1]$.

1 Trouver $\omega_0, \omega_1, \omega_2$ en fonction de α de sorte que $\mathcal{I}_2(g)$ soit exacte sur $\mathbb{P}_2$.

2 Montrer qu'avec de tels poids $\mathcal{I}_2(g)$ est exacte pour tout polynôme de degré 3.

3 Existe-t-il α tel que $\mathcal{I}_2(g)$ est exacte pour tout polynôme p de degré 5? Si oui, calculer ce α.

02 Soit $0 < \alpha < 1$ un nombre réel, soit $\tau_0 = -1$, $\tau_1 = -\alpha$, $\tau_2 = \alpha$, $\tau_3 = 1$ et soit $\omega_0, \omega_1, \omega_2, \omega_3$ quatre nombres réels. On considère la formule de quadrature définie par

$$\int_{-1}^{1} g(t)dt \simeq \sum_{j=0}^{3} \omega_j g(\tau_j) = \mathcal{I}_3(g),$$

où g est une fonction continue sur $[-1,1]$.

1 Trouver w_j, $0 \leq j \leq 3$ en fonction de α de sorte que $\mathcal{I}_3(g)$ soit exacte sur les polynômes de degré inférieur ou égal a 3.

2 Existe-t-il α tel que $\mathcal{I}_3(g)$ est exacte pour tout polynôme p de degré $r > 3$? Si oui, quelle est la valeur maximale de r et que valent alors α et $\omega_0, \omega_1, \omega_2, \omega_3$?

03 Soient x_1 et x_2 deux points de $[-1,1]$ et λ_1 et $\lambda_2 \in \mathbb{R}$. On désigne par $\mathcal{C}[-1,1]$ l'espace vectoriel des fonctions continues sur $[-1,1]$ et à valeurs réelles et on définit

$$T : \mathcal{C}[-1,1] \to \mathbb{R} \quad \text{par} \quad T(f) = \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

1 Quelles conditions doivent vérifier $x_1, x_2, \lambda_1, \lambda_2$ pour que T soit une méthode d'intégration sur $[-1,1]$ exacte pour

(α) Les fonctions constantes?

(β) Les fonctions affines?

(γ) Les polynômes de degré inférieur ou égale à 2?

2 Parmi les méthodes exactes pour les polynômes de degré ≤ 2 une seul vérifie $x_1 = -x_2$. Montrer que ce choix de x_1 et x_2 (et des λ_1 et λ_2 correspondants) fournit une méthode exacte pour les polynômes de degré ≤ 3 et qu'il s'agit de la seule méthode d'intégration exacte pour les polynômes de degré ≤ 3 qui soit du type étudié dans le problème. Quelle est cette méthode?

04 On considère la formule de quadrature suivante

$$\mathcal{I}(f) = a_0[f(x_0) + f(-x_0)] + a_1[f(1) + f(-1)] \simeq \int_{-1}^{1} f(x)dx.$$

1 Déterminer a_0, a_1 et x_0 pour que la formule soit exacte sur les polynômes de degré 5.

2 Appliquer ce résultat à la fonction $f(x) = \frac{1}{1+x^2}$. Quelle approximation de π obtient-on?

3 Comparer le résultat obtenu avec la formule de *Gauss* a trois points.

Corrigés des exercices

Exercice 01

1 la formule $\mathcal{I}_Q$ est exacte sur $\mathbb{P}_2$ ssi elle est exacte sur $1, x, x^2$ et cela s'exprime sous la forme suivante

$$\begin{cases} \omega_1+\omega_2+\omega_3 & = \int_{-1}^{1} 1dx = 2 \\ -\alpha\omega_1+0\omega_2+\alpha\omega_3 & = \int_{-1}^{1} xdx = 0 \\ \alpha^2\omega_1+0\omega_2+\alpha^2\omega_3 & = \int_{-1}^{1} x^2dx = \frac{2}{3}. \end{cases}$$

c-à-d

$$\begin{cases} \omega_2 & = 2-\omega_1-\omega_3 \\ \omega_1 & = \omega_3 \\ \omega_1 & = \frac{1}{3\alpha^2}. \end{cases}$$

On déduit que, la formule $\mathcal{I}_Q$ est exacte sur $\mathbb{P}_2$ si et seulement si $\omega_1=\omega_3=\frac{1}{3\alpha^2}$ et $\omega_2=2-\frac{2}{3\alpha^2}$.

2 On a

$$\mathcal{I}_Q(x^3)=\omega_1(-\alpha)^3+\omega_1 0+\omega_3\alpha^3=0=\int_{-1}^{1}x^3dx.$$

D'où l'exactitude sur $\mathbb{P}_3$.

3 La formule $\mathcal{I}_Q$ est exacte sur $\mathbb{P}_4$ est équivalent

$$\mathcal{I}_Q(x^4)=\int_{-1}^{1}x^4dx,$$

c-à-d

$$\omega_1(-\alpha)^4+\omega_1 0+\omega_3\alpha^4=\frac{2}{5} \Leftrightarrow \alpha=\sqrt{\frac{3}{5}}.$$

Finalement, on vérifie facilement que :

$$\mathcal{I}_Q(x^5)=\int_{-1}^{1}x^5dx=0.$$

D'où l'exactitude sur $\mathbb{P}_5$.

Exercice 02

Similaire à l'exercice précédent, seulement on augmente le degré.

Exercice 03

$T:\mathcal{C}[-1,1]\to\mathbb{R}$ par $T(f)=\lambda_1 f(x_1)+\lambda_2 f(x_2)$.

1 (α) La formule T est exacte sur les constantes ssi

$$T(1)=\lambda_1+\lambda_2=\int_{-1}^{1}1dx \Longleftrightarrow \lambda_1+\lambda_2=2.$$

(β) La formule T est exacte sur les fonctions affines ssi

$$\begin{cases} T(1)=\lambda_1+\lambda_2=\int_{-1}^{1}1dx \\ T(x)=\lambda_1x_1+\lambda_2x_2=\int_{-1}^{1}xdx \end{cases} \Longleftrightarrow \begin{cases} \lambda_1+\lambda_2=2 \\ \lambda_1x_1+\lambda_2x_2=0 \end{cases}$$

(γ) La formule T est exacte sur les polynômes ≤ 2 ssi

$$\begin{cases} T(1)=\lambda_1+\lambda_2=\int_{-1}^{1}1dx \\ T(x)=\lambda_1x_1+\lambda_2x_2=\int_{-1}^{1}xdx \\ T(x^2)=\lambda_1x_1^2+\lambda_2x_2^2=\int_{-1}^{1}x^2dx \end{cases} \Longleftrightarrow \begin{cases} \lambda_1+\lambda_2=2 \\ \lambda_1x_1+\lambda_2x_2=0 \\ \lambda_1x_1^2+\lambda_2x_2^2=\frac{2}{3}. \end{cases}$$

2 Si on prend, $x_1=-x_2$, on déduit que

$$\begin{cases} \lambda_1=\lambda_2=1 \\ x_2=\frac{1}{\sqrt{3}}. \end{cases}$$

On vérifie facilement que

$$T(x^3)=\lambda_1x_1^3+\lambda_2x_2^3=0=\int_{-1}^{1}x^3dx.$$

Cette formule est dite formule de quadrature de Gauss à deux points

Exercice 04

$$\mathcal{I}_Q(f)=a_0[f(x_0)+f(-x_0)]+a_1[f(1)+f(-1)].$$

1 Cette formule est exacte sur $\mathbb{P}_5$, cela est équivalent à :

$$\begin{cases} a_0+a_0+a_1+a_1=2, \\ a_0x_0-a_0x_0+a_1-a_1=0, \\ a_0x_0^2+a_0x_0^2+a_1+a_1=\frac{2}{3}, \\ a_0x_0^3-a_0x_0^3+a_1-a_1=0, \\ a_0x_0^4+a_0x_0^4+a_1+a_1=\frac{2}{5}, \\ a_0x_0^5-a_0x_0^5+a_1-a_1=0. \end{cases} \Leftrightarrow \begin{cases} a_0+a_1=1, \\ a_0x_0^2+a_1=\frac{1}{3}, \\ a_0x_0^4+a_1=\frac{1}{5}. \end{cases}$$

$$\Leftrightarrow \begin{cases} a_1=1-a_0, \\ a_0-a_0x_0^2=\frac{2}{3}, \\ a_0-a_0x_0^4=\frac{4}{5}. \end{cases} \Leftrightarrow \begin{cases} a_1=1-a_0, \\ a_0(1-x_0^2)=\frac{2}{3}, \\ 1+x_0^2=\frac{6}{5}. \end{cases} \Leftrightarrow \begin{cases} a_0=\frac{5}{6}, \\ a_1=\frac{1}{6}, \\ x_0=\pm\frac{1}{\sqrt{5}}. \end{cases}$$

2 Si on applique cette formule à la fonction $f(x)=\frac{1}{1+x^2}$, on obtient

$$\int_{-1}^{1}\frac{1}{1+x^2}dx=\frac{\pi}{2}\simeq\frac{1}{6}\left(f(-1)+5f(-\frac{1}{\sqrt{5}})+5f(\frac{1}{\sqrt{5}})+f(1)\right)$$
$$=\frac{14}{9}.$$

D'où

$$\pi\simeq\frac{28}{9}=3.1111.$$