Analyse Numérique

Cours: SMP3 et TEER3

Prof: A. Ou-yassine
Université Ibn Zohr, Faculté Polydisciplinaire de Ouarzazate, Maroc

2020/2021

Chapitre 1

Résolution d'un système d'équations linéaires (Partie 1) : méthodes directes

Systèmes linéaires

- Beaucoup de problèmes se réduisent à la résolution numérique d'un système d'équations linéaires

- Deux grandes classes de méthodes :
  1. Méthodes directes : déterminent explicitement la solution après un nombre fini d'opérations arithmétiques
  2. Méthodes itératives (sur $\mathbb{R}$ ou $\mathbb{C}$) : consistent à générer une suite qui converge vers la solution du système

- Autres méthodes non abordées dans ce cours :
  - Méthodes intermédiaires : Splitting, décomposition incomplètes
  - Méthodes probabilistes comme celle de Monte-Carlo

Objet de l'étude

$$(S)\begin{cases} a_{1,1}\,x_1 & + & a_{1,2}\,x_2 & + & \cdots & + & a_{1,n}\,x_n & = & b_1 \\ a_{2,1}\,x_1 & + & a_{2,2}\,x_2 & + & \cdots & + & a_{2,n}\,x_n & = & b_2 \\ \vdots & & & & & & & & \vdots \\ a_{n,1}\,x_1 & + & a_{n,2}\,x_2 & + & \cdots & + & a_{n,n}\,x_n & = & b_n \end{cases}$$

- Données : les $a_{i,j}$ et $b_1, \ldots, b_n$ dans $\mathbb{K}$ avec $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$
- Inconnues : $x_1, \ldots, x_n$ dans $\mathbb{K}$

Écriture matricielle

$$(S) \quad Ax = b,$$

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ a_{n,1} & \cdots & \cdots & a_{n,n} \end{pmatrix} \in \mathbb{M}_{n\times n}(\mathbb{K})$$

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{K}^n, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} \in \mathbb{K}^n$$

- Dans ce chapitre, A **est inversible !**

Motivation (1)

- Pourquoi ce problème se pose-t-il ?
- En effet, les formules de Cramer donnent la solution :

$$\forall i \in \{1, \ldots, n\}, \quad x_i = \frac{\begin{vmatrix} a_{1,1} & \cdots & a_{1,(i-1)} & b_1 & a_{1,(i+1)} & \cdots & a_{1,n} \\ \vdots & & & \vdots & & & \vdots \\ a_{n,1} & \cdots & a_{n,(i-1)} & b_n & a_{n,(i+1)} & \cdots & a_{n,n} \end{vmatrix}}{\det(A)}.$$

- Regardons le nombre d'opérations nécessaires !

Motivation (2)

- Regardons le nombre d'opérations nécessaires !

Lemme

Le nombre d'opérations nécessaires pour résoudre le système à l'aide des formules de Cramer est de $(n+1)(n\,n!-1)$ *opérations à virgule flottante.*

- Lorsque $n = 100$, nombre d'opérations de l'ordre de $9,4.10^{161}$!

$\rightsquigarrow$ Ordi. fonctionnant à 100 megaflops, environ 3.10^{146} années !

$\rightsquigarrow$ Impossible d'utiliser Cramer pour résoudre de grands systèmes !

Résolution d'un système triangulaire

- Idée des méthodes directes : se ramener à la résolution d'1 (ou 2) système triangulaire

- A triangulaire supérieure : (S) s'écrit :

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

- A inversible $\Rightarrow$ les $a_{i,i}$ sont non nuls

$\rightsquigarrow$ Système facile à résoudre : algorithme de substitution rétrograde

Résolution d'un système triangulaire : exemple

- On considère le système triangulaire supérieur :

$$(S)\begin{cases} x_1 + 2x_2 + 5x_3 = 1 \\ -4x_2 - 16x_3 = -\frac{5}{2} \\ -17x_3 = -\frac{17}{8} \end{cases}$$

  - 3ième équation : $x_3 = \frac{1}{8}$
  - 2ième équation : $x_2 = \dfrac{-5/2 + 16x_3}{-4} = \frac{1}{8}$
  - 1ière équation : $x_1 = \dfrac{1 - 2x_2 - 5x_3}{1} = \frac{1}{8}$

- Idem si A triang. inf. : algorithme de substitution progressive

Système triangulaire : opérations et propriétés

Lemme

La résolution d'un système d'équations linéaires triangulaire se fait en n^2 opérations à virgule flottante.

Lemme (Propriétés)

Soient $A, B \in \mathbb{M}_{n\times n}(\mathbb{K})$ deux matrices triangulaires supérieures. On a alors les résultats suivants :

1. *AB est triangulaire supérieur*
2. *Si A et B sont à diagonale unité (i.e., n'ont que des 1 sur la diagonale), alors AB est à diagonale unité*
3. *Si A est inversible, alors A^{-1} est aussi triangulaire supérieure*
4. *Si A est inversible et à diagonale unité, alors A^{-1} est aussi à diagonale unité.*

2 méthodes directes étudiées dans la suite

1. Méthode de Gauss : système $\rightsquigarrow (M\,A)\,x = M\,b$ avec $M\,A$ triang. sup. (sans calculer explicitement M).
   - Associée à la factorisation $A = L\,U$ de la matrice A avec L triang. inf. et U triang. sup., $A\,x = b \Leftrightarrow L\,y = b,\ U\,x = y$
2. Méthode de Cholesky
   - Associée à la factorisation de Cholesky $A = R^T\,R$ avec R triang. sup., $A\,x = b \Leftrightarrow R^T\,y = b,\ R\,x = y$
   - Méthode valable pour A symétrique et définie positive

Méthode de Gauss : description

- (S) : $Ax = b$ avec A inversible
- On pose $b^{(1)} = b$ et $A^{(1)} = A = (a_{i,j}^{(1)}) \rightsquigarrow (S^{(1)})$: $A^{(1)}x = b^{(1)}$

Étape 1

- A inversible $\Rightarrow$ on suppose (quitte à permuter lignes) $a_{1,1}^{(1)} \neq 0$.
C'est le premier pivot de l'élimination de Gauss
- Pour $i = 2, \ldots, n$, on remplace L_i par $L_i - g_{i,1} L_1$ où $g_{i,1} = \frac{a_{i,1}^{(1)}}{a_{1,1}^{(1)}}$

Méthode de Gauss : description

- On obtient alors $(S^{(2)})$: $A^{(2)}x = b^{(2)}$ avec :

$$\begin{cases} a_{1,j}^{(2)} = a_{1,j}^{(1)}, & j = 1, \ldots, n \\ a_{i,1}^{(2)} = 0, & i = 2, \ldots, n \\ a_{i,j}^{(2)} = a_{i,j}^{(1)} - g_{i,1}\, a_{1,j}^{(1)}, & i, j = 2, \ldots, n \\ b_1^{(2)} = b_1^{(1)} & \\ b_i^{(2)} = b_i^{(1)} - g_{i,1}\, b_1^{(1)}, & i = 2, \ldots, n \end{cases}$$

- La matrice $A^{(2)}$ et le vecteur $b^{(2)}$ sont donc de la forme :

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

Méthode de Gauss : description

Étape k

- On a ramené le système à $(S^{(k)})$: $A^{(k)}x = b^{(k)}$ avec

$$A^{(k)} = \begin{pmatrix} a_{1,1}^{(1)} & & \cdots & \cdots & a_{1,k}^{(1)} & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & & & a_{2,k}^{(2)} & \cdots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & & a_{3,k}^{(3)} & \cdots & a_{3,n}^{(3)} \\ \vdots & \ddots & \ddots & \ddots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & a_{k,k}^{(k)} & \cdots & a_{k,n}^{(k)} \\ \vdots & & \vdots & 0 & a_{k+1,k}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & a_{n,k}^{(k)} & \cdots & a_{n,n}^{(k)} \end{pmatrix}$$

Méthode de Gauss : description

- A inversible $\Rightarrow$ on suppose (quitte à permuter lignes) $a_{k,k}^{(k)} \neq 0$.
C'est le *kième pivot* de l'élimination de Gauss

- Par le même principe qu'à l'étape 1 et en utilisant $g_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}$ pour $i > k$, on obtient alors $(S^{(k+1)})$: $A^{(k+1)} x = b^{(k+1)}$ avec

$$A^{(k+1)} = \begin{pmatrix} a_{1,1}^{(1)} & & \cdots & \cdots & a_{1,k+1}^{(1)} & \cdots & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & & & a_{2,k}^{(2)} & \cdots & \cdots & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & & a_{3,k}^{(3)} & \cdots & \cdots & a_{3,n}^{(3)} \\ \vdots & \ddots & \ddots & \ddots & \vdots & & & \vdots \\ 0 & \cdots & 0 & 0 & a_{k,k}^{(k)} & \cdots & \cdots & a_{k,n}^{(k)} \\ \vdots & & \vdots & 0 & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} \\ \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & 0 & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} \end{pmatrix}$$

Méthode de Gauss : description

Étape $n-1$

- Le système $(S^{(n)})$: $A^{(n)}x = b^{(n)}$ obtenu est triangulaire supérieure avec

$$A^{(n)} = \begin{pmatrix} a_{1,1}^{(1)} & & \cdots & \cdots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & & & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & & a_{3,n}^{(3)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & a_{n,n}^{(n)} \end{pmatrix}$$

- On peut le résoudre par l'algorithme de substitution rétrograde

Méthode de Gauss : exemple

$$(S) = (S^{(1)}) \begin{cases} x_1 + 2x_2 + 5x_3 = 1, \\ 3x_1 + 2x_2 - x_3 = \frac{1}{2}, \\ 5x_2 + 3x_3 = 1. \end{cases}$$

- Le premier pivot de l'élimination de Gauss est donc $a_{1,1}^{(1)} = 1$ et on a $g_{2,1}^{(1)} = 3$, $g_{3,1}^{(1)} = 0$. La première étape fournit donc

$$(S^{(2)}) \begin{cases} x_1 + 2x_2 + 5x_3 = 1, \\ -4x_2 - 16x_3 = -\frac{5}{2}, \\ 5x_2 + 3x_3 = 1. \end{cases}$$

Méthode de Gauss : exemple

- Le second pivot de l'élimination de Gauss est donc $a_{2,2}^{(2)} = -4$ et on a $g_{3,2}^{(2)} = -\frac{5}{4}$. On obtient donc le système

$$(S^{(3)}) \begin{cases} x_1 + 2x_2 + 5x_3 = 1, \\ -4x_2 - 16x_3 = -\frac{5}{2}, \\ -17x_3 = -\frac{17}{8}. \end{cases}$$

- Algorithme de substitution rétrograde $\rightsquigarrow x_1 = x_2 = x_3 = \frac{1}{8}$

Remarque

- Au cours de l'exécution de l'élimination de Gauss, si on tombe sur un pivot nul, alors on permute la ligne en question avec une ligne en dessous pour se ramener à un pivot non nul (ceci est toujours possible car A est supposée inversible).

Certains choix de pivots peuvent s'avérer plus judicieux que d'autres.

Lien avec la factorisation LU d'une matrice

Définition

On appelle factorisation LU de A une facto. $A = L\,U$ avec L triang. inf. et U triang. sup. (de la même taille que A).

Lemme

À l'étape k de l'élimination de Gauss, on a $A^{(k+1)} = G_k\,A^{(k)}$ où

$$G_k = \begin{pmatrix} 1 & & (0) & & 0 & \dots & 0 \\ & \ddots & & & \vdots & & \vdots \\ & (0) & & 1 & 0 & \dots & 0 \\ 0 & \dots & 0 & -g_{k+1,k} & 1 & & (0) \\ \vdots & & \vdots & \vdots & & \ddots & \\ 0 & \dots & 0 & -g_{n,k} & (0) & & 1 \end{pmatrix}, \quad g_{i,k} = \frac{a_{i,k}^{(k)}}{a_{k,k}^{(k)}}$$

On a de plus $b^{(k+1)} = G_k\,b^{(k)}$.

Lien avec la factorisation LU d'une matrice

Définition

Soit $A \in \mathbb{M}_{n\times n}(\mathbb{K})$. Les mineurs fondamentaux D_k, $k = 1, \ldots, n$ de A sont les déterminants des sous-matrices de A formées par les k premières lignes et les k premières colonnes de A :

$$D_k = \det\left((a_{i,j})_{1\leq i,j\leq k}\right), \quad k = 1, \ldots, n.$$

Théorème

Soit $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ une matrice carrée inversible. Les propriétés suivantes sont équivalentes :

(i) *L'élimination de Gauss s'effectue sans permutation de lignes ;*

(ii) *Il existe $L \in \mathbb{M}_{n\times n}(\mathbb{K})$ triangulaire inférieure inversible et $U \in \mathbb{M}_{n\times n}(\mathbb{K})$ triangulaire supérieure inversible telles que $A = LU$;*

(iii) *Tous les mineurs fondamentaux de A sont non nuls.*

Lien avec la factorisation LU d'une matrice

Lemme

Avec les notations précédentes, on a

$$(G_{n-1}\, G_{n-2} \cdots G_1)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & \cdots & 0 \\ g_{2,1} & 1 & \ddots & & \vdots \\ g_{3,1} & g_{3,2} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ g_{n,1} & g_{n,2} & \cdots & g_{n,n-1} & 1 \end{pmatrix}.$$

Lien avec la factorisation LU d'une matrice

Corollaire

Soit $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ une matrice carrée inversible. Si tous les mineurs fondamentaux de A sont non nuls, alors avec les notations précédentes, l'élimination de Gauss fournit la factorisation LU de A suivante :

$$A = \begin{pmatrix} 1 & 0 & \dots & \dots & 0 \\ g_{2,1} & 1 & \ddots & & \vdots \\ g_{3,1} & g_{3,2} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ g_{n,1} & g_{n,2} & \dots & g_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} a_{1,1}^{(1)} & & \dots & \dots & a_{1,n}^{(1)} \\ 0 & a_{2,2}^{(2)} & & & a_{2,n}^{(2)} \\ 0 & 0 & a_{3,3}^{(3)} & & a_{3,n}^{(3)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & 0 & a_{n,n}^{(n)} \end{pmatrix}.$$

- Remarque : la matrice L obtenue est à diagonale unité.

Factorisation LU : exemple

Pour la matrice du système

$$(S) : \begin{cases} x_1 + 2x_2 + 5x_3 = 1 \\ 3x_1 + 2x_2 - x_3 = \frac{1}{2} \\ 5x_2 + 3x_3 = 1 \end{cases}$$

on a :

$$\underbrace{\begin{pmatrix} 1 & 2 & 5 \\ 3 & 2 & -1 \\ 0 & 5 & 3 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ 0 & -\frac{5}{4} & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 1 & 2 & 5 \\ 0 & -4 & -16 \\ 0 & 0 & -17 \end{pmatrix}}_{U}$$

Lien avec la factorisation LU d'une matrice

Proposition

Soit $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ une matrice carrée inversible admettant une factorisation LU. Alors il existe une unique factorisation LU de A avec L à diagonale unité.

• Lorsque A admet une factorisation LU, la résolution du système d'équations linéaires (S) : $Ax = b$ se ramène à la résolution de deux systèmes linéaires triangulaires. En effet :

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

• En pratique, on résout donc d'abord $Ly = b$ puis connaissant y on résout $Ux = y$.

Coût de l'algorithme de Gauss

Lemme

Soit $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ une matrice carrée inversible. Résoudre un système linéaire (S) : $Ax = b$ via l'élimination de Gauss nécessite un nombre d'opérations à virgule flottante équivalent à $\frac{2n^3}{3}$ lorsque n tend vers l'infini. Ce coût asymptotique est aussi celui du calcul de la factorisation LU de A.

- Pour $n = 100$, cela donne $6,6.10^5$ opérations à virgule flottante à comparer à $9,4.10^{161}$ avec Cramer

- Avec un ordinateur fonctionnant à 100 megaflops, cela prendra moins de 7 millièmes de secondes. À comparer avec 3.10^{146} années pour Cramer

Méthode de Cholesky

- Alternative à Gauss pour matrices symétriques et définies positives

Définition

Une matrice $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ est dite symétrique si elle est égale à sa transposée, i.e., $A^T = A$.

Définition

Soit $\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$. Le produit scalaire canonique sur $\mathbb{K}^n$ est défini comme l'application $\langle .\,,\, . \rangle : \mathbb{K}^n \times \mathbb{K}^n \to \mathbb{K}$, $(u, v) \mapsto \langle u, v \rangle$ qui vérifie :

- *Si $\mathbb{K} = \mathbb{R}$, $\langle u, v \rangle = v^T\, u = \sum_{i=1}^{n} u_i\, v_i$ (produit scalaire euclidien),*
- *Si $\mathbb{K} = \mathbb{C}$, $\langle u, v \rangle = \overline{v}^T\, u = \sum_{i=1}^{n} u_i\, \overline{v_i}$ (produit scalaire hermitien).*

Méthode de Cholesky

Définition

Une matrice $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ est dite définie positive, resp. semi définie positive si pour tout $x \in \mathbb{R}^n$ non nul, on a $\langle Ax, x\rangle > 0$, resp. $\langle Ax, x\rangle \geq 0$.

1. Une matrice définie positive est inversible ;
2. Si $A \in \mathbb{M}_{n\times n}(\mathbb{K})$ est inversible, alors $A^T A$ est symétrique et définie positive ;
3. Si $A = (a_{i,j}) \in \mathbb{M}_{n\times n}(\mathbb{K})$ est définie positive, alors $a_{i,i} > 0$ pour tout $i = 1, \ldots, n$.

Théorème

Une matrice réelle $A \in \mathbb{M}_{n\times n}(\mathbb{R})$ est symétrique définie positive ssi il existe une matrice $L = (l_{i,j})_{1\leq i,j\leq n} \in \mathbb{M}_{n\times n}(\mathbb{R})$ triangulaire inférieure inversible telle que $A = L\,L^T$. De plus, si pour tout $i = 1, \ldots, n$, $l_{i,i} \geq 0$, alors L est unique.

Algorithme de Cholesky

Entrée : $A = (a_{i,j})_{1 \leq i,j \leq n} \in \mathbb{M}_{n \times n}(\mathbb{R})$ symétrique et définie positive.

Sortie : $L = (l_{i,j})_{1 \leq i,j \leq n} \in \mathbb{M}_{n \times n}(\mathbb{R})$ tel que $A = L\,L^T$.

1. $l_{1,1} = \sqrt{a_{1,1}}$;
2. Pour i de 2 à n par pas de 1, faire :
   - $l_{i,1} = \frac{a_{i,1}}{l_{1,1}}$;
3. Pour j de 2 à n par pas de 1, faire :
   - Pour i de 1 à $j-1$ par pas de 1, faire :
     $$l_{i,j} = 0 ;$$
   - $l_{j,j} = \sqrt{a_{j,j} - \sum_{k=1}^{j-1} l_{j,k}^2}$;
   - Pour i de $j+1$ à n par pas de 1, faire :
     $$l_{i,j} = \frac{a_{i,j} - \sum_{k=1}^{j-1} l_{i,k}\, l_{j,k}}{l_{j,j}} ;$$
4. Retourner $L = (l_{i,j})_{1 \leq i,j \leq n} \in \mathbb{M}_{n \times n}(\mathbb{R})$.

Coût de l'algorithme de Cholesky

Proposition

L'algorithme de Cholesky décrit ci-dessus nécessite n extractions de racines carrées et un nombre d'opérations à virgule flottante équivalent à $\frac{n^3}{3}$ lorsque n tend vers l'infini.

- Asymptotiquement, presque deux fois moins d'opérations à virgule flottante que pour LU

$\rightsquigarrow$ Il est conseillé de l'utiliser lorsque A est réelle symétrique et définie positive

Chapitre 2

Résolution d'un système d'équations linéaires (Partie 2) : méthodes itératives

Modèle général d'un schéma itératif

- $A \in \mathbb{M}_{n\times n}(\mathbb{K})$, $b \in \mathbb{K}^n$ et (S) : $Ax = b$

- Principe général : générer une suite de vecteurs qui converge vers la solution $A^{-1}\,b$

- Idée : écrire (S) sous une forme équivalente permettant de voir la solution comme un point fixe :

$$(S) \Longleftrightarrow Bx + c = x$$

$B \in \mathbb{M}_{n\times n}(\mathbb{K})$ et $c \in \mathbb{K}^n$ bien choisis cad $\mathbb{I} - B$ inversible et $c = (\mathbb{I} - B)\,A^{-1}\,b$

- Exemple : $A = M - N$ (M inversible), $B = M^{-1}\,N$ et $c = M^{-1}\,b$

Modèle général d'un schéma itératif

• On se donne alors $x^{(0)} \in \mathbb{K}^n$ et on construit une suite de vecteurs $x^{(k)} \in \mathbb{K}^n$ à l'aide du schéma itératif

$$(\star) \qquad x^{(k+1)} = B\,x^{(k)} + c, \quad k = 1, 2, \ldots.$$

• Si $(x^{(k)})_{k\in\mathbb{N}}$ est convergente, alors elle converge vers la solution $A^{-1}\,b$ de (S)

Convergence

Définition

Une méthode itérative définie par $(x^{(k)})_{k\in\mathbb{N}}$ pour résoudre un système $Ax = b$ est dite convergente si pour toute valeur initiale $x^{(0)} \in \mathbb{K}^n$, on a $\lim_{k\to+\infty} x^{(k)} = A^{-1}\, b$.

Lemme

Si la méthode itérative est convergente et si on note $x = A^{-1}\, b$ la solution, alors

$$x^{(k)} - x = B^k(x^{(0)} - x).$$

• $x^{(k)} - x$ erreur à la k-ième itération $\rightsquigarrow$ estimation de cette erreur en fonction de l'erreur initiale

Convergence

Théorème

Les assertions suivantes sont équivalentes :

(i) $(\star)$ *est convergente ;*

(ii) *Pour tout* $y \in \mathbb{K}^n$, $\lim_{k\to+\infty} B^k y = 0$;

(iv) *Pour toute norme matricielle* $\|.\|$ *sur* $\mathbb{M}_{n\times n}(\mathbb{K})$, *on a* $\lim_{k\to+\infty} \|B^k\| = 0$.

- En pratique, caractérisations difficiles à vérifier $\rightsquigarrow$

Théorème

Les assertions suivantes sont équivalentes :

(i) $(\star)$ *est convergente ;*

(ii) $\rho(B) < 1$, *où* $\rho(B)$ *désigne* le rayon spectral de la matrice B ;

(iii) *Il existe une norme matricielle* $\|.\|$ *sur* $\mathbb{M}_{n\times n}(\mathbb{K})$ *subordonnée à une norme vectorielle* $\|.\|$ *sur* $\mathbb{K}^n$ *telle que* $\|B\| < 1$.

Vitesse de convergence

Définition

Considérons un schéma itératif $(\star)$ convergent. Soit $\|.\|$ une norme matricielle sur $\mathbb{M}_{n\times n}(\mathbb{K})$ et soit k un entier tel que $\|B^k\| < 1$. On appelle taux moyen de convergence associé à la norme $\|.\|$ pour k itérations de $x^{(k+1)} = B\,x^{(k)} + c$ le nombre positif

$$R_k(B) = -\ln\left(\left[\|B^k\|\right]^{\frac{1}{k}}\right).$$

Définition

Considérons deux méthodes itératives convergentes

$$(1) \quad x^{(k+1)} = B_1\,x^{(k)} + c_1, \qquad k = 1, 2, \ldots,$$
$$(2) \quad x^{(k+1)} = B_2\,x^{(k)} + c_2, \qquad k = 1, 2, \ldots.$$

Soit k un entier tel que $\|B_1^k\| < 1$ et $\|B_2^k\| < 1$. On dit que (1) est plus rapide que (2) relativement à la norme $\|.\|$ si $R_k(B_1) \geq R_k(B_2)$.

Les méthodes itératives classiques

- (S) : $A\,x = b$ avec A inversible

- Idée : déduire un schéma itératif d'une décomposition $A = M - N$, M inversible

- En pratique, on suppose que les systèmes de matrice M sont faciles à résoudre (par ex. M diagonale, triangulaire, ...)

- (S) s'écrit alors $M\,x = N\,x + b$ cad $x = B\,x + c$ avec $B = M^{-1}\,N$ et $c = M^{-1}\,b$ et on considère le schéma itératif associé :

$$x^{(0)} \in \mathbb{K}^n, \quad M\,x^{(k+1)} = N\,x^{(k)} + b.$$

- On montre alors que $\mathbb{I} - B$ inversible et $c = (\mathbb{I} - B)\,A^{-1}\,b$

Trois exemples classiques

- Dans ce cours, 3 exemples classiques : les méthodes de Jacobi, Gauss-Seidel et de relaxation

- Point de départ : décomposition de $A = (a_{i,j})_{1 \leq i,j \leq n}$ sous la forme $A = D - E - F$ avec :
  - $D = (d_{i,j})_{1 \leq i,j \leq n}$ diagonale, telle que $d_{i,i} = a_{i,i}$ et $d_{i,j} = 0$ pour $i \neq j$;
  - $E = (e_{i,j})_{1 \leq i,j \leq n}$ triangulaire inférieure **stricte** telle que $e_{i,j} = -a_{i,j}$ si $i > j$ et $e_{i,j} = 0$ si $i \leq j$;
  - $F = (f_{i,j})_{1 \leq i,j \leq n}$ triangulaire supérieure **stricte** telle que $f_{i,j} = -a_{i,j}$ si $i < j$ et $f_{i,j} = 0$ si $i \geq j$;

Exemple de décomposition $A = D - E - F$

$$\underbrace{\begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{pmatrix}}_{D} - \underbrace{\begin{pmatrix} 0 & 0 & 0 \\ -2 & 0 & 0 \\ 1 & 1 & 0 \end{pmatrix}}_{E} - \underbrace{\begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix}}_{F}$$

Trois exemples classiques

- On suppose D inversible
  - Méthode de Jacobi : $M = D$, $N = E + F$;
  - Méthode de Gauss-Seidel : $M = D - E$, $N = F$;
  - Méthode de relaxation : $M = \frac{1}{\omega}(D - \omega E)$, $N = \left(\frac{1-\omega}{\omega}\right) D + F$ avec ω paramètre réel non nul.
- Gauss-Seidel est un cas particulier de relaxation pour $\omega = 1$.

Méthode de Jacobi : description

- (S) : $A\,x = b$ avec A inversible
- $A = M - N$ avec $M = D$ inversible et $N = E + F$
- Le schéma itératif s'écrit alors

$$D\,x^{(k+1)} = (E+F)\,x^{(k)} + b \Longleftrightarrow x^{(k+1)} = D^{-1}\,(E+F)\,x^{(k)} + D^{-1}\,b$$

Définition

La matrice $B_J = D^{-1}\,(E+F)$ s'appelle la matrice de Jacobi associée à A.

Jacobi :

Pour calculer $x^{(k+1)}$ à partir de $x^{(k)}$

- On a $D\,x^{(k+1)} = (E+F)\,x^{(k)} + b$ donc pour tout $i = 1, \ldots, n$, $(D\,x^{(k+1)})_i = ((E+F)\,x^{(k)})_i + b_i$ cad

$$a_{i,i}\,x_i^{(k+1)} = -\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j}\,x_j^{(k)} + b_i$$

$$\Longleftrightarrow x_i^{(k+1)} = \frac{1}{a_{i,i}}\left[(-\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j}\,x_j^{(k)}) + b_i\right].$$

Jacobi : convergence et exemple

Théorème

La méthode de Jacobi converge si et seulement si $\rho(B_J) < 1$.

- Exemple : pour la matrice $A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$ précédente :

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

- Valeurs propres : 0 et $\pm \frac{i\sqrt{5}}{2}$ donc $\rho(B_J) = \frac{\sqrt{5}}{2} > 1$ et la méthode de Jacobi diverge

Méthode de Gauss-Seidel : description

- (S) : $Ax = b$ avec A inversible
- $A = M - N$ avec $M = D - E$ inversible et $N = F$
- Le schéma itératif s'écrit alors

$$(D-E)\,x^{(k+1)} = F\,x^{(k)}+b \iff x^{(k+1)} = (D-E)^{-1}\,F\,x^{(k)}+(D-E)^{-1}\,b$$

Définition

La matrice $B_{GS} = (D-E)^{-1}\,F$ s'appelle la matrice de Gauss-Seidel associée à A.

Gauss-Seidel : description

Pour calculer $x^{(k+1)}$ à partir de $x^{(k)}$

- On a $(D-E)\,x^{(k+1)} = F\,x^{(k)} + b$ donc pour tout $i = 1, \ldots, n$, $\left((D-E)\,x^{(k+1)}\right)_i = \left(F\,x^{(k)}\right)_i + b_i$ c'est-à-dire

$$a_{i,i}\,x_i^{(k+1)} + \sum_{j=1}^{i-1} a_{i,j}\,x_j^{(k+1)} = -\sum_{j=i+1}^{n} a_{i,j}\,x_j^{(k)} + b_i,$$

Gauss-Seidel : description

ce qui entraîne

$$x_1^{(k+1)} = \frac{1}{a_{1,1}} \left[- \sum_{j=2}^{n} a_{1,j}\, x_j^{(k)} + b_1 \right],$$

et pour $i = 2, \ldots, n$,

$$x_i^{(k+1)} = \frac{1}{a_{i,i}} \left[- \sum_{j=1}^{i-1} a_{i,j}\, x_j^{(k)} + b_i - \sum_{j=i+1}^{n} a_{i,j}\, x_j^{(k)} \right].$$

Gauss-Seidel : convergence et exemple

Théorème

La méthode de Gauss-Seidel converge si et seulement si $\rho(B_{GS}) < 1$.

- Exemple : pour la matrice $A = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$ précédente :

$$B_{GS} = \begin{pmatrix} 2 & 0 & 0 \\ 2 & 2 & 0 \\ -1 & -1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix},$$

$$B_{GS} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ -\frac{1}{2} & \frac{1}{2} & 0 \\ 0 & \frac{1}{4} & \frac{1}{2} \end{pmatrix} \begin{pmatrix} 0 & 1 & -1 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}.$$

- Valeurs propres : 0 et $-\frac{1}{2}$ (mult. 2) donc $\rho(B_{GS}) = \frac{1}{2} < 1$ et Gauss-Seidel converge

Cas particulier : matrice symétrique définie positive

Théorème

Soit A une matrice symétrique définie positive et écrivons $A = M - N$ avec M inversible et $M^T + N$ définie positive. Alors la méthode itérative

$$x^{(0)} \in \mathbb{K}^n, \quad x^{(k+1)} = M^{-1} N x^{(k)} + M^{-1} b,$$

converge.

Corollaire

Soit A une matrice symétrique définie positive. Alors la méthode de Gauss-Seidel converge.

Cas particulier : matrice à diagonale strictement dominante

Définition

Une matrice $A = (a_{i,j})_{1 \leq i,j \leq n}$ est dite à diagonale strictement dominante si :

$$\forall\, i = 1, \ldots, n, \quad |a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}|.$$

Théorème

Soit A une matrice à diagonale strictement dominante. Alors A est inversible et les méthodes de Jacobi et de Gauss-Seidel convergent toutes les deux.

Chapitre 3

Résolution d'équations et de systèmes d'équations non linéaires

Problème

- $f : \mathbb{R} \to \mathbb{R}$ fonction d'une seule variable réelle

- On cherche à résoudre l'équation $f(x) = 0$ = trouver une valeur approchée $\overline{x}$ d'un réel $\tilde{x}$ vérifiant $f(\tilde{x}) = 0$.

- Mise en oeuvre pratique : on se donne une tolérance sur la solution cherchée. L'algorithme numérique utilisé doit alors avoir un critère d'arrêt dépendant de cette tolérance et nous assurant que la solution calculée a bien la précision recherchée

- 2 possibilités :
  - on sait à l'avance combien d'étapes de l'algorithme sont nécessaires
  - à chaque étape, on vérifie une condition nous permettant d'arrêter le processus après un nombre fini d'étapes

Vitesse de convergence

Définition

Soit $(x_n)_{n\in\mathbb{N}}$ une suite convergente et soit $\tilde{x}$ sa limite.

1. *On dit que la convergence de $(x_n)_{n\in\mathbb{N}}$ est linéaire de facteur $K \in]0,1[$ s'il existe $n_0 \in \mathbb{N}$ tel que, pour tout $n \geq n_0$, $|x_{n+1} - \tilde{x}| \leq K\,|x_n - \tilde{x}|$.*
2. *On dit que la convergence de $(x_n)_{n\in\mathbb{N}}$ est superlinéaire d'ordre $p \in \mathbb{N}$, $p > 1$ s'il existe $n_0 \in \mathbb{N}$ et $K > 0$ tels que, pour tout $n \geq n_0$, $|x_{n+1} - \tilde{x}| \leq K\,|x_n - \tilde{x}|^p$. Si $p = 2$, on parle de convergence quadratique et si $p = 3$ on parle de convergence cubique.*

- Remarque : K n'est pas unique.

- En pratique il peut être difficile de prouver la convergence d'une méthode d'autant plus qu'il faut tenir compte des erreurs d'arrondis.

Vitesse de convergence

Définition

Soit $(x_n)_{n\in\mathbb{N}}$ une suite convergent vers une limite $\tilde{x}$. On dit que la convergence de $(x_n)_{n\in\mathbb{N}}$ est linéaire de facteur K (resp. superlinéaire d'ordre p) s'il existe une suite $(y_n)_{n\in\mathbb{N}}$ convergent vers 0, linéaire de facteur K (resp. superlinéaire d'ordre p) au sens de la définition précédente telle que $|x_n - \tilde{x}| \leq y_n$.

- $d_n = -\log_{10}(|x_n - \tilde{x}|)$ *mesure* du nbre de décimales exactes de x_n.

$\rightsquigarrow$ Convergence d'ordre $p \Rightarrow$ asymptotiquement, on a $|x_{n+1} - \tilde{x}| \sim K\,|x_n - \tilde{x}|^p$ d'où $-d_{n+1} \sim \log_{10}(K) - p\,d_n$ et donc asymptotiquement x_{n+1} a p fois plus de décimales exactes que x_n

$\rightsquigarrow$ l'ordre p représente asymptotiquement le facteur multiplicatif du nombre de décimales exactes que l'on gagne à chaque itération

$\rightsquigarrow$ Nous avons donc intérêt à ce qu'il soit le plus grand possible.

Méthode de dichotomie : principe

- Méthode de localisation des racines d'une équation $f(x) = 0$ basée sur le théorème des valeurs intermédiaires

Si f est continue sur $[a, b]$ et $f(a)\,f(b) < 0$, alors il existe $\tilde{x} \in]a, b[$ tel que $f(\tilde{x}) = 0$.

- Principe :
  1. On part d'un intervalle $[a, b]$ vérifiant la propriété $f(a)\,f(b) < 0$
  2. On le scinde en deux intervalles $[a, c]$ et $[c, b]$ avec $c = \frac{a+b}{2}$
  3. On teste les bornes des nouveaux intervalles (on calcule $f(a)\,f(c)$ et $f(c)\,f(b)$) pour en trouver un (au moins) qui vérifie encore la propriété, *i.e.*, $f(a)\,f(c) < 0$ ou/et $f(c)\,f(b) < 0$.
  4. On itère ensuite ce procédé un certain nombre de fois dépendant de la précision que l'on recherche sur la solution.

Méthode de dichotomie : algorithme

Entrées : la fonction f, $(a, b) \in \mathbb{R}^2$ tels que f est continue sur $[a, b]$ et $f(a)\,f(b) < 0$ et la précision ϵ.
Sortie : x_{k+1} valeur approchée de $\tilde{x}$ solution de $f(\tilde{x}) = 0$ à ϵ près.

1. $x_0 \leftarrow a$, $y_0 \leftarrow b$;
2. Pour k de 0 à $E\left(\dfrac{\ln(b-a) - \ln(\epsilon)}{\ln(2)}\right)$ par pas de 1, faire :
   - Si $f(x_k)\,f\left(\dfrac{x_k + y_k}{2}\right) > 0$, alors $x_{k+1} \leftarrow \dfrac{x_k + y_k}{2}$, $y_{k+1} \leftarrow y_k$;
   - Si $f(x_k)\,f\left(\dfrac{x_k + y_k}{2}\right) < 0$, alors $x_{k+1} \leftarrow x_k$, $y_{k+1} \leftarrow \dfrac{x_k + y_k}{2}$;
   - Sinon retourner $\dfrac{x_k + y_k}{2}$;
3. Retourner x_{k+1}.

Méthode de dichotomie

- Remarques sur l'algorithme précédent :
  - Il construit une suite de segments emboîtés contenant tous $\tilde{x}$
  - À chaque passage dans la boucle : une évaluation de f
  - En pratique, avec les arrondis, > 0 et < 0 ne veulent rien dire !

Théorème

Le nombre minimum d'itérations de la méthode de dichotomie nécessaire pour approcher $\tilde{x}$ à ϵ près est $E\left(\frac{\ln(b-a)-\ln(\epsilon)}{\ln(2)}\right)+1$, où $E(x)$ désigne la partie entière d'un réel x.

Proof.

$\frac{b-a}{2^n} \leq \epsilon \Leftrightarrow n \geq \frac{\ln(b-a)-\ln(\epsilon)}{\ln(2)}$. □

Proposition

La convergence de la dichotomie est linéaire de facteur $\frac{1}{2}$.

Méthode du point fixe

Définition

Soit $g : \mathbb{R} \to \mathbb{R}$. *On dit que* $x \in \mathbb{R}$ *est un point fixe de* g *si* $g(x) = x$.

• Principe : associer à l'équation $f(x) = 0$ une équation de point fixe $g(x) = x$ de sorte que trouver une solution de $f(x) = 0$ équivaut à trouver un point fixe de g.

Lemme

Soit $(x_n)_{n\in\mathbb{N}}$ *la suite définie par* $x_0 \in \mathbb{R}$ *donné et* $x_{n+1} = g(x_n)$. *Si* $(x_n)_{n\in\mathbb{N}}$ *est convergente et* g *est continue, alors la limite de* $(x_n)_{n\in\mathbb{N}}$ *est un point fixe de* g.

Fonctions contractantes

Définition

Soit $g : \Omega \subseteq \mathbb{R} \to \mathbb{R}$. *On dit que* g *est lipschitzienne sur* Ω *de constante de Lipschitz* γ *(ou* γ*-lipschitzienne) si pour tout* $(x, y) \in \Omega^2$, *on a* $|g(x) - g(y)| \leq \gamma |x - y|$. *On dit que* g *est strictement contractante sur* Ω *si* g *est* γ*-lipschitzienne sur* Ω *avec* $\gamma < 1$.

Proposition

Soit g *une fonction dérivable sur l'intervalle* $[a, b]$. *Si sa dérivée* g' *vérifie* $\max_{x \in [a,b]} |g'(x)| = L < 1$, *alors* g *est strictement contractante sur* $[a, b]$ *de constante de Lipschitz* L.

Théorème du point fixe

Théorème

Soit g une application strictement contractante sur un intervalle $[a, b] \subset \mathbb{R}$ de constante de Lipschitz $\gamma < 1$. Supposons que l'intervalle $[a, b]$ soit stable sous g, i.e., $g([a, b]) \subseteq [a, b]$ ou encore pour tout $x \in [a, b]$, $g(x) \in [a, b]$. Alors g admet un unique point fixe $x^* \in [a, b]$ et la suite définie par $x_{n+1} = g(x_n)$ converge linéairement de facteur γ vers x^* pour tout point initial $x_0 \in [a, b]$. De plus,

$$\forall\, n \in \mathbb{N},\ |x_n - x^*| \leq \frac{\gamma^n}{1-\gamma} |x_1 - x_0|.$$

- Erreur d'autant plus petite que γ est proche de 0
- De plus $\forall\, n \in \mathbb{N},\ |x_n - x^*| \leq \frac{\gamma}{1-\gamma} |x_n - x_{n-1}|$

Si $\gamma \leq \frac{1}{2}$, alors $|x_n - x^*| \leq |x_n - x_{n-1}| \rightsquigarrow$ test d'arrêt

$|x_n - x_{n-1}| < \epsilon$ qui certifiera une précision ϵ sur le résultat

Mise en oeuvre

Proposition

Soit $x^* \in [a, b]$ *un point fixe d'une fonction* $g \in C^1([a, b])$.

- *Si* $|g'(x^*)| < 1$, *alors il existe un intervalle* $[\alpha, \beta] \subseteq [a, b]$ *contenant* x^* *pour lequel la suite définie par* $x_0 \in [\alpha, \beta]$ *et* $x_{n+1} = g(x_n)$ *converge vers* x^* ;
- *Si* $|g'(x^*)| > 1$, *alors pour tout* $x_0 \neq x^*$, *la suite définie par* x_0 *et* $x_{n+1} = g(x_n)$ *ne converge pas vers* x^* ;
- *Si* $|g'(x^*)| = 1$, *on ne peut pas conclure.*

Convergence

Proposition

On considère l'équation $g(x) = x$ *où* g *est une fonction au moins* $p+1$ *fois dérivable avec* $p \geq 1$. *Supposons que les hypothèses du théorème du point fixe soient vérifiées de sorte que* g *admette un unique point fixe* $x^* \in [a, b]$. *Si* $g'(x^*) = g''(x^*) = \cdots = g^{(p)}(x^*) = 0$ *et* $g^{(p+1)}(x^*) \neq 0$, *alors la convergence de la méthode* $x_{n+1} = g(x_n)$ *est superlinéaire d'ordre* $p+1$.

Méthode de Newton

- Revenons à

$$\forall x \in [a,b], \quad |1 - \lambda f'(x)| < \gamma < 1$$

- La méthode convergera d'autant plus vite que γ est petite

$\rightsquigarrow$ Idée : poser $\lambda = \frac{1}{f'(x)}$ cad $g(x) = x - \frac{f(x)}{f'(x)}$.

Définition

La fonction d'itération de Newton associée à l'équation $f(x) = 0$ sur $[a,b]$ est

$$\mathcal{N} : \begin{cases} [a,b] & \to & \mathbb{R}, \\ x & \mapsto & \mathcal{N}(x) = x - \dfrac{f(x)}{f'(x)}. \end{cases}$$

Cette fonction est définie pour f dérivable sur $[a,b]$ et telle que f' ne s'annule pas sur $[a,b]$.

Convergence locale

Théorème

Soit f une fonction de classe C^2 sur un intervalle $[a, b]$ de $\mathbb{R}$. On suppose qu'il existe $\tilde{x} \in [a, b]$ tel que $f(\tilde{x}) = 0$ et $f'(\tilde{x}) \neq 0$ ($\tilde{x}$ est un zéro simple de f). Alors il existe $\epsilon > 0$, tel que pour tout $x_0 \in [\tilde{x} - \epsilon, \tilde{x} + \epsilon]$, la suite des itérés de Newton donnée par $x_{n+1} = \mathcal{N}(x_n)$ pour $n \geq 1$ est bien définie, reste dans l'intervalle $[\tilde{x} - \epsilon, \tilde{x} + \epsilon]$ et converge vers $\tilde{x}$ quand n tend vers l'infini. De plus, cette convergence est (au moins) quadratique.

Zéro multiple et convergence globale

Théorème

Avec les notations, précédentes, si $\tilde{x}$ est un zéro de multiplicité m de f, i.e., $f(x^*) = f'(x^*) = \cdots = f^{(m-1)}(x^*) = 0$ et $f^{(m)}(x^*) \neq 0$, alors la méthode itérative définie par $x_{n+1} = \mathcal{N}_m(x_n)$ avec $\mathcal{N}_m(x_n) = x - m\dfrac{f(x)}{f'(x)}$ est d'ordre supérieure ou égal à 2.

Théorème

Soit $f \in \mathcal{C}^2([a,b])$ vérifiant :

- $f(a)\,f(b) < 0$,
- $\forall\, x \in [a,b],\ f'(x) \neq 0$,
- $\forall\, x \in [a,b],\ f''(x) \neq 0$.

Alors, en choisissant $x_0 \in [a,b]$ tel que $f(x_0)\,f''(x_0) > 0$, la suite $(x_n)_{n\in\mathbb{N}}$ définie par x_0 et $x_{n+1} = \mathcal{N}(x_n)$ converge vers l'unique solution de $f(x) = 0$ dans $[a,b]$.

Méthode de la sécante

- Newton nécessite le calcul de la dérivée de la fonction f qui peut s'avérer difficile

- Idée : remplacer la dérivée f' de f qui apparait dans la méthode de Newton par une différence divisée

Définition

Pour tout $k \in \mathbb{N}$, on appelle différence divisée d'ordre k de f associée à la suite de points deux à deux distincts $(x_j)_{j\in\mathbb{N}}$ la quantité $f[x_0, x_1, \ldots, x_k]$ définie par :

$$f[x_0] = f(x_0), \quad \forall k \in \mathbb{N}^*, f[x_0, x_1, \ldots, x_k] = \frac{f[x_0, x_1, \ldots, x_{k-1}] - f[x_1, x_2, \ldots, x_k]}{x_0 - x_k}.$$

$$\rightsquigarrow x_{n+1} = x_n - \frac{f(x_n)}{f[x_n, x_{n-1}]}, \quad \text{où} \quad f[x_n, x_{n-1}] = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

Convergence

Théorème

Soit f une fonction de classe C^2 sur un intervalle $[a, b]$ de $\mathbb{R}$. On suppose qu'il existe $\tilde{x} \in [a, b]$ tel que $f(\tilde{x}) = 0$ et $f'(\tilde{x}) \neq 0$ ($\tilde{x}$ est un zéro simple de f). Alors il existe $\epsilon > 0$, tel que pour tout $x_0, x_1 \in [\tilde{x} - \epsilon, \tilde{x} + \epsilon]$, la suite des itérés de la méthode de la sécante est bien définie, reste dans l'intervalle $[\tilde{x} - \epsilon, \tilde{x} + \epsilon]$ et converge vers $\tilde{x}$ quand n tend vers l'infini. De plus, cette convergence est d'ordre $p = \frac{1+\sqrt{5}}{2} \approx 1,618$ (nombre d'or).

Systèmes d'équations non linéaires

$$f : \begin{cases} \mathbb{R}^n & \to & \mathbb{R}^n \\ x = (x_1 \dots x_n)^T & \mapsto & f(x) = (f_1(x_1, \dots, x_n), \dots, f_n(x_1, \dots, x_n))^T. \end{cases}$$

On cherche donc un vecteur $x = (x_1 \dots x_n)^T \in \mathbb{R}^n$ tel que

$$f(x) = 0_{\mathbb{R}^n} \iff \begin{cases} f_1(x_1, \dots, x_n) & = & 0, \\ \vdots & & \\ f_n(x_1, \dots, x_n) & = & 0. \end{cases}$$

- Méthode 1 vu précédemment se généralise :

$$x^{(n+1)} = x^{(n)} + M^{-1} f(x^{(n)}),$$

où M est une certaine matrice, et nous avons les mêmes résultats de convergence que dans le cas d'une seule équation.

Matrice Jacobienne

Définition

La matrice jacobienne d'une fonction $f : \mathbb{R}^n \to \mathbb{R}^n$ notée J_f est définie (lorsqu'elle existe) par :

$$\forall x = (x_1 \dots x_n)^T \in \mathbb{R}^n, \quad J_f(x) = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \frac{\partial f_n}{\partial x_2}(x) & \dots & \frac{\partial f_n}{\partial x_n}(x) \end{pmatrix}.$$

Méthode de Newton

- Méthode de Newton se généralise : $x^{(0)} \in \mathbb{R}^n$ et

$$x^{(n+1)} = x^{(n)} - J_f\left(x^{(n)}\right)^{-1} f(x^{(n)}),$$

où $J_f\left(x^{(n)}\right)^{-1}$ désigne l'inverse de la matrice jacobienne de f évaluée en $x^{(n)}$.

Théorème

Soit $f : \mathbb{R}^n \to \mathbb{R}^n$ une fonction de classe $\mathcal{C}^2$ sur une boule fermée B de $\mathbb{R}^n$. On suppose qu'il existe un zéro $\tilde{x}$ de f dans B et que $J_f(\tilde{x})$ est inversible. Alors il existe $\epsilon > 0$ tel que pour tout $x^{(0)} \in B$ tel que $\|x^{(0)} - \tilde{x}\| \leq \epsilon$, la suite des itérés de la méthode de Newton ci-dessus est bien définie et converge vers $\tilde{x}$ quand n tend vers l'infini.

Méthode de Newton

- Calculer l'itéré $n+1$ à partir de l'itéré n : on a besoin d'inverser la matrice $J_f\left(x^{(n)}\right)$

- Pour éviter ce calcul d'inverse :

$$J_f\left(x^{(n)}\right)\left(x^{(n+1)}-x^{(n)}\right)=-f\left(x^{(n)}\right).$$

- À chaque itération, calcul de l'inverse remplacé par la résolution d'un système d'équations linéaires ce qui est asymptotiquement moins couteux .

Exemple (1)

- Considérons le système d'équations non linéaires :

$$(S) : \begin{cases} x_1^2 + 2\,x_1 - x_2^2 - 2 &= 0, \\ x_1^3 + 3\,x_1\,x_2^2 - x_2^3 - 3 &= 0. \end{cases}$$

- Notations précédentes : $n = 2$, $f_1(x_1, x_2) = x_1^2 + 2\,x_1 - x_2^2 - 2$, et $f_2(x_1, x_2) = x_1^3 + 3\,x_1\,x_2^2 - x_2^3 - 3$

- Matrice jacobienne de f :

$$J_f(x_1, x_2) = \begin{pmatrix} 2\,x_1 + 2 & -2\,x_2 \\ 3\,(x_1^2 + x_2^2) & 6\,x_1\,x_2 - 3\,x_2^2 \end{pmatrix}.$$

Exemple (2)

• Point de départ : $x^{(0)} = (1 \quad -1)^T$. Calculons le premier itéré de la méthode de Newton

• Formule d'itération pour $n = 1$:

$$J_f\left(x^{(0)}\right)\left(x^{(1)} - x^{(0)}\right) = -f\left(x^{(0)}\right),$$

c'est-à-dire

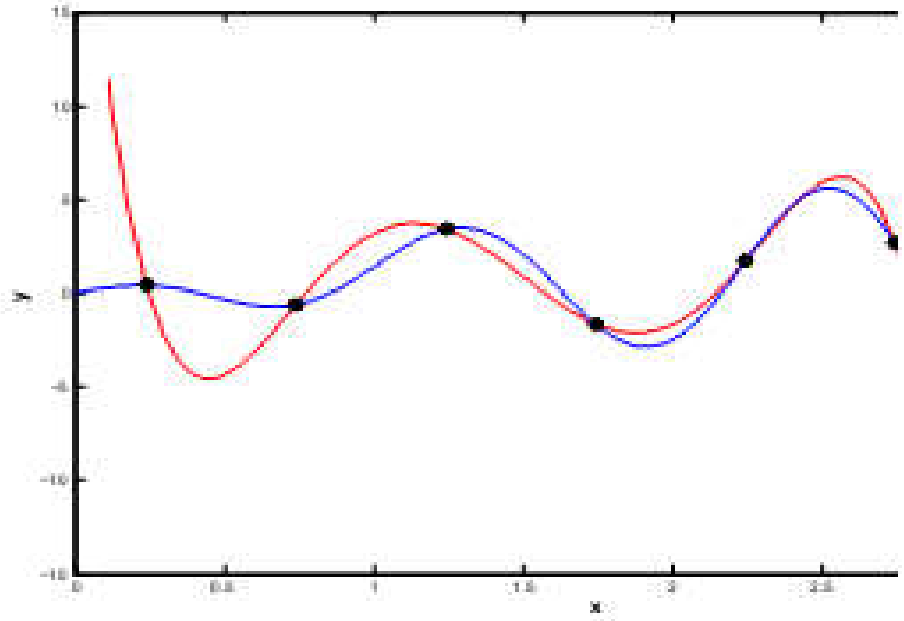
$$\begin{pmatrix} 4 & 2 \\ 6 & -9 \end{pmatrix}\begin{pmatrix} x_1^{(1)} - 1 \\ x_2^{(1)} + 1 \end{pmatrix} = -\begin{pmatrix} 0 \\ 2 \end{pmatrix}.$$

• En résolvant ce système linéaire, on trouve $x_1^{(1)} - 1 = -\frac{1}{12}$ et $x_2^{(1)} + 1 = \frac{1}{6} \rightsquigarrow x^{(1)} = \left(\frac{11}{12} \quad -\frac{5}{6}\right)^T$.

Chapitre 4

Interpolation polynomiale

Problème de l'interpolation



Problème de l'interpolation

- $\mathcal{P}_n = \mathbb{R}_n[x]$: ensemble des poly. de degré $\leq n$ et à coeffs dans $\mathbb{R}$. (espace vect. de dimension $n+1$ sur $\mathbb{R}$)
- $(a, b) \in \mathbb{R}^2$ $(a < b)$ et $f : [a, b] \to \mathbb{R}$ continue sur $[a, b]$
- On considère $n+1$ points $x_0, \ldots, x_n$ de l'intervalle $[a, b]$ tels que $a \leq x_0 \leq x_1 \leq \cdots \leq x_n \leq b$.
- Problème $(I)^f_{m,n}$: ? existe $P_m \in \mathcal{P}_m$ tel que $P_m(x_i) = f(x_i)$, $\forall\, i$.
- $P_m(x) = \lambda_0 + \lambda_1\, x + \cdots + \lambda_m\, x^m$ avec les λ_i dans $\mathbb{R}$, alors ? $\lambda_0, \ldots, \lambda_m$ tels que :

$$(S) : \begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^m \\ 1 & x_1 & x_1^2 & \ldots & x_1^m \\ \vdots & & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^m \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_m \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_m) \end{pmatrix}.$$

Problème de l'interpolation

$\rightsquigarrow$ Système linéaire $n+1$ équations en $m+1$ inconnues

Proposition

Le problème d'interpolation $(I)^f_{m,n}$ admet une unique solution ssi $m = n$ et les nœuds $(x_i)_{0\leq i\leq n}$ sont deux à deux distincts.

- Dans la suite, on s'intéresse au cas où le problème admet une unique solution et on le note $(I)^f_n$: la solution notée $P_n(x\,;f)$ s'appelle polynôme d'interpolation de f aux nœuds $(x_i)_{0\leq i\leq n}$.

- Problème qui apparait dans un contexte expérimental : calcul des valeurs d'une fonction f inconnue.

- Il est naturel de supposer que l'on connait un minimum d'information sur la fonction f à interpoler.

Problème de l'interpolation

En pratique, résoudre directement le système (S) n'est pas forcément une bonne idée, car :

- méthode couteuse $(\mathcal{O}(n^3))$,
- le système est souvent mal conditionné,
- il n'est pas indispensable de calculer les coefficients de $P_n(x; f)$ en base monomiale; il y a d'autres bases de $\mathcal{P}_n$ qui se prêtent mieux à résoudre le problème de l'interpolation.

Remarque: dans plusieurs applications on est surtout intéressé à évaluer $P_n(\tilde{x}; f)$ pour $\tilde{x}$ donné.

Base d'interpolation de Lagrange

Définition

Pour $j \in \{0, \ldots, n\}$, le polynôme $L_j^{(n)}$ défini par

$$L_j^{(n)}(x) = \prod_{\substack{i=0 \\ i \neq j}}^{n} \frac{x - x_i}{x_j - x_i} = \frac{(x - x_0) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)},$$

est appelé interpolant de base de Lagrange ou polynôme de base de Lagrange associé à la suite $(x_i)_{0 \leq i \leq n}$ et relatif au point x_j.

Proposition

Pour $n \in \mathbb{N}$ fixé, les $(L_j^{(n)}(x))_{0 \leq j \leq n}$ forment une base de l'espace vectoriel $\mathcal{P}_n$ que l'on appelle base de Lagrange.

Base d'interpolation de Lagrange

Proposition

Les interpolants de base de Lagrange vérifient les propriétés suivantes :

1. *Pour tout* $j = 0, \ldots, n$, *si on note* g_j *la fonction de* $[a, b]$ *dans* $\mathbb{R}$ *définie par* $\forall\, i = 0, \ldots, n,\ g_j(x_i) = \delta_{ij}$, *alors* $P_n(x\,; g_j) = L_j^{(n)}(x)$;
2. *Si on pose* $\pi_{n+1}(x) = \prod_{j=0}^{n}(x - x_j) \in \mathcal{P}_{n+1}$, *alors, pour tout* $j = 0, \ldots, n,\ L_j^{(n)}(x) = \frac{\pi_{n+1}(x)}{(x - x_j)\,\pi'_{n+1}(x_j)}$.
3. *Pour tout* $k = 0, \ldots, n,\ x^k = \sum_{j=0}^{n} x_j^k\, L_j^{(n)}(x)$.

Méthode de Lagrange

• La méthode d'interpolation de Lagrange consiste à écrire le polynôme d'interpolation sur la base de Lagrange.

Théorème

Soit $f : [a, b] \to \mathbb{R}$ et $n+1$ nœuds $(x_i)_{0 \leq i \leq n}$ deux à deux distincts. Le polynôme d'interpolation de f aux nœuds $(x_i)_{0 \leq i \leq n}$ s'écrit alors :

$$P_n(x ; f) = \sum_{j=0}^{n} f(x_j) \, L_j^{(n)}(x).$$

Base d'interpolation de Newton

Définition

Les polynômes $N_j^{(n)}$ *définis pour* $j = 0, \ldots, n$ *par :*

$$\begin{cases} N_0^{(n)}(x) &= 1, \\ N_1^{(n)}(x) &= (x - x_0), \\ N_2^{(n)}(x) &= (x - x_0)(x - x_1), \\ \quad\vdots & \\ N_j^{(n)}(x) &= (x - x_0)(x - x_1)\cdots(x - x_{j-1}), \\ \quad\vdots & \\ N_n^{(n)}(x) &= (x - x_0)(x - x_1)\cdots(x - x_{n-1}), \end{cases}$$

sont appelés polynômes de base de Newton relatifs à la suite de points $(x_i)_{i=0,\ldots,n-1}$.

Base d'interpolation de Newton

- Remarque : là où on avait besoin de $n+1$ points pour définir les $L_j^{(n)}(x)$, $j = 0, \ldots, n$, la définition des $N_j^{(n)}(x)$, $j = 0, \ldots, n$, ne nécessite que n points.

Proposition

Pour $n \in \mathbb{N}$ fixé, les $(N_j^{(n)}(x))_{0 \leq j \leq n}$ forment une base de l'espace vectoriel $\mathcal{P}_n$.

Expression de l'interpolant de Newton

- $f : [a, b] \longrightarrow \mathbb{R}$ et n nœuds $(x_i)_{0 \leq i \leq n-1}$
- ? α_i, $i = 0, \ldots, n$ tels que $P_n(x ; f) = \sum_{i=0}^{n} \alpha_i N_i^{(n)}(x)$. On a :

$$P_n(x_0 ; f) = \alpha_0 = f(x_0) \Longrightarrow \alpha_0 = f(x_0)$$

$$P_n(x_1 ; f) = f(x_0) + \alpha_1 (x_1 - x_0) = f(x_1) \Longrightarrow \alpha_1 = \frac{f(x_0) - f(x_1)}{x_0 - x_1}$$

$$P_n(x_2 ; f) = f(x_0) + \frac{f(x_0) - f(x_1)}{x_0 - x_1} (x_2 - x_0) + \alpha_2 (x_2 - x_0)(x_2 - x_1) = f(x_2)$$

$$\Longrightarrow \alpha_2 = \frac{\dfrac{f(x_0) - f(x_2)}{x_0 - x_2} - \dfrac{f(x_0) - f(x_1)}{x_0 - x_1}}{x_2 - x_1}$$

En posant

$$f[u, v] = \frac{f(u) - f(v)}{u - v},$$

on a alors

$$\alpha_1 = f[x_0, x_1], \quad \alpha_2 = \frac{f[x_0, x_2] - f[x_0, x_1]}{x_2 - x_1} = \frac{f[x_0, x_1] - f[x_1, x_2]}{x_0 - x_2}.$$

Différence divisée

Définition

Pour tout $k \in \mathbb{N}$, on appelle différence divisée d'ordre k de f associée à la suite de points deux à deux distincts $(x_j)_{j\in\mathbb{N}}$ la quantité $f[x_0, x_1, \ldots, x_k]$ définie par :

$$f[x_0] = f(x_0). \quad \forall\, k \in \mathbb{N}^*,\ f[x_0, x_1, \ldots, x_k] = \frac{f[x_0, x_1, \ldots, x_{k-1}] - f[x_1, x_2, \ldots, x_k]}{x_0 - x_k}.$$

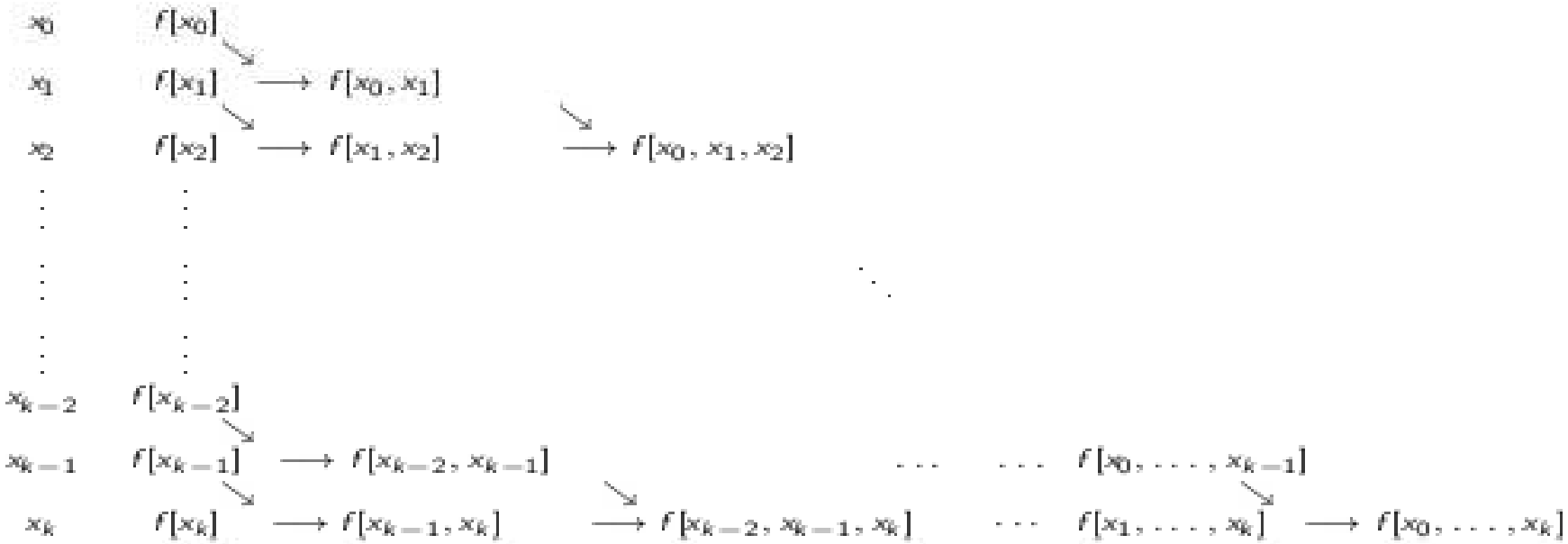
Théorème

$$P_n(x\,; f) = \sum_{k=0}^{n} f[x_0, x_1, \ldots, x_k]\, N_k^{(n)}(x).$$

Corollaire

$$P_n(x\,; f) = P_{n-1}(x\,; f) + f[x_0, x_1, \ldots, x_n]\, N_n^{(n)}(x).$$

Algorithme de calcul des différences divisées



- Contrairement à Lagrange, l'ajout d'un nouveau nœud n'oblige pas à recalculer toutes les différences divisées : passer de n à $n+1$ nœuds demande simplement le calcul de n différences divisées.

Erreur d'interpolation

Lemme

Soit $(x_i)_{0\leq i\leq n}$ tels que, pour tout $i = 0, \ldots, n$, $x_i \in [a, b]$ et soit $P_n(x\,; f)$ le polynôme d'interpolation de f aux nœuds $(x_i)_{0\leq i\leq n}$. Alors, avec les notations précédentes, pour tout $x \in [a, b]$ tel que, pour tout $i = 0, \ldots, n$, $x \neq x_i$, on a :

$$f(x) - P_n(x\,; f) = f[x_0, x_1, \ldots, x_n, x]\,\pi_{n+1}(x).$$

Lemme

Si $f \in C^n([a, b])$ est de classe C^n sur $[a, b]$, alors :

$$\exists\, \xi \in]a, b[, \quad f[x_0, x_1, \ldots, x_n] = \frac{1}{n!}\, f^{(n)}(\xi).$$

Erreur d'interpolation

Théorème

Soit $(x_i)_{0\leq i\leq n}$ *tels que, pour tout* $i = 0, \ldots, n$, $x_i \in [a, b]$ *et soit* $P_n(x\,; f)$ *le polynôme d'interpolation de* f *aux nœuds* $(x_i)_{0\leq i\leq n}$. *Si* $f \in C^{n+1}([a, b])$, *alors :*

$$\forall x \in [a, b],\ \exists \xi_x \in]a, b[, \quad f(x) - P_n(x\,; f) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x)\, \pi_{n+1}(x).$$

Corollaire

Avec les mêmes hypothèses, on a :

$$\forall x \in [a, b], \quad |f(x) - P_n(x\,; f)| \leq \frac{|\pi_{n+1}(x)|}{(n+1)!} \sup_{y\in[a,b]} |f^{(n+1)}(y)|.$$

Chapitre 5

Intégration numérique

Objectif

- On veut approcher de façon numérique la valeur d'intégrales de la forme

$$I(f) = \int_a^b f(x)\,\mathrm{d}x$$

- Remarques :
  - En pratique, on ne connait pas forcément l'expression symbolique de f ;
  - La plupart des fonctions n'admettent pas de primitives pouvant s'exprimer à l'aide de fonctions élémentaires.

Introduction

• Hypothèse : fonctions que l'on cherche à intégrer numériquement sont continues sur l'intervalle $[a, b]$.

• Soit $x_0 = a < x_1 < x_2 < \cdots < x_{n-1} < x_n = b$ une subdivision de l'intervalle $[a, b]$.

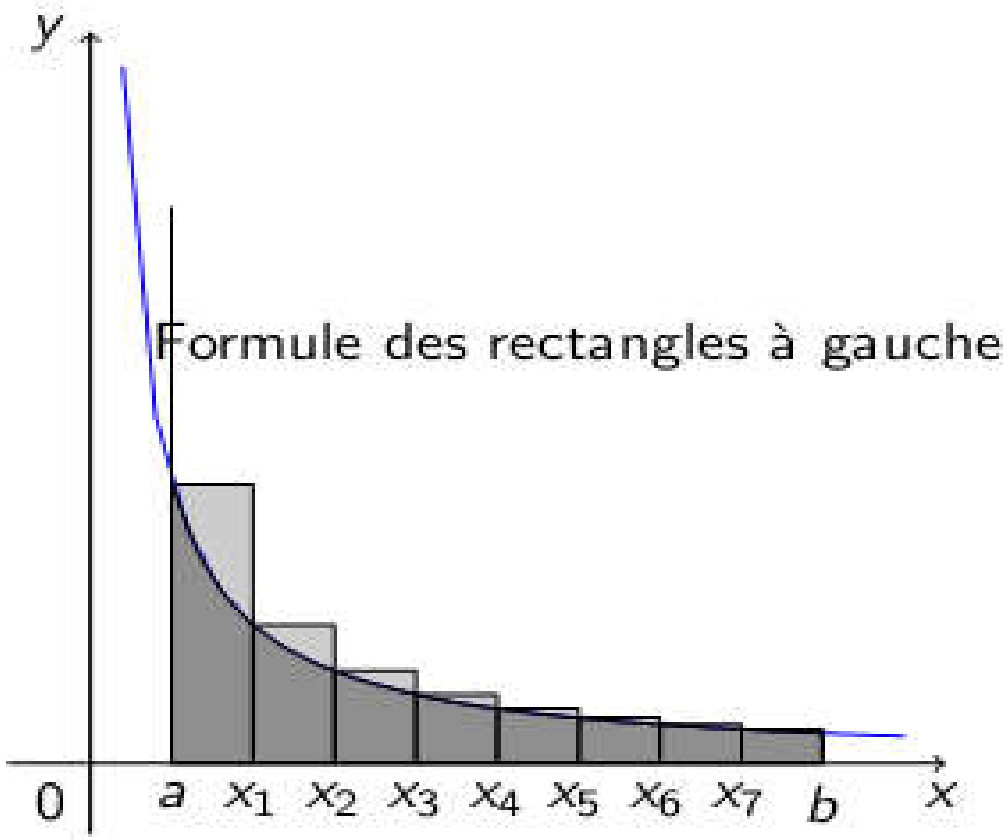
• Théorie élémentaire de l'intégration $\rightsquigarrow$

$$I(f) = \int_a^b f(x)\mathrm{d}x = \lim_{n\to+\infty} \underbrace{\sum_{j=0}^{n-1} f(\xi_j)(x_{j+1} - x_j)}_{\text{Somme de Riemann}}, \quad \forall j,\ \xi_j \in [x_j, x_{j+1}].$$

• Différents choix des ξ_j mènent aux méthodes classiques

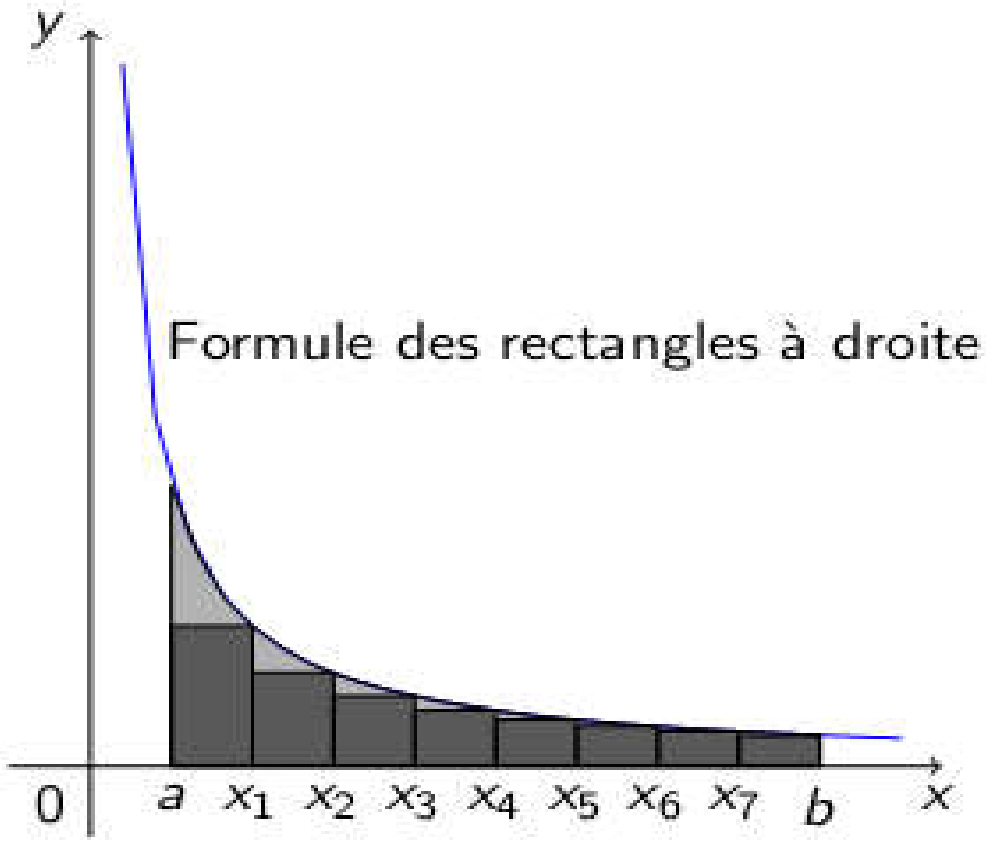
Formule des rectangles à gauche

$$\xi_j = x_j \rightsquigarrow I_{rg}(f) = \sum_{j=0}^{n-1} f(x_j)\,(x_{j+1} - x_j)$$



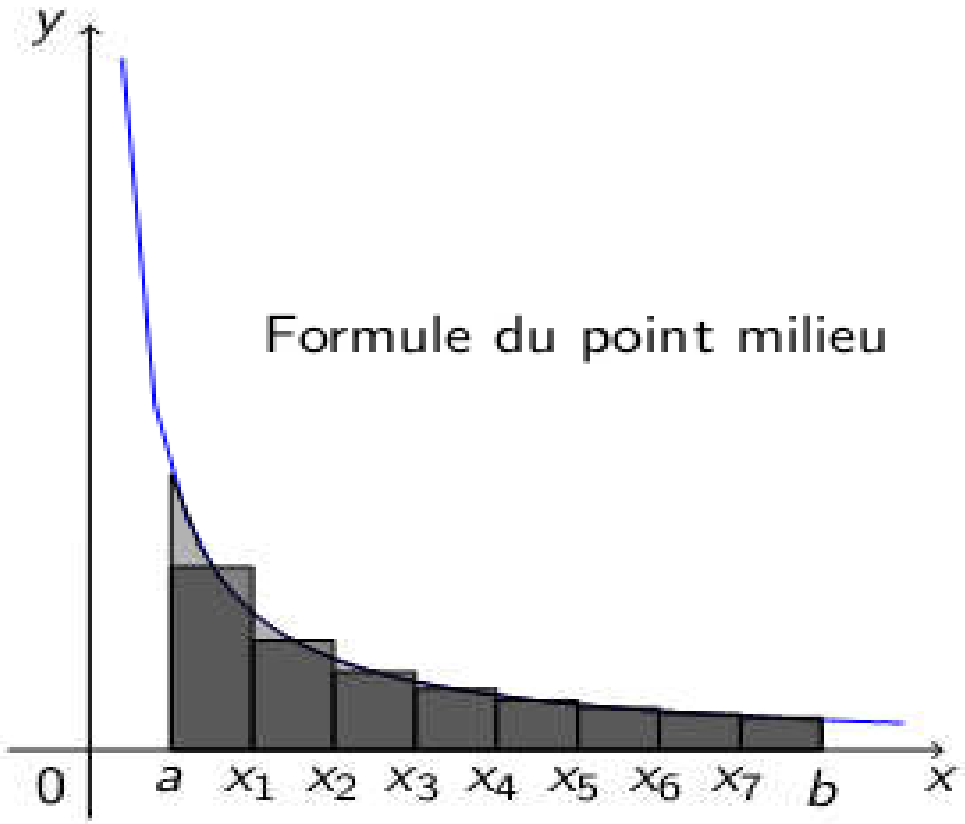
Formule des rectangles à droite

$$\xi_j = x_{j+1} \rightsquigarrow I_{rd}(f) = \sum_{j=0}^{n-1} f(x_{j+1})\,(x_{j+1} - x_j)$$



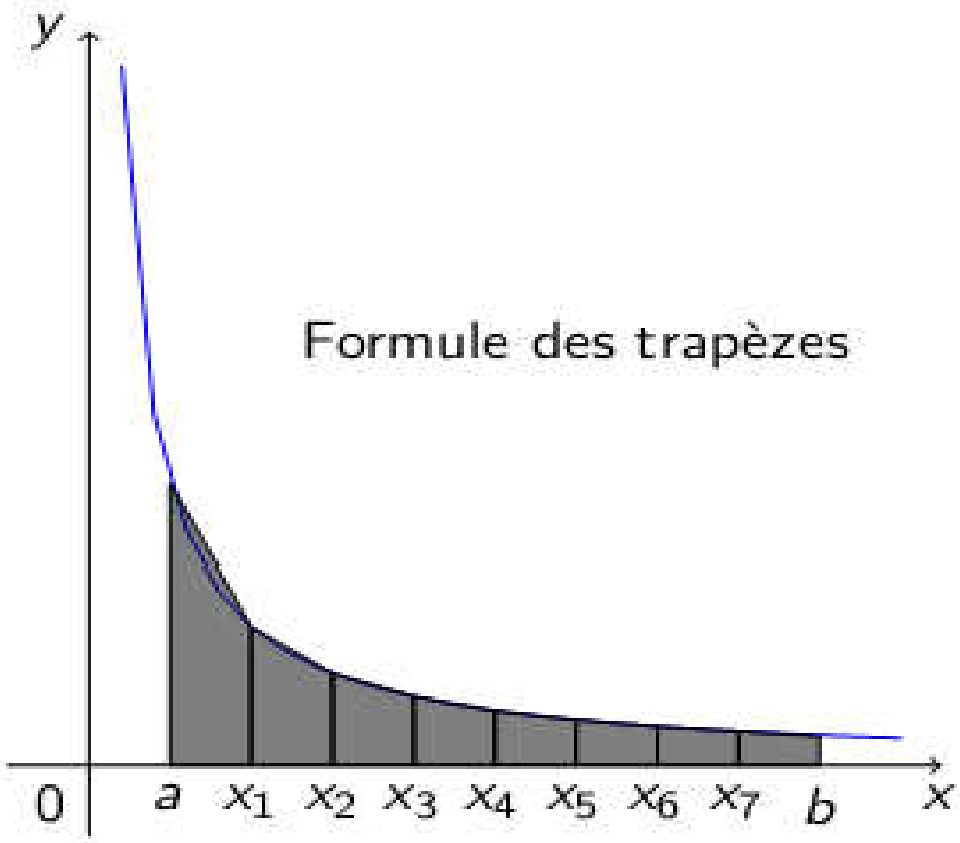
Formule du point milieu

$$\xi_j = \frac{x_j + x_{j+1}}{2} \rightsquigarrow I_{pm}(f) = \sum_{j=0}^{n-1} f\left(\frac{x_j + x_{j+1}}{2}\right)(x_{j+1} - x_j)$$



Méthode des trapèzes

$$I_t(f) = \sum_{j=0}^{n-1} \frac{f(x_j)+f(x_{j+1})}{2} (x_{j+1} - x_j)$$



Fin de Cours