Module AP-41 :
Mathématiques Appliquées.
Elément de module AP-41-2 : Analyse Numérique.
Polycopié du cours

Mohammed Heyouni [1]

1. ENSA d'Al-Hoceima, Université Mohammed Premier, Oujda.
   Email: mohammed.heyouni@gmail.com

Contenu du polycopié

Ce polycopié de cours de l'élément de Module ***AP-45-2 : ANALYSE NUMERIQUE*** s'adresse aux élèves de la deuxième année du cycle préparatoire de l'Ecole Nationale des Sciences Appliquées d'Al-Hoceima (ENASH). Le cours traite, autant que possible, le contenu des chapitres ci-dessous. Ce contenu a été adopté lors de la demande de renouvellement d'accréditation, selon le nouveau CNPN (Cahier des Normes Pédagogiques Nationales) pour l'année 2014.

- **Chapitre 1. Résolution de systèmes linéaires.**
  Généralités sur la résolution de systèmes linéaires. Méthodes directes : Cas des systèmes triangulaires. Méthode d'élimination de Gauss. Décomposition LU. Décomposition de Cholesky. Décomposition QR. Méthodes itératives : Méthodes de Jacobi, de Gauss-Seidel et méthode de relaxation.

- **Chapitre 2. Interpolation polynomiale.**
  Interpolation dans la base de Lagrange. Interpolation dans la base de Newton. Schéma de Neville-Aitken. Erreur d'interpolation..

- **Chapitre 3. Quadrature Numérique.**
  Méthode des rectangles. Méthode des trapèzes. Méthode de Simpson. Obtention des formules de quadrature. Formules composites. Formules de Gauss..

- **Chapitre 4. Résolution d'équations non linéaires.**
  Résultats généraux sur la localisation des racines d'une fonction. Méthodes de points fixes. Méthodes d'encadrement : dichotomie, fausse position, Newton, Sécante. Convergence et ordre de convergence d'une méthode.

Table des matières

Chapitre 1

Résolution de systèmes d'équations linéaires

1.1 Généralités

Soit à résoudre le système linéaire

$$A\,x = b, \tag{1.1}$$

où $A \in \mathbb{R}^{n\times n}$ est une matrice réelle inversible de taille n et $x,\ b \in \mathbb{R}^n$. Dans la suite, $x^* = A^{-1}\,b$ désignera la solution exacte de ce système.

On rappelle que théoriquement, la méthode de Cramer permet de donner chaque composante x_i^* de la solution x^* sous la forme d'un quotient de deux déterminants de taille n chacun, i.e.,

$$x_i^* = \frac{\Delta_i}{\Delta} = \frac{det\left(A^{(1)},\ldots,A^{(i-1)},b,A^{(i+1)},\ldots,A^{(n)}\right)}{det(A)},\ \text{pour } i = 1,2,\ldots,n. \tag{1.2}$$

où $\Delta = det(A)$ est le déterminant de la matrice A et Δ_i est le déterminant obtenu à partir de Δ en remplaçant $A^{(i)}$ la i-ème colonne de A par b le vecteur second membre.

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 $(n+1)!$ *flops*. Ainsi, en supposant qu'on dispose d'un ordinateur capable d'effectuer 10^9 *flops* par seconde, alors il nous faudrait $9.6\,10^{47}$ années pour résoudre un système linéaire ayant 50 inconnues comme le montre le tableau ci-dessous !

Taille n	nombre de *flops*	Temps
10	$11! \approx 3.99\,10^7$	0,04 seconde
20	$21! \approx 5.1\,10^{19}$	1620 années
50	$51! \approx 1.55\,10^{66}$	$9,6\,10^{47}$ années

Ajoutons à cela que même pour des systèmes linéaire de tailles "modestes", le calcul, sur des ordinateur, de la solution x^* par la méthode de Cramer est entaché par la propagation des erreurs d'arrondi et de troncature.

Pour toutes ses raisons, de nombreuses méthodes qui permettent de résoudre (1.1) ont été développées. Ces méthodes sont généralement subdivisées en :

- **Méthodes directes.** Ces méthodes fournissent la solution du système en un nombre fini d'étapes.

  - Elles sont généralement utilisées pour des systèmes ayant une matrice dense.

  - Elles sont basées sur une décomposition de la matrice A sous forme d'un produit de matrices plus "simples" à manipuler : $A = \mathscr{L}\mathscr{U}$, $A = \mathscr{Q}\mathscr{R}$, $A = \mathscr{L}\mathscr{D}\mathscr{L}^T$, $A = \mathscr{S}\mathscr{S}^T$, ...

- **Méthodes itératives.** Ces méthodes déterminent, en théorie, la solution x^* du système (1.1) après un nombre infini d'itérations. Cette famille de méthodes est elle même subdivisée en deux sous classes de méthodes

  - Les méthodes **semi-itératives** telle que *la méthode de Jacobi* ou *la méthode de Gauss-Seidel* sont basées sur la décomposition de la matrice A sous forme d'une somme de matrices et construisent une suite de solutions approchées de la forme :

    $$x^{(k+1)} = B\,x^{(k)} + d, \text{ pour } k = 0, 1, \ldots, \tag{1.3}$$

    où l'approximation initiale $x^{(0)}$ est donnée. La matrice $B \in \mathbb{R}^{n\times n}$ dépend de A et est appelée matrice d'itération et le vecteur $d \in \mathbb{R}^n$ dépend du vecteur second membre b.

  - Les méthodes de type **Krylov** telle que *la méthode d'orthogonalization complète* (**FOM**), *la méthode du résidu minimal généralisé* (**GMRES**), *la méthode du résidu quasi-minimal* (**QMR**), *la méthode de changement du résidu minimal* (**CMRH**), ... etc. Ces méthodes sont généralement appliquées à des systèmes de très grande taille et dont la matrice est creuse.

    En partant d'une approximation initiale $x^{(0)}$ dont le résidu est $r^{(0)} = b - A\,x^{(0)}$, les méthodes de Krylov construisent des solutions approchées de la forme

    $$x^{(k+1)} = V_k\,x^{(k)} + d^{(k)}, \text{ pour } k = 0, 1, \ldots, \tag{1.4}$$

    où $V_k \in \mathbb{R}^k$ est une matrice qui dépend de l'itération courante et dont les vecteurs colonnes forment une base de $K_k(r^{(0)}, A) = span\{r^{(0)}, A\,r^{(0)}, \ldots, A^{k-1}\,r^{(0)}\}$ (le sous espace de Krylov engendré par $r^{(0)}, A\,r^{(0)}, \ldots, A^{k-1}\,r^{(0)}$) et $d^{(k)} \in \mathbb{R}^k$ est déterminé par une condition d'orthogonalité de la forme $r^{(k)} \perp K_k(r^{(0)}, A)$, ou $r^{(k)} \perp A\,K_k(r^{(0)}, A)$. Il est à noter qu'en arithmétique exacte, la suite des itérées $x^{(k)}$ converge vers x^* en moins de n itérations.

Remarque : La convergence des méthodes itératives se teste généralement en utilisant un paramètre de tolérance ϵ. On arrête ainsi les calculs dès que l'erreur relative (respectivement absolue) est assez petite :

$$\frac{\|x^{(k+1)} - x^{(k)}\|}{\|x^{(k)}\|} < \epsilon \qquad (\|x^{(k+1)} - x^{(k)}\| < \epsilon)$$

Une autre alternative est de tester le résidu relatif (respectivement absolu) :

$$\frac{\|r^{(k)}\|}{\|r^{(0)}\|} < \epsilon \qquad (\|r^{(k)}\| < \epsilon),$$

où $r^{(k)} := b - A\,x^{(k)}$.

Avant de passer à la description des méthodes citées ci-dessus, signalons qu'un des points les plus essentiels dans l'efficacité des méthodes de résolution de systèmes linéaires concerne la taille des systèmes à

résoudre. Entre 1980 et 2000, la taille de la mémoire des ordinateurs a augmenté. La taille des systèmes qu'on peut résoudre sur ordinateur a donc également augmenté, selon l'ordre de grandeur suivant :

	1980	2000
matrice "pleine" (tous les termes sont non nuls)	$n = 100$	$n = 10^6$
matrice "creuse" (peu d'éléments non nuls)	$n = 10^6$	$n = 10^8$

1.2 Rappels et compléments d'algèbre linéaire

Définition 1.1.
Soit $A = (a_{i,j})_{i,j=1,\dots,n}$ une matrice réelle de taille n. On dit que A est fortement inversible si les n sous matrices principales de A sont toutes inversibles, i.e., si $det(A_k) \neq 0$, pour $k = 1,\dots,n$ et où $A_k = (a_{i,j})_{i,j=1,\dots,k} \in \mathbb{R}^{k\times k}$.

Définition 1.2.
Soit $A = (a_{i,j})_{i,j=1,\dots,n}$ une matrice réelle de taille n. On dit que A est à diagonale dominante (respectivement strictement dominante) si et seulement si

$$\begin{cases} |a_{i,i}| \geq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{i,j}| \quad (\textit{ respectivement } |a_{i,i}| > \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{i,j}|) \quad \forall i = 1,\dots,n. \\ |a_{j,j}| \geq \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{i,j}| \quad (\textit{ respectivement } |a_{j,j}| > \sum_{\substack{i=1\\ i\neq j}}^{n} |a_{i,j}|) \quad \forall j = 1,\dots,n. \end{cases}$$

Définition 1.3.
Soit A une matrice réelle de taille n. On dit que A est définie positive si pour tout $x \in \mathbb{R}^n$, $x \neq 0 \Longrightarrow x^T A x > 0$.

Définition 1.4.
Soit A une matrice réelle de taille n. On appelle rayon spectral de A qu'on note $\rho(A)$ la plus grande valeur propre en module, i.e.,

$$\rho(A) = \max_{i=1,\dots,n} \{|\lambda_i|;\ \lambda_i \textit{ valeur propre de } A\}.$$

Proposition 1.1.
Soit $A \in \mathbb{R}^{n\times n}$ une matrice réelle de taille n alors :

- *Si A est à diagonale strictement dominante alors A est inversible.*
- *Si A est symétrique alors toutes ses valeurs propres sont réelles.*
- *Si A est symétrique définie positive alors toutes ses valeurs propres sont strictement positives.*

1.3 Méthodes directes

1.3.1 Cas des systèmes triangulaires

Les systèmes linéaires dont la matrice A est triangulaire inférieure ou triangulaire supérieure sont parmi les systèmes les plus faciles à résoudre. En effet, la matrice A étant supposée inversible, alors ses coefficients diagonaux sont non nuls et donc pour un

- **système triangulaire inférieur** tel que celui décrit par la système $(\mathcal{T}_i)$ ci dessous,

$$(\mathcal{T}_i) \quad \begin{cases} a_{1,1}x_1 & & & & = & b_1 \\ a_{2,1}x_1 + & a_{2,2}x_2 & & & = & b_2 \\ & & \ddots & & & \vdots \\ a_{n,1}x_1 + & a_{n,2}x_2 + & \dots + & a_{n,n}x_n & = & b_n \end{cases},$$

la résolution se fait via les formules de *descente* (ou encore *substitution directe*) résumées dans

Algorithme : *Résolution d'un système triangulaire inférieur*

- $x_1 = \dfrac{b_1}{a_{1,1}}$;
- *pour* $i = 2, 3, \ldots, n$ *faire*
$$x_i = \frac{1}{a_{i,i}}(b_i - \sum_{j=1}^{i-1} a_{i,j}x_j)$$
fin pour.

- **système triangulaire supérieur** tel que celui décrit par la système $(\mathscr{T}_s)$ ci dessous,

$$(\mathscr{T}_s) \quad \begin{cases} a_{1,1}x_1 + & a_{1,2}x_2 + & \ldots + & a_{1,n}x_n & = & b_1 \\ & a_{2,2}x_2 + & \ldots + & a_{2,n}x_n & = & b_2 \\ & & \ddots & \vdots & & \vdots \\ & & & a_{n,n}x_n & = & b_n \end{cases}$$

la résolution se fait via les formules de *remontée* (ou encore *substitution rétrograde*) résumées dans l'algorithme ci dessous

Algorithme : *Résolution d'un système triangulaire supérieur*

- $x_n = \dfrac{b_n}{a_{n,n}}$;
- *pour* $i = n-1, n-2, \ldots, 1$ *faire*
$$x_i = \frac{1}{a_{i,i}}(b_i - \sum_{j=i+1}^{n} a_{i,j}x_j)$$
fin pour.

Notons que pour les deux systèmes le coût opératoire est de l'ordre de n^2 flops.

1.3.2 Méthode de Gauss

Cette méthode est basée essentiellement sur l'utilisation de combinaisons linéaires entre lignes (et/ou de permutations de lignes ou de colonnes) du système dans le but de transformer progressivement le système de départ en un système triangulaire supérieure plus facile à résoudre. En effet, rappelons que la solution d'un système linéaire ne change pas quand on ajoute à une équation donnée une combinaison linéaire des autres équations.

La simplicité et l'efficacité de l'algorithme de cette méthode font que la méthode de Gauss est très utilisée dans la pratique et qu'elle est à la base de plusieurs algorithmes de résolution de systèmes linéaires. Dans la suite, nous abordons le principe de cette méthode à l'aide de quelques exemples.

— **Exemple 1.** Soit le système

$$(\mathscr{S}) \left\{ \begin{array}{rrrrcr} 2x_1+ & x_2+ & x_3 & & = & -3 \\ 4x_1+ & 3x_2+ & 3x_3+ & x_4 & = & -5 \\ 8x_1+ & 7x_2+ & 9x_3+ & 5x_4 & = & -7 \\ 6x_1+ & 7x_2+ & 9x_3+ & 8x_4 & = & 1 \end{array} \right.$$

Soit respectivement $A^{(0)} = A$ et $b^{(0)} = b$ la matrice et le vecteur second membre du système. Appliquons alors la méthode d'élimination de Gauss au système précédent en utilisant le tableau 4 lignes 5 colonnes ci-dessous

$$\left[\ A^{(0)} \ \middle|\ b^{(0)}\ \right] = \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & -3 \\ 4 & 3 & 4 & 1 & -5 \\ 8 & 7 & 9 & 5 & -7 \\ 6 & 7 & 9 & 8 & 1 \end{array}\right] \begin{array}{l} l_1 \\ l_2 \\ l_3 \\ l_4 \end{array}$$

$$\left[\ A^{(1)} \ \middle|\ b^{(1)}\ \right] = \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & -3 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 3 & 5 & 5 & 5 \\ 0 & 4 & 6 & 8 & 10 \end{array}\right] \begin{array}{l} \\ l_2 \leftarrow l_2 - 2l_1 = l_2 - l_{2,1} l_1 \\ l_3 \leftarrow l_3 - 4l_1 = l_3 - l_{3,1} l_1 \\ l_4 \leftarrow l_4 - 3l_1 = l_4 - l_{4,1} l_1 \end{array}$$

$$\left[\ A^{(2)} \ \middle|\ b^{(2)}\ \right] = \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & -3 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & 2 & 4 & 6 \end{array}\right] \begin{array}{l} \\ \\ l_3 \leftarrow l_3 - 3l_2 = l_3 - l_{3,2} l_2 \\ l_4 \leftarrow l_4 - 4l_2 = l_4 - l_{4,2} l_2 \end{array}$$

$$\left[\ A^{(3)} \ \middle|\ b^{(3)}\ \right] = \left[\begin{array}{cccc|c} 2 & 1 & 1 & 0 & -1 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 & 2 \\ 0 & 0 & 0 & 2 & 4 \end{array}\right] \begin{array}{l} \\ \\ \\ l_4 \leftarrow l_4 - l_3 = l_4 - l_{4,3} l_3, \end{array}$$

où $l_{i,k} = \dfrac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}$ pour $k = 1,2,3$ et $i = k+1,\ldots,4$. Ainsi

$$(\mathscr{S}) \Longleftrightarrow \left\{ \begin{array}{rrrrcr} 2x_1+ & x_2+ & x_3 & & = & -3 \\ & x_2+ & x_3+ & x_4 & = & 1 \\ & & 2x_3+ & 2x_4 & = & 2 \\ & & & 2x_4 & = & 4 \end{array} \right.$$

et on obtient alors $x_1 = -1,\ x_2 = 0,\ x_3 = -1$ et $x_4 = 2$.

— **Exemple 2.** Soit les systèmes

$$(\mathscr{S}_1) \left\{ \begin{array}{rrrcr} x_1+ & x_2+ & 2x_3 & = & 2 \\ 2x_1+ & x_2- & x_3 & = & 3 \\ 3x_1+ & 2x_2+ & x_3 & = & 7 \end{array} \right. \quad \text{et } (\mathscr{S}_2) \left\{ \begin{array}{rrrcr} x_1+ & x_2+ & 2x_3 & = & 1 \\ 2x_1+ & x_2- & x_3 & = & 4 \\ 3x_1+ & 2x_2+ & x_3 & = & 5 \end{array} \right.$$

Les deux systèmes précédents ayant la même matrice A, nous allons appliquer la méthode d'élimination de Gauss en utilisant le tableau 3 lignes 5 colonnes formé par les éléments de la matrice

$A^{(0)} = A$ et des vecteurs second membres $b_1^{(0)} = b_1$ et $b_2^{(0)} = b_2$. Nous avons alors

$$\left[\begin{array}{c|c|c} A^{(0)} & b_1^{(0)} & b_2^{(0)} \end{array}\right] = \left[\begin{array}{ccc|c|c} 1 & 1 & 2 & 2 & 1 \\ 2 & 1 & -1 & 3 & 4 \\ 3 & 2 & 1 & 7 & 5 \end{array}\right] \begin{array}{l} l_1 \\ l_2 \\ l_3 \end{array}$$

$$\left[\begin{array}{c|c|c} A^{(1)} & b_1^{(1)} & b_2^{(1)} \end{array}\right] = \left[\begin{array}{ccc|c|c} 1 & 1 & 2 & 2 & 1 \\ 0 & -1 & -5 & -1 & 2 \\ 0 & -1 & -5 & 1 & 2 \end{array}\right] \begin{array}{l} \\ l_2 \leftarrow l_2 - 2l_1 = l_2 - l_{2,1} l_1 \\ l_3 \leftarrow l_3 - 3l_1 = l_3 - l_{3,1} l_1 \end{array}$$

$$\left[\begin{array}{c|c|c} A^{(2)} & b_1^{(2)} & b_2^{(2)} \end{array}\right] = \left[\begin{array}{ccc|c|c} 1 & 1 & 2 & 2 & 1 \\ 0 & -1 & -5 & 1 & 2 \\ 0 & 0 & 0 & 2 & 0 \end{array}\right] \begin{array}{l} \\ \\ l_3 \leftarrow l_3 - l_2 = l_3 - l_{3,2} l_2, \end{array}$$

avec $l_{i,k} = \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}$ pour $k = 1,2$ et $i = k+1,\ldots,3$. Ainsi

$$(\mathscr{S}_1) \Longleftrightarrow \left\{\begin{array}{rcl} x_1 + x_2 + 2x_3 &=& 2 \\ -x_2 - 5x_3 &=& 1 \\ 0 &=& 2 \end{array}\right. \text{ et } (\mathscr{S}_2) \Longleftrightarrow \left\{\begin{array}{rcl} x_1 + x_2 + 2x_3 &=& 2 \\ -x_2 - 5x_3 &=& 2 \\ 0 &=& 0 \end{array}\right.$$

On en déduit que le système $(\mathscr{S}_1)$ n'a pas de solutions et n'est donc pas compatible. Quant au système $(\mathscr{S}_2)$, il possède une infinité de solutions de la forme $x_3 = \alpha, x_2 = -2 - 5\alpha$ et $x_1 = 3 + 3\alpha$, où $\alpha \in \mathbb{R}$.

Mise en Œuvre de l'algorithme.

Soit le système linéaire $Ax = b$ où $A \in \mathbb{R}^{n\times n}$ est telle que le premier terme diagonal $a_{1,1}$ soit non nul.

Posons $A^{(0)} = A$, $b^{(0)} = b$, alors comme $a_{1,1}^{(0)} \neq 0$, on peut éliminer l'inconnue x_1 de la ligne $l_i^{(0)}$, $i = 2,3,\ldots,n$, en retranchant à cette ligne la quantité $l_{i,1} l_1^{(0)}$ où $l_{i,1} = \frac{a_{i,1}^{(0)}}{a_{1,1}^{(0)}}$. Ainsi, en définissant

$$\left\{\begin{array}{rcll} a_{i,j}^{(1)} &=& a_{i,j}^{(0)} - l_{i,1} a_{1,j}^{(0)} & i,j = 2,3,\ldots,n, \\ b_i^{(1)} &=& b_i^{(0)} - l_{i,1} b_1^{(0)} & i = 2,3,\ldots,n, \end{array}\right.$$

on obtient le système $A^{(1)} x = b^{(1)}$ équivalent suivant

$$\underbrace{\begin{pmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} & \ldots & a_{1,n}^{(0)} \\ 0 & a_{2,2}^{(1)} & \ldots & a_{2,n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n,2}^{(1)} & \ldots & a_{n,n}^{(1)} \end{pmatrix}}_{=:A^{(1)}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \underbrace{\begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ \vdots \\ b_n^{(1)} \end{pmatrix}}_{=:b^{(1)}}$$

De la même manière, pour $k = 2,3,\ldots,n$ et si $a_{k,k}^{(k-1)} \neq 0$, on peut éliminer x_k de la ligne $l_i^{(k-1)}$, $i = k+1,k+2,\ldots,n$, en retranchant à cette ligne la quantité $l_{i,k} l_k^{(k-1)}$ où $l_{i,k} = \frac{a_{i,k}^{(k-1)}}{a_{k,k}^{(k-1)}}$ Ainsi, en définissant les quantités :

$$\left\{\begin{array}{rcll} a_{i,j}^{(k)} &=& a_{i,j}^{(k-1)} - l_{i,k} a_{k,j}^{(k-1)} & i,j = k+1,k+2,\ldots,n, \\ b_i^{(k)} &=& b_i^{(k-1)} - l_{i,k} b_k^{(k-1)} & i = k+1,k+2,\ldots,n, \end{array}\right.$$

on construit une suite finie de systèmes

$$A^{(k)}x = b^{(k)}, \quad \text{pour } k = 0,2,\ldots,n-1,$$

où la matrice $A^{(k)}$ pour $1 \le k \le n-1$ à la forme suivante

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

et où on a supposé que $a_{i,i}^{(i-1)} \neq 0$ pour $i = 1,\ldots,k$.

Finalement, notons que pour $k = n-1$, on obtient le système triangulaire supérieur $Ux = y$ suivant :

$$\underbrace{\begin{pmatrix} a_{1,1}^{(0)} & a_{1,2}^{(0)} & \cdots & \cdots & a_{1,n}^{(0)} \\ 0 & a_{2,2}^{(1)} & & & a_{2,n}^{(1)} \\ \vdots & 0 & \ddots & & \vdots \\ 0 & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & a_{n,n}^{(n-1)} \end{pmatrix}}_{=:A^{(n-1)}} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ \vdots \\ x_n \end{pmatrix} = \underbrace{\begin{pmatrix} b_1^{(0)} \\ b_2^{(1)} \\ \vdots \\ \vdots \\ b_n^{(n-1)} \end{pmatrix}}_{=:b^{(n-1)}}$$

où $U := A^{(n-1)}$ et $y := b^{(n-1)}$, et qu'on a établi que la résolution du système $Ax = b$ équivaut à celle du système $Ux = y$. Finalement, la méthode d'élimination de Gauss, dont le coût opératoire est de l'ordre de $\frac{2}{3}n^3$, est résumé par l'algorithme si dessous :

Algorithme : *Méthode de Gauss sans pivotage*

1. *Phase de triangularisation.*
   Pour $k = 1,2,\ldots,n-1$
   Pour $i = k+1,k+2,\ldots,n$
   Pour $j = k+1,k+2,\ldots,n$
   $a_{i,j} = a_{i,j} - \dfrac{a_{i,k}}{a_{k,k}} a_{k,j}$;
   Fin pour
   $b_i = b_i - \dfrac{a_{i,k}}{a_{k,k}} b_k$;
   $a_{i,k} = 0$;
   Fin pour
   Fin pour
2. *Phase de résolution du système triangulaire supérieur.*
   $x_n = \dfrac{b_n}{a_{n,n}}$;
   Pour $i = n-1,n-2,\ldots,1$
   $x_i = \dfrac{1}{a_{i,i}}(b_i - \sum_{j=i+1}^{n} a_{i,j}x_j)$;
   Fin pour

Stratégie de pivotage.

Exemple 1. Soit à résoudre le système $Ax = b$, où $b \in \mathbb{R}^3$ est quelconque et $A \in \mathbb{R}^{3\times 3}$ 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 à la matrice A, on obtient

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

et on remarque qu'on ne peut pas continuer, et pourtant la matrice A est inversible ! Par contre si on permute les ligne 2 et 3 de la matrice A, et que l'on applique la méthode de Gauss à la nouvelle matrice obtenu

$$\tilde{A} = PA \begin{pmatrix} 1 & 2 & 3 \\ 7 & 8 & 9 \\ 2 & 4 & 5 \end{pmatrix}, \text{ avec } P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}.$$

on obtient

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

qui est déjà une matrice triangulaire supérieure et donc le système $Ax = b$ peut être résolu.

Ce premier exemple nous montre que dans certains cas, la permutation de quelques équations d'un système linéaire donné -et donc la permutation de certaines lignes de la matrice A- est nécessaire afin de résoudre ce système.

Exemple 2. Considérons le système

$$(\mathscr{S}) \quad \begin{cases} \epsilon x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases},$$

où $\epsilon \neq 0$. De plus, on suppose que $\epsilon \neq 1$, et ainsi le système $(\mathscr{S})$ est inversible. La solution de ce système est alors donnée par $x^* = (x_1^*, x_2^*)$ avec $x_1^* = \frac{1}{1-\epsilon}$ et $x_2^* = \frac{1-2\epsilon}{1-\epsilon}$.

Supposons maintenant que $\epsilon \neq 1$ est "proche de 0" et appliquons l'élimination de Gauss décrite précédemment. La première étape transforme donc le système $(\mathscr{S})$ en $(\mathscr{S}^{(1)})$

$$(\mathscr{S}^{(1)}) \quad \begin{cases} \epsilon x_1 + x_2 = 1 \\ x_1 + 1 - \frac{1}{\epsilon} x_2 = 2 - \frac{1}{\epsilon} \end{cases}.$$

Comme les valeurs 1 et 2 sont très petites par rapport à $-\frac{1}{\epsilon}$, on voit alors que la deuxième équation nous donne $-\frac{1}{\epsilon} x_2 \simeq \frac{1}{\epsilon}$, d'où $x_2 \simeq 1$ et ensuite la première équation nous fournit $x_1 \simeq 0$. Mais ce résultat est faux !
L'erreur ne provient pas seulement du fait que ϵ est "très petit" car si on multiplie la première ligne par une puissance de 10 quelconque, on va trouver la même erreur !

L'anomalie provient du déséquilibre entre les coefficients de x_1 et x_2 de la ligne du pivot. Pour y remédier, échangeons maintenant les deux lignes de notre système et appliquons l'élimination de Gauss avec la valeur 1 comme premier pivot. On obtient alors

$$(\mathscr{S}^{(1)}) \quad \begin{cases} x_1 & + & x_2 & = & 2 \\ & + & (1-\epsilon)x_2 & = & 1-2\epsilon \end{cases},$$

ce qui entraine alors que $x_2 \simeq 1$ et $x_1 \simeq 1$. Et ce résultat est correct !

De même, ce deuxième exemple montre qu'il ne faut pas utiliser des pivots "trop petits" car les erreurs d'arrondi et/ou de troncature peuvent donner entrainer des solutions fausses. Et donc, il est nécessaire de recourir à des permutations de lignes voire des colonnes de la matrice du système.

Ainsi, pour éviter les problèmes rencontrés dans les deux exemples précédents, on peut appliquer l'une des deux 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 ligne l tel que $a_{l,k} = \max_{k\leq i\leq n} |a_{i,k}|$. 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 $a_{l,m} = \max_{k\leq i,j\leq n} |a_{i,j}|$. 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.

Remarques.

- Les pivots ayant une valeur petite ne disparaissent pas, que l'on utilise la stratégie du pivot total ou celle du pivot partiel. En fait, l'apparition des petits pivots est juste rejetée à la fin de la méthode de Gauss et dans ce cas l'influence de ces petits pivots est moins importante.

- Généralement, dans la pratique, la méthode de Gauss est implémentée avec une stratégie de pivot partiel.

1.3.3 Décomposition LU

Nous commençons cette section en donnant quelques résultats théoriques concernant l'existence et l'unicité de la décomposition LU.

Théorème 1.1.
Une matrice $A \in \mathbb{R}^{n\times n}$ possède une décomposition LU, où la matrice L est triangulaire inférieure à diagonale unité et U triangulaire supérieure si et seulement si A est fortement inversible. Dans ce cas la décomposition LU est unique.

Ce théorème permet d'affirmer que si la matrice A est fortement inversible, alors les n étapes de la méthode de Gauss peuvent être appliquées à la matrice A sans rencontrer un pivot nul.

Théorème 1.2.
Soit $A \in \mathbb{R}^{n\times n}$ une matrice inversible, alors il existe au moins une matrice de permutation P telle que la matrice PA admette une décomposition LU.

Dans la suite, nous introduisons la décomposition LU d'une matrice A en donnant une interprétation matricielle des différentes étapes de la méthode d'élimination de Gauss.

Décomposition $A = LU$.

Reprenons le système $Ax = b$ vu dans l'exemple 1 du paragraphe précédent. Soient alors la matrice A et le vecteur second membre b

$$A^{(0)} \leftarrow A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}, \quad b^{(0)} \leftarrow b = \begin{pmatrix} -3 \\ -5 \\ -7 \\ 1 \end{pmatrix},$$

et notons l_1, l_2, l_3 et l_4 les vecteurs lignes des différentes matrices $A^{(k)}$ et des vecteurs $b^{(k)}$ obtenus durant la méthode d'élimination de Gauss.

— **Etape 1 :** Par des combinaisons linéaires ci-dessous :

$$l_2 \leftarrow l_2 - l_{2,1} l_1, \text{ avec } l_{2,1} = 2,$$
$$l_3 \leftarrow l_3 - l_{3,1} l_1, \text{ avec } l_{3,1} = 4,$$
$$l_4 \leftarrow l_4 - l_{4,1} l_1, \text{ avec } l_{4,1} = 3,$$

nous avons transformé la matrice $A^{(0)}$ et le vecteur $b^{(0)}$ respectivement en la matrice $A^{(1)} = L_1 A^{(0)}$ et en le vecteur $b^{(1)} = L_1 b^{(0)}$, où

$$L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -4 & 0 & 1 & 0 \\ -3 & 0 & 0 & 1 \end{pmatrix}, \quad A^{(1)} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 3 & 5 & 5 \\ 0 & 4 & 6 & 8 \end{pmatrix}, \quad b^{(1)} = \begin{pmatrix} -3 \\ 1 \\ 5 \\ 10 \end{pmatrix}.$$

Notons que la matrice L_1 n'est autre que la matrice identité où les 0 de la première colonne, au niveau des positions 2, 3 et 4, ont été remplacés par les coefficients $-l_{2,1}$, $-l_{3,1}$, $-l_{4,1}$.

— **Etape 2 :** Comme lors de la précédente étape, les combinaisons linéaires entre les lignes ci-dessous :

$$l_3 \leftarrow l_3 - l_{3,2} l_2, \text{ avec } l_{3,2} = 3,$$
$$l_4 \leftarrow l_4 - l_{4,2} l_2, \text{ avec } l_{4,2} = 4,$$

ont permis de transformer $A^{(1)}$ en $A^{(2)} = L_2 A^{(1)}$ et $b^{(1)}$ en $b^{(2)} = L_2 b^{(1)}$, où

$$L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & -4 & 0 & 1 \end{pmatrix}, \quad A^{(2)} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 2 & 4 \end{pmatrix}, \quad b^{(2)} = \begin{pmatrix} -3 \\ 1 \\ 2 \\ 6 \end{pmatrix}.$$

Notons également que la matrice L_2 n'est autre que la matrice identité où les 0, au niveau des positions 3 et 4, de la deuxième colonne ont été remplacés par les coefficients $-l_{3,2}$, $-l_{4,2}$.

— **Etape 3 :** A l'aide de la combinaison linéaire

$$l_4 \leftarrow l_4 - l_{4,3} l_3, \text{ avec } l_{4,3} = 1,$$

on transforme $A^{(2)}$ en $A^{(3)} = L_3 A^{(2)}$ et $b^{(2)}$ en $b^{(3)} = L_3 b^{(2)}$, où

$$L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \quad A^{(3)} = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}, \quad b^{(3)} = \begin{pmatrix} -3 \\ 1 \\ 2 \\ 4 \end{pmatrix}.$$

Remarquons que la matrice L_3 n'est autre que la matrice identité où le 0, au niveau de la position 4, de la troisième colonne a été remplacé par le coefficient $-l_{4,3}$.

Ainsi, en posant $U = A^{(3)}$, on voit que $L_3 L_2 L_1 A = U$. D'ou

$$A = LU \text{ avec } L = (L_3 L_2 L_1)^{-1} = L_1^{-1} L_2^{-1} L_3^{-1}.$$

Notons qu'il est facile de vérifier que

$$L_1^{-1} = -L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 0 & 1 & 0 \\ 3 & 0 & 0 & 1 \end{pmatrix}, \; L_2^{-1} = -L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 0 & 4 & 0 & 1 \end{pmatrix}, \quad \text{et} \quad L_3^{-1} = -L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix},$$

et que de plus la matrice L est une matrice triangulaire inférieure à diagonale unité dont les éléments non nuls ne sont autre que les coefficients $l_{i,j}$ utilisés précédemment, i.e.,

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{pmatrix}.$$

Finalement, on a obtenu

$$\underbrace{\begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}}_{A} = \underbrace{\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 4 & 3 & 1 & 0 \\ 3 & 4 & 1 & 1 \end{pmatrix}}_{L} \underbrace{\begin{pmatrix} 2 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 2 \end{pmatrix}}_{U}.$$

Décomposition $PA = LU$.

Dans certains cas, la décomposition LU d'une matrice A donnée n'existe pas. Par contre, il est possible de trouver une matrice de permutation P telle que la décomposition LU de PA existe.

L'exemple ci-dessous montre comment trouver P, L et U en utilisant la stratégie de pivotage partiel. Soit donc

$$A^{(0)} \leftarrow A = \begin{pmatrix} 2 & 1 & 1 & 0 \\ 4 & 3 & 3 & 1 \\ 8 & 7 & 9 & 5 \\ 6 & 7 & 9 & 8 \end{pmatrix}.$$

— **Etape 1 :** Examinons les composantes de la première colonne de A pour déterminer la composante maximale en valeur absolue. On voit alors que $|a_{3,1}^{(0)}| = \max_{i=1\ldots,n} \{|a_{i,1}^{(0)}|\}$. Permutons alors, la 1ère et la 3ème ligne de $A^{(0)}$. Matriciellement, cette opération se traduit par une multiplication à gauche de $A^{(0)}$ par la matrice de permutation P_1, i.e.,

$$\tilde{A}^{(0)} = P_1 A^{(0)} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 4 & 3 & 3 & 1 \\ 2 & 1 & 1 & 0 \\ 6 & 7 & 9 & 8 \end{pmatrix} \text{ avec } P_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

En utilisant les combinaisons linéaires

$$l_2 \leftarrow l_2 - l_{2,1} l_1, \text{ avec } l_{2,1} = \tfrac{1}{2},$$
$$l_3 \leftarrow l_3 - l_{3,1} l_1, \text{ avec } l_{3,1} = \tfrac{1}{4},$$
$$l_4 \leftarrow l_4 - l_{4,1} l_1, \text{ avec } l_{4,1} = \tfrac{3}{4},$$

transformons $\tilde{A}^{(0)}$ en $A^{(1)} = L_1 \tilde{A}^{(0)} = L_1 P_1 A^{(0)}$, où

$$A^{(1)} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \\ 0 & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \end{pmatrix}, \quad L_1 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ -\frac{1}{4} & 0 & 1 & 0 \\ -\frac{3}{4} & 0 & 0 & 1 \end{pmatrix}.$$

— **Etape 2 :** Comme lors de la précédente étape, déterminons la composante maximale, en valeur absolue, de la deuxième colonne de $A^{(1)}$. On voit alors que $|a^{(1)}_{4,2}| = \max_{i=2\ldots,n} \{|a^{(1)}_{i,2}|\}$. Permutons alors, la $2^{\text{ème}}$ et la $4^{\text{ème}}$ ligne de $A^{(1)}$. On a alors les relations :

$$\tilde{A}^{(1)} = P_2 A^{(1)} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & -\frac{3}{4} & -\frac{5}{4} & -\frac{5}{4} \\ 0 & -\frac{1}{2} & -\frac{3}{2} & -\frac{3}{2} \end{pmatrix} \text{ avec } P_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Maintenant, si on utilise les combinaisons linéaires suivantes :

$$l_3 \leftarrow l_3 - l_{3,2} l_2, \text{ avec } l_{3,2} = -\tfrac{3}{7},$$
$$l_4 \leftarrow l_4 - l_{4,2} l_2, \text{ avec } l_{4,2} = -\tfrac{2}{7},$$

alors $\tilde{A}^{(1)}$ se transforme en $A^{(2)} = L_2 \tilde{A}^{(1)}$, où

$$A^{(2)} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & 0 & -\frac{2}{7} & \frac{4}{7} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \end{pmatrix}, \quad L_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{3}{7} & 1 & 0 \\ 0 & \frac{2}{7} & 0 & 1 \end{pmatrix}.$$

— **Etape 3 :** Maintenant, déterminons la composante maximale, en valeur absolue, de la troisième colonne de $A^{(2)}$. On voit alors que $|a^{(2)}_{4,3}| = \max_{i=3\ldots,n} \{|a^{(2)}_{i,3}|\}$. Permutons alors, la $3^{\text{ème}}$ et la $4^{\text{ème}}$ ligne de $A^{(2)}$. Ce qui se traduit par les relations matricielles :

$$\tilde{A}^{(2)} = P_3 A^{(2)} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & -\frac{2}{7} & \frac{4}{7} \end{pmatrix} \text{ avec } P_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Finalement, en utilisant la combinaison linéaire

$$l_4 \leftarrow l_4 - l_{4,3} l_3, \text{ avec } l_{4,3} = \tfrac{1}{3},$$

la matrice $\tilde{A}^{(2)}$ est transformée en $A^{(3)} = L_3 \tilde{A}^{(2)}$, où

$$A^{(3)} = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix}, \quad L_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{3} & 1 \end{pmatrix}.$$

Ainsi, en posant $U = A^{(3)}$, on voit que $L_3 P_3 L_2 P_2 L_1 P_1 A = U$. Cette dernière relation peut être reformulée de la manière suivante

$$U = (L_3 P_3 L_2 P_2 L_1 P_1) A = \left(L_3' L_2' L_1'\right) (P_3 P_2 P_1) A,$$

où $L_k^{'}$ est égale à L_k à des permutations près. Plus précisément, on définit

$$L_3^{'} = L_3,\ \ L_2^{'} = P_3 L_2 P_3,\ \ L_1^{'} = P_3 P_2 L_1 P_2 P_3.$$

Finalement, on a

$$\underbrace{(P_3 P_2 P_1)}_{P} A = \underbrace{\left(L_3^{'} L_2^{'} L_1^{'}\right)^{-1}}_{L} U,$$

avec

$$L_3^{'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{3} & 1 \end{pmatrix},\ L_2^{'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{7} & 1 & 0 \\ 0 & \frac{3}{7} & 0 & 1 \end{pmatrix} \text{ et } L_1^{'} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{3}{4} & 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 & 0 \\ -\frac{1}{4} & 0 & 0 & 1 \end{pmatrix}.$$

c'est à dire $PA = LU$, où

$$P = P_3 P_2 P_1 = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix},\ L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{3}{4} & 1 & 0 & 0 \\ \frac{1}{2} & -\frac{2}{7} & 1 & 0 \\ \frac{1}{4} & -\frac{3}{7} & \frac{1}{3} & 1 \end{pmatrix} \text{ et } U = \begin{pmatrix} 8 & 7 & 9 & 5 \\ 0 & \frac{7}{4} & \frac{9}{4} & \frac{17}{4} \\ 0 & 0 & -\frac{6}{7} & -\frac{2}{7} \\ 0 & 0 & 0 & \frac{2}{3} \end{pmatrix}.$$

Remarques :

1. Au cours de la recherche du pivot, il se peut que le maximum soit atteint en deux (ou plus) lignes. Dans ce cas, on retiendra comme ligne du pivot la première ligne où est atteint l'élément maximal.

2. La décomposition $PA = LU$ peut être utilisée pour résoudre un système linéaire $Ax = b$. En effet :

$$\begin{aligned} Ax = b &\Leftrightarrow PAx = \tilde{b} \quad (\text{ avec } \tilde{b} = Pb), \\ &\Leftrightarrow LUx = \tilde{b}, \\ &\Leftrightarrow Ly = \tilde{b} \quad (\text{ avec } y = Ux) \end{aligned}$$

On voit alors que pour obtenir x la solution de $Ax = b$, il suffit de résoudre respectivement les deux systèmes triangulaires (inférieur et supérieur) suivants :

$$Ly = \tilde{b} \text{ où } y \text{ est l'inconnu, avec } \tilde{b} = Pb,$$

et

$$Ux = y.$$

Exemple : Résoudre (S) $\left\{ \begin{array}{lcl} 2x_1 + x_2 + 4x_4 & = & 1 \\ -4x_1 - 2x_2 + 3x_3 - 7x_4 & = & -3 \\ 4x_1 + x_2 - 2x_3 + 8x_4 & = & 1 \\ -3x_2 - 12x_3 - x_4 & = & -2 \end{array} \right.$

Matriciellement, (S) s'écrit $\underbrace{\begin{pmatrix} 2 & 1 & 0 & 4 \\ -4 & -2 & 3 & -7 \\ 4 & 1 & -2 & 8 \\ 0 & -3 & -12 & -1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}}_{x} = \underbrace{\begin{pmatrix} 1 \\ -3 \\ 1 \\ -2 \end{pmatrix}}_{b}$.

En effectuant la factorisation LU de A avec stratégie de pivotage partiel, on trouve que $PA = LU$,

avec

$$P=\begin{pmatrix}0&1&0&0\\0&0&0&1\\0&0&1&0\\1&0&0&0\end{pmatrix},\ L=\begin{pmatrix}1&0&0&0\\0&1&0&0\\-1&\frac{1}{3}&1&0\\-\frac{1}{2}&0&\frac{3}{10}&1\end{pmatrix}\ \text{et}\ U=\begin{pmatrix}-4&-2&3&-7\\0&-3&-12&-1\\0&0&5&\frac{4}{3}\\0&0&0&\frac{1}{10}\end{pmatrix}.$$

Ainsi

$$\begin{aligned}Ax=b &\Leftrightarrow PAx=\tilde{b}\quad(\text{ avec }\tilde{b}=Pb=(-3,-2,1,1)^T),\\ &\Leftrightarrow LUx=\tilde{b},\\ &\Leftrightarrow Ly=\tilde{b}\quad(\text{ avec }y=Ux)\end{aligned}$$

En résolvant le système triangulaire inférieur $Ly=\tilde{b}$, on trouve $y=(-3,-2,-\frac{4}{3},-\frac{1}{10})^T$. Finalement, en résolvant le système triangulaire supérieur $Ux=y$, on trouve $x=(2,1,0,-1)^T$.

Mise en œuvre de la décomposition LU.

En supposant que la décomposition LU de la matrice A existe et en identifiant le terme général $a_{i,j}$ de la matrice A avec les éléments du produit scalaire $l_i\,u^j$, où $l_i=(l_{i,1},\dots,l_{i,i-1},1,0\dots,0)$ est la i-ème ligne de la matrice L et $u^j=(u_{1,j},\dots,l_{j,j},0\dots,0)^t$ est la j-ème colonne de la matrice U, on a

$$a_{i,j}=\sum_{k=1}^{n}l_{i,k}\,u_{k,j}=\sum_{k=1}^{\min(i,j)}l_{i,k}\,u_{k,j},$$

car $l_{i,k}=0$ lorsque $k>i$ et $u_{k,j}=0$ pour $k>j$. Ainsi, on obtient :

- si $i\le j$, alors $a_{i,j}=\sum_{k=1}^{i}l_{i,k}\,u_{k,j}$ et donc $u_{i,j}=a_{i,j}-\sum_{k=1}^{i-1}l_{i,k}\,u_{k,j}$, (1),
- si $i>j$, alors $a_{i,j}=\sum_{k=1}^{j}l_{i,k}\,u_{k,j}$ et donc $l_{i,j}=\frac{1}{u_{j,j}}(a_{i,j}-\sum_{k=1}^{j-1}l_{i,k}\,u_{k,j})$, (2).

Les deux égalités (1) et (2) permettent de former les éléments non nuls de la i-ème ligne de la matrice U ainsi que les éléments non nuls de la i-ème colonne de la matrice L. En effet, en prenant $j=1$ dans l'égalité (1), nous avons $u_{1,i}=a_{1,i}$ pour $i=1,\dots,n$, et puis l'égalité (2) nous donne $l_{i,1}=\frac{a_{i,1}}{u_{1,1}}$.

En supposant avoir déterminé les colonnes (respectivement les lignes) $1,\dots,i-1$ de la matrice U (respectivement L), nous obtenons à partir de (1) (respectivement à partir de (2)) les éléments $j=i,\dots,n$ de la i-ème ligne de U (respectivement colonne de L), i.e., nous avons

$$\begin{cases}u_{i,j} &= a_{i,j}-\sum_{k=1}^{i-1}l_{i,k}\,u_{k,j} &\text{pour } j=i,i+1,\dots,n\\ l_{j,i} &= \frac{1}{u_{i,i}}(a_{j,i}-\sum_{k=1}^{i-1}l_{i,k}\,u_{k,j}) &\text{pour } j=i+1,\dots,n.\end{cases}$$

Remarquons que la division par $u_{i,i}$ est possible puisque par hypothèse on suppose que la factorisation LU existe. De plus, toute nouveau coefficient $u_{i,j}$ ou $l_{j,i}$ est obtenu en fonction des éléments de L et U déjà calculés. Ainsi, les relations précédentes et les propriétés vérifiées par les matrices L et U nous permettent de donner l'algorithme suivant :

Algorithme. *Décomposition LU.*

1. *Initialisation des matrices L et U.*
   $L = I_n$; $U = 0_n$;
2. *Calcul de la première colonne de L et de la première ligne de U.*
   Pour $i = 1, \ldots, n$
   $u_{1,i} = a_{1,i}$;
   $l_{i,1} = \dfrac{a_{i,1}}{a_{1,1}}$;
   Fin pour
3. *Calcul des éléments non nuls de la i-ème ligne de U et de la i-ème colonne de L.*
   Pour $i = 2, \ldots, n$
   Pour $j = i, \ldots, n$
   $u_{i,j} = a_{i,j} - \sum_{k=1}^{i-1} l_{i,k}\, u_{k,j}$;
   Fin pour
   Pour $j = i+1, \ldots, n$
   $l_{j,i} = \dfrac{1}{u_{i,i}} (a_{j,i} - \sum_{k=1}^{i-1} l_{j,k}\, u_{k,i})$;
   Fin pour
   Fin pour
4. *Phase de résolution du système triangulaire inférieur $Ly = b$.*
5. *Phase de résolution du système triangulaire supérieur $Ux = y$.*

Méthode de Cholesky et décomposition $A = LL^T$.

Théorème 1.3.
Soi $A \in \mathbb{R}^{n\times n}$ une matrice symétrique définie positive, alors il existe une matrice inférieure unique L ayant des éléments diagonaux positifs telle que $A = LL^T$.

La matrice L est appelée matrice de Cholesky associée à A.

Mise en œuvre de la décomposition LL^T.

Soit A une matrice symétrique définie positive et L une matrice triangulaire inférieure,
On cherche la matrice :

$$L = \begin{bmatrix} l_{1,1} & & & \\ l_{2,1} & l_{2,2} & & \\ \vdots & \vdots & \ddots & \\ l_{n,1} & l_{n,2} & \ldots & l_{n,n} \end{bmatrix}$$

De l'égalité $A = LLT$ et comme puisque $l_{p,q} = 0$ si $1 \le p < q \le n$. Alors, et identifiant les éléments de A avec ceux de $L\text{Ł}^T$, on a :

$$a_{i,j} = \left(LL^T\right)_{i,j} = \sum_{k=1}^{n} l_{i,k} l_{j,k} = \sum_{k=1}^{\min\{i,j\}} l_{i,k} l_{j,k},\ 1 \le i, j \le n.$$

Comme la matrice A est symétrique, il suffit que les relations précédentes soient vérifiées pour $i \le j$,

c'est-à-dire que les éléments $l_{i,j}$ de la matrice L doivent satisfaire :

$$a_{i,j} = \sum_{k=1}^{i} l_{i,k}\, l_{j,k},\ 1 \le i \le j \le n.$$

Ainsi, pour $i = 1$, on détermine la première colonne de L :

— $a_{1,1} = l_{1,1}\, l_{1,1}$, d'où $l_{1,1} = \sqrt{a_{1,1}}$.

— $a_{1,j} = l_{1,1}\, l_{j,1}$, d'où $l_{j,1} = \frac{a_{1,j}}{l_{1,1}},\ 2 \le j \le n$.

On détermine la i-ème colonne de $L (2 \le i \le n)$, après avoir calculé les $(i-1)$ premières colonnes :

— $a_{i,i} = l_{i,1}\, l_{i,1} + \ldots + l_{i,i}\, l_{i,i}$ d'où $l_{i,i} = \sqrt{a_{i,i} - \sum_{k=1}^{i-1} l_{i,k}^2}$

— $a_{j,i} = l_{j,1}\, l_{i,1} + l_{j,2}\, l_{i,2} + \ldots + l_{j,i}\, l_{i,i}$ d'où $l_{j,i} = \frac{a_{j,i} - \sum_{k=1}^{i-1} l_{j,k}\, l_{i,k}}{l_{i,i}},\ i+1 \le j \le n$.

On voit alors qu'il y a unicité de la décomposition L^T si on choisit tous les éléments $l_{i,i} > 0$.
En résumant ce qui précède, on peut établir l'algorithme suivant :

Algorithme. *Décomposition $L L^T$.*

1. *Initialisation de la matrice L .*
   $L = 0_n$;
2. *Calcul de la première colonne de L.*
   $l_{1,1} = \sqrt{a_{1,1}}$
   Pour $i = 2, \ldots, n$
   $$l_{i,1} = \frac{a_{i,1}}{l_{1,1}}\ ;$$
   Fin pour
3. *Calcul des éléments non nuls de la i-ème colonne de L.*
   Pour $i = 2, \ldots, n$
   $$l_{i,i} = \sqrt{a_{i,i} - \sum_{k=1}^{i-1} l_{i,k}^2}\ ;$$
   Pour $j = i+1, \ldots, n$
   $$l_{j,i} = \frac{1}{l_{i,i}} (a_{j,i} - \sum_{k=1}^{i-1} l_{j,k}\, l_{i,k})\ ;$$
   Fin pour
   Fin pour
4. *Phase de résolution du système triangulaire inférieur $L\, y = b$.*
5. *Phase de résolution du système triangulaire supérieur $L^T x = y$.*

1.4 Méthodes semi-itératives

Lorsque la taille des systèmes à résoudre est grande, et vu que la résolution d'un système linéaire $A\, x = b$ de taille n requiert en général $\mathcal{O}(n^3)$ opérations, les méthodes directes sont coûteuses en terme de nombre d'opérations et donc aussi en terme de temps d'exécution sur ordinateur. Dans ce cas, il est préférable d'utiliser des méthodes itératives qui construisent des solutions approchées à la solution exacte du système.

Dans cette section, nous allons décrire une classe de méthodes itératives basées sur la décomposition de la matrice A sous la forme de matrices plus faciles à manipuler ! Plus précisément, étant donnée A est une matrice carrée inversible, on suppose qu'on puisse écrire $A = M - N$ où les matrices M et N sont "convenablement choisies".

Afin de résoudre le système $Ax = b$, on remplace l'équation du système par la nouvelle équation :

$$Mx = Nx + b,$$

et en supposant que la matrice M est facilement inversible, nous obtenons

$$x = Bx + c,$$

où $B = M^{-1}N$ et $c = M^{-1}b$. On voit alors qu'on peut définir le schéma itératif suivant :

$$\begin{cases} x^{(k+1)} = Bx^{(k)} + c \\ x^{(0)} \text{ donné ou à choisir} \end{cases}$$

La convergence ou la non convergence de la suite $x^{(k)}$ vers la solution exacte $x^* = A^{-1}b$ du système dépend du rayon spectrale de la matrice B. Plus précisément, nous avons le résultat suivant

Théorème 1.4.
La suite $x^{(k)}$ définie par $x^{(k+1)} = Bx^{(k)} + c$ converge pour tout $x^{(0)} \in \mathbb{R}^n$ si et seulement si $\rho(B) < 1$.

Dans la suite, nous décomposons la matrice A sous la forme

$$A = \begin{pmatrix} \ddots & & -F \\ & D & \\ -E & & \ddots \end{pmatrix},$$

c'est à dire $A = D - E - F$ où

- $D = diag(a_{1,1}, \ldots, a_{n,n})$ est la matrice diagonale formée par les éléments diagonaux de A.
- $-E = tril(A)$ est la matrice triangulaire inférieure formée par la partie triangulaire inférieure stricte de A.
- $-F = triu(A)$ est la matrice triangulaire supérieure formée par la partie triangulaire supérieure stricte de A.

De plus, nous supposons que les éléments diagonaux de A sont tous non nuls et ainsi D est inversible.

1.4.1 Méthode de Jacobi

Le schéma itératif de la méthode de Jacobi s'obtient en prenant $M = D$ et $N = E + F$. Ainsi, ce schéma s'écrit

$$\begin{cases} x^{(k+1)} = B_J x^{(k)} + c_J & (J) \\ x^{(0)} \text{ donné ou à choisir} \end{cases},$$

avec $B_J = M^{-1}N = D^{-1}(E + F)$ et $c_J = D^{-1}b$.

Notons que l'égalité (J) entraine que $D\,x^{(k+1)} = (E = F)x^{(k)} + b$, et donc on a :

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

Ainsi, le schéma itératif de la méthode de Jacobi s'écrit :

$$\left\{ \begin{array}{l} x_i^{(k+1)} = \dfrac{1}{a_{i,i}}\left(b_i - \displaystyle\sum_{\substack{j=1\\ j\neq i}}^{n} a_{i,j}\,x_j^{(k)} \right), \qquad \text{pour } i = 1,\ldots,n. \\ x^{(0)} \text{ donné ou à choisir} \end{array} \right. ,$$

et la matrice B_J est telle que

$$B_J = \begin{pmatrix} 0 & -\dfrac{a_{1,2}}{a_{1,1}} & \cdots & \cdots & -\dfrac{a_{1,n}}{a_{1,1}} \\ -\dfrac{a_{2,1}}{a_{2,2}} & 0 & -\dfrac{a_{2,3}}{a_{2,2}} & \cdots & -\dfrac{a_{2,n}}{a_{2,2}} \\ -\dfrac{a_{2,1}}{a_{3,3}} & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & -\dfrac{a_{n-1,n}}{a_{n-1,n-1}} \\ -\dfrac{a_{n,1}}{a_{n,n}} & \cdots & \cdots & -\dfrac{a_{n,n-1}}{a_{n,n}} & 0 \end{pmatrix}.$$

Exemples :

- *Exemple 1.* Soit à résoudre le système

$$(S_1) \left\{ \begin{array}{rcl} 4x_1 + x_2 & = & 3 \\ x_1 - 2x_2 & = & -15. \end{array} \right.$$

Pour obtenir le schéma itératif de la méthode de Jacobi associé au système (S_1) ainsi que la matrice B_J et le vecteur c_J correspondants, il suffit d'écrire :

$$\left\{ \begin{array}{rcl} 4x_1^{(k+1)} + x_2^{(k)} & = & 3 \\ x_1^{(k)} - 2x_2^{(k+1)} & = & -15. \end{array} \right. ,$$

ce qui donne

$$\left\{ \begin{array}{rcl} x_1^{(k+1)} & = & \dfrac{1}{4}(3 - x_2^{(k)}) \\ x_2^{(k+1)} & = & \dfrac{1}{2}(15 + x_1^{(k)}), \end{array} \right. ,$$

et ainsi $x^{(k+1)} = B_J\,x^{(k)} + c_J$, avec

$$B_J = \begin{pmatrix} 0 & -\dfrac{1}{4} \\ \dfrac{1}{2} & 0 \end{pmatrix} \text{ et } c_J = \begin{pmatrix} \dfrac{3}{4} \\ \dfrac{15}{2} \end{pmatrix}.$$

Pour savoir si la méthode de Jacobi appliquée au système (S_1) est convergente ou non, il suffit de détérminer le rayon spectrale de la matrice B_J ; Et pour cela, nous pouvons calculer le polynôme

caractéristique de B_J :

$$P_{B_J}(\lambda) = det(b_{B_J} - \lambda I) = \begin{vmatrix} -\lambda & -\frac{1}{4} \\ \frac{1}{2} & -\lambda \end{vmatrix} = \lambda^2 + \frac{1}{8}.$$

Ce qui donne que les valeurs propres de B_J sont $\lambda_1 = \bar{\lambda}_2 = \frac{\imath}{\sqrt{8}}$ et donc $\rho(B_J) = \frac{1}{\sqrt{8}} < 1$.
Ainsi, la méthode de Jacobi appliquée au système (S_1) est convergente pour tout $x^{(0)} \in \mathbb{R}^2$.

En partant de $x^{(0)} = (0,0)^T$, nous obtenons $x^{(1)} = (\frac{3}{4},\ \frac{15}{2})^T$, $x^{(2)} = (-1.125\cdots,\ 7.875\cdots)^T$, ..., $x^{(10)} = (-1.000\cdots,\ 7.000\cdots)^T$. Et, on remarque alors que $x^{(k)} \longrightarrow_{k\to+\infty} (-1,\ 7)^T$.

- *Exemple 2.* Dans cet exemple, nous nous intéressons encore au système (S_1) de l'exemple précédent que nous réécrivons comme suit

$$(S_2) \begin{cases} y_1 + 4y_2 &= 3 \\ -2y_1 + y_2 &= -15. \end{cases}$$

C'est à dire qu'on a $y_1 = x_2$ et $y_2 = x_1$.

En suivant la même démarche suivie pour l'exemple précédent, nous obtenons le schéma itératif associé au système (S_2) ainsi que la matrice B_J et le vecteur c_J correspondants en écrivant

$$\begin{cases} y_1^{(k+1)} + 4y_2^{(k)} &= 3 \\ -2y_2^{(k)} + y_1^{(k+1)} &= -15. \end{cases},$$

et ainsi,

$$\begin{cases} y_1^{(k+1)} &= 3 - 4y_2^{(k)} \\ y_2^{(k+1)} &= -15 + 2y_1^{(k)}, \end{cases}.$$

ce qui donne $y^{(k+1)} = B_J^{'} x^{(k)} + c_J^{'}$, avec

$$B_J^{'} = \begin{pmatrix} 0 & -4 \\ 2 & 0 \end{pmatrix} \text{ et } c_J^{'} = \begin{pmatrix} 3 \\ -15 \end{pmatrix}.$$

Et, en calculant le polynôme caractéristique de la matrice $B_J^{'}$, nous trouvons que $P_{B_J^{'}}(\lambda) = \lambda^2 - 8$, c'est à dire que $\rho(B_J^{'}) = \sqrt{8} > 1$.

Ainsi, la méthode de Jacobi appliquée au système (S_2) n'est pas convergente. Ce résultat est confirmé numériquement, puisqu'on constate que $x^{(k)} \to (\infty,\infty)^T$ lorsque $k \to +\infty$.

Nous terminons ce paragraphe en donnant l'algorithme de la méthode de Jacobi

Algorithme. *Méthode de Jacobi.*

1. *Initialisation :*
   Choisir $x^{(0)}$ une solution de départ.
2. *Itération :*
   Pour $k = 0, 1, \ldots$ jusqu'a convergence
   Pour $i = 1, \ldots, n$

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

   Fin pour
   Fin pour

1.4.2 Méthode de Gauss-Seidel

Le choix $M = D - E$ et $N = F$ permet de définir la méthode de Gauss-Seidel et dans ce case le schéma itératif s'écrit

$$\begin{cases} x^{(k+1)} = B_G x^{(k)} + c_G & (J) \\ x^{(0)} \text{ donné ou à choisir} \end{cases},$$

avec $B_G = M^{-1}N = (D-E)^{-1}F$ et $c_G = (D-E)^{-1}b$.

Notons que dans la pratique, on ne calcule pas $(D-E)^{-1}$, mais on résout le système $(D-E)x^{(k+1)} = Fx^{(k)} + Fb$. Ce dernier système est un système triangulaire inférieure, donc facile à résoudre numériquement. On obtient ainsi,

$$\sum_{j=1}^{i} a_{i,j} x_j^{(k+1)} + \sum_{j=i+1}^{n} a_{i,j} x_j^{(k)} = b_i, \text{ pour } i = 1, \ldots, n,$$

ou encore :

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

Ceci permet d'obtenir le schéma itératif de la méthode de Gauss-Seidel qui s'écrit

$$\begin{cases} x_i^{(k+1)} = \dfrac{1}{a_{i,i}} \left(b_i - \displaystyle\sum_{j=1}^{i-1} a_{i,j} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j} x_j^{(k)} \right), & \text{pour } i = 1, \ldots, n. \\ x^{(0)} \text{ donné ou à choisir} \end{cases},$$

Il est à remarquer que le calcul de $x_i^{(k+1)}$ -la i-ème composante de la nouvelle solution approchée $x^{(k+1)}$- fait intervenir les valeurs des $x_j^{(k)}$ pour $j > i$ et des $x_j^{(k+1)}$ pour $j < i$. Il est donc nécessaire d'avoir calculé $x_1^{(k+1)}, \ldots, x_{i-1}^{(k+1)}$ pour pouvoir calculer $x_i^{(k+1)}$. Ceci n'est pas nécessaire pour la méthode de Jacobi où le calcul de $x_i^{(k+1)}$ ne fait intervenir que les composantes de la solution approchée précédente $x^{(k)}$.

Exemples :

- *Exemple 1.* Soit à résoudre le système

$$(S_1) \begin{cases} 4x_1 + x_2 &= 3 \\ x_1 - 2x_2 &= -15. \end{cases}$$

Pour obtenir le schéma itératif de Gauss-Seidel associé au système (S_1) ainsi que la matrice B_G et le vecteur c_G correspondants, il suffit d'écrire :

$$\left\{ \begin{array}{rcl} 4x_1^{(k+1)} + x_2^{(k)} & = & 3 \\ x_1^{(k+1)} - 2x_2^{(k+1)} & = & -15. \end{array} \right. ,$$

ce qui donne

$$\left\{ \begin{array}{rcl} x_1^{(k+1)} & = & \frac{1}{4}(3 - x_2^{(k)}) \\ x_2^{(k+1)} & = & \frac{1}{2}(15 + x_1^{(k+1)}) = \frac{1}{2}(15 + \frac{1}{4}(3 - x_2^{(k)})) = \ldots = \frac{63}{8} - \frac{1}{8}x_2^{(k)}. \end{array} \right.$$

et donc $x^{(k+1)} = B_G\, x^{(k)} + c_G$, avec

$$B_G = \begin{pmatrix} 0 & -\frac{1}{4} \\ 0 & -\frac{1}{8} \end{pmatrix} \text{ et } c_G = \begin{pmatrix} \frac{3}{4} \\ \frac{63}{8} \end{pmatrix}.$$

Maintenant, calculons le polynôme caractéristique de B_G pour déterminer son rayon spectrale.

$$P_{B_G}(\lambda) = det(b_{B_G} - \lambda I) = \begin{vmatrix} -\lambda & -\frac{1}{4} \\ 0 & -\frac{1}{8} - \lambda \end{vmatrix} = \lambda(\frac{1}{8} + \lambda).$$

Ce qui donne que les valeurs propres de B_G sont $\lambda_1 = \frac{-1}{8}$ et $\lambda_2 = 0$ et donc $\rho(B_G) = \frac{1}{8} < 1$. Ainsi, la méthode de Gauss-Seidel appliquée au système (S_1) est convergente pour tout $x^{(0)} \in \mathbb{R}^2$.

En partant de $x^{(0)} = (0,0)^T$, nous obtenons $x^{(1)} = (3,\ -15)^T$, $x^{(2)} = (-1.218\cdots,\ 6.890\cdots)^T$, …, $x^{(7)} = (-1.000\cdots,\ 7.000\cdots)^T$. Et, comme pour la méthode de Jacobi on remarque alors que $x^{(k)} \longrightarrow_{k\to+\infty} (-1,\ 7)^T$. De plus, pour cet exemple, la méthode de Gauss-Seidel converge plus vite que la méthode de Jacobi puisque $\rho(B_G) < \rho(B_J)$.

- *Exemple 2.* Si on applique la méthode de Gauss-Seidel au système (S_2) ci dessous

$$(S_2) \quad \left\{ \begin{array}{rcl} y_1 + 4y_2 & = & 3 \\ -2y_1 + y_2 & = & -15. \end{array} \right.$$

on trouve que $y^{(k+1)} = B_G^{'}\, x^{(k)} + c_G^{'}$, avec

$$B_G^{'} = \begin{pmatrix} 0 & -4 \\ 0 & -8 \end{pmatrix} \text{ et } c_G^{'} = \begin{pmatrix} -15 \\ -9 \end{pmatrix}.$$

De plus, on peut vérifier que $\rho(B_G^{'}) = 8 > 1$. Et donc la méthode de Gauss-Seidel (comme la méthode de Jacobi) appliquée au système (S_2) n'est pas convergente.

Nous terminons ce paragraphe en donnant l'algorithme de la méthode de Gauss-Seidel.

Algorithme. *Méthode de Gauss-Seidel.*

1. *Initialisation :*
   Choisir $x^{(0)}$ une solution de départ.
2. *Itération :*
   Pour $k = 0, 1, \ldots$ jusqu'a convergence
   Pour $i = 1, \ldots, n$
   $$x_i^{(k+1)} = \frac{1}{a_{i,i}} \left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{i,j} x_j^{(k)} \right).$$
   Fin pour
   Fin pour

1.4.3 Convergence des méthodes de Jacobi et de Gauss-Seidel

Dans cette section, nous allons donner des résultats concernant la convergence des méthodes de Jacobi et de Gauss-Seidel dans le cas où la matrice A est une matrice symétrique définie positive, ainsi que dans le cas où la matrice A est à diagonale strictement dominante.

Théorème 1.5.
Si la matrice A est symétrique définie positive, alors la méthode de Gauss-Seidel est convergente. ***(Attention : La méthode de Jacobi, ne converge pas forcément !)***

Théorème 1.6.
Si la matrice A est une matrice à diagonale strictement dominante, alors A est inversible et de plus les méthodes de Jacobi et de Gauss-Seidel sont convergentes.

Enfin, on peut également établir le résultat suivant :

Théorème 1.7.
Soit A une matrice tridiagonale. Alors
$$\rho(B_{GS}) = \rho((B_J)^2),$$
et donc les méthodes de Jacobi et Gauss-Seidel convergent ou divergent simultanément. Si elles convergent, la méthode de Gauss-Seidel est la plus rapide.

Chapitre 2

Interpolation polynomiale

2.1 Introduction

Soient $[a,\ b]$ un intervalle de $\mathbb{R}$, $S=(x_i)_{0\leq i\leq n}$ une subdivision de $[a,\ b]$ et $f\ :\ [a,\ b]\longrightarrow\mathbb{R}$ une fonction connue aux $(n+1)$ points x_i $(i=0,1,\ldots,n)$ de la subdivision S, c'est à dire qu'on connaît les valeurs

$$y_i=f(x_i)\text{ pour }i=0,1\ldots,n.$$

Définition 2.1.
On dit qu'un polynôme p de degré inférieur ou égal à n (i.e., $deg(p)\leq n$) interpole f (ou encore interpole les valeurs $y_0,y_1,\ldots,y_n$ aux $(n+1)$ points $x_0,x_1,\ldots,x_n$ s'il vérifie les conditions d'interpolation suivantes :

$$p(x_i)=f(x_i)\quad(\textit{ou encore }p(x_i)=y_i)\quad\textit{pour }i=0,1,\ldots,n.$$

Questions.

1. un tel polynôme interpolant existe t'il ?
2. s'il existe, est-il unique ?
3. comment trouver ce polynôme ?
4. comment estimer l'erreur $e(x):=f(x)-p(x)$ (dite erreur d'interpolation) pour $x\neq x_i$?

Notation : Si $k\in\mathbb{N}$, on désigne par $\mathscr{P}_k$ ou $\mathbb{R}_k[X]$ l'espace vectoriel des polynômes à une indéterminée (variable X) de degré inférieur ou égal à k.

2.2 Détermination du polynôme d'interpolation

2.2.1 Cas $n=2$

On écrit le système :

$$(\mathscr{S})\quad\left\{\begin{array}{rcl}p(x_0)&=&y_0\\p(x_1)&=&y_1\\p(x_2)&=&y_2\end{array}\right.$$

Rappelons qu'on cherche $p\in\mathscr{P}_2=\mathbb{R}_2[X]$, c'est à dire qu'on cherche trois réels $a_0,\ a_1$ et a_2 tels que :

$$p=a_0+a_1X+a_2X^2.$$

Ainsi,le système $(\mathscr{S})$ est équivalent au système suivant dons les inconnus sont $a_0,\ a_1$ et a_2

$$(\mathscr{S}) \quad \left\{ \begin{array}{rcl} a_0+a_1x_0+a_2x_0^2 & = & y_0 \\ a_0+a_1x_1+a_2x_1^2 & = & y_1 \\ a_0+a_1x_2+a_2x_2^2 & = & y_2 \end{array} \right.$$

Matriciellement, le système $(\mathscr{S})$ s'écrit :

$$\underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}}_{=V(x_0,x_1,x_2)(=V)} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix}}_{=a} = \underbrace{\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}}_{=y}.$$

Rappelons que le système précédent admet une unique solution si $\Delta \neq 0$ où Δ est le déterminant du système $(\mathscr{S})$ et qui est donné par :

$$\begin{array}{rcl} \Delta = det(V) & = & \begin{vmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{vmatrix} \\ & = & \begin{vmatrix} 1 & 0 & 0 \\ 1 & x_1-x_0 & x_1^2-x_0x_1 \\ 1 & x_2-x_0 & x_2^2-x_0x_2 \end{vmatrix} \\ & = & (x_1-x_0)(x_2-x_0)\begin{vmatrix} 1 & x_1 \\ 1 & x_2 \end{vmatrix} \\ & = & (x_1-x_0)(x_2-x_0)(x_2-x_1). \end{array}$$

Ainsi, pour $n=2$, le problème d'interpolation une solution unique si x_0, x_1 et x_2. Dans ce cas, on a :

$$a = \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = V^{-1} \begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix},$$

où V^{-1} est la matrice inverse de V.
Une fois les coefficients $a_0,\ a_1$et a_2 sont calculés, on a :

$$\begin{array}{rcl} p(x)=a_0+a_1x+a_2x^2 & = & (1,\ x,\ x^2)\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \\ & = & (1,\ x,\ x^2)\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}^{-1}\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix} \\ & = & (t_0,\ t_1,\ t_2)\begin{pmatrix} y_0 \\ y_1 \\ y_2 \end{pmatrix}, \end{array}$$

avec $(t_0,\ t_1,\ t_2)=(1,\ x,\ x^2)\begin{pmatrix} 1 & x_0 & x_0^2 \\ 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \end{pmatrix}^{-1}$, c'est à dire en transposant :

$$\begin{pmatrix} t_0 \\ t_1 \\ t_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{pmatrix}^{-1} \begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix},$$

ou encore

$$\underbrace{\begin{pmatrix} 1 \\ x \\ x^2 \end{pmatrix}}_{=X} = \underbrace{\begin{pmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{pmatrix}}_{=M} \underbrace{\begin{pmatrix} t_0 \\ t_1 \\ t_2 \end{pmatrix}}_{=t}.$$

En résolvant le système $Mt = X$ (d'inconnue le vecteur colonne t), on obtient

$$t_0 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x & x_1 & x_2 \\ x^2 & x_1^2 & x_2^2 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{vmatrix}} = \frac{(x_1-x)(x_2-x)(x_2-x_1)}{(x_1-x_0)(x_2-x_0)(x_2-x_1)} = \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)},$$

ainsi que

$$t_1 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x & x_2 \\ x_0^2 & x^2 & x_2^2 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{vmatrix}} = \ldots = \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

et

$$t_2 = \frac{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x \\ x_0^2 & x_1^2 & x^2 \end{vmatrix}}{\begin{vmatrix} 1 & 1 & 1 \\ x_0 & x_1 & x_2 \\ x_0^2 & x_1^2 & x_2^2 \end{vmatrix}} = \ldots = \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}.$$

On pose alors $l_0(x) = t_0$, $l_1(x) = t_1$ et $l_2(x) = t_2$ (qu'on appelle polynômes de Lagrange associés aux points x_0, x_1 et x_2). Dans ce cas, on a :

$$p(x) = y_0 l_0(x) + y_1 l_1(x) + y_2 l_2(x).$$

2.2.2 Cas général

On cherche $p \in \mathscr{P}_n$ vérifiant le problème d'interpolation $p(x_i) = y_1$ pour $i = 0, 1, \ldots, n$, c'est à dire, on cherche $a_0, a_1, \ldots, a_n$ tels que $p = a_0 + a_1 X + \ldots + a_n X^n$ et vérifiant les $(n+1)$ équations ci-dessous :

$$(\mathscr{S}) \quad \begin{cases} a_0 + a_1 x_0 + \ldots + a_n x_0^n &= y_0 \\ a_0 + a_1 x_1 + \ldots + a_n x_1^n &= y_1 \\ \vdots & = \vdots \\ a_0 + a_1 x_n + \ldots + a_n x_n^n &= y_n \end{cases}$$

La forme matricielle de ce système est :

$$\underbrace{\begin{pmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{pmatrix}}_{V(x_0,x_1,\ldots,x_n)(=V)} \underbrace{\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix}}_{=a} = \underbrace{\begin{pmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}}_{=y}.$$

Le déterminant Δ du système $(\mathscr{S})$ précédent est appelé déterminant de Vandermonde associé aux coefficients $x_0, x_1, \ldots, x_n$. Ce déterminant s'écrit :

$$\Delta = det(V) = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix},$$

et on montre que :

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

D'ou le résultat suivant :

Théorème 2.1.
Le polynôme d'interpolation $p = a_0 + a_1 x + \ldots + a_n x^n$ interpolant les valeurs $y_0, y_1, \ldots, y_n$ existe et est unique si et seulement si $x_i \neq x_j, \ \forall i \neq j$.

Exemple.

- *Question.* On considère la subdivision $x_0 = 0$, $x_1 = 1$, $x_2 = 4$, $x_3 = 9$. Déterminer le polynôme d'interpolation de la fonction f définie pour tour tout $x \in \mathbb{R}^+$ par $f(x) = \sqrt{x}$ en x_0, x_1, x_2 et x_3.
- *Réponse.* En utilisant la base canonique, le polynôme recherché est sous la forme $p = a_0 + a_1 + a_2 x^2 + a_3 x^3$, où a_0, a_1, a_2 et a_3 sont quatre réels qu'on détermine comme ci-dessous :
  En écrivant

$$(\mathscr{S}) \quad \begin{cases} p(x_0) = (x_0) = y_0 \\ p(x_1) = f(x_1) = y_1 \\ p(x_2) = f(x_2) = y_2 \\ p(x_3) = f(x_3) = y_3 \end{cases},$$

  on obtient le système

$$(\mathscr{S}) \quad \begin{cases} a_0 = 0 \\ a_0 + a_1 + a_2 + a_3 = 1 \\ a_0 + 4a_1 + 16a_2 + 64a^3 = 2 \\ a_0 + 9a_1 + 81a_2 + 729a^3 = 3 \end{cases},$$

  qui s'écrit matriciellement sous la forme

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 4 & 16 & 64 \\ 1 & 9 & 81 & 729 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}.$$

Ainsi

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 4 & 16 & 64 \\ 1 & 9 & 81 & 729 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0 & 0 & 0 \\ -\frac{49}{36} & \frac{3}{2} & -\frac{3}{20} & \frac{1}{90} \\ \frac{7}{18} & -\frac{13}{24} & \frac{1}{6} & -\frac{1}{72} \\ -\frac{1}{36} & \frac{1}{24} & -\frac{1}{60} & \frac{1}{360} \end{pmatrix} \begin{pmatrix} 0 \\ 1 \\ 2 \\ 3 \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ \frac{37}{30} \\ -\frac{1}{4} \\ \frac{1}{60} \end{pmatrix}$$

Ainsi le polynome p recherche vérifie $p(x) = \frac{37}{30}x - \frac{1}{4}x^2 + \frac{1}{60}x^3$.

Remarque 2.1.

— *La taille du système $\mathscr{S}$ augment avec le nombre de points d'interpolation.*

— *la résolution du système $\mathscr{S}$ étant difficile à réaliser, on préfère utiliser les polynômes de Lagrange (ou encore les polynômes de Newton) pour déterminer le polynôme d'interpolation.*

2.3 Les polynômes de Lagrange

Définition 2.2.

On appelle polynômes de Lagrange relatifs (ou associés) aux $(n+1)$ points $x_0, x_1, \ldots, x_n$ (qu'on suppose deux à deux distincts) les $(n+1)$ polynômes notés $L_0, L_1, \ldots, L_n$ de degré n chacun et vérifiant :

$$L_i(x_j) = \delta_{i,j} := \begin{cases} 1 & si \quad i = j \\ 0 & si \quad i \neq j. \end{cases}$$

2.3.1 Expression des polynômes de Lagrange

Soit $i \in \{0, 1, \ldots, n\}$, le polynôme L_i, étant de degré n et s'annulant aux n points $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, alors, il existe $\alpha_i \in \mathbb{R}$ tel que ce polynôme s'écrit sous la forme :

$$L_i(x) = \alpha_i \prod_{\substack{j=0 \\ j \neq i}}^{n} (x - x_j).$$

De plus, comme $L_i(x_i) = 1$, il vient que $L_i(x_i) = \alpha_i \prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j) = 1$, ce qui entraîne que

$$\alpha_i = \frac{1}{\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)}.$$

Et finalement le i-ème polynôme de Lagrange est donnée par :

$$L_i(x) = \prod_{\substack{j=0 \\ j \neq i}}^{n} \left(\frac{x - x_j}{x_i - x_j} \right).$$

Exercice. Montrer que la famille $\mathscr{L} = \{L_0, L_1, \ldots, L_n\}$ est une base de l'espace vectoriel $\mathscr{P}_n$.

2.3.2 Le polynôme d'interpolation dans la base de Lagrange

La famille $L_0, L_1, \ldots, L_n$ des polynômes de Lagrange étant une base de l'espace $\mathscr{P}_n$, alors

$$\forall P \in \mathscr{P}_n,\ \exists!(\alpha_0, \alpha_1, \ldots, \alpha_n) \in \mathbb{R}^{n+1} \text{ tel que } P(x) = \sum_{i=0}^{n} \alpha_i L_i(x),\ \text{pour tout } x \in \mathbb{R}.$$

En particulier, pour $x = x_j$ ($j = 0, 1, \ldots, n$), on a :

$$P(x_j) = \sum_{i=0}^{n} \alpha_i L_i(x_j) = \alpha_j \quad \text{pour } j = 0, \ldots, n.$$

D'où la formule de Lagrange :

$$\forall P \in \mathscr{P}_n,\ \forall x \in \mathbb{R},\ P(x) = \sum_{i=0}^{n} P(x_i) L_i(x).$$

Ainsi, on a le résultat suivant :

Théorème 2.2.
Dans la base de Lagrange, le polynôme p_n interpolant une fonction f (ou des valeurs $y_0, y_1, \ldots, y_n$) aux points $x_0, x_1, \ldots, x_n$, est donné par :

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x) = \sum_{i=0}^{n} y_i L_i(x).$$

Exemple.

— *Question.* Cherchons p le polynôme interpolant la fonction $f(x) = \sqrt{x}$ aux points 0, 1, 4 et 9.

— *Réponse.* En notant $x_0 = 0$, $x_1 = 1$ $x_2 = 4$ et $x_3 = 9$ et en utilisant la base de Lagrange, on peut affirmer que :

$$p(x) := p_3(x) = f(x_0)L_0(x) + f(x_1)L_1(x) + f(x_2)L_2) + f(x_3)L_3(x),$$

où

$$L_0(x) = \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = -\frac{1}{36}(x-1)(x-4)(x-9),$$

$$L_1(x) = \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{1}{24}x(x-4)(x-9),$$

$$L_2(x) = \frac{(x-x_0)(x-x_1)(x-x_3)}{(x_2-x_0)(x_2-x_1)(x_2-x_3)} = -\frac{1}{60}x(x-1)(x-9),$$

et

$$L_3(x) = \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{1}{360}x(x-1)(x-4).$$

Et comme $f(x_0) = 0$, $f(x_1) = 1$, $f(x_2) = 2$ et $f(x_3) = 3$, alors

$$p(x) = L_1(x) + 2L_2(x) + 3L_3(x).$$

En retournant à la base canonique de $\mathscr{P}_3$, on obtient $p(x) = \frac{37}{30}x - \frac{1}{4}x^2 + \frac{1}{60}x^3$.

2.4 Les polynômes de Newton

Définition 2.3.
On appelle polynômes de Newton relatifs (ou associés) aux $(n+1)$ *points* $x_0, x_1, \ldots, x_n$ *(qu'on suppose deux à deux distincts) les* $(n+1)$ *polynômes notés* $N_0, N_1, \ldots, N_n$ *définis pour tout* $x \in \mathbb{R}$ *par :*

$$N_0(x) = 1,\ N_1(x) = x - x_0,\ \ldots,\ N_n(x) = (x-x_0)(x-x_1)\ldots(x-x_{n-1}),$$

c'est à dire

$$N_i(x) = (x-x_0)(x-x_1)\ldots(x-x_{i-1}) = \prod_{j=0}^{i-1}(x-x_j) \text{ pour } i = 0,1,\ldots,n.$$

Remarque 2.2.
Le dernier point x_n *n'intervient pas ans la définition des polynômes* $N_0, N_1, \ldots, N_n$.

Exercice. Montrer que la famille $\mathcal{N} = \{N_0, N_1, \ldots, N_n\}$ est une base de l'espace vectoriel $\mathscr{P}_n$.

2.4.1 Différences divisées

Définition 2.4.
Soient $x_0, x_1, \ldots, x_n$, $n+1$ *réels deux à deux distincts. On appelle différence divisée d'ordre k de la fonction* f *aux points* $x_{i_0}, x_{i_1}, \ldots, x_{i_k}$ *(où* $i_j \in \{0,1,\ldots,n\}$ *et* $i_j \neq i_l$ *pour* $j \neq l$*) l'expression notée* $[f, x_{i_0}, \ldots, x_{i_k}]$ *et définie par la récurrence :*

$$[f, x_{i_0}, \ldots, x_{i_k}] = \frac{[f, x_{i_1}, \ldots, x_{i_k}] - [f, x_{i_0}, \ldots, x_{i_{k-1}}]}{x_{i_k} - x_{i_0}},$$

Sachant que la différence divisée d'ordre 0 est définie par

$$[f, x_i] = f(x_i). \text{ pour } i = 0,1,\ldots,n.$$

Illustration.

- ordre 0 : $[f, x_i] = f(x_i)$, pour $i = 0,1,\ldots,n$.
- ordre 1 : $[f, x_i, x_j] = \dfrac{[f, x_j] - [f, x_i]}{x_j - x_i}$, pour $i, j = 0,1,\ldots,n$ et $i \neq j$.
- ordre 2 : $[f, x_i, x_j, x_k] = \dfrac{[f, x_j, x_k] - [f, x_i, x_j]}{x_k - x_i}$, pour $i, j, k = 0,1,\ldots,n$ et $i \neq j,\ i \neq k,\ j \neq k$.

2.4.2 Le polynôme d'interpolation dans la base de Newton

La famille des polynômes de Newton étant une base de l'espace vectoriel $\mathscr{P}_n$, alors on peux ut écrire le polynôme d'interpolation p d'une fonction f aux points $x_0, x_1, \ldots, x_n$ dans cette base. En effet, on montre par récurrence le résultat suivant : Ainsi, on a le résultat suivant :

Théorème 2.3.
Dans la base de Newton, le polynôme p_n *interpolant une fonction* f *aux points* $x_0, x_1, \ldots, x_n$, *est donné par :*

$$p_n(x) = \sum_{i=0}^{n} [f, x_0, x_1, \ldots, x_i] N_i(x).$$

Exemple.

— *Question.* Reprenons le précédent exemple et cherchons p le polynôme interpolant la fonction $f(x) = \sqrt{x}$ aux points 0, 1, 4 et 9.

— *Réponse.* En notant $x_0 = 0$, $x_1 = 1$ $x_2 = 4$ et $x_3 = 9$ et en utilisant la base de Newton, on peut affirmer que :

$$p(x) := p_3(x) = [f,x_0]\,N_0(x) + [f,x_0,x_1]\,N_1(x) + [f,x_0,x_1,x_2]\,N_2(x) + [f,x_0,x_1,x_2,x_3]\,N_3(x),$$

où

$$N_0(x) = 1,\ N_1(x) = (x-x_0) = x,\ N_2(x) = (x-x_0)(x-x_1) = x(x-1) \text{ et } N_3(x) = (x-x_0)(x-x_1)(x-x_2) = x(x-1)(x-4),$$

et les différences divisées sont calculés par le schéma suivant :

x_i	ordre 0	ordre 1	ordre 2	ordre 3
0	$0^{=[f,x_0]}$			
		$1^{=[f,x_0,x_1]}$		
1	1		$-\frac{1}{6}^{=[f,x_0,x_1,x_2]}$	
		$\frac{1}{3}$		$\frac{1}{60}^{=[f,x_0,x_1,x_2,x_3]}$
4	2		$-\frac{1}{60}$	
		$\frac{1}{5}$		
9	3			

Ainsi, dans la base de Newton, le polynôme d'interpolation p de f s'écrit :

$$p(x) = 0\,N_0(x) + 1\,N_1(x) - \frac{1}{6}N_2(x) + \frac{1}{60}N_3(x) = N_1(x) - \frac{1}{6}N_2(x) + \frac{1}{60}N_3(x).$$

De même, on vérifie que ce polynôme s'écrit dans la base canonique sous la forme $p(x) = \frac{37}{30}x - \frac{1}{4}x^2 + \frac{1}{60}x^3$.

Questions.

1. Que se passe t'il si on rajoute un point x_4 aux abscisses x_0, x_1 x_2 et x_3 lorsqu'on utilise :
   . la base de Lagrange ?
   . la base de Newton ?
2. Faut-il refaire tous les calculs ?
3. Laquelle des deux bases a plus d'intérêt ?

2.5 Erreur d'interpolation

Théorème 2.4.
Soit f une fonction de classe C^{n+1} sur $[a,\ b]$ où $a = \min\limits_{0\le i\le n}\{x_i\}$ et $b = \max\limits_{0\le i\le n}\{x_i\}$. Alors

$$\forall\ x \in [a,\ b],\ \exists\ \xi_x \in]a,\ b[,\ \textit{tel que}\ e(x) := f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x-x_i).$$

Preuve.
Comme $f(x_i) = p_n(x_i)$ pour $i = 0,1,\ldots,n$, alors il existe une fonction réelle h telle que

$$f(x) - p_n(x) = h(x)\prod_{i=0}^{n}(x-x_i).$$

Pour x fixé et $x \neq x_i$ $(i = 0, 1, \ldots, n)$, posons $g(t) = f(t) - p_n(t) - c\prod_{i=0}^{n}(x - x_i)$ où c est un paramètre réel choisi tel que $g(x) = 0$. Dans ce cas, la condition $g(x) = 0$ entraîne que

$$c = \frac{f(x) - p_n(x)}{\prod_{i=0}^{n}(x - x_i)}.$$

Maintenant, remarquons que la fonction g est de classe C^{n+1} sur $[a,\ b]$ et de plus

$$\begin{cases} g(x) = 0 \\ g(x_i) = 0 \quad \text{pour } i = 0, 1, \ldots, n, \end{cases}$$

c'est à dire que g admet $n+2$ racines distinctes deux à deux et est de de classe C^{n+1} sur $[a,\ b]$.
Cela entraîne que $g^{'}$ admet $n+1$ racines distinctes deux à deux et est de de classe C^{n} sur $[a,\ b]$.
Et en réitérant, on a $g^{''}$ admet n racines distinctes deux à deux et est de de classe C^{n-1} sur $[a,\ b]$ et finalement, on arrive à vérifier que $g^{(n+1)}$ admet une racine sur $]a,\ b[$. Ainsi,

$$\exists\ \xi_x \in]a,\ b[,\ \text{tel que } g^{(n+1)}(\xi_x) = 0.$$

Or, $g^{(n+1)}(t) = f^{(n+1)}(t) - p_n^{(n+1)}(t) - c(n+1)!$, et puisque $p_n^{(n+1)}(t) = 0$ (car $deg(p_n) \leq n$) et $g^{(n+1)}(\xi_x) = 0$, alors on vérifie que

$$c = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x - x_i),$$

c'est à dire que

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x - x_i).$$

Remarque 2.3.

— *En posant* $M_{n+1} = \sup_{t\in[a,\ b]}\left|f^{(n+1)}(t)\right|$, *on a*

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

— *La majoration de l'erreur d'interpolation donnée ci-dessus peut laisser croire que si le nombre d'abscisses d'interpolation n est grand, alors le polynôme d'interpolation* p_n *de* f *aux points d'interpolation* $x_0, x_1, \ldots, x_n$ *tend vers* f. *En fait, on n'a pas nécessairement* $\lim_{n\to+\infty} f(x) - p_n(x) = 0$ *pour tout* $x \in [a,\ b]$ *car cette limite dépend aussi de la façon dont la quantité* M_n *se comporte lorsque la valeur de n devient large. L'exemple ci dessous permet d'illustrer cette remarque.*

Exemple. *Interpolation de la fonction* $f(x) = \dfrac{1}{1+25x^2}$

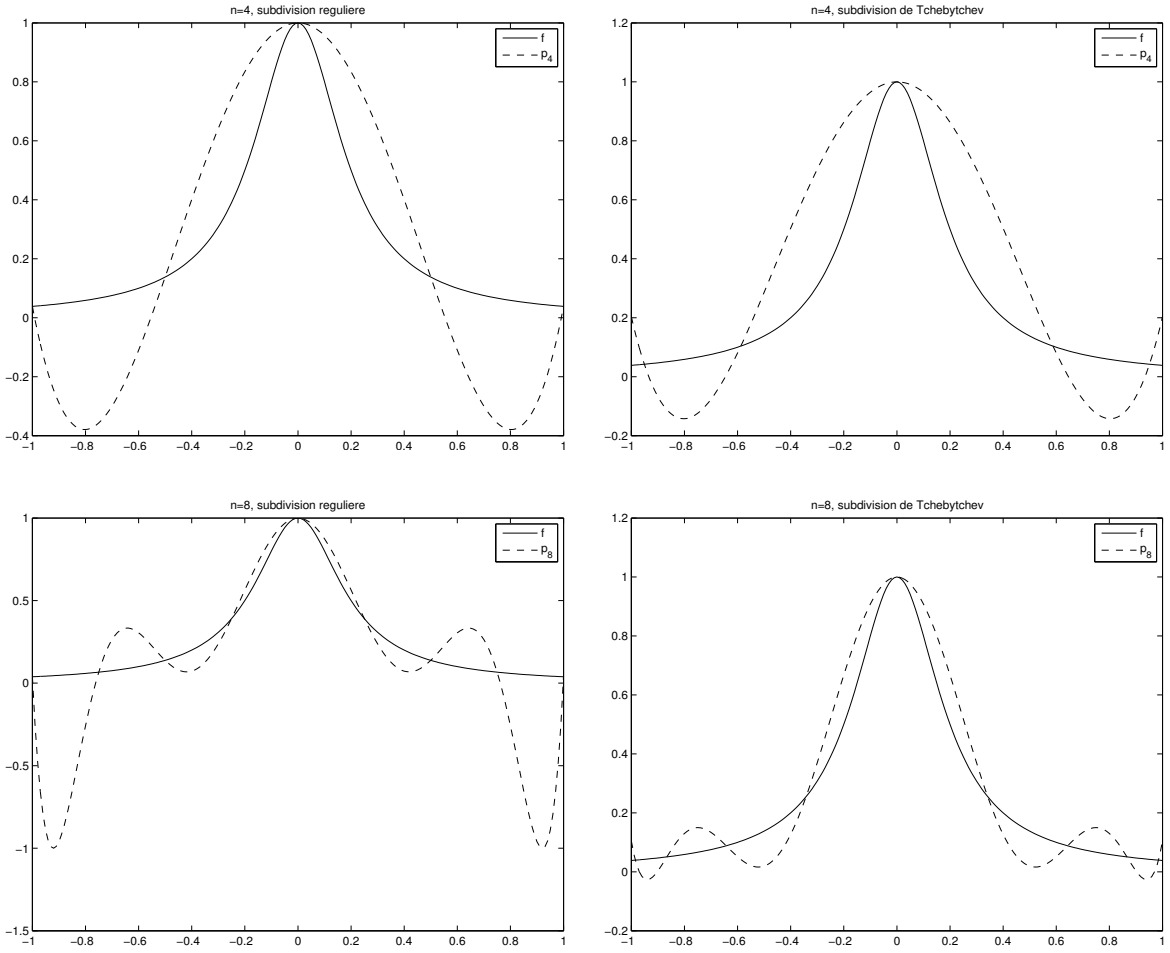
Soit f la fonction réelle définie sur $\mathbb{R}$ par $f(x) = \dfrac{1}{1+25x^2}$ pour tout $x \in \mathbb{R}$. On cet exemple, on illustre l'interpolation de cette fonction par des polynômes de degré n ($n = 4,\ 8,\ 10$) associés à deux subdivisions de l'intervalle $[a,\ b]$ ($a = -1$et $b = 1$). Pour cela, on considère $(x_i)_{i=0,1,\ldots,n}$ et $(x_i^{'})_{i=0,1,\ldots,n}$ les deux subdivisions suivantes :

— la première subdivision $(x_i)_{i=0,1,\ldots,n}$ est la subdivision régulière associée à $[a,\ b]$ et est définie par :

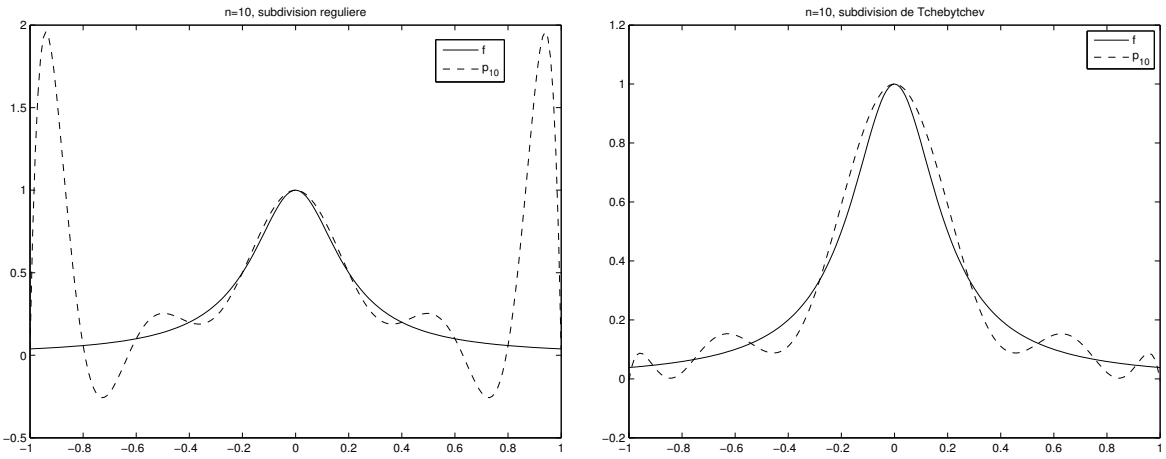
$$x_i = a + i\,h, \quad \text{pour } i0,1,\ldots,n, \text{et où } h = \frac{b-a}{n}.$$

— la deuxième subdivision $(x_i^{'})_{i=0,1,\ldots,n}$ est la subdivision de Tchebytchev associée à $[a,\ b]$ et est définie par :

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



Nous observons que, au voisinage des extrémités de l'intervalle $[-1,\ 1]$, le polynôme associé à la subdivision régulière $x_0, x_1, \ldots, x_n$ présente de grandes oscillations (instabilités numériques). Il n'en est pas de même pour le polynôme associé à la subdivision de Tchebytcehv $x_0^{'}, x_1^{'}, \ldots, x_n^{'}$. Ainsi, il n'est pas recommandé d'interpoler une fonction f par un polynôme de degré n élevé en des points équidistants.



2.6 l'algorithme de Neville-Aitken

Etant donnée une fonction f connue aux points d'une subdivision $x_0, x_1, \ldots, x_n$ de l'intervalle $[a,\ b]$. Si les $(n+1)$ abscisses $x_0, x_1, \ldots, x_n$ sont distinctes deux à deux, alors il existe un unique polynôme P de degré inférieur ou égale à n interpolant f aux $(n+1)$ points $x_0, x_2, \ldots, x_n$. Un tel polynôme peut être calculé en utilisant soit la base de Lagrange soit la base de Newton ou encore en utilisant de le procédé de Neville-Aitken. Dans ce dernier cas, le polynôme d'interpolation P est calculé de façon récursif.

2.6.1 Description de l'algorithme

Soit $k \in \{0,1,\ldots,n\}$ et $m \in \{0,1,\ldots,n-k\}$. Désignons par $P_{m,k}$ le polynôme de degré inférieur ou égal à k (i.e., $deg(P_{m,k}) \leq k$) interpolant f aux $(k+1)$ points $x_m, x_{m+1}, \ldots, x_{m+k}$. Le polynôme $P_{m,k}$ existe et est unique car les points $x_m, x_{m+1}, \ldots, x_{m+k}$ sont distinctes deux à deux et on a

$$\forall x \in \mathbb{R},\ \ P_{m,0}(x) = f(x_m),\ \text{pour } m = 0,1,\ldots,n-k.$$

En effet, comme $deg(P_{m,0}) = 0$, alors $deg(P_{m,0})$ est un polynôme constant et ainsi $\forall x \in \mathbb{R},\ \ P_{m,0}(x) = c$ (où c est constante réelle). Et comme $P_{m,0}$ interpole f en un seul point x_m, alors $c = P_{m,0}(x_m) = f(x_m)$.

On peut également établir le résultat suivant :

Proposition 2.1.
Pour tout $k \in \{1,2,\ldots,n\}$, on a

$$\forall x \in \mathbb{R},\ \ P_{m,k}(x) = \frac{(x-x_m)P_{m+1,k-1}(x) - (x-x_{m+k})P_{m,k-1}(x)}{x_{m+k} - x_m},\ \textit{pour } m = 0,1,\ldots,n-k.$$

Preuve :
Posons

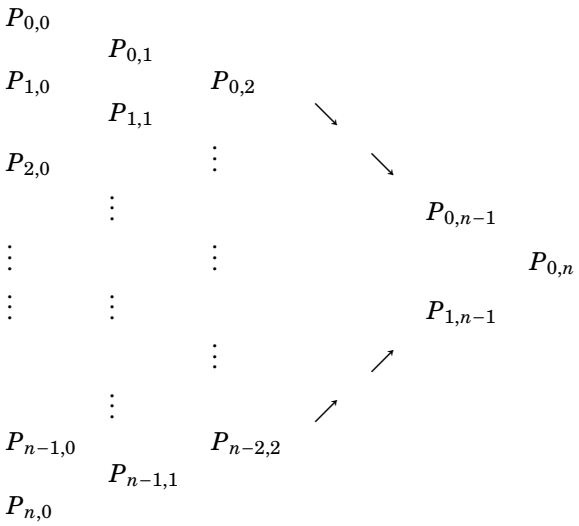
$$Q(x) = \frac{(x-x_m)P_{m+1,k-1}(x) - (x-x_{m+k})P_{m,k-1}(x)}{x_{m+k} - x_m},$$

alors Q est un polynôme de degré inférieur ou égale à k vérifiant

$$Q(x_m) = f(x_m),\ \ Q(x_{m+k}) = f(x_{m+k}),\ \text{et } Q(x_j) = f(x_j) \text{ pour } j = m+1, m+2, \ldots, m+k-1.$$

Ainsi $Q = P_{k,m}$ d'après l'unicité du polynôme d'interpolation.

Le résultat précédent montre donc que l'on peut calculer le polynôme P de degré inférieur ou égale à n interpolant f aux $(n+1)$ points $x_0, x_2, \ldots, x_n$ puisque $P = P_{0,n}$. En fait, on a le schéma itératif suivant :



2.6.2 Mise en oeuvre de l'algorithme

Soit $x \in \mathbb{R}$ et P le polynôme interpolant une fonction f en $x_0, x_1, \ldots, x_n$. Pour calculer $P(x)$, on peut utiliser l'algorithme suivant :

Algorithme : *Schéma de Neville-Aitken*

- *Initialisation : pour* $m = 0, 1 \ldots, n$ *faire*
  $\alpha_m = f(x_m)$;
  fin pour.

- *Itération : pour* $k = 1, 2, \ldots, n$ *faire*
  pour $m = 0, 1, \ldots, n-k$ *faire*
  $$\alpha_m = \frac{(x - x_m)\,\alpha_{m+1} - (x - x_{m+k})\,\alpha_m}{x_{m+k} - x_m} ;$$
  fin pour.
  fin pour.

Chapitre 3

Intégration numérique

3.1 Quelques outils de base

Avant d'aborder la sujet de ce chapitre, rappelons ou donnons quelques outils mathématiques de base qui sont nécessaires pour aborder ce chapitre.

Théorème 3.1. *(1ère formule de la moyenne -cas continu-)*
Soient u et v deux fonctions continues sur $[a,\ b]$ telles que u est de signe constant dans $[a,\ b]$. Alors

$$\exists \eta \in]a,\ b[\quad \textit{tel que} \int_a^b u(t)v(t)dt = v(\eta)\int_a^b u(t)dt.$$

Théorème 3.2. *(1ère formule de la moyenne -cas discret-)*
Soient v une fonction continue sur $[a,\ b]$, $t_1, t_2, \ldots, t_s$, $s+1$ points de l'intervalle $[a,\ b]$ et $u_1, u_2, \ldots, u_s$, $s+1$ constantes, toutes de même signe. Alors

$$\exists \eta \in [a,\ b] \quad \textit{tel que} \sum_{k=0}^{s} u_k\, v(t_k) = v(\eta) \sum_{k=0}^{s} u_k.$$

Théorème 3.3. *(1ère formule de l'erreur d'interpolation)*
Soient f une fonction de classe C^{n+1} sur $[a,\ b]$ et p_n le polynôme d'interpolation de la fonction f aux abscisses $x_0, x_1, \ldots, x_n$ ($x_i \in [a,\ b]$ pour $i = 0, 1, \ldots, n$). Alors, pour tout $x \in \mathbb{R}$, on a

$$\exists\ \xi_x \in]a,\ b[, \textit{ tel que } e(x) := f(x) - p_n(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!} \prod_{i=0}^{n} (x - x_i).$$

Théorème 3.4. *(2ème formule de l'erreur d'interpolation)*
Soient f une fonction de classe C^{n+1} sur $[a,\ b]$ et p_n le polynôme d'interpolation de la fonction f aux abscisses $x_0, x_1, \ldots, x_n$ ($x_i \in [a,\ b]$ pour $i = 0, 1, \ldots, n$). Alors, pour tout $x \in \mathbb{R}$, on a

$$e(x) := f(x) - p_n(x) = [f, x_0, x_1, \ldots, x_n, x] \prod_{i=0}^{n} (x - x_i).$$

Théorème 3.5. *(Théorème de Cauchy) Soit f une fonction de classe C^{n+1} sur l'intervalle $[a,\ b]$ contenant les points deux à deux distincts x_i, $i = 0, 1, \ldots, n$. Alors pour tout $x \in [\min_i x_i,\ \max_i x_i]$*

$$\exists \xi_x \in [\min_i x_i,\ \max_i x_i] \textit{ tel que } [f, x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}.$$

3.2 Introduction

— On désire calculer les intégrales suivantes :

$$\int_0^1 e^{-x^2}\,dx, \quad \int_0^{\frac{\pi}{2}} \sqrt{1+\cos^2 x}\,dx, \quad \int_{-1}^1 e^{\sin x}\,dx, \quad \ldots$$

— **Problème :** on n'a pas d'expression analytique de la primitive des fonctions : $x \longmapsto e^{-x^2}$, $x \longmapsto \sqrt{1+\cos^2 x}$, $x \longmapsto e^{\sin x}$, ...

— **Solution :** on va appliquer des méthodes numériques pour évaluer (approcher, approximer) la valeur de l'intégrale donnée

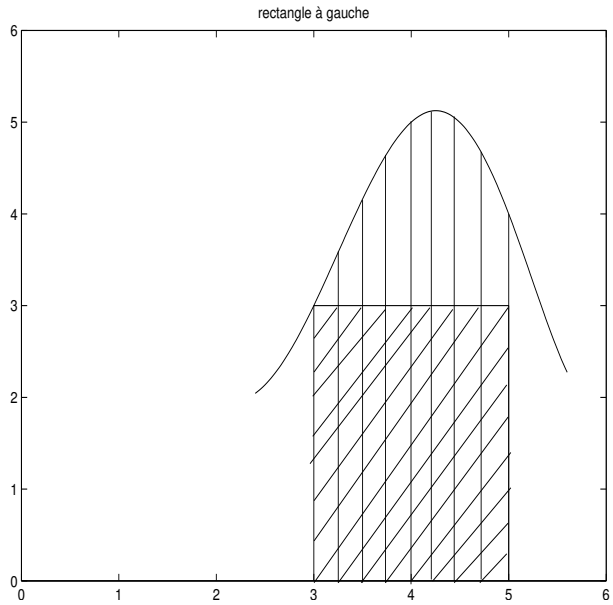
Ainsi, le problème posé peut être formulé de la façon suivante : étant donnée une fonction $f : [a, b] \to \mathbb{R}$ continue (ou dérivable, de classe $\mathbb{C}^\infty$, ...), on se propose de calculer numériquement la quantité

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

3.3 Quelques formules d'intégration "simples"

3.3.1 Formule du rectangle

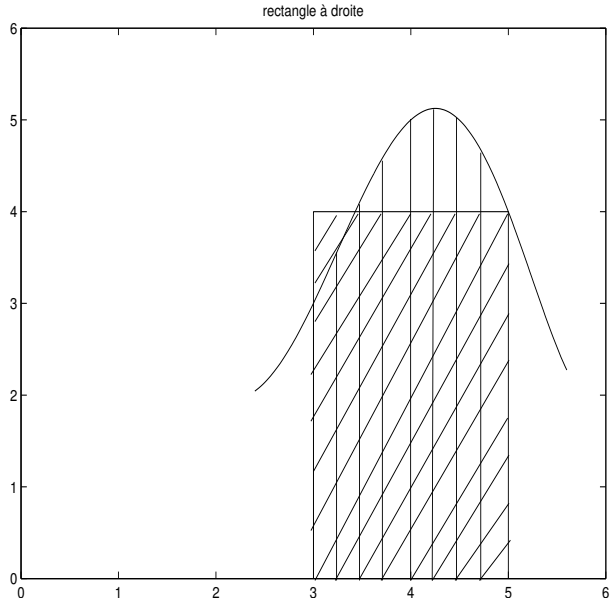
- **Rectangle à gauche**. La valeur de $I(f)$ est approchée par l'aire du rectangle $\mathscr{R}$ de sommets $(a,0)$, $(a,f(a))$, $(b,f(a))$, $(b,0)$ comme illustré par la figure suivante :



Ainsi on a

$$I(f) \approx (b-a)f(a).$$

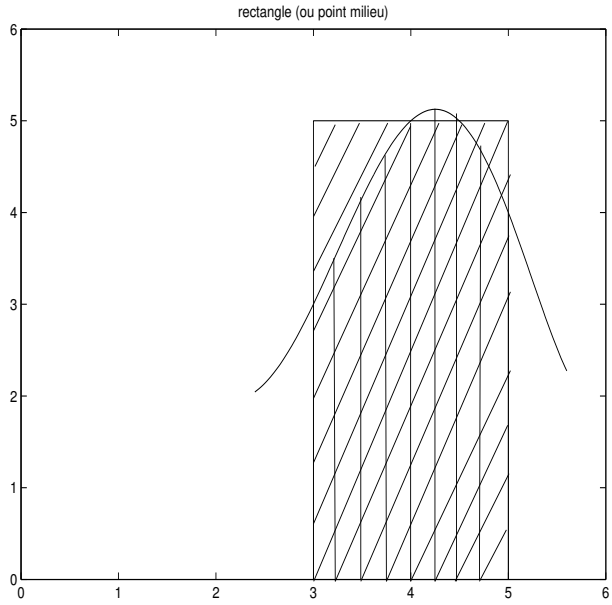
- **Rectangle à droite**. La valeur de $I(f)$ est approchée par l'aire du rectangle $\mathscr{R}$ de sommets $(b,0)$, $(b,f(b))$, $(b,f(b))$, $(a,0)$ comme illustré par la figure suivante :



Ainsi on a

$$I(f) \approx (b-a) f(b).$$

- **Rectangle** (ou **point milieu**). La valeur de $I(f)$ est approchée par l'aire du rectangle $\mathscr{R}$ de sommets $(a,0)$, $(a, f(\frac{a+b}{2}))$, $(b, f(\frac{a+b}{2})))$, $(b,0)$ comme illustré par la figure suivante :

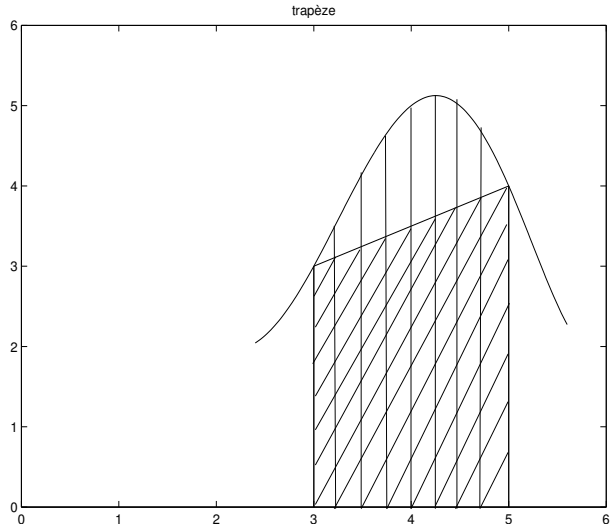


Ainsi on a

$$I(f) \approx (b-a) f(\frac{a+b}{2}).$$

3.3.2 Formule des trapèzes

La valeur de $I(f)$ est approchée par l'aire du trapèze $\mathscr{T}$ de sommets $(a,0)$, $(a,f(a))$, $(b,f(b))$, $(b,0)$ comme illustré par la figure suivante :

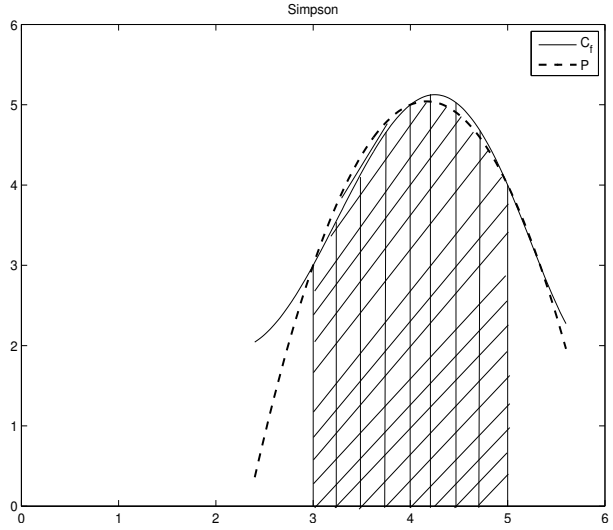


Ainsi on a

$$I(f) \approx \frac{1}{2}(b-a)(f(a)+f(b)).$$

3.3.3 Formule de Simpson

La valeur de $I(f)$ est approchée par l'aire de la parabole $\mathscr{P}$ passant par les points $(a,f(a))$, $(\frac{a+b}{2},f(\frac{a+b}{2}))$ et $(b,f(b))$ comme illustré par la figure suivante :



Ainsi on a

$$I(f) \approx \frac{1}{6}(b-a)(f(a)+4f(\frac{a+b}{2})+f(b)).$$

3.4 Obtention des formules de quadrature

Soit f une fonction réelle définie sur $[a,\ b]$, on désire calculer une valeur approchée de l'intégrale

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

3.4.1 L'idée

L'idée de base est d'écrire

$$f(x) = p(x) + e(x), \text{ pour tout } x \in [a,\ b].$$

où p est le polynôme interpolant f en des abscisses $x_0, x_1, \ldots, x_n$ ($x_i \in [a,\ b]$) et $e(x)$ étant l'erreur d'interpolation.
Ainsi, en intégrant on a :

$$I(f) = \int_a^b f(x)dx = \underbrace{\int_a^b p(x)dx}_{=I_Q(f)} + \underbrace{\int_a^b e(x)dx}_{=E_Q(f)}.$$

En utilisant la base de Lagrange, le polynôme d'interpolation $p := p_n$ s'écrit :

$$p(x) = p_n(x) = \sum_{i=0}^{n} f(x_i)L_i(x),$$

et l'erreur d'interpolation e s'écrit

$$e(x) = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\prod_{i=0}^{n}(x-x_i).$$

Ainsi

$$I_Q(f) = \int_a^b p_n(x)dx = \sum_{i=0}^{n} f(x_i)\int_a^b L_i(x)dx,$$

c'est à dire

$$I_Q(f) = \sum_{i=0}^{n} A_i f(x_i), \text{ où } A_i = \int_a^b L_i(x)dx.$$

Remarque 3.1.

- *Les coefficients A_i (appelés poids) ne dépendent pas de la fonction f.*
- *La quantité $I_Q(f) = \sum_{i=0}^{n} A_i f(x_i)$ représente la valeur approchée de $I(f)$, on écrit alors : $I(f) \approx I_Q(f)$.*
- *Les abscisses x_i ($i = 0, 1, \ldots, n$) sont appelés les noeuds.*
- *Les coefficients A_i sont déterminés de telle sorte que l'erreur de quadrature $E_Q(f)$ soit nulle lorsque $f \in E$ où E est un ensemble à préciser. En général E est l'espace des polynômes de degré inférieur ou égal à n, i.e., $E = \mathscr{P}_n = \mathbb{R}_n[X]$.*

Définition 3.1.
On dit que la formule de quadrature

$$\int_a^b f(x)dx = \sum_{i=0}^{n} A_i f(x_i) + E_Q(f),$$

est exacte sur l'ensemble E si et seulement si $E_Q(g) = 0$ pour tout $g \in E$.

3.4.2 Etude de quelques exemples classiques

- **Cas $n = 0$ et $x_0 = a$.** La formule de quadrature s'écrit :

$$\int_a^b f(x)dx = A_0 f(a) + E_Q(f) \qquad (Q),$$

où $E_Q(g) = 0$ pour tout $g \in \mathbb{R}_0[X]$.

Ainsi, en prenant $g \equiv 1$ (i.e. $g(x) = 1$ pour tout $x \in [a,\ b]$), alors la formule de quadrature (Q) donne que $b - a = A_0$. et pour déterminer l'erreur $E_Q(f)$, on sait que

$$E_Q(f) = \int_a^b \underbrace{(x-a)}_{=:u(x)\geq 0}\ \underbrace{f^{'}(\xi_x)}_{=:v(x)}\, dx, \quad \text{où } \xi_x \in]a,\ b[.$$

Ainsi, en appliquant la première formule de la moyenne, on a $\exists \alpha \in]a,\ b[$ tel que

$$E_Q(f) = f^{'}(\alpha)\left[\frac{1}{2}(x-a)^2\right]_a^b = \frac{(b-a)^2}{2} f^{'}(\alpha).$$

Finalement, on a

$$I(f) = (b-a)f(a) + \frac{(b-a)^2}{2} f^{'}(\alpha), \quad \text{où } \alpha \in]a,\ b[$$

et $I(f)$ est approchée par $(b-a)f(a)$, i.e., $I(f) \approx (b-a)f(a)$. On remarque alors que l'on retrouve la formule du rectangle à gauche qui est exacte sur $\mathscr{P}_0$.

- **Cas $n = 0$ et $x_0 = b$.** De la même façon que précédemment, la formule de quadrature s'écrit :

$$\int_a^b f(x)dx = A_0 f(b) + E_Q(f) \qquad (Q),$$

où $E_Q(g) = 0$ pour tout $g \in \mathbb{R}_0[X]$. Et on vérifie que l'on retrouve la formule du rectangle à droite qui est exacte également sur $\mathscr{P}_0$ et qui est donnée par.

$$I(f) = (b-a)f(b) + \frac{(b-a)^2}{2} f^{'}(\beta), \quad \text{où } \beta \in]a,\ b[.$$

- **Cas $n = 0$ et $x_0 = \frac{a+b}{2}$.** Dans ce cas, la formule de quadrature s'écrit :

$$\int_a^b f(x)dx = A_0 f(\frac{a+b}{2}) + E_Q(f) \qquad (Q),$$

En écrivant que $I_Q(g)$ est exacte sur $\mathscr{P}_0$, i.e., $E_Q(g) = 0$ pour $g \equiv 1$, on vérifie que $A_0 = b - a$. Ainsi, $I(f) = (b-a)f(\frac{a+b}{2}) + E_Q(f)$, avec

$$E_Q(f) = \int_a^b (x - \frac{a+b}{2}) f^{'}(\xi_x)dx, \quad (\xi_x \in]a,\ b[).$$

Notons que la fonction $x \longmapsto x - \frac{a+b}{2}$ change de signe dans $[a,\ b]$ et donc on ne peut pas appliquer la première formule de la moyenne. Pour déterminer l'erreur de quadrature, on va utiliser la deuxième expression de l'erreur d'interpolation à savoir que :

$$E_Q(f) = \int_a^b (x - x_0)[f, x_0, x]dx \quad \text{avec } x_0 = \frac{a+b}{2}.$$

En utilisant la décomposition suivante :

$$(x-x_0)[f,x_0,x_0,x]=[f,x_0,x]-[f,x_0,x_0],$$

on a

$$\begin{aligned}
E_Q(f) &= \int_a^b ([f,x_0,x_0]+(x-x_0)[f,x_0,x_0,x])\,(x-x_0)\,dx \\
&= \int_a^b [f,x_0,x_0](x-x_0)\,dx+\int_a^b [f,x_0,x_0,x](x-x_0)^2\,dx \\
&= [f,x_0,x_0]\underbrace{\int_a^b (x-x_0)\,dx}_{=0}+\int_a^b \underbrace{[f,x_0,x_0,x]}_{=:v(x)}\underbrace{(x-x_0)^2}_{=:u(x)\geq 0}\,dx \\
&= [f,x_0,x_0,\gamma]\int_a^b (x-x_0)^2\,dx \quad \text{où } \gamma\in]a,\,b[
\end{aligned}$$

où on a appliqué la première formule de la moyenne. Finalement, en utilisant le théorème de Cauchy, on obtient la formule de quadrature du point milieu qui s'écrit :

$$\int_a^b f(x)\,dx=(b-a)f(\frac{a+b}{2})+\frac{(b-a)^3}{24}f^{''}(\lambda) \quad \text{où } \lambda\in]a,\,b[.$$

Remarque 3.2. *On note que la formule du rectangle est encore exacte sur $\mathscr{P}_1$ et pas simplement sur $\mathscr{P}_0$ alors que les formules du rectangle à gauche et à droites sont exactes uniquement sur $\mathscr{P}_0$.*

- **Cas $n=1$ et $x_0=a$, $x_1=b$.** Dans ce cas, la formule de quadrature s'écrit :

$$\int_a^b f(x)\,dx=A_0f(a)+A_1f(b)+E_Q(f) \qquad (Q),$$

où $E_Q(g)=0$ pour tout $g\in\mathbb{R}_1[X]$.

En prenant respectivement $g\equiv 1$ puis $g\equiv x$ dans la formule de quadrature (Q), alors par un simple calcul, on trouve que $A_0=A_1=\frac{1}{2}(b-a)$. Ainsi, on a $I(f)=\frac{1}{2}(b-a)(f(a)+f(b))+E_Q(f)$, avec

$$E_Q(f)=\int_a^b \underbrace{(x-x_0)(x-x_1)}_{=:u(x)}\underbrace{[f,x_0,x_1,x]}_{=:v(x)}\,dx \quad \text{avec } x_0=a \text{ et } x_1=b.$$

En utilisant respectivement la 1ère formule de la moyenne et le théorème de Cauchy, on a :

$$\begin{aligned}
E_Q(f) &= [f,x_0,x_1,\xi]\int_a^b (x-a)(x-b)\,dx \quad \text{où } \xi\in]a,\,b[, \\
&= \frac{f^{''}(\eta)}{2}\int_a^b (x-a)(x-b)\,dx, \\
&= \frac{-1}{12}(b-a)^3f^{''}(\eta).
\end{aligned}$$

Finalement, on voit que la formule obtenue est celle du trapèze. Cette formule s'écrit

$$\int_a^b f(x)\,dx=\frac{1}{2}(b-a)(f(a)+f(b))+\frac{-1}{12}(b-a)^3f^{''}(\eta) \quad \text{où } \eta\in]a,\,b[.$$

Remarque 3.3. *On note que la formule du trapèze est exacte seulement sur* $\mathscr{P}_1$.

- **Cas** $n=2$ **et** $x_0=a$**,** $x_1=\frac{a+b}{2}$**,** $x_2=b$**.** La formule de quadrature s'écrit :

$$\int_a^b f(x)dx = A_0 f(a) + A_1 f(a+b2) + A_2 f(b) + E_Q(f) \qquad (Q),$$

où $E_Q(g)=0$ pour tout $g \in \mathbb{R}_2[X]$.

En écrivant que $E_Q(g)=0$ pour $g\equiv 1$, $g\equiv x$ et $g\equiv x^2$, on obtient la système

$$\begin{cases} \int_a^b 1\,dx &= A_0+A_1+A_2 \\ \int_a^b x\,dx &= x_0A_0+x_1A_1+x_2A_2 \\ \int_a^b x^2\,dx &= x_0^2A_0+x_1^2A_1+x_2^2A_2 \end{cases}$$

où $x_0=a$, $x_1=\frac{a+b}{2}$ et $x_1=b$. Ainsi, en développant les calculs, on obtient le système

$$\begin{cases} A_0+A_1+A_2 &= b-a \\ aA_0+\frac{a+b}{2}A_1+bA_2 &= \frac{b^2-a^2}{2} \\ a^2A_0+(\frac{a+b}{2})^2A_1+b^2A_2 &= \frac{b^3-a^3}{3} \end{cases},$$

dont la solution est $A_0=A_1=\frac{b-a}{6}$ et $A_2=\frac{2}{3}f(\frac{(a+b)}{2})$.

En utilisant une démarche similaire à celle des cas précédents, on montre que l'erreur de quadrature $E_Q(f)$ est donnée par $E_Q(f)=-\frac{(b-a)^5}{2880}f^{(4)}(v)$ où $v\in]a,\ b[$.

Finalement, la formule qu'on retrouve pour le cas $n=2$, $x_0=a$, $x_1=\frac{a+b}{2}$ et $x_2=b$ est celle de Simpson. Cette formule s'écrit :

$$\int_a^b f(x)dx = \frac{1}{6}(b-a)(f(a)+4f(\frac{a+b}{2})+f(b)) - \frac{(b-a)^5}{2880}f^{(4)}(v) \quad \text{où } v\in]a,\ b[.$$

Remarque 3.4. *On note que la formule de Simpson est exacte sur* $\mathscr{P}_3$ *et pas simplement sur* $\mathscr{P}_2$.

3.5 Les formules composites

L'idée des méthodes composites est simple et repose sur la linéarité de l'intégrale. En effet, on peut écrire

$$\int_a^b f(x)dx = \int_{\alpha_0}^{\alpha_1} f(x)dx + \int_{\alpha_1}^{\alpha_2} f(x)dx + \ldots \int_{\alpha_{N-1}}^{\alpha_N} f(x)dx,$$

où $\alpha_0=a$, $\alpha_N=b$, et $\alpha_1,\alpha_2,\ldots,\alpha_{N-1}$ sont des points de l'intervalle $[a,\ b]$. Généralement, ces points sont équidistants mais dans certains cas le choix de ces points peut tenir compte de l'allure de la courbe de la fonction g à intégrer (si cette allure est connue).

Dans la suite, on considère que les points $\alpha_0,\alpha_1,\ldots,\alpha_N$ sont équidistants, i.e., $\alpha_i=a+iH$ pour $i=0,1,\ldots,N$ où $H=\frac{b-a}{N}$. Ainsi,

$$\int_a^b f(x)dx = \sum_{k=0}^{N-1} \underbrace{\int_{\alpha_k}^{\alpha_{k+1}} f(x)dx}_{=:I_k},$$

et alors, on doit calculer une valeur approchée de l'intégrale I_k à l'aide d'une formule de quadrature (de type Newton-Cotes) avec n généralement faible, $n = 0, 1, 2$.

3.5.1 Formule composite du rectangle

Sachant que

$$\begin{aligned} I_k &= (\alpha_{k+1}-\alpha_k) f(\frac{\alpha_k+\alpha_{k+1}}{2}) + \frac{(\alpha_{k+1}-\alpha_k)^3}{24} f^{"}(\lambda_k) \quad \text{où } \lambda_k \in]\alpha_k,\ \alpha_{k+1}[, \\ &= H f(\alpha_k + \frac{H}{2}) + \frac{H^3}{24} f^{"}(\lambda_k). \end{aligned}$$

il vient, en utilisant la première formule de la moyenne pour le cas discret, que

$$\begin{aligned} \int_a^b f(x)dx &= H \sum_{k=0}^{N-1} f(\alpha_k + \frac{H}{2}) + \frac{H^3}{24} \sum_{k=0}^{N-1} \underbrace{1}_{=:v_k} \underbrace{f^{"}(\lambda_k)}_{=:u(\lambda_k)}, \\ &= H \sum_{k=0}^{N-1} f(a+(2k+1)\frac{H}{2}) + \frac{H^3}{24} f^{"}(\lambda) \underbrace{\sum_{k=0}^{N-1} 1}_{=N} \quad \text{où } \lambda \in]a,\ b[. \\ &= \underbrace{H \sum_{k=0}^{N-1} f(a+(2k+1)\frac{H}{2})}_{=:I_{0,N}(f)} + \underbrace{\frac{(b-a)}{24} H^2 f^{"}(\lambda)}_{=:E_{0,N}(f)}. \end{aligned}$$

3.5.2 Formule composite du trapèze

Dans la cas de la formule du trapèze, on a

$$\begin{aligned} I_k &= \frac{1}{2}(\alpha_{k+1}-\alpha_k)(f(\alpha_k)+f(\alpha_{k+1})) - \frac{1}{12}(\alpha_{k+1}-\alpha_k)^3 f^{"}(\eta_k) \quad \text{où } \eta_k \in]\alpha_k,\ \alpha_{k+1}[, \\ &= \frac{1}{2} H\,(f(a+kH)+f(a+(k+1)H)) - \frac{1}{12} H^3 f^{"}(\eta_k). \end{aligned}$$

De même, en utilisant la première formule de la moyenne pour le cas discret, on vérifie que

$$\begin{aligned} \int_a^b f(x)dx &= \frac{H}{2}\left(\sum_{k=0}^{N-1} f(a+kH) + \sum_{k=0}^{N-1} f(a+(k+1)H)\right) - \frac{1}{12} H^3 \sum_{k=0}^{N-1} f^{"}(\eta_k) \\ &= \frac{H}{2}\left(f(a) + 2\sum_{k=0}^{N-1} f(a+kH) + f(b)\right) - \frac{1}{12} H^3 N f^{"}(\eta), \quad \text{où } \eta \in]a,\ b[\\ &= \underbrace{H\left(\frac{1}{2}f(a) + \sum_{k=0}^{N-1} f(a+kH) + \frac{1}{2}f(b)\right)}_{=:I_{1,N}(f)} \underbrace{- \frac{(b-a)}{12} H^2 f^{"}(\eta)}_{=:E_{1,N}(f)}. \end{aligned}$$

3.5.3 Formule composite de Simpson

Dans la cas, on rappelle que

$$\begin{aligned} I_k &= \frac{1}{6}(\alpha_{k+1}-\alpha_k)(f(\alpha_k)+4f(\frac{\alpha_k+\alpha_{k+1}}{2})+f(\alpha_{k+1})) - \frac{1}{2880}(\alpha_{k+1}-\alpha_k)^5 f^{(4)}(\nu_k) \quad \text{où } \nu_k \in]\alpha_k,\ \alpha_{k+1}[, \\ &= \frac{1}{6} H\left(f(a+kH)+4f(a+(k+1)\frac{H}{2})+f(a+(k+1)H)\right) - \frac{1}{2880} H^5 f^{(4)}(\nu_k). \end{aligned}$$

Ainsi, en utilisant des calculs et critères similaires à ceux des formules composites du rectangle ou du trapèze, on montre que

$$\int_a^b f(x)dx = \underbrace{\frac{H}{6}\left(f(a)+2\sum_{k=0}^{N-1} f(a+kH)+4\sum_{k=0}^{N-1} f((2k+1)\frac{H}{2})+f(b)\right)}_{=:I_{2,N}} \underbrace{-\frac{(b-a)}{2880}H^4 f^{(4)}(\nu)}_{=:E_{2,N}} \quad \text{où } \eta \in]a,\ b[.$$

3.6 Formules de Gauss

Toutes les formules de quadrature vues précédemment sont de la forme (dite interpolatoire). En effet, si P_n interpole f aux abscisses **fixées** $x_0, x_1, \ldots, x_n$ alors

$$\int_a^b f(x)dx = \sum_{i=0}^{n} A_i f(x_i) + E_Q(f), \qquad (Q)$$

où les $A_i = \int_a^b L_i(x)dx$ ne dépendent pas de f. De plus la formule (Q) est exacte sur $\mathscr{P}_n$.

Maintenant, on considère, le problème suivant : étant donnée la formule de quadrature (Q), on cherche à déterminer les coefficients A_i ainsi que les abscisses x_i, $i = 0, \ldots, n$ pour que cette formule de quadrature soit exacte sur $\mathscr{P}_{2n+1}$.

3.6.1 Cas $n = 0$

La forme de quadrature que l'on cherche à la forme suivante :

$$\int_a^b f(x)dx \approx A_0 f(x_0), \qquad (Q)$$

où les inconnues A_0 et x_0 sont à déterminer. En écrivant que (Q) est exacte sur $\mathscr{P}_1$, c'est à dire que

$$\int_a^b g(x)dx = A_0 g(x_0) \text{ pour } g \in \{1, x\},$$

on obtient les deux équations

$$\begin{cases} \int_a^b 1\,dx &= A_0 \\ \int_a^b x\,dx &= A_0 x_0 \end{cases}$$

c'est à dire

$$\begin{cases} A_0 &= b-a \\ A_0 x_0 &= \frac{b^2-a^2}{2} \end{cases}$$

et donc $A_0 = b-a$ et $x_0 = \frac{a+b}{2}$. On voit alors que l'on retrouve la formule du point milieu.

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

3.6.2 Cas $n = 1$

La formule de quadrature recherchée à la forme

$$\int_a^b f(x)dx \approx A_0 f(x_0) + A_1 f(x_1), \qquad (Q)$$

où les coefficients A_0, A_1 et les abscisses x_0, x_1 sont à déterminer en imposant que (Q) doit être exacte sur $\mathscr{P}_3$.

Pour résoudre ce problème tout en simplifiant les calculs, on va travailler sur l'intervalle $[-1,1]$, puis on se ramènera sur l'intervalle $[a,\ b]$ à l'aide du changement de variable

$$x(t) = \alpha\, t + \beta, \qquad \text{avec } x := x(t) \in [a,b] \text{ et } t \in [-1,\ 1],$$

et où $\alpha = \frac{b-a}{2}$ et $\beta = \frac{a+b}{2}$. Ainsi, comme $dx = \alpha\, dt =$, $x(-1) = a$ et $x(1) = b$, il vient que :

$$\int_a^b f(x)\,dx = \frac{b-a}{2}\int_{-1}^{1} h(t)\,dt,$$

où $h(t) = f(\alpha\, t + \beta) = (f \circ x)(t)$.

Ainsi, en posant

$$\int_{-1}^{1} h(t)\,dt \approx B_0\, h(t_0) + B_1\, h(t_1), \qquad (Q^{'})$$

on reformule le problème comme suit :

trouver $t_0,\ t_1$ et $B_0,\ B_1$ tels que la formule $(Q^{'})$ soit exacte pour $g \in \{1, t, t^2, t^3\}$. Cette dernière condition fournit le système non linéaire suivant :

$$\left\{\begin{array}{lcll} B_0 + B_1 & = & 2 & (1)\\ B_0\, t_0 + B_1\, t_1 & = & 0 & (2)\\ B_0\, t_0^2 + B_1\, t_1^2 & = & \frac{2}{3} & (3)\\ B_0\, t_0^3 + B_1\, t_1^3 & = & 0 & (4) \end{array}\right.$$

En multipliant les équations (1), (2) et (3) par t_0 et en retranchant respectivement les équations (2), (3) et (4), on trouve que

$$t_0\, t_1 = -\frac{1}{3}, \quad t_0 = -t_1 \quad \text{et} \quad B_0 + B_1 = 2, \quad B_0 - B_1 = 0.$$

Finalement, on prend $t_0 = -t_1 = \frac{-1}{\sqrt{3}}$ et $B_0 = B_1 = 1$, et la formule de quadrature $(Q^{'})$ recherchée est donc

$$\int_{-1}^{1} h(t)\,dt \approx h(-\frac{1}{\sqrt{3}}) + h(\frac{1}{\sqrt{3}}), \qquad (Q^{'})$$

et la formule de quadrature (Q) s'obtient en écrivant

$$\begin{aligned} \int_a^b f(x)\,dx &\approx \frac{b-a}{2}\int_{-1}^{1} h(t)\,dt\\ &\approx \frac{b-a}{2}\int_{-1}^{1} (f\circ x)(t)\,dt\\ &\approx \frac{b-a}{2}\left[(f\circ x)(-\frac{1}{\sqrt{3}}) + (f\circ x)(\frac{1}{\sqrt{3}})\right]\\ &\approx \frac{b-a}{2}\left[f(\frac{a-b}{\sqrt{3}} + \frac{a+b}{2}) + f(\frac{b-a}{\sqrt{3}} + \frac{a+b}{2})\right] \end{aligned}$$

Chapitre 4

Résolution d'équations non linéaires

Etant donnée une fonction $f : D \longrightarrow \mathbb{R}$ supposée continue sur $D \subset \mathbb{R}$, on s'intéresse dans ce chapitre au problème de recherche d'une solution (ou des solutions) de l'équation $f(x) = 0$. Il est à noter que dans la suite, on suppose que f est non linéaire !

4.1 Introduction

Avant de décrire quelques méthodes permettant de rechercher numériquement un ou plusieurs zéros x^* de f, c'est à dire rechercher les nombres réels x^* vérifiant $f(x^*) = 0$ (ou -sur ordinateur- $|f(x^*)| < \epsilon$ avec ϵ très proche de 0), donnons quelques outils mathématiques qui nous seront utiles dans la suite.

4.1.1 Localisation des racines

Les premiers outils permettent de localiser grossièrement le zéro (ou les zéros) de f. En effet, après étude de la fonction f, les résultats donnés ci dessous permettent de trouver un intervalle qui contient un et un seul zéro x^*. On dit dans ce cas que la racine x^* est séparable.

Théorème 4.1. *(TVI : Théorème des Valeurs Intermédiaires)*
Etant donnée une fonction f continue sur un intervalle $[a,\ b]$ de $\mathbb{R}$, alors f atteint toutes les valeurs intermédiaires entre les images $f(a)$ et $f(b)$. Autrement dit : si $m = \min\{f(a), f(b)\}$ et $M = \max\{f(a), f(b)\}$, alors

$$\forall d \in [m,\ M],\ \exists c \in [a,\ b] \ \textit{tel que} \ \ f(c) = d.$$

Le résultat suivant est une application directe du théorème des valeurs intermédiaires.

Théorème 4.2. *(Théorème de BOLZANO)*
Si f est une fonction réelle continue sur $[a,\ b] \subset \mathbb{R}$ et si $f(a)f(b) < 0$, alors il existe $x \in]a,\ b[$ tel que $f(x) = 0$.

Remarquons que ce théorème garantit juste l'existence d'un zéro. Dans le cas où l'unicité de la racine doit être prouvée, on pourra appliquer le théorème de la bijection qu'on rappelle ci dessous

Théorème 4.3. *(Théorème de la bijection)*
Si la fonction f est continue et est strictement monotone (c'est à dire strictement croissante ou strictement décroissante) sur un intervalle I de $\mathbb{R}$, alors la fonction f induit une bijection de I dans $f(I)$. De plus, sa bijection réciproque est continue sur I, monotone sur I et de même sens de variation que f.

4.1.2 Construction d'une suite convergente vers la racine

Dès que la racine x^* est localisée dans un intervalle I, on choisit un point x_0 dans cet intervalle I et on applique une méthode itérative permettant de calculer une suite de points x_1, x_2,..., x_k, ... en ayant l'espoir que cette suite (x_k) sera convergente et que en plus elle convergera vers la racine x^*.

Définition 4.1. *(Ordre de convergence)*
On dit qu'une méthode itérative convergente est d'ordre p (où $p \in \mathbb{N}^*$) s'il existe un entier $k_0 \in \mathbb{N}$ et une constante $C > 0$ telle que $|x_{k+1} - x^*| \leq |x_k - x^*|^p$, pour tout $k \geq k_0$.
Notons que dans le cas où $p = 1$ et $C < 1$, on dit que la convergence est linéaire et la convergence est dite quadratique (respectivement cubique) dans le cas où $p = 2$ (respectivement $p = 3$).

La définition ci dessous permet de comparer la rapidité de convergence de deux suites (et donc éventuellement de deux méthodes).

Définition 4.2.
Soient (x_n) et (y_n) deux suites convergentes vers une même limite l. On dit que la suite (y_n) converge "plus vite" que (x_n) vers le réel l si le rapport $|y_n - l|/|x_n - l|$ a pour limite 0, i.e. si $\lim_n \dfrac{y_n - l}{x_n - l} = 0$.

4.1.3 Recherche d'un point fixe

Souvent, une équation de type $f(x) = 0$ peut être réécrite d'une manière équivalente sous la forme de la recherche d'un point fixe d'une fonction g, i.e, le réel x est tel que $g(x) = x$. La fonction g étant une fonction qui dépend de f mais qui n'est pas unique comme le montre les deux exemples suivants :

- **Exemple 1.** Soit l'équation $f(x) = 0$ avec $f(x) = x^3 + 2x^2 + 10x - 20$. Cette équation peut être transformée sous les formes :

$$g_1(x) = x, \quad \text{où } g_1(x) = \frac{20}{x^2 + 2x + 10}.$$

ou

$$g_2(x) = x, \quad \text{où } g_2(x) = \frac{20 - 2x^2 - x^3}{10}.$$

- **Exemple 2.** Pour $x \in [0,\ 2]$, l'équation $f(x) = 0$ où $f(x) = \sin(2x) + x - 1$ peut être réécrite sous les formes :

$$h_1(x) = x, \quad \text{où } h_1(x) = 1 - \sin(2x).$$

ou

$$h_2(x) = x, \quad \text{où } h_2(x) = \frac{1}{2}\arcsin(1 - x).$$

Ainsi, dans le cas où une équation $f(x) = 0$ est transformée de façon équivalente en une équation $g(x) = x$, il peut être intéressant d'utiliser l'un des résultats ci dessous et qui concernent l'étude de la convergence d'une suite définie à partir du problème de la recherche des points fixes de la fonction g.

Définition 4.3. *(fonction contractante)*
Soient $k \in]0,\ 1[$ et g une fonction réelle définie sur $D = [a,\ b] \subset \mathbb{R}$. On dit que g est k-contractante (ou contractante de rapport k) si

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

Théorème 4.4. *(théorème du point fixe)*
Soit $g : [a,\ b] \longrightarrow \mathbb{R}$ une fonction k-contractante. Alors, la fonction g admet un unique point fixe $l \in [a,\ b]$. De plus, quel que soit le réel $x_0 \in [a,\ b]$, la suite définie par $x_{n+1} = g(x_n)$, pour tout $n \geq 0$ est convergente de limite l quand $n \longrightarrow +\infty$.

Remarques :

- Une fonction g contractante sur $D = [a,\ b]$ est une fonction continue sur D.
- Souvent, pour montrer qu'une fonction $g \in C^1([a,\ b])$ est k-contractante sur $[a,\ b]$, il suffit de vérifier que $|g^{'}(x)| \leq k < 1$, pour tout $x \in [a,\ b]$.
- En posant $k = \max\limits_{x \in [a,\ b]} |g^{'}(x)|$, alors la suite (x_n) définie par $x_{n+1} = g(x_n)$, $\forall n \geq 0$ vérifie

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

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 : Méthode du point fixe.

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

Dans la suite, nous donnons les résultats obtenus pour la recherche des zéros de la fonction f définie par $f(x) = x^3 + 2x^2 + 10x - 20$.
Tout d'abord, comme $f^{'}(x) = 3x^2 + 4x + 10 < 0$ pour tout $x \in \mathbb{R}$ et $\lim\limits_{-\infty} f(x) = -\infty$, $\lim\limits_{+\infty} f(x) = +\infty$ alors f possède un unique zéro qui est situé dans l'intervalle $[1,\ 2]$ car $f(1) = -7 < 0$ et $f(2) = 36 > 0$.

- **Exemple 1.** On transforme l'équation $f(x) = 0$ en l'équation équivalente $g_1(x) = x$ où $g_1(x) = \dfrac{20}{x^2 + 2x + 10}$

On peut vérifier que la fonction dérivée de g_1 est donnée par $g_1'(x) = \frac{-40(x+1)}{(x^2+2x+1)^2}$ et que $|g_1'(x)| < 1$ pour tout $x \in [1,\ 2]$. Ainsi g_1 est contractante sur $[1,\ 2]$. Ainsi, grâce au théorème du point fixe, la convergence de la suite (x_n) définie $x_0 \in [1,\ 2]$ et $x_{n+1} = g_1(x_n) = \frac{20}{x_n^2+2x_n+10}$ vers la solution exacte x^* est assurée.

En prenant $x_0 = 1$, $g = g_1$ dans l'algorithme précédent et en considérant que l'on a obtenu x_n une bonne approximation de x^* lorsque $|x_n - x_{n+1}| < \epsilon$, où $\epsilon = 10^{-6}$ est la précision souhaitée, nous obtenons les résultats reportés dans la table ci dessous :

itération	x_n	itération	x_n
0	1.000000000000000	10	1.368696397555516
1	1.538461538461539	11	1.368857688628725
2	1.295019157088122	12	1.368786102577989
3	1.401825309448600	13	1.368817874396085
4	1.354209390404292	14	1.368803773143633
5	1.375298092487380	15	1.368810031675092
6	1.365929788170655	16	1.368807253960778
7	1.370086003401819	17	1.368808486788930
8	1.368241023612835	18	1.368807939624842
9	1.369059812007482		

Les résultats précédents montrent que la méthode du point fixe a nécessité 18 itérations pour converger et la solution déterminée par cette méthode est $x^* = 1.3688085$ avec $f(1.3688085) = 7.9947533\,10^{-6}$ et $g_1(1.3688085) = 1.3688079$.

— **Exemple 2.** On transforme l'équation $f(x) = 0$ en l'équation équivalente $g_2(x) = x$ où $g_2(x) = \dfrac{20 - 2x^2 - x^3}{10}$

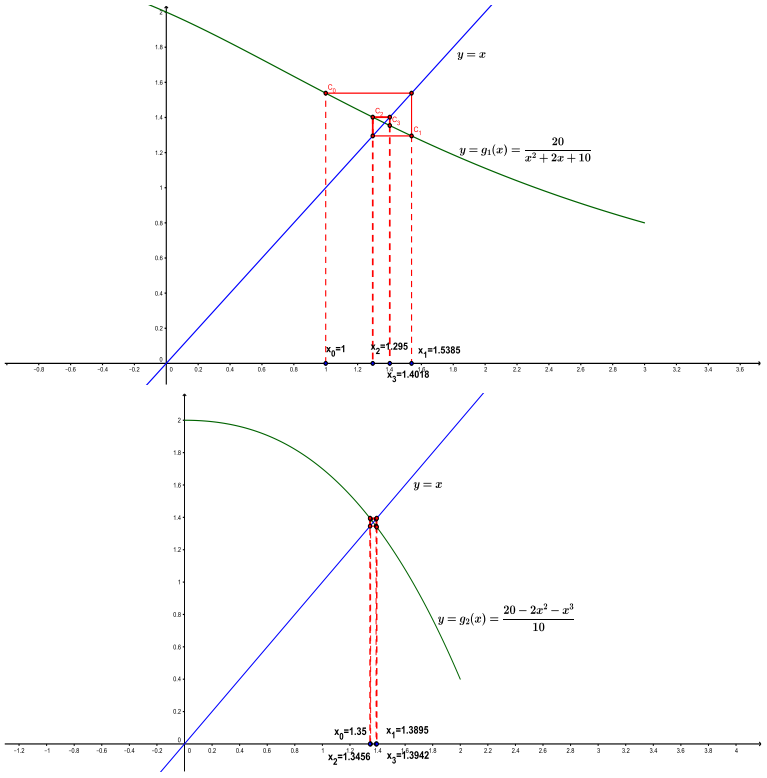
On peut vérifier que la fonction dérivée de g_2 est donnée par $g_2'(x) = \frac{-4x-3x^2}{10}$ et que $g_2'(x) < -1$ pour tout $x \in [1,\ 2]$. Ainsi, comme $|g_2'(x)| > 1$ pour tout $x \in [1,\ 2]$, le point fixe de g_2 est un point répulsif et donc quelque soit le chois de $x_0 \neq x^*$ la suite (x_n) définie par $x_{n+1} = g_2(x_n)$ ne convergera pas vers la solution exacte x^*.

En effet même en prenant $g = g_2$, et $x_0 = 1.35$ -assez proche de x^*- dans l'algorithme de la méthode du point fixe, les itérés construits s'éloignent du point fixe comme le montre les résultats reportés dans la table ci dessous.

itération	x_n	itération	x_n
0	1.350000000000000	131	0.548946478058069
1	1.389462500000000	132	1.923189476943791
2	1.345628322885400	133	0.548946478056689
3	1.394201865964998	134	1.923189476944218
4	1.340235433982419	⋮	⋮
5	1.400016550438115	⋮	⋮
6	1.333580999927215	147	1.923189476944801
7	1.407143192902503	148	0.548946478054790
8	1.325367942472636	149	1.923189476944807
9	1.415865806392921	150	0.548946478054780

La convergence de la méthode du point fixe dans le premier exemple et sa non convergence dans le

deuxième exemple sont respectivement illustrées pas les deux figures suivantes :



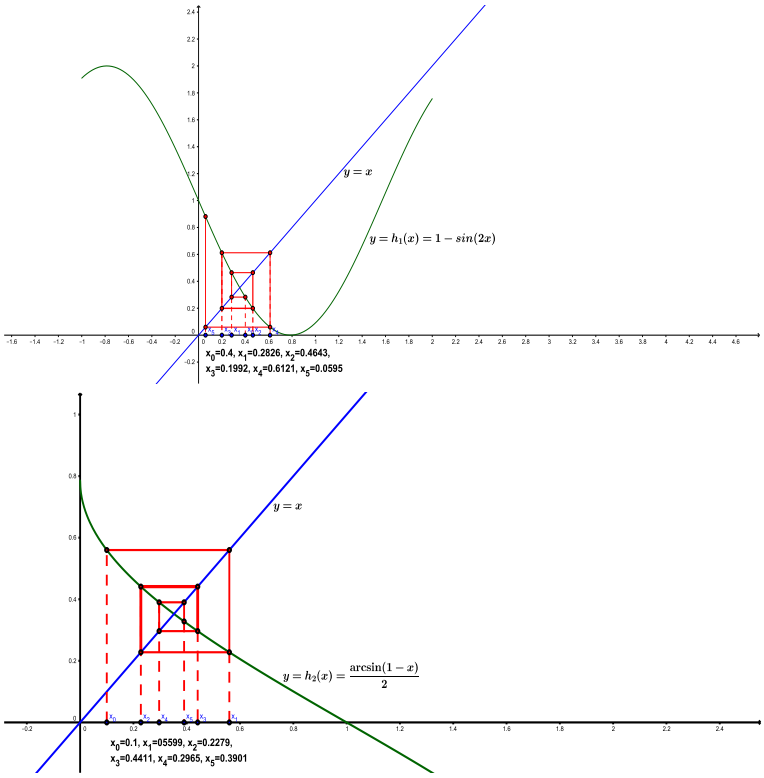
Exercice 1. Interpreter les figures ci dessous relatifs à la recherche des point fixes des fonctions h_1 et h_2 associées à la racine de l'équation $f(x) = 0$ où $f(x) = \sin(2x) + x - 1$ en tenant compte que les itérés associées aux fonctions h_1 et h_2 sont respectivement $x_0 = 0.4$ $x_1 = 0.2826$, $x_2 = 0.4643$, $x_3 = 0.1992$, $x_4 = 0.6121$, $x_5 = 0.0595$ et $x_0 = 0.1$ $x_1 = 0.5599$, $x_2 = 0.2279$, $x_3 = 0.4411$, $x_4 = 0.2965$, $x_5 = 0.3901$.

4.2 Méthodes d'encadrement

Soit f une fonction réelle d'une variable réelle, i.e.

$$\begin{array}{rccc} f : & D & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & f(x). \end{array}$$

On suppose que notre fonction est non linéaire (sinon c'est évident) et notre but est de résoudre l'équation $f(x) = 0$, et donc on recherche un nombre réel x^* vérifiant $f(x^*) = 0$ (ou -sur ordinateur- $|f(x^*)| < \epsilon$ avec ϵ très proche de 0). Dans la suite, on suppose que la racine x^* est séparable, c'est à dire qu'il existe une intervalle $[a,\ b]$ tel que x^* est l'unique racine dans cet intervalle.



4.2.1 Méthode de dichotomie

Le principe de la méthode de la dichotomie (ou de la bissection) est très simple. On suppose que $f : [a, b] \longrightarrow \mathbb{R}$ est continue et que $f(a)f(b) < 0$. Ainsi, d'après le théorème des valeurs intermédiaires, il existe au moins une racine réelle de f dans $[a, b]$. On prend alors le milieu de l'intervalle $m = \frac{a+b}{2}$.

Dans le cas $f(m) = 0$ (ou numériquement très proche de 0), on considère que le problème est résolu. Si non, deux cas se présentent :

— si $f(a)f(m) < 0$, alors f a une racine réelle dans $[a, m]$.
— si $f(m)f(b) < 0$, alors f a une racine réelle dans $[m, b]$.

On itère alors le processus avec l'intervalle qui contient la racine.

On voit ainsi que la méthode de dichotomie construit une suite décroissante d'intervalles contenant la racine x^* de la fonction f. A l'itération n, on considère comme nouvel intervalle $[a_{n+1}, b_{n+1}]$, celui construit à partir d'une des extrémités de l'intervalle $[a_n, b_n]$ et de son milieu $m = \frac{a_n+b_n}{2} : x_n$. L'algorithme de la méthode de dichotomie ou de bissection peut être résumé ainsi :

Algorithme : Méthode de dichotomie (ou bissection).

Initialisation : $a_0 = a$ *et* $b_0 = b$ *;*
Itération : Pour $n = 0, 1, 2, \ldots,$ *jusqu'à convergence :*
calculer $x_n = \dfrac{a_n + b_n}{2}$ *;*
Si $f(x_n) = 0$,
$x^* = x_n$, *stop ;*
fin du Si
Si $f(a_n)f(x_n) < 0$,
on pose $a_{n+1} = a_n$ *et* $b_{n+1} = x_n$ *;*
sinon
on pose $a_{n+1} = x_n$, $b_{n+1} = b_n$ *;*
fin du Si.
fin du Pour.

On constate que à chaque itération de la méthode de bissection, la longueur de l'intervalle contenant la racine x^* est réduite par un facteur de 2, (i.e., la longueur de l'intervalle contenant la racine est divisée par 2). On voit alors qu'il y'a convergence de la suite (x_n) vers x^* et qu'on peut obtenir une valeur approchée de x^* avec la précision souhaitée.

Exercice 2.

1. Montrer que la suite des intervalles $([a_n,\ b_n])$ vérifie $l_n = \frac{b-a}{2^n}$, où l_n est la longueur de l'intervalle $[a_n,\ b_n]$.

2. En déduire que N le nombre minimal d'itérations nécessaires pour calculer la racine x^* à ϵ près vérifie $N \geq \log_2\left(\frac{b-a}{\epsilon}\right)$.

4.2.2 Méthode de la fausse position

Comme la méthode de la bissection, la méthode de la fausse position (ou méthode regula falsi ou encore méthode de Lagrange) construit également une suite décroissante d'intervalles. Cependant, à l'itération n, au lieu de diviser l'intervalle $[a_n,\ b_n]$ en deux sous-intervalles de même longueur en considérant le milieu de l'intervalle $[a_n,\ b_n]$, on préfère découper l'intervalle $[a_n,\ b_n]$ en $[a_n,\ w_n]$ et $[w_n,\ b_n]$ où w_n est l'abscisse du point d'intersection de la droite passant par les points $(a_n, f(a_n))$ et $(b_n, f(b_n))$. Ainsi, w_n est donné par

$$w_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}.$$

Cette méthode est décrite par l'algorithme suivant :

Algorithme : Méthode de la fausse position (ou regula falsi).

Initialisation : $a_0 = a$ *et* $b_0 = b$ *;*
Itération : Pour $n = 0, 1, 2, \ldots,$ *jusqu'à convergence :*
calculer $w_n = \dfrac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)}$ *;*
Si $f(w_n) = 0$,
$x^* = w_n$, *stop ;*
fin du Si
Si $f(a_n) f(w_n) < 0$,
on pose $a_{n+1} = a_n$ *et* $b_{n+1} = w_n$ *;*
sinon
on pose $a_{n+1} = w_n$, $b_{n+1} = b_n$ *;*
fin du Si.
fin du Pour.

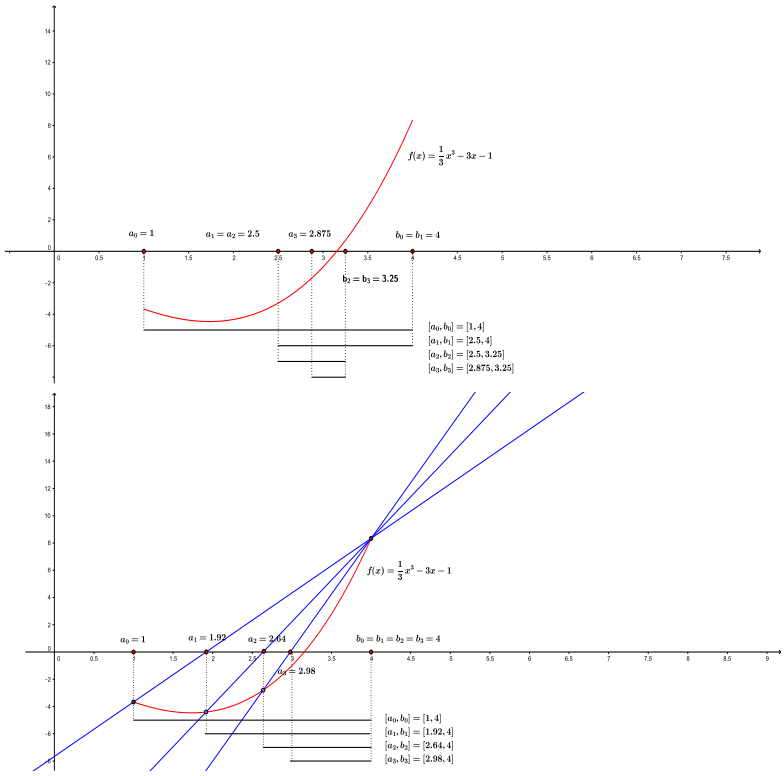
Cette méthode peut construire rapidement des itérées a_n, b_n pour lesquels $|f(x)|$ est très petite, cependant elle peut faillir dans certains cas en n'arrivant pas à déterminer un intervalle $[a_n,\ b_n]$ de longueur assez petite et contenant le zéro x^*.

4.2.3 Exemples

- **Exemple 1.** Dans cet exemple, la fonction test est notée f et est telle que $f(x) = \frac{1}{3}x^3 - 3x - 1$. La racine recherchée x^* est située dans l'intervalle $[a,\ b] = [1,\ 4]$. Les algorithmes s'arrêtent à l'itération n dès que : $b_n - a_n < \varepsilon$ ou dès que $|f(x_n)| < \varepsilon$ où $\varepsilon = 10^{-5}$.

  Les approximations successives construites par les méthodes de dichotomie et regula falsi sont reportées dans le tableau suivant :

Itération	Dichotomie	Regula falsi
	x_n	x_n
0	2.500000000000000	1.916666666666667
1	3.250000000000000	2.636879969708444
2	2.875000000000000	2.979619151850352
3	3.062500000000000	3.100609650227808
4	3.156250000000000	3.138412893531248
5	3.109375000000000	3.149754925521175
6	3.132812500000000	3.153115836701065
7	3.144531250000000	3.154108070854592
8	3.150390625000000	3.154400685040792
9	3.153320312500000	3.154486950343403
10	3.154785156250000	3.154512379708390
11	3.154052734375000	3.154519875589223
12	3.154418945312500	3.154522085151433
13	3.154602050781250	
14	3.154510498046875	
15	3.154556274414063	
16	3.154533386230469	
17	3.154521942138672	



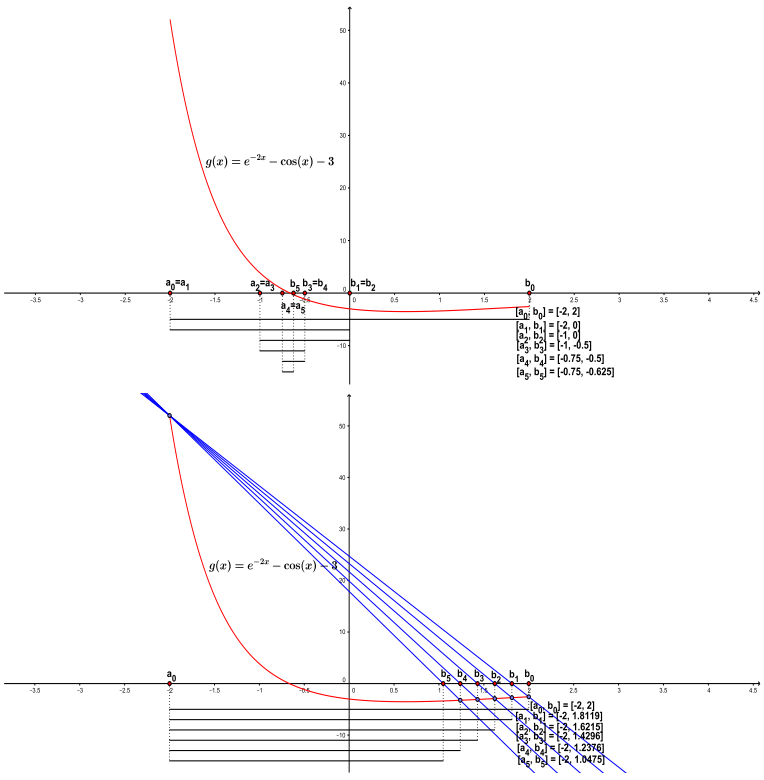
De même, les valeurs des extrémités des différents intervalles construits par les méthodes de dichotomie et regula falsi sont reportés dans le tableau ci dessous.

itération	Dichotomie		Regula falsi	
	a_n	b_n	a_n	b_n
0	1.000000000000000	4.000000000000000	1.000000000000000	4.000000000000000
1	2.500000000000000	4.000000000000000	1.916666666666667	4.000000000000000
2	2.500000000000000	3.250000000000000	2.636879969708444	4.000000000000000
3	2.875000000000000	3.250000000000000	2.979619151850352	4.000000000000000
4	3.062500000000000	3.250000000000000	3.100609650227808	4.000000000000000
5	3.062500000000000	3.156250000000000	3.138412893531248	4.000000000000000
6	3.109375000000000	3.156250000000000	3.149754925521175	4.000000000000000
7	3.132812500000000	3.156250000000000	3.153115836701065	4.000000000000000
8	3.144531250000000	3.156250000000000	3.154108070854592	4.000000000000000
9	3.150390625000000	3.156250000000000	3.154400685040792	4.000000000000000
10	3.153320312500000	3.156250000000000	3.154486950343403	4.000000000000000
11	3.153320312500000	3.154785156250000	3.154512379708390	4.000000000000000
12	3.154052734375000	3.154785156250000	3.154519875589223	4.000000000000000
13	3.154418945312500	3.154785156250000		
14	3.154418945312500	3.154602050781250		
15	3.154510498046875	3.154602050781250		
16	3.154510498046875	3.154556274414063		
17	3.154510498046875	3.154533386230469		

On voit que pour cet exemple, la méthode de dichotomie nécessite 17 itérations et fournit $x^* =$

3.1545219 avec $f(3.1545219) = -7.4136473\,10^{-6}$. Par contre, la méthode regula falsi nécessite 12 itérations et fournit $x^* = 3.1545221$ avec $f(3.1545219) = -6.4195643\,10^{-6}$.

- **Exemple 2.** Dans cet exemple, la fonction test est notée g. Cette fonction est définie par $g(x) = e^{-2x} - \cos(x) - 3$. La racine recherchée x^* est située dans l'intervalle $[a,\ b] = [-2,\ 2]$. De plus, on considère que les méthodes ont convergé à l'itération n dès que : $b_n - a_n < \varepsilon$ ou dès que $|f(x_n)| < \varepsilon$ où $\varepsilon = 10^{-6}$.



Les approximations successives et les bornes des intervalles construits par les méthodes de dichotomie et regula falsi sont reportées dans les deux tableaux ci-dessous.

Itération	Dichotomie	Regula falsi
	x_n	x_n
0	0.000000000000000	1.811979090589975
1	-1.000000000000000	1.621586693031262
2	-0.500000000000000	1.429695759399754
3	-0.750000000000000	1.237765872874417
4	-0.625000000000000	1.047755937945362
⋮	⋮	⋮
20	-0.665716171264648	-0.568539784247612
21	-0.665717124938965	-0.588218888839196
22	-0.665717601776123	-0.604030029928838
⋮	⋮	⋮
75		-0.665717356405127
76		-0.665717406147496
77		-0.665717445438678
78		-0.665717476474533

itération	Dichotomie		Regula falsi	
	a_n	b_n	a_n	b_n
0	-2.000000000000000	2.000000000000000	-2.000000000000000	2.000000000000000
1	-2.000000000000000	0.000000000000000	-2.000000000000000	1.811979090589975
2	-1.000000000000000	0.000000000000000	-2.000000000000000	1.621586693031262
3	-1.000000000000000	-0.500000000000000	-2.000000000000000	1.429695759399754
4	-0.750000000000000	-0.500000000000000	-2.000000000000000	1.237765872874417
⋮	⋮	⋮	⋮	⋮
20	-0.665718078613281	-0.665714263916016	-2.000000000000000	-0.544146472217116
21	-0.665718078613281	-0.665716171264648	-2.000000000000000	-0.568539784247612
22	-0.665718078613281	-0.665717124938965	-2.000000000000000	-0.588218888839196
⋮	⋮	⋮	⋮	⋮
75			-2.000000000000000	-0.665717293431628
76			-2.000000000000000	-0.665717356405127
77			-2.000000000000000	-0.665717406147496
78			-2.000000000000000	-0.665717445438678

Pour ce second exemple, la méthode de dichotomie a besoin de 22 itérations et la solution approchée obtenue est $x^* = -0.6657176$ avec $g(-0.6657176) = 7.0622965\,10^{-8}$. Par contre, la méthode regula falsi ne converge pas rapidement ! En effet, après 78 itération, la méthode donne $x^* = -0.6657174$, avec $g(-0.6657174) = -9.5566853\,10^{-7}$.

4.3 Méthodes de Newton et de la sécante

Une des méthodes classiques les plus utilisées pour la détermination d'une approximation de la racine d'une fonction est la méthode de Newton-Raphson (ou tout simplement méthode de Newton). Cette méthode est une méthode itérative et suppose que la fonction f est de classe C^1 sur l'intervalle où est situé le zéro recherché. En effet, la valeur de la dérivée au point courant x_n est nécessaire pour l'obtention du prochain point x_{n+1}.

Dans le cas où la fonction n'est pas de classe C^1 où si $f^{'}(x_n)$ n'est pas disponible ou ne peut être calculée, cette dérivée est approchée par une différence divisée et on obtient alors la méthode de la sécante.

4.3.1 Méthode de Newton

La méthode de newton peut être considérée comme une procédure de linéarisation. La fonction f est localement approchée par une fonction linéaire. En partant d'un bon point de départ x_0 (Le point x_0 peut être obtenu en faisant une étude préalable de la fonction f afin de situer le zéro recherché dans un intervalle assez petit), cette méthode construit une suite de valeurs $x_1, x_2, \ldots, x_n, \ldots$, qui sous certaines conditions converge de façon quadratique vers x^* la racine exacte de f.

En supposant avoir calculé x_n et en utilisant le développement de Taylor à l'ordre 1 afin d'approcher la fonction f localement au point x_n, on a,

$$f(x) = f(x_n) + (x - x_n) f'(x_n) + x\epsilon(x), \quad \text{ainsi} \quad f(x) \simeq f(x_n) + (x - x_n) f'(x_n).$$

En écrivant $0 = f(x_n) + (x - x_n) f'(x_n)$ (car on espère que $f(x)$ soit égale à 0), on obtient alors $x = x_n - \frac{f(x_n)}{f'(x_n)}$. Et donc le nouvel itéré x_{n+1} est donné par

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

Ainsi l'algorithme de la méthode de Newton peut être résumé comme suit :

Algorithme : Méthode de Newton.

Initialisation : Choisir x_0 une approximation initiale.
Itération : Pour $n = 0, 1, \ldots,$ jusqu'à convergence :

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

fin du Pour.

4.3.2 Méthode de la sécante

Comme cela a été déjà signalé, la méthode de la sécante est similaire à celle de la méthode de Newton sauf que le calcul de la dérivée $f'(x_n)$ est remplacé par la différence finie $\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$ et de plus l'initialisation nécessite deux points (voisins si possible) et proches de la solution exacte x^*. On obtient ainsi la récurrence suivante :

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

Ainsi, l'algorithme de la méthode de la sécante est comme suit :

Algorithme : Méthode de la sécante.

Initialisation : Choisir x_0 et x_1 deux approximations initiales.
Itération : Pour $n = 0, 1, \ldots,$ jusqu'à convergence :

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

fin du Pour.

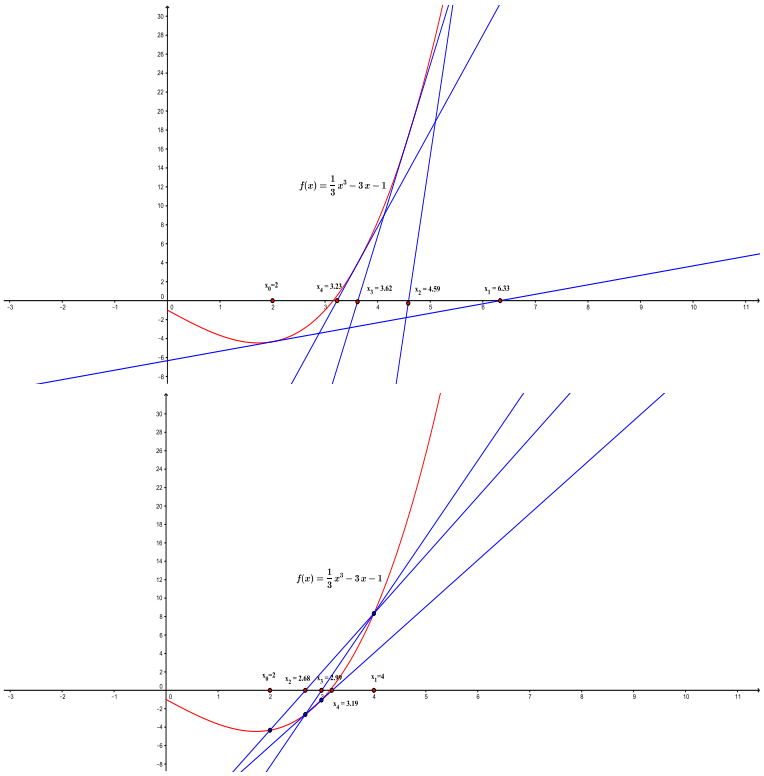
Notons également que la méthode de la sécante ressemble à la méthode de la fausse position (reégula falsi). En effet, le point w_n calculé lors de l'itération n de la méthode regula falsi est donné par

$$w_n = \frac{a_n f(b_n) - b_n f(a_n)}{f(b_n) - f(a_n)} = a_n - \frac{b_n - a_n}{f(b_n) - f(a_n)} f(a_n).$$

A noter cependant que dans la méthode de la sécante, il est impératif que les points a_n et b_n doivent encadrer la racine recherchée x^*. Par contre il n'est pas nécessaire dans la méthode de la sécante que les points x_n et x_{n-1} encadrent x^*.

4.3.3 Exemples

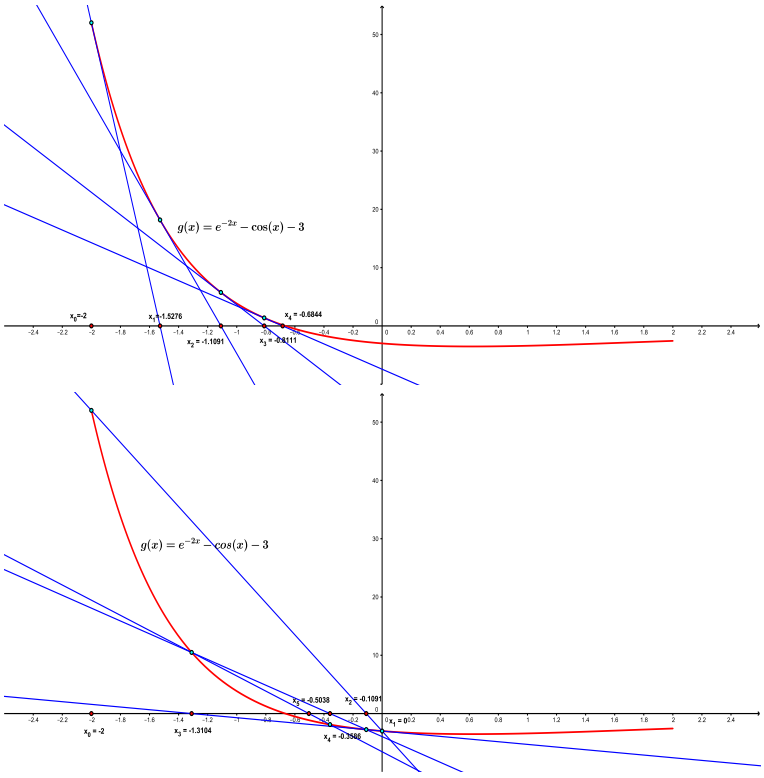
- **Exemple 3.** Dans cet exemple, la fonction test f est celle de l'exemple 1. Ainsi, $f(x) = \frac{1}{3}x^3 - 3x - 1$. Les solutions de départ sont respectivement $x_0 = 2$ pour la méthode de Newton et $x_0 = 2,\ x_1 = 4$ pour la méthode de la sécante. Les algorithmes sont stoppés dès que : $|x_n - x_{n-1}| < \varepsilon$ ou dès que $|f(x_n)| < \varepsilon$ où $\varepsilon = 10^{-5}$.



itération	Newton	Sécante
	x_n	x_n
0	2.000000000000000	2.000000000000000
1	6.333333333333334	4.000000000000000
2	4.590485695276115	2.684210526315789
3	3.623662471603157	2.997667703243003
4	3.229848830244545	3.197304031235400
5	3.156969701003765	3.151353280165889
6	3.154525720790307	3.154462307103545
7	3.154523008698545	

Dans ce test, l'algorithme de la méthode de Newton nécessite converge en 7 itérations et fournit $x^* = 3.1545230$ avec $f(3.1545230) = 2.3206325\,10^{-11}$. La méthode de la sécante fait légèrement mieux en convergeant en 6 itérations et donne $x^* = 3.1545231$ avec $f(3.1545231) = 6.0763745\,10^{-7}$.

- **Exemple 4.** La fonction testée dans cet exemple est la fonction g définie par $g(x) = e^{-2x} - \cos(x) - 3$. Les solutions de départ sont respectivement $x_0 = -2$ pour la méthode de Newton et $x_0 = -2$, $x_1 = 0$ pour la méthode de la sécante. Les algorithmes sont stoppés dès que : $|x_n - x_{n-1}| < \varepsilon$ ou dès que $|f(x_n)| < \varepsilon$ où $\varepsilon = 10^{-6}$.



itération	Newton	Sécante
	x_n	x_n
0	-2.000000000000000	-2.000000000000000
1	-1.527596252561599	0.000000000000000
2	-1.109132547403603	-0.109062559032824
3	-0.811076922352303	-1.310406534496917
4	-0.684359192809277	-0.358619704384055
5	-0.666051206702811	-0.503775826090157
6	-0.665717701377160	-0.723594093976699
7		-0.656288917145912
8		-0.665195908954252
9		-0.665722393967339

Pour cet exemple, la méthode de Newton a eu besoin de 6 itérations pour converger et la solution déterminée par cette méthode est $x^* = -0.6657177$ avec $f(-0.6657177) = 8.864124\,4\,10^{-7}$. La méthode de la sécante converge en 9 itérations et donne $x^* = -0.6657223$ avec $f(-0.6657223) = 3.932159\,10^{-5}$.