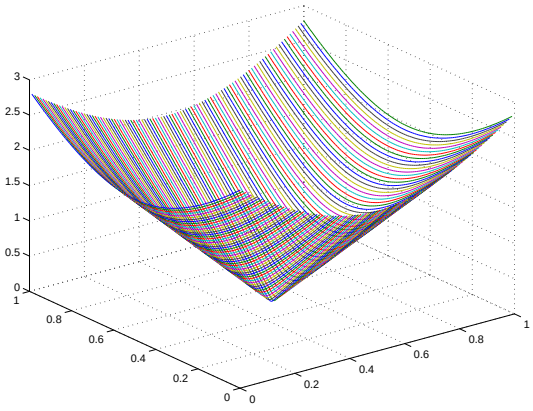
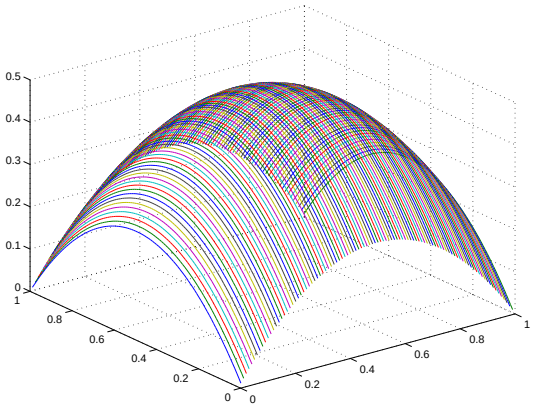
Université de Tlemcen
Faculté des Sciences

Cours d'Analyse Numérique





Année universitaire 2016/2017

Table des matières

Table des figures

Liste des tableaux

Préambule

Le développement technologique et plus particulièrement des ordinateurs a permis de résoudre des problèmes scientifiques de plus en plus difficiles et complexes en utilisant différentes techniques et méthodes numériques. Le chercheur d'aujourdhui doit impérativement maîtriser ces techniques. De ce fait, il doit avoir une formation suffisante en analyse numérique, qui représente la base d'une meilleure compréhension et résolution des problèmes à étudier. Les problèmes rencontrés par le futur chercheur sont souvent d'origine de branche fondamentale telle que la physique, la chimie, les mathématiques ; où les équations différentielles, les intégrales et les dérivées partielles jouent un rôle primordial dans leurs modélisations. Ce cours présente les éléments et les techniques fondamentaux de l'analyse numérique nécessaires pour résoudre un problème de façon convenable.

Les auteurs, en se basant sur leur expérience d'enseignement et de recherche multidisciplinaire nationale (Université de Tlemcen) et internationale (Université Paris 6 -Sorbonne-, Université de Pau, Université de Limoges, Université de Marseille et Université de Saint Joseph au Liban), ont exploité les interactions et différentes discussions avec des étudiants de différents niveaux et disciplines académiques afin de relever le défit dans la préparation de ce manuscrit.

Ce document se compose à la fois de cours magistraux, des exemples et des exercices. Un objectif que ce sont assigné les auteurs est de présenter ce cours de façon claire, didactique et accessible à un grand public à la recherche d'une référence dans le domaine, quelque soit leur discipline technique (mathématique, physique, mécanique,...).

Le cours comporte six chapitres. Dans le premier chapitre, nous introduisons un certain nombre de termes d'analyse numérique courants qui seront utilisés par la suite et qu'il convient de bien connaître tels que les opérations machines et les erreurs. Ensuite, dans le deuxième chapitre, nous présentons la résolution d'un système d'équations linéaires par des méthodes directes dans un premier temps et par des méthodes itératives dans le troisième chapitre. Dans le quatrième chapitre qui porte sur l'interpolation polynomiale, on s'intéresse à approcher une fonction connue par ses valeurs en certains points, par un polynôme en utilisant la méthode de Lagrange et celle de Newton ainsi que l'approximation par la méthode des moindres carrés. Dans le cinquième chapitre, nous nous intéressons aux méthodes d'intégration numérique dont le but est de déterminer l'intégrale d'une fonction sur un domaine fini délimité par des bornes finies. Nous exposons ici, les méthodes de Newton-Cotes simples et composites et les méthodes de Gauss-Legendre. Enfin, le sixième chapitre comporte la résolution d'une équation différentielle sous forme de problème de Cauchy en utilisant des schémas à un pas et à multi-pas.

Les solutions numériques ont été calculées à partir de programmes écrits par les auteurs en ***Fortran 95****. En utilisant le puissant et professionnel logiciel graphique* ***SigmaPlot****, une attention particulière a été réservée à la présentation graphique vu sa grande importance dans ce cours.*

Chapitre 1

Introduction à l'arithmétique flottante

Un nombre contenu dans un ordinateur est représenté par un nombre fini de caractères fixé qui dépend de l'architecture de la machine en terme de processeurs ou compilateurs et langages utilisés. De ce fait, les ordinateurs ne sont capables de représenter qu'un ensemble fini de nombres réels. Cela implique que les opérateurs qu'effectue un ordinateur, dite aussi "***opération machine***", ne sont qu'une approximation du calcul mathématique idéal.

Les conséquences sont alors très importantes, en particulier en terme de précision des résultats qui avèrent être dans certains cas grossièrement erronés, se présentant lors d'une simulation enchaînant plusieurs milliers d'opérations flottantes et entraînant par la suite des influences non négligeables tout au long de l'exécution d'un algorithme. Ici intervient alors la notion de "***compensation de la prorogation des erreurs***" qui reste le problème majeur du calcul flottant.

Les ordinateurs représentent les nombres réels sur un nombre fini de bits, ce qui ne permet la représentation exacte que d'un petit sous-ensemble des réels. Ainsi, la plupart des calculs conduisent à des résultats approchés qui résultent de la finitude de la représentation. L'analyse numérique essaie spécifiquement d'évaluer l'erreur lorsque sont utilisés des approximations de solutions d'équations ou des algorithmes numériques. La précision de la solution calculée par un algorithme numérique à un problème donné dépend non seulement du problème mais aussi de la stabilité de l'algorithme utilisé. Maintenir de bonnes performances pratiques tout en augmentant la précision est l'un des grands challenges en calcul scientifique.

Notons que dans un ordinateur, chaque nombre est codé par une séquence de bits, en général égal à 32 et prenant la valeur 0 ou 1. Cette séquence de bits est alors interprétée comme la représentation généralement en base 2 sur ordinateur d'un nombre réel, mais aussi 8 ou 16 sur certaines anciennes machines. Ainsi une donnée sur n bits est une combinaison linéaire de 2^{n-1} valeurs.

On parle d'***arithmétique en nombres entiers*** lorsque tous les calculs peuvent se faire en nombres entiers, évitant par la suite la propagation des erreurs qui peuvent induire des résultats aberrants parfois. Ces calculs sont alors exacts.

Dans le cas des calculs impliquant des nombres réels, on a recours à l'***arithmétique en virgule flottante*** pour représenter une approximation du résultat. Les nombres à virgule flottante, ap-

pelés plus simplement ***nombres flottants*** constituent le principal mode de représentation des nombres réels sur les calculateurs.

Il existe deux notations adoptées pour représenter les nombres réels sur ordinateur : ***système à virgule fixe*** et ***système à virgule flottante***.

1.1 Les chiffres significatifs

Dans un nombre comportant un séparateur décimal, on compte les chiffres significatifs à partir du premier chiffre non nul apparaissant à gauche et on procède de la manière suivante :

- Tous le chiffres non nuls sont significatifs,
- Les zéros à droite d'un chiffre significatif, sont significatifs quel que soit le nombre de rangs décimaux qui les séparent. Ils doivent être comptés lorsqu'on dénombre les chiffres significatifs et conservés lors d'un changement d'unité,
- les zéros à gauche ne sont pas significatifs,
- Lorsque le nombre est écrit en notation scientifique, tous les chiffres de ce dernier écrit devant la puissance de la base sont significatifs. La puissance quant à elle n'intervient pas dans le décompte.

Exemple :

- le nombre $\underset{1}{\boxed{1}}\,\underset{2}{\boxed{0}}\,\underset{3}{\boxed{2}}\,,\,\underset{4}{\boxed{0}}\,\underset{5}{\boxed{3}}\,\underset{6}{\boxed{0}}$ compte 6 chiffres significatifs
- le nombre $0,0\,\underset{1}{\boxed{5}}\,\underset{2}{\boxed{3}}\,\underset{3}{\boxed{0}}$ compte 3 chiffres significatifs
- le nombre $\underset{1}{\boxed{5}}\,,\,\underset{2}{\boxed{5}}\,\underset{3}{\boxed{3}}\,\underset{4}{\boxed{0}} \times 10^2$ compte 4 chiffres significatifs

1.2 Représentation en virgule flottante

Soit x un nombre réel non nul. Les ***nombres à virgule flottante*** sont représentés de la façon suivante :

$$\boxed{x \;=\; (-1)^s \,.\, m \,.\, b^e}$$

où $(-1)^s$ représente le signe de x avec $s \in \{0, 1\}$, b est la base qui est un entier naturel supérieur ou égal à 2 (usuellement 2 ou 10), m la mantisse et e l'exposant de x.

Exemple : $23.456 = 0.23456.10^2$ où 0.23456 est la mantisse, 10 la base avec 5 digits, et 2 l'exposant.

Soit t l'entier naturel désignant le nombre de chiffres que compte la mantisse m en base b (t le nombre de bits disponibles pour coder la mantisse). Celle-ci s'écrit alors

$$m \;=\; \sum_{i=1}^{t} x_i \, b^{-i} \;=\; x_1 \, b^{-1} \;+\; +x_2 \, b^{-2} \;+\; \ldots \;+\; x_{t-1} \, b^{1-t} \;+\; x_t \, b^{-t}$$

avec $x_1 \neq 0$ et $x_i \in \{0, 1, ..., b\}$, pour $i \in \{2, ..., t\}$ en ***notation normalisée*** de la mantisse qui sert à maximiser le nombre de digits significatifs tout en éliminant les digits nuls à gauche, autrement dit lorsque le chiffre de poids fort de la mantisse est non nul. Même en fixant la place de la virgule dans la mantisse, seule la condition $x_1 \neq 0$ assure l'unicité de la représentation des nombres flottants normalisés, mais nous prive du nombre 0.

Ainsi, on partitionne le mot en deux parties, l'une contenant l'exposant e et l'autre contenant la mantisse m et le premier bit à gauche indique le signe.

La notation

$$\boxed{\{[x_1, x_2, ..., x_t], e, b, s\}}$$

est utilisée pour encoder le nombre réel

$$x \;=\; (-1)^s \,.\, m \,.\, b^e \;=\; (-1)^s \, b^e \, \sum_{i=1}^{t} x_i \, b^{-i}$$

Exemple : La notation $\{[1, 2, 3, 4, 5], e = 1, b = 10, s = 0\}$ désigne le réel

$$x = (-1)^s \, b^e \sum_{i=1}^{t} x_i \, b^{-i} = (-1)^0 \, 10^1 \left(1.10^{-1} + 2.10^{-2} + 3.10^{-3} + 4.10^{-4} + 5.10^{-5} \right) = 1.2345.$$

En faisant varier e, on fait flotter la virgule. Ceci s'oppose à la représentation dite en ***virgule fixe***, où l'exposant e est fixé pour un type de donnée correspondant à un nombre qui possède un nombre fixe de chiffres après la virgule.

Exemple : La notation $\{[1, 2, 3, 4, 5], e = 0, b = 10, s = 0\}$ désigne le réel

$$x = (-1)^s \, b^e \sum_{i=1}^{t} x_i \, b^{-i} = (-1)^0 \, 10^0 \left(1.10^{-1} + 2.10^{-2} + 3.10^{-3} + 4.10^{-4} + 5.10^{-5} \right) = 0.12345$$

et la notation $\{[1, 2, 3, 4, 5], e = -1, b = 10, s = 0\}$ désigne le réel

$$x = (-1)^s \, b^e \sum_{i=1}^{t} x_i \, b^{-i} = (-1)^0 \, 10^{-1} \left(1.10^{-1} + 2.10^{-2} + 3.10^{-3} + 4.10^{-4} + 5.10^{-5} \right) = 0.012345.$$

Sur un ordinateur, il existe deux formats disponibles pour les nombres à virgule flottante.

- ***En format à simple précision*** C'est l'une des représentations standard des nombre flottants les plus utilisés : la ***norme IEEE 754*** qui régit le comportement du calcul flottant sur de nombreux ordinateurs. Dans ce cas, l'ordinateur dispose de 32 cases mémoires pour stocker un nombre répartis comme suit : *1 bit pour le signe, 8 bits pour l'exposant et 23 bits pour la mantisse.*
- ***En format à double précision*** Dans ce cas, l'ordinateur dispose de 64 cases mémoires pour stocker un nombre répartis comme suit : *1 bit pour le signe, 11 bits pour l'exposant et 52 bits pour la mantisse.*

Remarque 1.1. *L'IEEE 754 est une norme conçue pour la représentation des nombres à virgule flottante en binaire qui sert à éviter une prolifération de systèmes de représentation, qui différent en base, en nombre de chiffres significatifs et en exposants. Ceci a été développé par le IEEE (Institute of Electrical and Electronic Engineers) en 1985, ensuite révisé en 2008 et approuvé par le IEC (International Electronic Commission). La norme IEEE 754 définit quant à elle 4 formats de flottants en base 2 : simple précision, simple précision étendue, double précision et double précision étendue. Il reste à noter qu'une nouvelle révision de la norme IEEE 754 définit aussi le format quadruple précision avec 113 bits de mantisse sur un total de 128 bits.*

1.3 Source d'erreurs dans le calcul numérique

Soit $\tilde{x}$ une approximation d'un nombre réel x non nul. On dit que $\tilde{x}$ est l'estimateur de la valeur exacte x.

Les quatre sources d'erreur, qui interviennent systématiquement lorsqu'on s'intéresse à la résolution d'un problème de nature physique ou autre à l'aide des différents schémas numériques, sont :

- Les ***erreurs du modèle*** dues au fait que le problème numérique n'approche pas convenablement le problème mathématique issu généralement d'un problème physique.
- Les ***erreurs de troncature*** qui proviennent des simplifications qu'on applique au modèle mathématique ou des processus infinis en analyse mathématique. C'est le cas par exemple lorsqu'on remplace une fonction analytique par un développement en série de Taylor limité qui servira à calculer une valeur approchée $\tilde{x}$ de la solution exacte x.
- Les ***erreurs des données*** qui proviennent en général des mesures physiques imprécises dont les valeurs ne peuvent être déterminées qu'approximativement suite à des mesures expérimentales.
- Les ***erreurs d'arrondi*** dues au fait qu'il n'est possible de représenter les réels exactement dans un ordinateur. L'ensemble des réels est ainsi approché à l'aide d'un ensemble de nombres flottants à l'origine d'une faible erreur d'arrondi cumulée à chaque opération arithmétique et influant considérablement sur la précision du résultat final. Dans ce cas, on parle d'un ***algorithme numériquement instable*** qui pourra produire de mauvaises solutions même pour des problèmes bien conditionnées.

Les erreurs de troncature et d'arrondi constituent l'erreur numérique. Quant à elles, les erreurs d'arrondi influent considérablement sur la précision d'une solution approchée calculée en arithmétique flottante, exigeant que l'erreur globale soit inférieure à une certaine tolérance fixée et garantissant par suite la convergence de la méthode numérique mais aussi la fiabilité et l'efficacité de cette dernière.

1.4 Calcul d'erreur et notions de précision

L'utilisation des méthodes numériques nécessite un conditionnement précis du problème afin d'éviter les instabilités numériques dues à la propagation des erreurs d'arrondi ou encore de troncatures. Comment s'assurer alors de la fiabilité d'un résultat calculé en arithmétique flottante obtenu à partir d'une simulation enchaînant plusieurs milliers d'opérations flottantes.

Comme il n'est pas possible d'éviter la propagation des erreurs d'arrondi, on préfère borner l'erreur.

Soit $\tilde{x}$ une approximation d'un nombre réel x non nul. Les deux principales façons de mesurer la précision de $\tilde{x}$ sont :

- ***Erreur absolue de*** $\tilde{x}$ Elle est définie comme

$$\boxed{\mathbf{E}_a(\tilde{x}) = |\tilde{x} - x|}$$

  Elle représente l'écart entre la valeur approchée et celle exacte. On l'utilise généralement quand on connaît une majoration a priori des valeurs intermédiaires calculées.

- ***Erreur relative à*** $\tilde{x}$ Par définition, l'erreur relative est le quotient de l'erreur absolue à la valeur exacte quand on dispose d'une, elle est égale à

$$\boxed{\mathbf{E}_r(\tilde{x}) = \frac{|\tilde{x} - x|}{|x|}}$$

  Cependant, elle n'est définie que si $x \neq 0$.

  Si $x = 0$, on utilise la mesure suivante qui n'est autre qu'une combinaison entre erreur absolue et erreur relative :

$$\boxed{\mathbf{E} = \frac{|\tilde{x} - x|}{|x|+1}}$$

  Cette mesure possède les caractéristiques de l'erreur absolue si $|x| \ll 1$ et les caractéristiques de l'erreur relative si $|x| \gg 1$.

  Remarque 1.2. *Dans le cas où x et $\tilde{x}$ sont deux vecteurs, les notions d'erreur absolue et relative deviennent respectivement :*

$$\boxed{\boldsymbol{E}_a(\tilde{x}) = \|\tilde{x} - x\|} \quad et \quad \boxed{\boldsymbol{E}_r(\tilde{x}) = \frac{\|\tilde{x} - x\|}{\|x\|}}$$

  où $\|.\|$ est une norme adéquate.

1.5 Incertitude des erreurs

Les sciences reposent en général sur la confrontation entre deux démarches numérique et théorique à l'origine de la notion d'erreur liée généralement aux résultats mis en œuvre grâce à de nombreux schémas numériques employant l'arithmétique flottante. Des résultats obtenus à partir de tels schémas sont souvent entachés des erreurs d'arrondi dont les causes sont multiples, ce qui constitue le problème majeur du calcul flottant. Pour une critique objective du résultat, on introduit la notion de l'erreur maximale que l'on appelle de façon plus appropriée ***incertitude***.

- ***Incertitude absolue*** On présente idéalement un résultat calculé sous la forme d'un intervalle

$$\boxed{x = \tilde{x} \pm \Delta x}$$

car l'indication complète d'un résultat obtenu à partir d'un algorithme généralement grâce à l'arithmétique flottante disponible sur les ordinateurs moderne, comporte la valeur approchant aux mieux la valeur exacte et l'intervalle dans lequel se situe cette solution exacte qui est en général le centre de cet intervalle.

$\triangle\, x$ est appelée l'incertitude absolue, qui n'est autre que la demi-longueur de cet intervalle, et on a

$$\boxed{\tilde{x} - \Delta x \leq x \leq \tilde{x} + \Delta x}$$

Remarque 1.3. *$\triangle\, x$ est aussi appelé le* ***majorant de l'erreur absolue*** *d'une valeur approchée $\tilde{x}$ puisque $\triangle\, x$ est le plus petit nombre réel positif vérifiant :*

$$|\tilde{x} - x| \leq \Delta x.$$

- ***Incertitude relative*** Afin de mieux juger la qualité et la fiabilité d'un résultat numérique, on se sert de l'incertitude relative :

$$\boxed{\frac{\Delta x}{\tilde{x}}}$$

Remarque 1.4. *L'incertitude relative ainsi que l'erreur relative sont des nombres sans unité que l'on exprime généralement en %.*

1.5.1 Règles des calculs classiques d'incertitude

Soient $\tilde{x}$, $\tilde{y}$ et $\tilde{z}$ respectivement des approximations des quantités exactes x, y et z. On note par Δx, Δy et Δz les incertitudes absolues de x, y et z.

- ***Incertitude sur une addition***

$$\boxed{\text{Si } z = x + y, \quad \text{alors} \quad \Delta z = \Delta x + \Delta y}$$

L'incertitude absolue d'une somme est égale à la somme des incertitudes absolues

Preuve 1.1. *On a*

$$\tilde{x} - \Delta x \leq x \leq \tilde{x} + \Delta x \quad et \quad \tilde{y} - \Delta y \leq y \leq \tilde{y} + \Delta y$$

Par suite,

$$(\tilde{x} + \tilde{y}) - (\Delta x + \Delta y) \leq x + y \leq (\tilde{x} + \tilde{y}) + (\Delta x + \Delta y)$$

Ainsi $(\Delta x + \Delta y)$ est l'incertitude absolue de $x + y$ et on obtient

$$\Delta z = \Delta(x + y) = \Delta x + \Delta y$$

- ***Incertitude sur une soustraction***

$$\boxed{\text{Si } z = x - y, \quad \text{alors} \quad \Delta z = \Delta x + \Delta y}$$

L'incertitude absolue d'une différence est égale à la somme des incertitudes absolues

Preuve 1.2. *On a*

$$\tilde{x} - \Delta x \leq x \leq \tilde{x} + \Delta x \quad et \quad -\tilde{y} - \Delta y \leq -y \leq -\tilde{y} + \Delta y$$

Par suite,

$$(\tilde{x} - \tilde{y}) - (\Delta x + \Delta y) \leq x - y \leq (\tilde{x} - \tilde{y}) + (\Delta x + \Delta y)$$

Ainsi $(\Delta x + \Delta y)$ *est l'incertitude absolue de* $x - y$ *et on obtient :*

$$\Delta z = \Delta(x - y) = \Delta x + \Delta y$$

Remarque 1.5. *Une somme ou une différence ne peut pas être plus précis que la donnée sur laquelle l'incertitude relative est la plus élevée, autrement dit le résultat d'une une somme ou d'une différence est arrondi au même nombre de décimales que la donnée qui en comporte le moins.*

Exemple : les nombres *3.4* et *2.46* comptent respectivement 2 et 3 chiffres après la virgule. La résultante de la somme des ces deux nombres, donnée par la calculatrice, est égale à *5.86* et le résultat s'arrondit à *5.9* afin d'avoir le même nombre de chiffres après la virgule que *3.4* qui en possède le moins.

En effet, chacune des données a une certaine précision indiquée par son nombre de chiffres après la virgule et celle qui en possède le moins influe considérablement sur le résultat qui doit avoir le même nombre de chiffres après la virgule que celui de la donnée qui en a le moins.

- ***Incertitude sur un produit***

$$\boxed{\text{Si } z = \frac{x}{y}, \quad \text{alors} \quad \Delta z = \tilde{y}\,\Delta x + \tilde{x}\,\Delta y \quad \text{et} \quad \frac{\Delta z}{\tilde{z}} = \frac{\Delta x}{\tilde{x}} + \frac{\Delta y}{\tilde{y}}}$$

L'incertitude relative d'un produit est plus au moins égale à la somme des incertitudes relatives

Preuve 1.3. *On a*

$$\tilde{x} - \Delta x \leq x \leq \tilde{x} + \Delta x \quad et \quad \tilde{y} - \Delta y \leq y \leq \tilde{y} + \Delta y$$

Par suite,

$$(\tilde{x} - \Delta x)(\tilde{y} - \Delta y) \leq x\,y \leq (\tilde{x} - \Delta x)(\tilde{y} - \Delta y)$$

$$\tilde{x}\,\tilde{y} - (\tilde{x}\,\Delta y + \tilde{y}\,\Delta x) + \Delta x\,\Delta y \leq x\,y \leq \tilde{x}\,\tilde{y} + (\tilde{x}\,\Delta y + \tilde{y}\,\Delta x) + \Delta x\,\Delta y$$

Si $\Delta x << x$ et $\Delta y << y$, on néglige le terme du second ordre $\Delta x \Delta y$ et on obtient

$$\tilde{x}\,\tilde{y} - (\tilde{x}\,\Delta y + \tilde{y}\,\Delta x) \leq xy \leq \tilde{x}\,\tilde{y} + (\tilde{x}\,\Delta y + \tilde{y}\,\Delta x)$$

Ainsi $(\tilde{x}\,\Delta y + \tilde{y}\,\Delta x)$ est l'incertitude absolue de xy et on obtient :

$$\Delta z = \Delta(xy) = \tilde{x}\,\Delta y + \tilde{y}\,\Delta x$$

D'autre part, l'incertitude relative est égale à

$$\frac{\Delta z}{\tilde{z}} = \frac{\Delta(xy)}{\tilde{x}\,\tilde{y}} = \frac{\tilde{x}\,\Delta y + \tilde{y}\,\Delta x}{\tilde{x}\,\tilde{y}} = \frac{\Delta x}{\tilde{x}} + \frac{\Delta y}{\tilde{y}}$$

- ***Incertitude sur un quotient***

$$\text{Si } z = \frac{x}{y}, \quad \text{alors} \quad \Delta z = \frac{\Delta x}{\tilde{y}} + \frac{\tilde{x}\,\Delta y}{\tilde{y}^2} \quad \text{et} \quad \frac{\Delta z}{\tilde{z}} = \frac{\Delta x}{\tilde{x}} + \frac{\Delta y}{\tilde{y}}$$

L'incertitude relative d'un quotient est plus au moins égale à la somme des incertitudes relatives

Preuve 1.4. *On a*

$$\tilde{x} - \Delta x \leq x \leq \tilde{x} + \Delta x \quad et \quad \tilde{y} - \Delta y \leq y \leq \tilde{y} + \Delta y$$

Par suite,

$$\frac{\tilde{x} - \Delta x}{\tilde{y} + \Delta y} \leq \frac{x}{y} \leq \frac{\tilde{x} + \Delta x}{\tilde{y} - \Delta y}$$

$$\frac{(\tilde{x} - \Delta x)(\tilde{y} - \Delta y)}{(\tilde{y} + \Delta y)(\tilde{y} - \Delta y)} \leq \frac{x}{y} \leq \frac{(\tilde{x} + \Delta x)(\tilde{y} + \Delta y)}{(\tilde{y} - \Delta y)(\tilde{y} + \Delta y)}$$

$$\frac{\tilde{x}\,\tilde{y} - (\tilde{x}\,\Delta y + \tilde{y}\,\Delta x) + \Delta x\,\Delta y}{\tilde{y}^2 - \Delta y^2} \leq \frac{x}{y} \leq \frac{\tilde{x}\,\tilde{y} + (\tilde{x}\,\Delta y + \tilde{y}\,\Delta x) + \Delta x\,\Delta y}{\tilde{y}^2 - \Delta y^2}$$

Si $\Delta x << x$ et $\Delta y << y$, on néglige les terme du second ordre $\Delta x \Delta y$ et Δy^2 et on obtient

$$\frac{\tilde{x}}{\tilde{y}} - \frac{\tilde{x}\,\Delta y + \tilde{y}\,\Delta x}{\tilde{y}^2} \leq \frac{x}{y} \leq \frac{\tilde{x}}{\tilde{y}} + \frac{\tilde{x}\,\Delta y + \tilde{y}\,\Delta x}{\tilde{y}^2}$$

Ainsi $\dfrac{\tilde{x}\,\Delta y + \tilde{y}\,\Delta x}{\tilde{y}^2}$ est l'incertitude absolue de $\dfrac{x}{y}$ et on obtient :

$$\Delta z = \Delta\left(\frac{x}{y}\right) = \frac{\tilde{x}\,\Delta y + \tilde{y}\,\Delta x}{\tilde{y}^2} = \frac{\Delta x}{\tilde{y}} + \frac{\tilde{x}\,\Delta y}{\tilde{y}^2}$$

D'autre part, l'incertitude relative est égale à

$$\frac{\Delta z}{\tilde{z}} = \frac{\Delta\left(\frac{x}{y}\right)}{\left(\frac{\tilde{x}}{\tilde{y}}\right)} = \frac{\tilde{x}\,\Delta y + \tilde{y}\Delta x}{\tilde{x}\tilde{y}} = \frac{\Delta x}{\tilde{x}} + \frac{\Delta y}{\tilde{y}}$$

Remarque 1.6. *Un produit et un quotient ne peut pas être plus précis que la donnée sur laquelle l'incertitude absolue est la plus élevée, autrement dit le résultat d'une multiplication ou une division est arrondi à autant de chiffres significatifs que la donnée qui en compte le moins.*

Exemple : les nombres *3.40* et *2.563* comptent respectivement 3 et 4 chiffres significatifs. La résultante du produit des ces deux nombres, donnée par la calculatrice, est égale à *8.7142* et le résultat s'arrondit à *8.71* afin d'avoir le même nombre de chiffres significatifs que *3.40* qui en possède le moins.

- ***Incertitude sur une puissance*** Soit r un nombre quelconque

$$\text{Si } z = x^r, \quad \text{alors} \quad \frac{\Delta z}{\tilde{z}} = |r| \frac{\Delta x}{\tilde{x}}$$

L'incertitude relative d'une puissance d'une variable est égale au produit de la valeur absolue de l'exposant par l'incertitude relative sur la variable

1.6 Troncature, arrondissement et chiffres significatifs d'un nombre

La troncature à l'unité d'un nombre décimal positif est sa partie entière. L'arrondi d'un nombre est la valeur approchée de ce nombre obtenue en réduisant le nombre de chiffres significatifs. Le résultat est moins précis, mais en de nombreuses circonstances, c'est plus facile à employer.

1.6.1 Règles d'arrondissement

Il existe plusieurs méthodes d'arrondissement parmi lesquelles on cite :

- la méthode ***d'arrondi au pair le plus proche*** est celle utilisée par défaut dans les microprocesseur et le calcul numérique suivant la norme IEEE 754. Elle suit les règles d'arrondissement suivantes :
  - Lorsque le chiffre suivant à droite du chiffre qu'on a choisi de conserver est inférieur à 5, le chiffre précédent reste inchangé,
    Exemple : ***1.0272*** arrondis aux millièmes devient ***1.027*** puisque le chiffre qui suit la décimale à laquelle le nombre doit être arrondi (2) est inférieur à 5
  - Lorsque le chiffre suivant à droite du chiffre qu'on a choisi de conserver est supérieur à 5, le chiffre précédent est augmenté d'une unité,
    Exemple : ***5.0637*** arrondis aux millièmes devient ***5.064*** puisque le chiffre qui suit la décimale à laquelle le nombre doit être arrondi (7) est supérieur à 5

- Lorsque le chiffre suivant à droite du chiffre qu'on a choisi de conserver est égal à 5 et si un des chiffres qui le suivent n'est pas nul, le chiffre précédent est augmenté d'une unité,
  Exemple : ***4.562501*** arrondis aux millièmes devient ***4.563*** puisque (1) est l'un des chiffres qui suivent le chiffre (5) et est non nul

- Tandis que si le chiffre suivant à droite du chiffre qu'on a choisi de conserver est égal à 5 et n'est suivi d'aucun chiffre ou que par des zéros, alors le chiffre précédent est augmenté d'une unité lorsqu'il est impair et reste inchangé sinon.
  Exemple : ***7.0135*** arrondis aux millièmes devient ***7.014*** puisque le dernier chiffre est 5, et le chiffre de la décimale à laquelle le nombre doit être arrondi (3) est impair
  Exemple : ***4.5625*** arrondis aux millièmes devient ***4.562*** puisque le dernier chiffre est 5, et le chiffre de la décimale à laquelle le nombre doit être arrondi(2) est pair

  On procède de la sorte afin d'éviter le biais qui surviendrait en arrondissant à chaque fois par excès les nombres dont le dernier chiffre est cinq.

- la méthode ***d'arrondi au plus proche ou arrondi arithmétique*** qui est la plus connue, mais pas la plus courante. Elle consiste à séparer les dix chiffres décimaux en deux parties :

  - Si le nombre à arrondir est positif : Lorsque le chiffre suivant à droite du chiffre qu'on a choisi de conserver vaut au moins 5, on augmente ce chiffre d'une unité, autrement dit on passe à la valeur inférieure,
    Exemple : ***2,752*** s'arrondit aux dixièmes à ***2.8*** car le chiffre suivant 7 vaut au moins 5

  - Si le nombre à arrondir est négatif : Lorsque le chiffre suivant à droite du chiffre qu'on a choisi de conserver vaut au moins 5, on augmente ce chiffre d'une unité, autrement dit on passe à la valeur supérieure,
    Exemple : ***-2.752*** s'arrondit aux dixièmes à ***-2.8*** car le chiffre suivant 7 vaut au moins 5

  - On conserve ce chiffre si le chiffre suivant est strictement inférieur à 5.
    Exemple : ***3.438*** et ***-3.438*** s'arrondissent aux dixièmes à ***3.4*** et ***-3.4*** respectivement car le chiffre suivant 4 vaut au moins 3

- la méthode ***d'arrondi stochastique*** qui consiste aussi à arrondir à l'entier le plus proche de manière aléatoire ou pseudo-aléatoire.

Ainsi, il sera utile lorsque l'on effectue un calcul avec plusieurs étapes qu'on n'arrondisse pas les résultats des étapes intermédiaires. Il n'y a que le résultat final qui doit comporter le bon nombre de chiffres significatifs. Cela permettra d'éviter de faire des erreurs d'arrondi.

Chapitre 2

Résolution des systèmes linéaires

2.1 Système linéaire

Une équation linéaire à n inconnues x_1, x_2, ..., x_n, où $n \in \mathbb{N}$ un entier naturel supérieur à 1, est une équation de la forme :

$$a_1\, x_1 \;+\; a_2\, x_2 \;+\; \ldots \;+\; a_n\, x_n \;=\; b,$$

où a_1, a_2,..., a_n et b sont des nombres réels donnés.

Soit $m \in \mathbb{N}^*$ un entier naturel supérieur à 1. Un *système de m équations à n inconnues* est un ensemble de m équations linéaires à n inconnues. On l'appelle aussi système linéaire donné sous la forme générale suivante :

$$\left\{\begin{array}{lllllll}
a_{11}\, x_1 & + a_{12}\, x_2 & + a_{13}\, x_3 & + \ldots & + a_{1j}\, x_j & + \ldots & + a_{1n}\, x_n \;=\; b_1,\\
a_{21}\, x_1 & + a_{22}\, x_2 & + a_{23}\, x_3 & + \ldots & + a_{2j}\, x_j & + \ldots & + a_{2n}\, x_n \;=\; b_2,\\
a_{31}\, x_1 & + a_{32}\, x_2 & + a_{33}\, x_3 & + \ldots & + a_{3j}\, x_j & + \ldots & + a_{3n}\, x_n \;=\; b_3,\\
 & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & \\
a_{i1}\, x_1 & + a_{i2}\, x_2 & + a_{i3}\, x_3 & + \ldots & + a_{ij}\, x_j & + \ldots & + a_{in}\, x_n \;=\; b_i,\\
 & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & \\
a_{m1}\, x_1 & + a_{m2}\, x_2 & + a_{m3}\, x_3 & + \ldots & + a_{mj}\, x_j & + \ldots & + a_{mn}\, x_n \;=\; b_m.
\end{array}\right.$$

Les nombres a_{ij}, pour $i = 1, 2, \ldots, m$ et $j = 1, 2, \ldots, n$ sont les coefficients du système et les nombres b_i pour $i = 1, 2, \ldots, m$ constituent le second membre du système.

2.2 Notation matricielle

Un système de m équations à n inconnues x_1, x_2, ..., x_n, où $n \in \mathbb{N}$ s'écrit aussi sous forme matricielle :

$$A\, x \;=\; b,$$

où

• A est une matrice comportant m lignes et n colonnes, notée aussi $A = a_{ij\,\{i=1,2,..,m;\ j=1,2,...,n\}}$, s'appelle la matrice du système linéaire :

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2j} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3j} & \dots & a_{3n} \\ & & \dots\dots\dots\dots\dots & & & & \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{ij} & \dots & a_{in} \\ & & \dots\dots\dots\dots\dots & & & & \\ a_{m1} & a_{m2} & a_{n3} & \dots & a_{mj} & \dots & a_{mn} \end{pmatrix}$$

• x est le vecteur colonne dont les composantes sont les inconnues $x_{i\{i=1,2,...,n\}}$. Le vecteur x est appelé *solution du système.*

• b aussi un vecteur colonne avec n composantes, noté $b_{i\{i=1,2,...,m\}}$. b s'appelle *le second membre.*

On introduit la matrice $\tilde{A}$:

$$\tilde{A} = \left(\begin{array}{ccccccc|c} a_{11} & a_{12} & a_{13} & \dots & a_{1j} & \dots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2j} & \dots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \dots & a_{3j} & \dots & a_{3n} & b_3 \\ & & \dots\dots\dots\dots\dots & & & & & \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{ij} & \dots & a_{in} & b_i \\ & & \dots\dots\dots\dots\dots & & & & & \\ a_{m1} & a_{m2} & a_{m3} & \dots & a_{mj} & \dots & a_{mn} & b_m \end{array}\right)$$

$\tilde{A}$ s'appelle la *matrice augmentée du système.* Elle contient la matrice du système linéaire A avec le second membre, le vecteur b, ajouté à sa droite.

Exemple : Considérons le système linéaire suivant :

$$\begin{cases} 2x_1 & -3x_2 & +x_3 & = 1, \\ & x_2 & -4x_3 & = 0, \\ x_1 & -x_2 & +2x_3 & = -4. \end{cases}$$

Sa matrice est :

$$A = \begin{pmatrix} 2 & -3 & 1 \\ 0 & 1 & -4 \\ 1 & -1 & 2 \end{pmatrix}$$

et sa matrice augmentée

$$\tilde{A} = \left(\begin{array}{ccc|c} 2 & -3 & 1 & 1 \\ 0 & 1 & -4 & 0 \\ 1 & -1 & 2 & -4 \end{array}\right).$$

2.3 Rappels d'algèbre linéaire

2.3.1 Quelques matrices particulières

Définition 2.1. *Soit $A = (a_{ij}) \in M_{m,n}(\mathbb{R})$. La matrice **transposée** d'une matrice A est la matrice obtenue en échangeant les lignes et les colonnes de A. Elle est notée ${}^tA$ et $\in M_{n,m}(\mathbb{R})$ et elle vérifie les propriétés suivantes :*

- ⋆ ${}^t\big({}^tA\big) = A$,
- ⋆ ${}^t\big(A+B\big) = {}^tA + {}^tB$, *où* $B = (b_{ij}) \in M_{n,m}(\mathbb{R})$,
- ⋆ ${}^t\big(AB\big) = {}^tA\,{}^tB$, *où* $B = (b_{ij}) \in M_{m,r}(\mathbb{R})$,
- ⋆ ${}^t\big(\alpha A\big) = \alpha\,{}^tA$, *où* α *est un nombre réel,*
- ⋆ $\big({}^tA\big)^{-1} = {}^t\big(A^{-1}\big)$, *où* A *est une matrice inversible.*

Définition 2.2. *Soit $A = (a_{ij})$ une matrice d'ordre n.*

- *La matrice **conjuguée** d'une matrice A à coefficients complexes est la matrice formée des éléments conjugués de A, elle est notée $\overline{A}$,*

- *La matrice **adjointe** ou **transconjuguée** d'une matrice A à coefficients complexes est égale à la matrice transposée de la matrice conjuguée de A, elle est notée A^*. Dans le cas particulier où A est à coefficients réels, sa matrice adjointe est donc simplement sa matrice transposée.*

- *Une matrice A est dite **symétrique**, si et seulement si elle est égale à sa propre transposée, i.e, elle vérifie ${}^tA = A$ ou encore $a_{ij} = a_{ji}$, $\forall\, i,j \in \{1,2,...,n\}$,*

  ***(i)** Une matrice A réelle est dite **symétrique positive** si et seulement si*

  $$\forall \boldsymbol{x} \in \mathbb{R}^n,\ \ {}^t\boldsymbol{x}\, A\, \boldsymbol{x} \geq 0,$$

  ***(ii)** Une matrice A réelle est dite **symétrique définie positive** si et seulement si elle vérifie l'une des trois propriétés équivalentes suivantes :*

  - ⋆ *La forme quadratique définie par la matrice A est strictement positive pour tout vecteur $\boldsymbol{x}$ non nul, i.e,*

    $$\forall \boldsymbol{x} \in \mathbb{R}^n\{0\},\ \ {}^t\boldsymbol{x}\, A\, \boldsymbol{x} > 0$$

  - ⋆ *Toutes les valeurs propres réelles de A sont strictement positives,*
  - ⋆ *La matrice A est congruente à la matrice identité, i.e, il existe une matrice $B \in M_n(\mathbb{R})$ inversible telle que $A = {}^tBB$.*

  ***(iii)** Une matrice A réelle est dite **symétrique définie négative** si et seulement si sa matrice opposée symétrique est définie positive.*

- *Une matrice A est dite **antisymétrique** si et seulement si sa transposée est égale à son opposée, i.e, elle vérifie ${}^tA = -A$ ou encore $a_{ij} = -a_{ji}$, $\forall\, i,j \in \{1,2,...,n\}$,*

- *Une matrice A à coefficients complexes est dite **hermitienne** si et seulement si elle est égale à sa propre transconjuguée, i.e, elle vérifie $A^* = A$ ou encore $a_{ij} = \overline{a_{ji}}$, $\forall\, i,j \in \{1,2,...,n\}$, et elle vérifie les propriétés suivantes :*

- $\left(A^*\right)^* = A$,
- $\left(A+B\right)^* = A^* + B^*$,
- $\left(AB\right)^* = B^*A^*$,
- $\left(\alpha A\right)^* = \alpha A^*$, *où* α *est un nombre complexe.*

(i) *Une matrice A complexe est dite* ***définie positive*** *si elle vérifie l'une des trois propriétés équivalentes suivantes :*

- *La forme quadratique définie par la matrice A est strictement positive pour tout vecteur $\boldsymbol{z}$ à coefficients complexes et non nul, i.e,*

$$\forall \boldsymbol{z} \in \mathbb{R}^n\{0\}, \ {}^t\boldsymbol{z}\, A\, \boldsymbol{z} > 0$$

- *La matrice A est hermitienne et toutes ses valeurs propres sont strictement positives,*
- *Il existe une matrice $B \in M_n(\mathbb{C})$ inversible telle que $A = B^*B$.*

(ii) *Une matrice A à coefficients complexes est dite* ***définie négative*** *si et seulement si sa matrice opposée est définie positive.*

- *Une matrice A à coefficients complexes est dite* ***antihermitienne*** *si et seulement si sa transconjuguée est égale à son opposée, i.e, elle vérifie $A^* = -A$ ou encore $a_{ij} = -\overline{a_{ji}}$, $\forall\, i,j \in \{1,2,...,n\}$,*

Remarque 2.1. *Une matrice définie positive vérifie les propriétés suivantes :*

- *Soit A une matrice définie positive. La matrice inverse de A est définie positive,*
- *Si A est définie positive et si α est un réel strictement positif, alors αA est définie positive,*
- *Si A et B sont positives et si l'une des deux est inversible, alors la matrice $A+B$ est définie positive.*

- *Une matrice A est dite* ***normale*** *si et seulement si elle vérifie $AA^* = A^*A$,*
- *Une matrice A à coefficients complexes est dite* ***unitaire*** *si et seulement si elle vérifie $AA^* = I_n$, où I_n est la matrice identité d'ordre n,*
- *Une matrice A* ***orthogonale*** *est une matrice unitaire à coefficients réels. Elle vérifie ${}^tAA = A^tA = I_n$, où I_n est la matrice identité d'ordre n ou encore $A^{-1} = {}^tA$,*
- *Une matrice A est dite* ***inversible*** *si et seulement si elle vérifie* $det\ A^{-1} = \dfrac{1}{det\ A}$.

2.3.2 Déterminant

Soit A une matrice carrée d'ordre n. Le déterminant se dénote $det\ A$ ou $|A|$. Il vérifie :

- $det\ {}^tA = det\ A$,
- $det\ AB = det\ A\ det\ B$,

⋆ Si A est une matrice inversible, alors $det\ A^{-1} = \dfrac{1}{det\ A}$ et $det\ A^* = \overline{det\ A}$,
⋆ Si A est une matrice triangulaire, alors $det\ A = \prod_{i=1}^{n} a_{ii}$,
⋆ Soit α un réel ou un complexe, alors $det\ (\alpha A) = \alpha\ det\ A$.

2.3.3 Trace d'une matrice

Soit $A = (a_{ij})$ une matrice d'ordre n. La trace de A est définie comme la somme de ses coefficients diagonaux et est notée $tr(A)$:

$$tr(A) = \sum_{i=1}^{n} a_{ii}$$

et elle vérifie les propriétés suivantes :

⋆ $tr\big(A+B\big) = tr(A) + tr(B)$, où $B = (b_{ij})$ une matrice d'ordre n,
⋆ $tr\big({}^tA\big) = tr(A)$,
⋆ $tr\big(\alpha A\big) = \alpha\, tr(A)$, où α est un nombre réel.

2.3.4 Rayon spectral et vecteurs propres

Définition 2.3. *Soit A une matrice réelle d'ordre n et $\lambda \in \mathbb{C}$. λ est une **valeur propre** de A s'il existe un vecteur $\boldsymbol{x} \in \mathbb{R}^n$ tel que $A\boldsymbol{x} = \lambda\boldsymbol{x}$. On dit alors que $\boldsymbol{x}$ est un **vecteur propre** de A associé à λ.*

Proposition 2.1. *λ est une **valeur propre** de A si et seulement si*

$$det\big(A - \lambda I\big) = 0$$

*Les valeurs propres de A sont donc les racines du **polynôme caractéristique** de A :*

$$P_A(\boldsymbol{x}) = det\big(A - \boldsymbol{x}I\big)$$

qui possède n racines $\lambda_1, \lambda_2, ..., \lambda_n$ dans $\mathbb{C}$ non nécessairement distincts telles que :

$$tr(A) = \sum_{i=1}^{n} \lambda_i \ \text{ et } \ detA = \prod_{i=1}^{n} \lambda_i$$

Définition 2.4. *Le **rayon spectral** d'une matrice A est définie comme étant :*

$$\rho(A) = \max\{|\lambda|;\ \lambda \in \mathbb{C} \text{ valeur propre de } A\}$$

2.3.5 Normes vectorielles et normes matricielles

Définition 2.5. *Soit F un $\mathbb{R}$-espace vectoriel. Une **norme vectorielle** sur $F = \mathbb{R}^n$, notée $\|.\|$ est une application de F à valeurs réelles positives et satisfaisant les hypothèses suivantes :*

- ***Séparation :*** $\forall \boldsymbol{x} \in F,\ \|\boldsymbol{x}\| = 0 \implies \boldsymbol{x} = 0_F$,
- ***Homogénéité :*** $\forall (\alpha, \boldsymbol{x}) \in \big(\mathbb{R}, F\big),\ \|\alpha\ \boldsymbol{x}\| = |\alpha|\|\boldsymbol{x}\|$,

- ***Inégalité triangulaire :*** $\forall(\boldsymbol{x},\boldsymbol{y})\in F^2,\ \|\boldsymbol{x}+\boldsymbol{y}\|\leq\ \|\boldsymbol{x}\|+\|\boldsymbol{y}\|$.

Un espace vectoriel muni d'une norme est appelé ***espace vectoriel normé***.

La norme vectorielle vérifie les propriétés suivantes :

⋆ $\forall(\alpha,\mathbf{x},\mathbf{y})\in\big(\mathbb{R},F^2\big),\ \|\alpha\ \mathbf{x}+\mathbf{y}\|\leq\ |\alpha|\|\mathbf{x}\|+\|\mathbf{y}\|$,

⋆ $\forall(\mathbf{x},\mathbf{y},\mathbf{z})\in F^3,\ \|\mathbf{x}-\mathbf{y}\|\leq\ \|\mathbf{x}-\mathbf{z}\|+\|\mathbf{z}-\mathbf{y}\|$,

⋆ $\forall(\mathbf{x},\mathbf{y})\in F^2,\ |\|\mathbf{x}\|-\|\mathbf{y}\||\leq\ \|\mathbf{x}-\mathbf{y}\|$.

Exemples de normes usuelles sur l'espace vectorielle $\mathbb{R}^n$ ***:***

Soit $\mathbf{x}\ =\ {}^t\big(x_1,x_2,...,x_n\big)\in\mathbb{R}^n$. Parmi les normes les plus couramment utilisées, on rappelle les normes vectorielles suivantes :

- ***Norme de Manhattan :*** $\mathbf{x}\implies\|\mathbf{x}\|_1=\ |x_1|+|x_2|+...+|x_n|=\ \sum_{i=1}^n|x_i|$
- ***Norme euclidienne :*** $\mathbf{x}\implies\|\mathbf{x}\|_2=\ \sqrt{|x_1|^2+|x_2|^2+...+|x_n|^2}\ =\ \left(\sum_{i=1}^n|x_i|^2\right)^{1/2}$
- ***Norme "infini" :*** $\mathbf{x}\implies\|\mathbf{x}\|_\infty=\ \max\big(|x_1|+|x_2|+...+|x_n|\big)\ =\ \max_i|x_i|$

Définition 2.6. *Soit F un espace vectoriel réel. Deux normes $\|.\|$ et $\|.\|'$ sur F sont dites* ***équivalentes*** *si et seulement si il existe deux constantes C_1 et C_2 positive telles que pour tout vecteur $\boldsymbol{x}\in F$, on a :*

$$C_1\ \|\mathbf{x}\|\leq\ \|\mathbf{x}\|'\leq\ C_2\ \|\mathbf{x}\|$$

Proposition 2.2. *Soit F un espace vectoriel de dimension finie. Alors, toutes les normes sont équivalentes.*

Définition 2.7. *Étant donné une norme vectorielle $\|.\|$ sur $\mathbb{K}^n$ où $\mathbb{K}=\mathbb{R}$ ou $\mathbb{K}=\mathbb{C}$. On appelle une* ***norme matricielle induite ou subordonnée****, par cette norme vectorielle, sur $M_n(\mathbb{K})$, l'application $\|.\|:M_n(\mathbb{K})\to\mathbb{R}^+$ définie par :*

$$\|A\|=\sup_{\mathbf{x}\in\mathbb{K}^n,\ \mathbf{x}\neq\mathbf{0}}\frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}\ =\sup_{\mathbf{x}\in\mathbb{K}^n,\ \|\mathbf{x}\|\leq1}\|A\mathbf{x}\|=\sup_{x\in\mathbb{K}^n,\ \|\mathbf{x}\|=1}\|A\mathbf{x}\|$$

Elle vérifie de plus les propriétés suivantes :

⋆ $\|A\mathbf{x}\|\leq\ \|A\|\ \|\mathbf{x}\|,\ \ \forall\mathbf{x}\in\mathbb{K}^n$,

⋆ $\|I\|=\ 1$,

2.4 Résolution d'un système triangulaire

Un système triangulaire est un système dont la matrice est triangulaire inférieure ou supérieure.

Définition 2.8. *Une matrice* ***triangulaire inférieure*** *(respectivement* ***triangulaire supérieure****) est une matrice carrée dont les éléments au dessus de la diagonale (respectivement dessous) sont nuls.*

$$\begin{pmatrix} a_{11} & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ a_{21} & a_{22} & 0 & \dots & \dots & \dots & \dots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \dots & \dots & \dots & 0 \\ & & & \dots\dots\dots\dots\dots\dots & & & & \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{ii} & 0 & \dots & 0 \\ & & & \dots\dots\dots\dots\dots\dots & & & & \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \dots & a_{n-1\,i} & \dots & a_{n-1\,n-1} & 0 \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{ni} & \dots & \dots & a_{nn} \end{pmatrix}$$

Allure d'une matrice triangulaire inférieure

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1i} & \dots & a_{1n} \\ 0 & a_{22} & a_{23} & \dots & a_{2i} & \dots & a_{2n} \\ 0 & 0 & a_{33} & \dots & a_{3i} & \dots & a_{3n} \\ & & & \dots\dots\dots\dots\dots & & & \\ 0 & \dots & \dots & 0 & a_{ii} & \dots & a_{in} \\ & & & \dots\dots\dots\dots\dots & & & \\ 0 & \dots & \dots & \dots & \dots & 0 & a_{nn} \end{pmatrix}$$

Allure d'une matrice triangulaire supérieure

2.4.1 Propriétés des matrices triangulaires

- La transposée d'une matrice triangulaire inférieure est triangulaire supérieure et réciproquement,
- Le produit de deux matrices triangulaires inférieures (réciproquement supérieures) est triangulaire inférieure (réciproquement supérieure),
- L'inverse d'une matrice triangulaire inférieure (réciproquement supérieure) est triangulaire inférieure (réciproquement supérieure).
- Le déterminant d'une matrice triangulaire est égal au produit de ses éléments diagonaux.

Soit $A \in M_n(\mathbb{R})$ qu'on supposera inversible et $\mathbf{b} \in \mathbb{R}^n$. On cherche à calculer $\mathbf{x} \in \mathbb{R}^n$ tel que $A\mathbf{x} = \mathbf{b}$, autrement dit on cherche à résoudre le système linéaire

$$A\,\mathbf{x} \;=\; \mathbf{b}.$$

Afin de fixer les idées, résolvons le système linéaire $A\mathbf{x} = \mathbf{b}$ où A est une matrice triangulaire supérieure, i.e :

$$a_{ij} \;=\; 0, \quad \text{pour } i > j.$$

De plus, comme A est une matrice inversible, on a nécessairement :

$$a_{ii} \;\neq\; 0, \quad \forall i.$$

Un système triangulaire supérieur se résout par substitution arrière :

$$\left\{\begin{array}{rrrrrrrrrrrrl} a_{11}\,x_1 & + a_{12}\,x_2 & + a_{13}\,x_3 & + \ldots & + a_{1i}\,x_i & + \ldots & + a_{1n}\,x_n & = & b_1, \\ & a_{22}\,x_2 & + a_{23}\,x_3 & + \ldots & + a_{2i}\,x_i & + \ldots & + a_{2n}\,x_n & = & b_2, \\ & & a_{33}\,x_3 & + \ldots & + a_{3i}\,x_i & + \ldots & + a_{3n}\,x_n & = & b_3, \\ & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & & \\ & & & & a_{ii}\,x_i & + \ldots & + a_{in}\,x_n & = & b_i, \\ & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & & \\ & & & & & & a_{nn}\,x_n & = & b_n. \end{array}\right.$$

La méthode de résolution est immédiate. On commence par calculer x_n dans la dernière équation :

$$x_n \;=\; \frac{b_n}{a_{nn}},$$

puis on peut calculer x_{n-1} à l'aide de x_n :

$$a_{n-1\,n-1}\,x_{n-1} \;+\; a_{n-1\,n}\,x_n \;=\; b_{n-1} \;\Longrightarrow\; x_{n-1} \;=\; \frac{b_{n-1} \;-\; a_{n-1\,n}\,x_n}{a_{n-1\,n-1}},$$

et ainsi de suite jusqu'à x_1. On obtient donc les valeurs des inconnues dans l'ordre décroissant des indices, ce qui donne l'algorithme suivant :

$$\boxed{\left\{\begin{array}{rcl} x_n & = & \dfrac{b_n}{a_{nn}}, \\ x_i & = & \dfrac{1}{a_{ii}}\left(b_i - \Sigma_{j=i+1}^{n} a_{ij}\,x_j\right), \;\; \forall\, i = n-1,\; n-2, \ldots,\; 1. \end{array}\right.}$$

Dans le cas où la matrice A est triangulaire inférieure, on obtient donc les valeurs dans l'ordre croissant des indices en effectuant des substitutions directes et l'algorithme est :

$$\boxed{\left\{\begin{array}{rcl} x_1 & = & \dfrac{b_1}{a_{11}}, \\ x_i & = & \dfrac{1}{a_{ii}}\left(b_i - \Sigma_{j=1}^{i-1} a_{ij}\,x_j\right), \;\; \forall\, i = 2,\; 3, \ldots,\; n. \end{array}\right.}$$

2.5 Résolution d'un système à matrice diagonale

Soit à résoudre le système $A\mathbf{x} = \mathbf{b}$ où $A \in M_n(\mathbb{R})$ une matrice diagonale et $\mathbf{b} \in \mathbb{R}^n$. On a

$$a_{ij} \;=\; 0, \quad \text{pour } i \neq j.$$

La solution $\mathbf{x} \in \mathbb{R}^n$ s'obtient en appliquant l'algorithme suivant :

$$\boxed{x_i \;=\; \frac{b_i}{a_{ii}}, \;\; \forall i = 1, 2, ..., n}$$

2.6 Système échelonné réduit

Définition 2.9. *Une matrice A est dite échelonnée si elle vérifie les propriétés suivantes :*

1 • Toutes les lignes non nulles de A se situent en dessus de ses lignes nulles.

2 • Le premier coefficient non nul d'une ligne se trouve à droite du premier coefficient non nul de la ligne précédente.

3 • Tous les coefficients de la colonne en dessous du premier coefficient non nul, sont nuls.

Autrement dit, soit dans la ligne i, a_{ij} le premier coefficient non nul, on a alors :

$$\begin{cases} a_{kj} = 0, & \text{pour } k > i, \\ a_{il} = 0, & \text{pour } l < j. \end{cases}$$

Exemple : Soit la matrice échelonnée A

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

Définition 2.10. *Une matrice A est dite échelonnée réduite si elle vérifie les propriétés suivantes :*

1 • La matrice A est échelonnée.

2 • Le premier coefficient non nul d'une ligne vaut 1.

3 • Le premier coefficient non nul d'une ligne est le seul élément non nul de sa colonne.

Exemple : Soit la matrice échelonnée réduite A

$$A = \begin{pmatrix} 1 & 0 & 2 & 0 & 4 \\ 0 & 1 & 3 & 0 & 5 \\ 0 & 0 & 0 & 1 & 6 \end{pmatrix}$$

2.7 Terminologie

Définition 2.11. *Soit A une matrice échelonnée :*

• *La colonne qui contient le premier coefficient non nul d'une ligne est dite colonne pivot.*

Exemple : *Revenons à l'exemple 2, les colonnes pivots sont les cinq premières colonnes.*

• *Les inconnues correspondant à une colonne pivot sont appelées inconnues ou variables essentielles. Les autres sont appelées inconnues ou variables libres.*

• *Une position pivot dans une matrice échelonnée réduite A, sont les emplacements des premiers coefficients non nuls d'une ligne, valant 1. Une colonne de A contient une position pivot est dite une colonne pivot.*

• *On dit que deux systèmes correspondant à deux différentes matrices augmentées sont équivalents s'ils ont le même ensemble de solutions.*

2.8 Résolution par la méthode de Gauss

Soit $A \in M_n(\mathbb{R})$ une matrice inversible. On cherche à calculer $\mathbf{x} \in \mathbb{R}^n$ tel que $A\mathbf{x} = \mathbf{b}$. On évite d'appliquer la méthode de Cramer, qui consiste à cherche la matrice A^{-1} et de la multiplier par le vecteur $\mathbf{b}$, pusique'elle reste coûteuse en terme de temps de calcul, de nombres d'opérations effectuées et de précision des résultats obtenus surtout pour les systèmes de grande taille ou pleins. Ainsi on préfère appliquer la méthode de Gauss dont le principe consiste à se ramener, en effectuant des combinaisons linéaires sur les lignes, à un système triangulaire équivalent

$$\tilde{A}\,\mathbf{x} \;=\; \tilde{\mathbf{b}},$$

où la matrice $\tilde{A}$ est une matrice triangulaire supérieure, issue de A, facile à résoudre comme on vient de le voir précédemment. Le principe de base est de rechercher une matrice régulière P, dite ***matrice de permutation***, déterminée à partir du produit de matrices élémentaires de permutation et telle que le produit $P.A$ soit une matrice triangulaire.

Elle comporte n étapes de transformation. La matrice $\tilde{A}$ recherchée correspond à la matrice $A^{(n)}$, où $A^{(n)}$ est l'état de la matrice A transformée à la n^{ime} étape, i.e :

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1j} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2j} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3j} & \dots & a_{3n} \\ & & \dots\dots\dots\dots & & & & \\ a_{i1} & a_{i2} & a_{i3} & \dots & a_{ii} & \dots & a_{in} \\ & & \dots\dots\dots\dots & & & & \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nj} & \dots & a_{nn} \end{pmatrix} \rightarrow \begin{pmatrix} \tilde{a}_{11} & \tilde{a}_{12} & \tilde{a}_{13} & \dots & \tilde{a}_{1j} & \dots & \tilde{a}_{1n} \\ 0 & \tilde{a}_{22} & \tilde{a}_{23} & \dots & \tilde{a}_{2j} & \dots & \tilde{a}_{2n} \\ 0 & 0 & \tilde{a}_{33} & \dots & \tilde{a}_{3j} & \dots & \tilde{a}_{3n} \\ & & \dots\dots\dots\dots & & & & \\ 0 & \dots & \dots & 0 & \tilde{a}_{ii} & \dots & \tilde{a}_{in} \\ & & \dots\dots\dots\dots & & & & \\ 0 & \dots & \dots & \dots & \dots & 0 & \tilde{a}_{nn} \end{pmatrix}$$

On obtient donc la solution par simple remontée.

La méthode de Gauss est fondée sur les notions suivantes. On appelle opérations élémentaires sur les lignes les trois opérations suivantes :

1 • Échanger deux lignes.

2 • Multiplier tous les éléments d'une ligne par une constante non nulle.

3 • Remplacer une ligne par la somme d'elle même et le multiple d'une autre ligne.

On cherche à résoudre le système suivant :

$$\left\{\begin{array}{lllllllll} a_{11}\,x_1 & +a_{12}\,x_2 & +a_{13}\,x_3 & +\ldots & +a_{1j}\,x_j & +\ldots & +a_{1n}\,x_n & = & b_1,\\ a_{21}\,x_1 & +a_{22}\,x_2 & +a_{23}\,x_3 & +\ldots & +a_{2j}\,x_j & +\ldots & +a_{2n}\,x_n & = & b_2,\\ a_{31}\,x_1 & +a_{32}\,x_2 & +a_{33}\,x_3 & +\ldots & +a_{3j}\,x_j & +\ldots & +a_{3n}\,x_n & = & b_3,\\ & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & & \\ a_{i1}\,x_1 & +a_{i2}\,x_2 & +a_{i3}\,x_3 & +\ldots & +a_{ij}\,x_j & +\ldots & +a_{in}\,x_n & = & b_i,\\ & & \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots & & & & & & \\ a_{n1}\,x_1 & +a_{n2}\,x_2 & +a_{n3}\,x_3 & +\ldots & +a_{nj}\,x_j & +\ldots & +a_{nn}\,x_n & = & b_n. \end{array}\right.$$

La méthode de Gauss consiste à éliminer successivement en remontant des inconnues entre les équations.

Étape 1 **:** On part de

$$A^{(1)} = A = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \ldots & a_{1j}^{(1)} & \ldots & a_{1n}^{(1)}\\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \ldots & a_{2j}^{(1)} & \ldots & a_{2n}^{(1)}\\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \ldots & a_{3j}^{(1)} & \ldots & a_{3n}^{(1)}\\ & & \ldots\ldots\ldots\ldots & & & & \\ a_{i1}^{(1)} & a_{i2}^{(1)} & a_{i3}^{(1)} & \ldots & a_{ij}^{(1)} & \ldots & a_{in}^{(1)}\\ & & \ldots\ldots\ldots\ldots & & & & \\ a_{n1}^{(1)} & a_{n2}^{(1)} & a_{n3}^{(1)} & \ldots & a_{nj}^{(1)} & \ldots & a_{nn}^{(1)} \end{pmatrix} \quad \text{et} \quad b^{(1)} = b = \begin{pmatrix} b_1^{(1)}\\ b_2^{(1)}\\ b_3^{(1)}\\ \vdots\\ b_i^{(1)}\\ \vdots\\ b_n^{(1)} \end{pmatrix}$$

La première étape consiste à éliminer l'inconnue x_1. Pour cela, on commence par inspecter la première colonne (correspondant aux coefficients de l'inconnue x_1 dans le système linéaire) et on choisit le terme non nul de la colonne qu'on appelle pivot. Deux cas se présentent :

1^{er} ***cas :*** Si $a_{11}^{(1)}$ est le pivot, on ne touche plus à la ligne 1 et on se sert du pivot pour mettre donc à zéro tous les éléments de la colonne 1 sous le pivot $a_{11}^{(1)}$ en remplaçant la ligne i, $i = 2,\ 3,\ \ldots, n$ par

$$\boxed{L_i^{(2)} \leftarrow L_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}\,.\,L_1^{(1)}, \qquad i = 2,\ 3,\ \ldots, n.}$$

Cela donne les formules suivantes :

$$\boxed{\left\{\begin{array}{lll} a_{ij}^{(1)} = a_{ij}, & \rightarrow a_{ij}^{(2)} = a_{ij}^{(1)} - \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}\,.\,a_{1j}^{(1)}, & \\ b_i^{(1)} = b_i, & \rightarrow b_i^{(2)} = b_i^{(1)} - \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}\,.\,b_1^{(1)}, & i = 2,\ 3,\ \ldots, n \text{ et } j = 1,\ 2,\ \ldots, n. \end{array}\right.}$$

On obtient alors un système linéaire sous la forme $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, équivalent à $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$, avec

$$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2j}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & a_{32}^{(2)} & a_{33}^{(2)} & \dots & a_{3j}^{(2)} & \dots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & a_{i2}^{(2)} & a_{i3}^{(2)} & \dots & a_{ij}^{(2)} & \dots & a_{in}^{(2)} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & a_{n2}^{(2)} & a_{n3}^{(2)} & \dots & a_{nj}^{(2)} & \dots & a_{nn}^{(2)} \end{pmatrix} \quad \text{et} \quad \mathbf{b}^{(2)} = \begin{pmatrix} b_1^{(2)} \\ b_2^{(2)} \\ b_3^{(2)} \\ \vdots \\ b_i^{(2)} \\ \vdots \\ b_n^{(2)} \end{pmatrix}$$

$\underline{2^{eme}}$ ***cas :*** Si l'élément $a_{11}^{(1)}$ est nul, veuillez vous référer au paragraphe *«Stratégie de résolution dans le cas des pivots nuls»* pour la représentation de la méthode de résolution.

Étape 2 ***:*** On cherche à éliminer la deuxième inconnue x_2. Pour cela, on commence par le choix du pivot qu'on notera $a_{22}^{(2)}$. On ne touche plus à la ligne 1 qui a servi déjà comme ligne de pivot à l'étape 1, ni à la ligne 2 qui joue le rôle de ligne de pivot $a_{22}^{(2)}$ supposé non nul.

Ensuite on se sert du pivot $a_{22}^{(2)}$ pour annuler tous les éléments de la colonne 2 sous le pivot $a_{22}^{(2)}$ en remplaçant la ligne i, $i = 3,\ 4,\ \dots, n$ par

$$\boxed{L_i^{(3)} \leftarrow L_i^{(2)} - \frac{a_{i2}^{(2)}}{a_{22}^{(2)}} \cdot L_2^{(2)}, \qquad i = 3,\ 4,\ \dots, n.}$$

Cela donne les formules suivantes :

$$\boxed{\begin{cases} a_{ij}^{(2)} \rightarrow a_{ij}^{(3)} = a_{ij}^{(2)} - \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}} \cdot a_{2j}^{(2)}, \\ b_i^{(2)} \rightarrow b_i^{(3)} = b_i^{(2)} - \dfrac{a_{i2}^{(2)}}{a_{22}^{(2)}} \cdot b_2^{(2)}, \quad i = 3,\ 4,\ \dots, n \text{ et } j = 2,\ 3,\ \dots, n. \end{cases}}$$

On obtient alors un système linéaire sous la forme $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$, équivalent à $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, avec

$$A^{(3)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \dots & a_{2j}^{(2)} & \dots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \dots & a_{3j}^{(3)} & \dots & a_{3n}^{(3)} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & 0 & a_{i3}^{(3)} & \dots & a_{ij}^{(3)} & \dots & a_{in}^{(3)} \\ \vdots & \vdots & \vdots & \dots & \vdots & \dots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \dots & a_{nj}^{(3)} & \dots & a_{nn}^{(3)} \end{pmatrix} \quad \text{et} \quad \mathbf{b}^{(3)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(3)} \\ \vdots \\ b_i^{(3)} \\ \vdots \\ b_n^{(3)} \end{pmatrix}$$

Étape k : En supposant qu'on a effectué jusqu'à présent $(k-1)$ étapes de l'élimination de Gauss, on a obtenu donc à l'étape $(k-1)$ de l'algorithme, un système linéaire $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$ dont les $(k-1)$ premières colonnes sont triangulaires supérieures.

Ainsi en supposant le pivot $a_{kk}^{(k)}$ non nul, on annule tous les éléments sous le pivot $a_{kk}^{(k)}$ en remplaçant la ligne i, $i = k+1,\ k+2,\ \ldots, n$ par

$$\boxed{L_i^{(k+1)} \leftarrow L_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} . L_k^{(k)}, \qquad i = k+1,\ k+2,\ \ldots, n.}$$

Cela donne les formules suivantes :

$$\boxed{\left\{\begin{array}{lll} a_{ij}^{(k)} & \rightarrow a_{ij}^{(k+1)} = a_{ij}^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} . a_{kj}^{(k)}, & \\ b_i^{(k)} & \rightarrow b_i^{(k+1)} = b_i^{(k)} - \dfrac{a_{ik}^{(k)}}{a_{kk}^{(k)}} . b_k^{(k)}, & i = k+1,\ k+2,\ \ldots, n \text{ et } j = 2,\ 3,\ \ldots, n. \end{array}\right.}$$

On obtient alors un système linéaire sous la forme $A^{(k+1)}\mathbf{x} = \mathbf{b}^{(k+1)}$, équivalent à $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$, avec

$$A^{(k+1)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \ldots & a_{1k}^{(1)} & a_{1\,k+1}^{(1)} & \ldots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & \ldots & a_{2k}^{(2)} & a_{2\,k+1}^{(2)} & \ldots & a_{2n}^{(2)} \\ \vdots & \ddots & \ddots & \vdots & \vdots & \vdots & \vdots \\ \vdots & & \ddots & a_{kk}^{(k)} & a_{k\,k+1}^{(k)} & \ldots & a_{kn}^{(k)} \\ \vdots & & & 0 & a_{k+1\,k+1}^{(k+1)} & \ldots & a_{k+1\,n}^{(k+1)} \\ \vdots & & & \vdots & \vdots & \ldots & \vdots \\ 0 & \ldots & \ldots & 0 & a_{n\,k+1}^{(k+1)} & \ldots & a_{nn}^{(k+1)} \end{pmatrix} \quad \text{et} \quad \mathbf{b}^{(k+1)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(2)} \\ b_3^{(3)} \\ \vdots \\ b_k^{(k)} \\ \vdots \\ b_n^{(3)} \end{pmatrix}$$

Ainsi, l'élimination de Gauss transforme la matrice A en une matrice triangulaire supérieure $\tilde{A}$. Le système obtenu après $(n-1)$ étapes, est ensuite résolu par substitution inverse ou remontée en commençant par la dernière équation.

Exemple : On cherche à résoudre le système suivant :

$$\left\{\begin{array}{rcrcrcrcl} 5 & x_1 & - & 4 & x_2 & + & & x_3 & = 2, \\ -4 & x_1 & + & 4 & x_2 & & & & = 0, \\ & x_1 & & & & + & 2 & x_3 & = 3. \end{array}\right.$$

Pour la mise en œuvre de la méthode de Gauss, nous procédons tout d'abord à la triangularisation du système linéaire $A\mathbf{x} = \mathbf{b}$ en le transformant en un système triangulaire supérieur $\tilde{A}\mathbf{x} = \tilde{\mathbf{b}}$. Elle comporte deux étapes.

Le système linéaire s'écrit sous forme matricielle $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\left(\begin{array}{ccc} \boxed{5} & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{array}\right)\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 2 \\ 0 \\ 3 \end{array}\right)$$

Étape 1 : On choisit $a_{11}^{(1)} = 5 \neq 0$ pivot de la première étape. La ligne $L_1^{(1)}$, servant de ligne de pivot, reste inchangée. On remplace la ligne i, $i = 2,\ 3$ par

$$L_i^{(2)} \leftarrow L_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)}, \qquad i = 2,\ 3,$$

i.e :

pour $i = 2$: $\quad L_2^{(2)} \leftarrow L_2^{(1)} - \dfrac{a_{21}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_2^{(1)} - \dfrac{(-4)}{5} \cdot L_1^{(1)} = L_2^{(1)} + \dfrac{4}{5} \cdot L_1^{(1)}.$

Les éléments $a_{2j}^{(2)}$ et $b_2^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{21}^{(2)} \leftarrow a_{21}^{(1)} + \dfrac{4}{5} \cdot a_{11}^{(1)} = -4 + \dfrac{4}{5} \cdot 5 = 0, \\ \bullet\ a_{22}^{(2)} \leftarrow a_{22}^{(1)} + \dfrac{4}{5} \cdot a_{12}^{(1)} = 4 + \dfrac{4}{5} \cdot (-4) = \dfrac{4}{5}, \\ \bullet\ a_{23}^{(2)} \leftarrow a_{23}^{(1)} + \dfrac{4}{5} \cdot a_{13}^{(1)} = 0 + \dfrac{4}{5} \cdot 1 = \dfrac{4}{5}, \\ \bullet\ b_2^{(2)} \leftarrow b_2^{(1)} + \dfrac{4}{5} \cdot b_1^{(1)} = 4 + \dfrac{4}{5} \cdot 2 = \dfrac{8}{5}. \end{cases}$$

pour $i = 3$: $\quad L_3^{(2)} \leftarrow L_3^{(1)} - \dfrac{a_{31}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_3^{(1)} - \dfrac{1}{5} \cdot L_1^{(1)} = L_3^{(1)} - \dfrac{4}{5} \cdot L_1^{(1)}.$

Les éléments $a_{3j}^{(2)}$ et $b_3^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{31}^{(2)} \leftarrow a_{31}^{(1)} - \dfrac{1}{5} \cdot a_{11}^{(1)} = 1 - \dfrac{1}{5} \cdot 5 = 0, \\ \bullet\ a_{32}^{(2)} \leftarrow a_{32}^{(1)} - \dfrac{1}{5} \cdot a_{12}^{(1)} = 0 - \dfrac{1}{5} \cdot (-4) = \dfrac{4}{5}, \\ \bullet\ a_{33}^{(2)} \leftarrow a_{33}^{(1)} - \dfrac{1}{5} \cdot a_{13}^{(1)} = 2 - \dfrac{1}{5} \cdot 1 = \dfrac{9}{5}, \\ \bullet\ b_3^{(2)} \leftarrow b_3^{(1)} - \dfrac{1}{5} \cdot b_1^{(1)} = 3 - \dfrac{1}{5} \cdot 2 = \dfrac{13}{5}. \end{cases}$$

On obtient alors le système linéaire $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, équivalent au système $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\left(\begin{array}{ccc} 5 & -4 & 1 \\ 0 & \boxed{\frac{4}{5}} & \frac{4}{5} \\ 0 & \frac{4}{5} & \frac{9}{5} \end{array}\right)\left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right) = \left(\begin{array}{c} 2 \\ \frac{8}{5} \\ \frac{13}{5} \end{array}\right)$$

Étape 2 : On choisit $a_{22}^{(2)} = \dfrac{4}{5} \neq 0$ pivot de la deuxième étape. On ne touche plus à la ligne $L_1^{(2)}$ qui a servi déjà comme ligne de pivot à l'étape 1, ni à la ligne $L_2^{(2)}$ servant de ligne de pivot de la deuxième étape. On remplace la ligne $L_3^{(2)}$ par

$$L_3^{(3)} \leftarrow L_3^{(2)} - \frac{a_{32}^{(2)}}{a_{22}^{(2)}} \cdot L_2^{(2)} = L_3^{(2)} - \frac{4/5}{4/5} \cdot L_2^{(2)} = L_3^{(2)} - L_2^{(2)}.$$

Les éléments $a_{3j}^{(3)}$ et $b_3^{(3)}$ deviennent :

$$\begin{cases} \bullet\ a_{31}^{(3)} \leftarrow a_{31}^{(2)} - a_{21}^{(2)} = 0 - 0 = 0, \\ \bullet\ a_{32}^{(3)} \leftarrow a_{32}^{(2)} - a_{22}^{(2)} = \dfrac{4}{5} - \dfrac{4}{5} = 0, \\ \bullet\ a_{33}^{(3)} \leftarrow a_{33}^{(2)} - a_{23}^{(2)} = \dfrac{9}{5} - \dfrac{4}{5} = 1, \\ \bullet\ b_3^{(3)} \leftarrow b_3^{(2)} - b_2^{(2)} = \dfrac{13}{5} - \dfrac{8}{5} = 1. \end{cases}$$

Ce qui donne le système linéaire $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$, équivalent au système $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$:

$$\begin{pmatrix} 5 & -4 & 1 \\ 0 & \frac{4}{5} & \frac{4}{5} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{8}{5} \\ 1 \end{pmatrix}$$

La matrice $A^{(3)}$ obtenue est, maintenant triangulaire supérieure, et on résout le système par remontée

$$\begin{cases} x_3 = 1, \\ \frac{4}{5} x_2 + \frac{4}{5} x_3 = \frac{8}{5} \iff x_2 = 1, \\ 5x_1 - 4x_2 + x_3 = 2 \iff x_1 = 1. \end{cases}$$

On trouve comme solution du système linéaire

$$S = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

2.8.1 Coût de la méthode d'élimination de Gauss

Afin d'effectuer l'élimination de de Gauss, la méthode requiert environ $\dfrac{2n^3}{3}$ opérations élémentaires, lorsque n est suffisamment grand.

En effet, il nous faut environ $\dfrac{4n3 + 2n^2 - 7n}{6}$ opérations élémentaires dont $\dfrac{n^2 - n}{2}$ additions et $\dfrac{n^3 - n}{3}$ multiplications et additions, auxquelles il fait ajouter n^2 opérations nécessaires pour la résolution par remontée du système triangulaire. Lorsque n est grand, les termes en n et n^2 sont négligeables devant n^3 et le nombre d'opérations est donc de l'ordre de $\dfrac{2n^3}{3}$.

2.8.2 Stratégie de résolution dans le cas des pivots nuls

Si le pivot de la colonne j à l'étape j est nul, on le remplace par un autre élément non nul parmi les éléments, de sa colonne ou des colonnes à droite en dessous.

En effet, on cherche un élément $a_{ij}^{(j)}$ sous le pivot dans la colonne j telle que $a_{ij}^{(j)} \neq 0$, $i = j+1, j+2, ..., \ \ldots, n$, puis on permute les lignes $L_j^{(j)}$ et $L_i^{(j)}$.

On obtient alors un système linéaire $\hat{A}^{(j)}\mathbf{x} = \hat{\mathbf{b}}^{(j)}$ avec $\hat{A}^{(j)} = P^{(j)}\ A^{(j)}$ et $\hat{\mathbf{b}}^{(j)} = P^{(j)}\mathbf{b}^{(j)}$, $P^{(j)}$ étant la matrice de permutation des lignes $L_j^{(j)}$ et $L_i^{(j)}$. Le système linéaire $\hat{A}^{(j)}\mathbf{x} = \hat{\mathbf{b}}^{(j)}$ est équivalent au système $A^{(j)}\mathbf{x} = \mathbf{b}^{(j)}$.

Par abus de simplification, on notera les éléments de la matrice $\hat{A}^{(j)}$ par $a_{ij}^{(j)}$ et ceux du vecteur $\hat{\mathbf{b}}^{(j)}$ par $b_i^{(j)}$, $i = j+1,\ j+2,\ \ldots, m$ et $j = 1,\ 2,\ \ldots, n$, auxquels on applique les mêmes transformations que ceux dans le cas de la résolution par la méthode de Gauss.

Exemple : On cherche à résoudre le système suivant :

$$\left\{\begin{array}{rcrcrcrcrcl} 2 & x_1 & + & 2 & x_2 & + & 3 & x_3 & + & x_4 & = 8, \\ 2 & x_1 & + & 2 & x_2 & - & & x_3 & - & x_4 & = 2, \\ - & x_1 & + & & x_2 & - & 2 & x_3 & + & 3\ x_4 & = 1, \\ 4 & x_1 & - & & x_2 & + & & x_3 & - & x_4 & = 3. \end{array}\right.$$

Pour la mise en œuvre de la méthode de Gauss, nous procédons tout d'abord à la triangularisation du système linéaire $A\ x\ =\ b$ en le transformant en un système triangulaire supérieur $\tilde{A}\mathbf{x} = \tilde{\mathbf{b}}$. Elle comporte trois étapes.

Le système linéaire s'écrit sous forme matricielle $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\begin{pmatrix} \boxed{2} & 2 & 3 & 1 \\ 2 & 2 & -1 & -1 \\ -1 & 1 & -2 & 3 \\ 4 & -1 & 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 2 \\ 1 \\ 3 \end{pmatrix}$$

Étape 1 : On choisit $a_{11}^{(1)} = 2 \neq 0$ pivot de la première étape. La ligne $L_1^{(1)}$, servant de ligne de pivot, reste inchangée. On remplace la ligne i, $i = 2,\ 3,\ 4$ par

$$L_i^{(2)} \leftarrow L_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} . L_1^{(1)}, \qquad i = 2,\ 3,\ 4,$$

i.e :

pour $i = 2$: $\quad L_2^{(2)} \leftarrow L_2^{(1)} - \dfrac{a_{21}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_2^{(1)} - L_1^{(1)}$.

Les éléments $a_{2j}^{(2)}$ et $b_2^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{21}^{(2)} \leftarrow a_{21}^{(1)} - a_{11}^{(1)} = 2 - 2 = 0, \\ \bullet\ a_{22}^{(2)} \leftarrow a_{22}^{(1)} - a_{12}^{(1)} = 2 - 2 = 0, \\ \bullet\ a_{23}^{(2)} \leftarrow a_{23}^{(1)} - a_{13}^{(1)} = -1 - 3 = -4, \\ \bullet\ a_{24}^{(2)} \leftarrow a_{24}^{(1)} - a_{14}^{(1)} = -1 - 1 = -2, \\ \bullet\ b_2^{(2)} \leftarrow b_2^{(1)} - b_1^{(1)} = 2 - 8 = -6. \end{cases}$$

pour $i = 3$: $\quad L_3^{(2)} \leftarrow L_3^{(1)} - \dfrac{a_{31}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_3^{(1)} + \dfrac{1}{2} \cdot L_1^{(1)}$.

Les éléments $a_{3j}^{(2)}$ et $b_3^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{31}^{(2)} \leftarrow a_{31}^{(1)} + \dfrac{1}{2} \cdot a_{11}^{(1)} = -1 + \dfrac{1}{2} \cdot 2 = 0, \\ \bullet\ a_{32}^{(2)} \leftarrow a_{32}^{(1)} + \dfrac{1}{2} \cdot a_{12}^{(1)} = 1 + \dfrac{1}{2} \cdot 2 = 2, \\ \bullet\ a_{33}^{(2)} \leftarrow a_{33}^{(1)} + \dfrac{1}{2} \cdot a_{13}^{(1)} = -2 + \dfrac{1}{2} \cdot 3 = -\dfrac{1}{2}, \\ \bullet\ a_{34}^{(2)} \leftarrow a_{34}^{(1)} + \dfrac{1}{2} \cdot a_{14}^{(1)} = 3 + \dfrac{1}{2} \cdot 1 = \dfrac{7}{2}, \\ \bullet\ b_3^{(2)} \leftarrow b_3^{(1)} + \dfrac{1}{2} \cdot b_1^{(1)} = 1 + \dfrac{1}{2} \cdot 8 = 5. \end{cases}$$

pour $i = 4$: $\quad L_4^{(2)} \leftarrow L_4^{(1)} - \dfrac{a_{41}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_4^{(1)} - 2 \cdot L_1^{(1)}$.

Les éléments $a_{4j}^{(2)}$ et $b_4^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(2)} \leftarrow a_{41}^{(1)} - 2 \cdot a_{11}^{(1)} = 4 - 2 \cdot 2 = 0, \\ \bullet\ a_{42}^{(2)} \leftarrow a_{42}^{(1)} - 2 \cdot a_{12}^{(1)} = -1 - 2 \cdot 2 = -5, \\ \bullet\ a_{43}^{(2)} \leftarrow a_{43}^{(1)} - 2 \cdot a_{13}^{(1)} = 1 - 2 \cdot 3 = -5, \\ \bullet\ a_{44}^{(2)} \leftarrow a_{44}^{(1)} - 2 \cdot a_{14}^{(1)} = -1 - 2 \cdot 1 = -3, \\ \bullet\ b_4^{(2)} \leftarrow b_4^{(1)} - 2 \cdot b_1^{(1)} = 3 - 2 \cdot 8 = -13. \end{cases}$$

On obtient alors le système linéaire $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, équivalent au système $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\begin{pmatrix} 2 & 2 & 3 & 1 \\ 0 & \boxed{0} & -4 & -2 \\ 0 & 2 & -\frac{1}{2} & \frac{7}{2} \\ 0 & -5 & -5 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ -6 \\ 5 \\ -13 \end{pmatrix}$$

Étape 2 : *Maintenant le pivot $a_{22}^{(2)}$ est nul. On permute les lignes $L_2^{(2)}$ et $L_3^{(2)}$*, on obtient le système linéaire :

$$\begin{pmatrix} 2 & 2 & 3 & 1 \\ 0 & \boxed{2} & -\frac{1}{2} & \frac{7}{2} \\ 0 & 0 & -4 & -2 \\ 0 & -5 & -5 & -3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \\ -6 \\ -13 \end{pmatrix}$$

On garde toujours la même notation. Le pivot est maintenant $a_{22}^{(2)} = 2 \neq 0$. On ne touche plus à la ligne $L_1^{(2)}$ qui a servi déjà comme ligne de pivot à l'étape 1, ni à la ligne $L_2^{(2)}$ servant de ligne de pivot de la deuxième étape. On remplace la ligne $L_4^{(2)}$ par

$$L_4^{(3)} \leftarrow L_4^{(2)} - \frac{a_{42}^{(2)}}{a_{22}^{(2)}} \cdot L_2^{(2)} = L_4^{(2)} + \frac{5}{2} \cdot L_2^{(2)}.$$

Les éléments $a_{4j}^{(3)}$ et $b_4^{(3)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(3)} \leftarrow a_{41}^{(2)} - \frac{5}{2} \cdot a_{21}^{(2)} = 0 - \frac{5}{2} \cdot 0 = 0, \\ \bullet\ a_{42}^{(3)} \leftarrow a_{42}^{(2)} - \frac{5}{2} \cdot a_{22}^{(2)} = -5 - \frac{5}{2} \cdot 2 = 0, \\ \bullet\ a_{43}^{(3)} \leftarrow a_{43}^{(2)} - \frac{5}{2} \cdot a_{23}^{(2)} = -5 - \frac{5}{2} \cdot -\frac{1}{2} = -\frac{25}{4}, \\ \bullet\ a_{44}^{(3)} \leftarrow a_{44}^{(2)} - \frac{5}{2} \cdot a_{24}^{(2)} = -3 - \frac{5}{2} \cdot \frac{7}{2} = \frac{23}{4}, \\ \bullet\ b_4^{(3)} \leftarrow b_4^{(2)} - \frac{5}{2} \cdot b_2^{(2)} = -13 - \frac{5}{2} \cdot 5 = -\frac{1}{2}. \end{cases}$$

Ce qui donne le système linéaire $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$, équivalent au système $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$:

$$\begin{pmatrix} 2 & 2 & 3 & 1 \\ 0 & 2 & -\frac{1}{2} & \frac{7}{2} \\ 0 & 0 & \boxed{-4} & -2 \\ 0 & 0 & -\frac{25}{4} & \frac{23}{4} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \\ -6 \\ -\frac{1}{2} \end{pmatrix}$$

Étape 3 : On choisit $a_{33}^{(3)} = -4 \neq 0$ pivot de la troisième étape. La ligne $L_3^{(3)}$, servant de ligne de pivot, reste inchangée. On remplace la ligne $L_4^{(3)}$ par

$$L_4^{(4)} \leftarrow L_4^{(3)} - \frac{a_{43}^{(13)}}{a_{33}^{(3)}} \cdot L_3^{(3)} = L_4^{(3)} - \frac{25}{16} \cdot L_3^{(3)}.$$

Les éléments $a_{4j}^{(4)}$ et $b_4^{(4)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(4)} \leftarrow a_{41}^{(3)} - \frac{25}{16} \cdot a_{31}^{(3)} = 0 - \frac{25}{16} \cdot 0 = 0, \\ \bullet\ a_{42}^{(4)} \leftarrow a_{42}^{(3)} - \frac{25}{16} \cdot a_{32}^{(3)} = 0 - \frac{25}{16} \cdot 0 = 0, \\ \bullet\ a_{43}^{(4)} \leftarrow a_{43}^{(3)} - \frac{25}{16} \cdot a_{33}^{(3)} = -\frac{25}{4} - \frac{25}{16} \cdot (-4) = 0, \\ \bullet\ a_{44}^{(4)} \leftarrow a_{44}^{(3)} - \frac{25}{16} \cdot a_{34}^{(3)} = \frac{23}{4} - \frac{25}{16} \cdot (-2) = \frac{71}{8}, \\ \bullet\ b_4^{(4)} \leftarrow b_4^{(3)} - \frac{25}{16} \cdot b_3^{(3)} = -\frac{1}{2} - \frac{25}{16} \cdot (-6) = \frac{71}{8}. \end{cases}$$

Ce qui donne le système linéaire $A^{(4)}\mathbf{x} = \mathbf{b}^{(4)}$, équivalent au système $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$:

$$\begin{pmatrix} 2 & 2 & 3 & 1 \\ 0 & 2 & -\frac{1}{2} & \frac{7}{2} \\ 0 & 0 & \boxed{-4} & -2 \\ 0 & 0 & 0 & \frac{71}{8} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 8 \\ 5 \\ -6 \\ \frac{71}{8} \end{pmatrix}$$

La matrice $A^{(4)}$ obtenue est, maintenant triangulaire supérieure, et on résout le système par

remontée

$$\left\{\begin{array}{rcrcrcrcrcl} & & & & & & \frac{71}{8} & x_4 & = & \frac{71}{8} & \Longleftrightarrow \quad x_4 = 1, \\ & & & & - & 4 \; x_3 & - & 2 \; x_4 & = & -6 & \Longleftrightarrow \quad x_3 = 1, \\ & & & 2 \; x_2 & - & \frac{1}{2} \; x_3 & + & \frac{7}{2} \; x_4 & = & 5 & \Longleftrightarrow \quad x_2 = 1, \\ 2 \; x_1 & + & 2 \; x_2 & + & 3 \; x_3 & + & & x_4 & = & 8 & \Longleftrightarrow \quad x_1 = 1. \end{array}\right.$$

On trouve comme solution du système linéaire

$$S = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

Exemple : On cherche à résoudre le système suivant :

$$\left\{\begin{array}{rcrcrcrcrcl} 2 \; x_1 & + & & & x_2 & + & & x_3 & + & & x_4 & = & 5, \\ 4 \; x_1 & + & 2 \; x_2 & - & 2 \; x_3 & - & 2 \; x_4 & = & 2, \\ 6 \; x_1 & + & 3 \; x_2 & - & 9 \; x_3 & + & 3 \; x_4 & = & 3, \\ 8 \; x_1 & + & 4 \; x_2 & + & 8 \; x_3 & - & 8 \; x_4 & = & 12. \end{array}\right.$$

Pour la mise en œuvre de la méthode de Gauss, nous procédons tout d'abord à la triangularisation du système linéaire $A\mathbf{x} = \mathbf{b}$ en le transformant en un système triangulaire supérieur $\tilde{A}\mathbf{x} = \tilde{\mathbf{b}}$. Elle comporte trois étapes.

Le système linéaire s'écrit sous forme matricielle $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\begin{pmatrix} \boxed{2} & 2 & 3 & 1 \\ 2 & 1 & 1 & 1 \\ 6 & 3 & -9 & 3 \\ 8 & 4 & 8 & -8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5 \\ 2 \\ 3 \\ 12 \end{pmatrix}$$

Étape 1 : On choisit $a_{11}^{(1)} = 2 \neq 0$ pivot de la première étape. La ligne $L_1^{(1)}$, servant de ligne de pivot, reste inchangée. On remplace la ligne i, $i = 2,\ 3,\ 4$ par

$$L_i^{(2)} \leftarrow L_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)}, \qquad i = 2,\ 3,\ 4,$$

i.e :

pour $i = 2$: $\quad L_2^{(2)} \leftarrow L_2^{(1)} - \dfrac{a_{21}^{(1)}}{a_{11}^{(1)}} . L_1^{(1)} = L_2^{(1)} - 2\, L_1^{(1)}.$

Les éléments $a_{2j}^{(2)}$ et $b_2^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{21}^{(2)} \leftarrow a_{21}^{(1)} - 2\ .a_{11}^{(1)} = 4 - 2\ .\ 2 = 0, \\ \bullet\ a_{22}^{(2)} \leftarrow a_{22}^{(1)} - 2\ .a_{12}^{(1)} = 2 - 2\ .\ 1 = 0, \\ \bullet\ a_{23}^{(2)} \leftarrow a_{23}^{(1)} - 2\ .a_{13}^{(1)} = -2 - 2\ .\ 1 = -4, \\ \bullet\ a_{24}^{(2)} \leftarrow a_{24}^{(1)} - 2\ .a_{14}^{(1)} = -2 - 2\ .\ 1 = -4, \\ \bullet\ b_2^{(2)} \leftarrow b_2^{(1)} - 2\ .b_1^{(1)} = 2 - 2\ .\ 5 = -8. \end{cases}$$

pour $i = 3$: $\quad L_3^{(2)} \leftarrow L_3^{(1)} - \dfrac{a_{31}^{(1)}}{a_{11}^{(1)}} . L_1^{(1)} = L_3^{(1)} - 3\, .\, L_1^{(1)}.$

Les éléments $a_{3j}^{(2)}$ et $b_3^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{31}^{(2)} \leftarrow a_{31}^{(1)} - 3\ .\ a_{11}^{(1)} = 6 - 3\ .\ 2 = 0, \\ \bullet\ a_{32}^{(2)} \leftarrow a_{32}^{(1)} - 3\ .\ a_{12}^{(1)} = 3 - 3\ .\ 1 = 0, \\ \bullet\ a_{33}^{(2)} \leftarrow a_{33}^{(1)} - 3\ .\ a_{13}^{(1)} = -9 - 3\ .\ 1 = -12, \\ \bullet\ a_{34}^{(2)} \leftarrow a_{34}^{(1)} - 3\ .\ a_{14}^{(1)} = 3 - 3\ .\ 1 = 0, \\ \bullet\ b_3^{(2)} \leftarrow b_3^{(1)} - 3\ .\ b_1^{(1)} = 3 - 3\ .\ 5 = -12. \end{cases}$$

pour $i = 4$: $\quad L_4^{(2)} \leftarrow L_4^{(1)} - \dfrac{a_{41}^{(1)}}{a_{11}^{(1)}} . L_1^{(1)} = L_4^{(1)} - 4\, .\, L_1^{(1)}.$

Les éléments $a_{4j}^{(2)}$ et $b_4^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(2)} \leftarrow a_{41}^{(1)} - 4\ .\ a_{11}^{(1)} = 8 - 4\ .\ 2 = 0, \\ \bullet\ a_{42}^{(2)} \leftarrow a_{42}^{(1)} - 4\ .\ a_{12}^{(1)} = 4 - 4\ .\ 1 = 0, \\ \bullet\ a_{43}^{(2)} \leftarrow a_{43}^{(1)} - 4\ .\ a_{13}^{(1)} = 8 - 4\ .\ 1 = 4, \\ \bullet\ a_{44}^{(2)} \leftarrow a_{44}^{(1)} - 4\ .\ a_{14}^{(1)} = -8 - 4\ .\ 1 = -12, \\ \bullet\ b_4^{(2)} \leftarrow b_4^{(1)} - 4\ .\ b_1^{(1)} = 12 - 4\ .\ 5 = -8. \end{cases}$$

On obtient alors le système linéaire $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, équivalent au système $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\begin{pmatrix} 2 & 1 & 1 & 1 \\ 0 & 0 & \boxed{-4} & -4 \\ 0 & 0 & -12 & 0 \\ 0 & 0 & 4 & -12 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5 \\ -8 \\ -12 \\ -8 \end{pmatrix}$$

Étape 2 : *Maintenant le pivot* $a_{22}^{(2)}$ *est nul ainsi que* $a_{32}^{(2)}$ *et* $a_{42}^{(2)}$. On choisit $a_{23}^{(2)} = -4 \neq 0$

comme pivot de la deuxième étape, on obtient le système linéaire : On ne touche plus à la ligne $L_1^{(2)}$ qui a servi déjà comme ligne de pivot à l'étape 1, ni à la ligne $L_2^{(2)}$ servant de ligne de pivot de la deuxième étape. On remplace les lignes $L_i^{(2)}$, $i = 3,\ 4$ par

$$L_i^{(3)} \leftarrow L_i^{(2)} - \frac{a_{i3}^{(2)}}{a_{23}^{(2)}} \,.\, L_2^{(2)}.$$

i.e :

pour $i = 3$: $\quad L_3^{(3)} \leftarrow L_3^{(2)} - \dfrac{a_{33}^{(2)}}{a_{32}^{(2)}} \,.\, L_2^{(2)} = L_3^{(2)} - 3 \,.\, L_2^{(2)}.$

Les éléments $a_{3j}^{(3)}$ et $b_3^{(3)}$ deviennent :

$$\begin{cases} \bullet\ a_{31}^{(3)} \leftarrow a_{31}^{(2)} - 3 \,.\, a_{21}^{(2)} = 0 - 3 \,.\, 0 = 0, \\ \bullet\ a_{32}^{(3)} \leftarrow a_{32}^{(2)} - 3 \,.\, a_{22}^{(2)} = 0 - 3 \,.\, 0 = 0, \\ \bullet\ a_{33}^{(3)} \leftarrow a_{33}^{(2)} - 3 \,.\, a_{23}^{(2)} = -12 - 3 \,.\, (-4) = 0, \\ \bullet\ a_{34}^{(3)} \leftarrow a_{34}^{(2)} - 3 \,.\, a_{24}^{(2)} = 0 - 3 \,.\, (-4) = 12, \\ \bullet\ b_3^{(3)} \leftarrow b_3^{(2)} - 3 \,.\, b_2^{(2)} = -12 - 3 \,.\, (-8) = 12. \end{cases}$$

pour $i = 4$: $\quad L_4^{(3)} \leftarrow L_4^{(2)} - \dfrac{a_{43}^{(2)}}{a_{23}^{(2)}} \,.\, L_2^{(2)} = L_4^{(2)} + L_2^{(2)}.$

Les éléments $a_{4j}^{(3)}$ et $b_4^{(3)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(3)} \leftarrow a_{41}^{(2)} + a_{21}^{(2)} = 0 + 0 = 0, \\ \bullet\ a_{42}^{(3)} \leftarrow a_{42}^{(2)} + a_{22}^{(2)} = 0 + 0 = 0, \\ \bullet\ a_{43}^{(3)} \leftarrow a_{43}^{(2)} + a_{23}^{(2)} = 4 + (-4) = 0, \\ \bullet\ a_{44}^{(3)} \leftarrow a_{44}^{(2)} + a_{24}^{(2)} = -12 + (-4) = -16, \\ \bullet\ b_4^{(3)} \leftarrow b_4^{(2)} + b_2^{(2)} = -8 + (-8) = -16. \end{cases}$$

Ce qui donne le système linéaire $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$, équivalent au système $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$:

$$\begin{pmatrix} 2 & 1 & 1 & 1 \\ 0 & 0 & -4 & -4 \\ 0 & 0 & 0 & \boxed{12} \\ 0 & 0 & 0 & -16 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 5 \\ -8 \\ 12 \\ -16 \end{pmatrix}$$

Ce qui donne :

$$\begin{cases} -16\, x_4 = -16 \iff x_4 = 1, \\ 12\, x_4 = 12 \iff x_4 = 1, \\ -4\, x_3 - 4\, x_4 = -8 \iff x_3 = 1, \\ 2\, x_1 + x_2 + x_3 + x_4 = 5 \iff x_1 = \dfrac{3 - x_2}{2}. \end{cases}$$

Par suite, on conclut que le système possède une infinité de solutions

$$\mathbf{x} = \left\{ \begin{pmatrix} \frac{3 - x_2}{2} \\ x_2 \\ 1 \\ 1 \end{pmatrix} \right\} = \left\{ \begin{pmatrix} \frac{3}{2} \\ 0 \\ 1 \\ 1 \end{pmatrix} \right\} + x_2 \left\{ \begin{pmatrix} -\frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix} \right\}$$

où x_2 est l'inconnue auxiliaire ou libre parcourant indépendamment $\mathbb{R}$. L'ensemble de ces solutions constitue un espace vectoriel de dimension 2 engendré par par les vecteurs $\begin{pmatrix} \frac{3}{2} \\ 0 \\ 1 \\ 1 \end{pmatrix}$ et $\begin{pmatrix} -\frac{1}{2} \\ 1 \\ 0 \\ 0 \end{pmatrix}$. C'est aussi le plan d'équation $2x_1 + x_2 = 3$.

Dans ce cas x_1, x_3 et x_4 sont dites variables essentielles et x_2 variable libre (vu qu'on a pas obtenu un pivot dans la deuxième colonne associée aux coefficients de la variable x_2 dans le système). En d'autres termes, pour toute valeur réelle de x_2, la valeur de x_1 calculée fournit une solution du système. Par suite, l'ensemble des solutions est :

$$S = \left\{ \begin{pmatrix} \frac{3 - x_2}{2} \\ x_2 \\ 1 \\ 1 \end{pmatrix}; \quad x_2 \in \mathbb{R} \right\}.$$

Définition 2.12. *Les inconnues correspondant à une colonne de pivot sont dites inconuues ou vraiables essentielles. Les autres sont dites inconnues ou variables libres.*

Définition 2.13. ***Nombre de solutions*** *Un système possède zéro, une ou une infinité de solutions. Dans le cas d'un système $m \times n$, on a :*

1 • *Soit le système n'admet aucune solution s'il y a une ligne de la forme :*

$$0\,x_1 + 0\,x_2 + 0\,x_3 + \ldots + 0\,x_{n-1} + 0\,x_n = b, \quad avec \ b \neq 0.$$

2 • *Soit il admet une solution unique.*

3 • *Soit il admet une infinité de solutions s'il y a une ligne de la forme :*

$$0\,x_1 + 0\,x_2 + 0\,x_3 + \ldots + 0\,x_{n-1} + 0\,x_n = 0$$

ou s'il y a moins d'équations que d'inconnues, dans ce cas il existe au moins une variable libre. Le système est dit sous-déterminé et possèdera une infinité de solutions que l'on pourra expliciter en fonctions d'inconnues libres.

Exemple : On cherche à résoudre le système suivant :

$$\left\{\begin{array}{rcrcrcrcrcl} & x_1 & + & & x_2 & - & 2\ x_3 & & & = & 7,\\ 2 & x_1 & + & 2 & x_2 & - & 4\ x_3 & + & 6\ x_4 & = & 50,\\ & x_1 & - & & x_2 & - & 2\ x_3 & - & x_4 & = & -1,\\ 3 & x_1 & + & 2 & x_2 & - & 6\ x_3 & + & x_4 & = & 23. \end{array}\right.$$

Le système linéaire s'écrit sous forme matricielle $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\begin{pmatrix} \boxed{1} & 1 & -2 & 0 \\ 2 & 2 & -4 & 6 \\ 1 & -1 & -2 & -1 \\ 3 & 2 & -6 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ 50 \\ -1 \\ 23 \end{pmatrix}$$

Étape 1 : On choisit $a_{11}^{(1)} = 21 \neq 0$ pivot de la première étape. La ligne $L_1^{(1)}$, servant de ligne de pivot, reste inchangée. On remplace la ligne i, $i = 2,\ 3,\ 4$ par

$$L_i^{(2)} \leftarrow L_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)}, \qquad i = 2,\ 3,\ 4,$$

i.e :

pour $i = 2$: $\quad L_2^{(2)} \leftarrow L_2^{(1)} - \dfrac{a_{21}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_2^{(1)} - 2\ L_1^{(1)}.$

Les éléments $a_{2j}^{(2)}$ et $b_2^{(2)}$ deviennent :

$$\left\{\begin{array}{lcrcrcrcrcl} \bullet\ a_{21}^{(2)} & \leftarrow & a_{21}^{(1)} & - & 2 & . & a_{11}^{(1)} & = & 2 & - & 2 . 1 & = & 0,\\ \bullet\ a_{22}^{(2)} & \leftarrow & a_{22}^{(1)} & - & 2 & . & a_{12}^{(1)} & = & 2 & - & 2 . 1 & = & 0,\\ \bullet\ a_{23}^{(2)} & \leftarrow & a_{23}^{(1)} & - & 2 & . & a_{13}^{(1)} & = & -4 & - & 2 . (-2) & = & 0,\\ \bullet\ a_{24}^{(2)} & \leftarrow & a_{24}^{(1)} & - & 2 & . & a_{14}^{(1)} & = & 6 & - & 2 . 0 & = & 6,\\ \bullet\ b_2^{(2)} & \leftarrow & b_2^{(1)} & - & 2 & . & b_1^{(1)} & = & 50 & - & 2 . 7 & = & 36. \end{array}\right.$$

pour $i = 3$: $\quad L_3^{(2)} \leftarrow L_3^{(1)} - \dfrac{a_{31}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_3^{(1)} - L_1^{(1)}.$

Les éléments $a_{3j}^{(2)}$ et $b_3^{(2)}$ deviennent :

$$\left\{\begin{array}{lcrcrcrcrcl} \bullet\ a_{31}^{(2)} & \leftarrow & a_{31}^{(1)} & - & a_{11}^{(1)} & = & 1 & - & 1 & = & 0,\\ \bullet\ a_{32}^{(2)} & \leftarrow & a_{32}^{(1)} & - & a_{12}^{(1)} & = & -1 & - & 1 & = & -2,\\ \bullet\ a_{33}^{(2)} & \leftarrow & a_{33}^{(1)} & - & a_{13}^{(1)} & = & -2 & - & (-2) & = & 0,\\ \bullet\ a_{34}^{(2)} & \leftarrow & a_{34}^{(1)} & - & a_{14}^{(1)} & = & -1 & - & 0 & = & -1,\\ \bullet\ b_3^{(2)} & \leftarrow & b_3^{(1)} & - & b_1^{(1)} & = & -1 & - & 7 & = & -8. \end{array}\right.$$

pour $i = 4$: $\quad L_4^{(2)} \leftarrow L_4^{(1)} - \dfrac{a_{41}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_4^{(1)} - 3 \cdot L_1^{(1)}$.

Les éléments $a_{4j}^{(2)}$ et $b_4^{(2)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(2)} \leftarrow a_{41}^{(1)} - 3 \cdot a_{11}^{(1)} = 3 - 3 \cdot 1 = 0, \\ \bullet\ a_{42}^{(2)} \leftarrow a_{42}^{(1)} - 3 \cdot a_{12}^{(1)} = 2 - 3 \cdot 1 = -1, \\ \bullet\ a_{43}^{(2)} \leftarrow a_{43}^{(1)} - 3 \cdot a_{13}^{(1)} = -6 - 3 \cdot (-2) = 0, \\ \bullet\ a_{44}^{(2)} \leftarrow a_{44}^{(1)} - 3 \cdot a_{14}^{(1)} = 1 - 3 \cdot 1 = 1, \\ \bullet\ b_4^{(2)} \leftarrow b_4^{(1)} - 3 \cdot b_1^{(1)} = 23 - 3 \cdot 7 = 2. \end{cases}$$

On obtient alors le système linéaire $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, équivalent au système $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$:

$$\begin{pmatrix} 1 & 1 & -2 & 0 \\ 0 & \boxed{0} & 0 & 6 \\ 0 & -2 & 0 & -1 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ 36 \\ -8 \\ 2 \end{pmatrix}$$

Étape 2 : *Maintenant le pivot $a_{22}^{(2)}$ est nul. On permute les lignes $L_2^{(2)}$ et $L_3^{(2)}$*, on obtient le système linéaire :

$$\begin{pmatrix} 1 & 1 & -2 & 0 \\ 0 & \boxed{-2} & 0 & -1 \\ 0 & 0 & 0 & 6 \\ 0 & -1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ -8 \\ 36 \\ 2 \end{pmatrix}$$

On garde toujours la même notation. Le pivot est maintenant $a_{22}^{(2)} = -2 \neq 0$. On remplace la ligne $L_4^{(2)}$ par

$$L_4^{(3)} \leftarrow L_4^{(2)} - \frac{a_{42}^{(2)}}{a_{22}^{(2)}} \cdot L_2^{(2)} = L_4^{(2)} - \frac{1}{2} \cdot L_2^{(2)}.$$

Les éléments $a_{4j}^{(3)}$ et $b_4^{(3)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(3)} \leftarrow a_{41}^{(2)} - \dfrac{1}{2} \cdot a_{21}^{(2)} = 0 - \dfrac{1}{2} \cdot 0 = 0, \\ \bullet\ a_{42}^{(3)} \leftarrow a_{42}^{(2)} - \dfrac{1}{2} \cdot a_{22}^{(2)} = -1 - \dfrac{1}{2} \cdot (-2) = 0, \\ \bullet\ a_{43}^{(3)} \leftarrow a_{43}^{(2)} - \dfrac{1}{2} \cdot a_{23}^{(2)} = 0 - \dfrac{1}{2} \cdot 0 = 0, \\ \bullet\ a_{44}^{(3)} \leftarrow a_{44}^{(2)} - \dfrac{1}{2} \cdot a_{24}^{(2)} = 1 - \dfrac{1}{2} \cdot (-1) = \dfrac{3}{2}, \\ \bullet\ b_4^{(3)} \leftarrow b_4^{(2)} - \dfrac{1}{2} \cdot b_2^{(2)} = 2 - \dfrac{1}{2} \cdot (-8) = 6. \end{cases}$$

Ce qui donne le système linéaire $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$, équivalent au système $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$:

$$\begin{pmatrix} 1 & 1 & -2 & 0 \\ 0 & -2 & 0 & -1 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & \frac{3}{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ -8 \\ 36 \\ 6 \end{pmatrix}$$

Étape 3 : On choisit $a_{33}^{(3)} = 6 \neq 0$ pivot de la troisième étape et on remplace la ligne $L_4^{(3)}$ par

$$L_4^{(4)} \leftarrow L_4^{(3)} - \frac{a_{44}^{(13)}}{a_{34}^{(3)}} \cdot L_3^{(3)} = L_4^{(3)} - \frac{1}{4} \cdot L_3^{(3)}.$$

Les éléments $a_{4j}^{(4)}$ et $b_4^{(4)}$ deviennent :

$$\begin{cases} \bullet\ a_{41}^{(4)} \leftarrow a_{41}^{(3)} - \frac{1}{4} \cdot a_{31}^{(3)} = 0 - \frac{1}{4} \cdot 0 = 0, \\ \bullet\ a_{42}^{(4)} \leftarrow a_{42}^{(3)} - \frac{1}{4} \cdot a_{32}^{(3)} = 0 - \frac{1}{4} \cdot 0 = 0, \\ \bullet\ a_{43}^{(4)} \leftarrow a_{43}^{(3)} - \frac{1}{4} \cdot a_{33}^{(3)} = 0 - \frac{1}{4} \cdot 0 = 0, \\ \bullet\ a_{44}^{(4)} \leftarrow a_{44}^{(3)} - \frac{1}{4} \cdot a_{34}^{(3)} = \frac{3}{2} - \frac{1}{4} \cdot 6 = 0, \\ \bullet\ b_4^{(4)} \leftarrow b_4^{(3)} - \frac{1}{4} \cdot b_3^{(3)} = 6 - \frac{1}{4} \cdot 36 = -3. \end{cases}$$

Ce qui donne le système linéaire $A^{(4)}\mathbf{x} = \mathbf{b}^{(4)}$, équivalent au système $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$:

$$\begin{pmatrix} 1 & 1 & -2 & 0 \\ 0 & -2 & 0 & -1 \\ 0 & 0 & 0 & 6 \\ 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 7 \\ -8 \\ 36 \\ -3 \end{pmatrix}$$

équivalent à

$$\begin{cases} 0\,x_1 + 0\,x_2 + 0\,x_3 + x_4 = -3, \\ 6\,x_4 = 36, \\ -2\,x_2 - x_4 = -8, \\ x_1 + x_2 - 2\,x_3 = 7. \end{cases}$$

On remarque d'après la première équation que le système S est un système impossible à résoudre, par suite

$$S = \{\phi\}.$$

Exercice 1 : Trouver les trois réels a, b et c tels que $\forall x \in \mathbb{R}\backslash\{1,2,3\}$,

$$\frac{x^2 + x + 2}{(x-1)(x-2)(x-3)} = \frac{a}{(x-1)} + \frac{b}{(x-2)} + \frac{c}{(x-3)}.$$

Solution : On a

$$\frac{a}{(x-1)}+\frac{b}{(x-2)}+\frac{c}{(x-3)} = \frac{x^2\,(a\,+\,b\,+\,c)\,+\,x\,(-5\,a\,-\,4\,b\,-\,3\,c)\,+\,6\,a\,+\,3\,b\,+\,2\,c}{(x-1)\,(x-2)\,(x-3)},$$

d'où

$$\left\{\begin{array}{rcrcrcl} & a & + & b & + & c & = 1,\\ -5 & a & - 4 & b & - 3 & c & = 1,\\ 6 & a & + 3 & b & + 2 & c & = 2. \end{array}\right.$$

On résout ce système par la méthode de Gauss et on obtient

$$\left\{\begin{array}{rcrcrcl} a & + & b & + & c & = 1,\\ & & b & + 2 & c & = 6,\\ & - 3 & b & - 4 & c & = -4. \end{array}\right. \approx \left\{\begin{array}{rcrcrcl} a & + & b & + & c & = 1,\\ & & b & + 2 & c & = 6,\\ & & & 2 & c & = 14. \end{array}\right. \Longleftrightarrow \left\{\begin{array}{rcl} a & = & 2,\\ b & = & -8,\\ c & = & 7. \end{array}\right.$$

Par suite, on écrit :

$$\frac{x^2\,+\,x\,+\,2}{(x-1)\,(x-2)\,(x-3)} = \frac{2}{(x-1)} + \frac{-8}{(x-2)} + \frac{7}{(x-3)}.$$

Exercice 2 : Résoudre dans $\mathbb{R}$, le système non linéaire suivant :

$$\left\{\begin{array}{rcl} \dfrac{4}{x\,-\,2} - \dfrac{1}{y\,+\,2} & = & 5,\\[2ex] \dfrac{2}{x\,-\,2} + \dfrac{2}{y\,+\,2} & = & 10. \end{array}\right.$$

Solution : On pose

$$\left\{\begin{array}{lll} X = \dfrac{1}{x\,-\,2}, & \text{à condition que} & x \neq 2,\\[2ex] Y = \dfrac{1}{y\,+\,2}, & \text{à condition que} & y \neq -2. \end{array}\right.$$

Le système s'écrit alors sous la forme matricielle suivante :

$$\left\{\begin{array}{rcrcl} 4\;X & - & Y & = & 5,\\ 2\;X & + 2 & Y & = & 10, \end{array}\right.$$

qu'on résout par la méthode de Gauss :

$$\left\{\begin{array}{rcrcl} 4\;X & - & Y & = & 5,\\ 2\;X & + 2 & Y & = & 10, \end{array}\right. \approx \left\{\begin{array}{rcrcl} 4\;X & - & Y & = & 5,\\ & & \dfrac{5}{2}\;Y & = & \dfrac{15}{2}, \end{array}\right.$$

$$\Longleftrightarrow \left\{\begin{array}{lllll} X = 2 & \Longleftrightarrow & x = \dfrac{5}{2} & \neq 2, & \text{acceptable},\\[2ex] Y = 3 & \Longleftrightarrow & y = \dfrac{-5}{3} & \neq -2, & \text{acceptable}. \end{array}\right.$$

On trouve alors :

$$S = \left\{ \begin{pmatrix} 5 \\ \frac{2}{} \\ \frac{-5}{3} \end{pmatrix} \right\}.$$

2.9 Dérivées de la méthode de Gauss : Méthodes directes

Nous présenterons dans ce paragraphe plusieurs variantes de la méthode de Gauss à qui chacune son intérêt propre et ses diverses applications.

2.9.1 Factorisation LU

La méthode de factorisation LU, appelée aussi ***méthode décomposition LU (Lower Upper)***, appliquée à une matrice inversible A consiste à décomposer la matrice en le produit de deux matrices triangulaires l'une inférieure L avec des 1 sur la diagonale et l'autre supérieure U telles que

$$A = LU$$

Ainsi, résoudre le système de départ $A\mathbf{x} = \mathbf{b}$ consiste à résoudre le système équivalent suivant :

$$LU\mathbf{x} = \mathbf{b}$$

Une fois calculée les deux matrices L et U, on résout successivement les deux systèmes triangulaires

$$\begin{cases} L\mathbf{y} = \mathbf{b} & \text{d'inconnue } \mathbf{y} \\ U\mathbf{x} = \mathbf{y} & \text{d'inconnue } \mathbf{x} \end{cases}$$

Définition 2.14. *Le **mineur principal d'ordre k d'une matrice A** est défini comme étant le déterminant des k premières lignes et colonnes de la matrice A.*

Théorème 2.1. *Soit A une matrice inversible d'ordre n dont les mineurs principaux sont non nuls. Alors, il existe une unique matrice L triangulaire inférieure avec des 1 sur la diagonale, et une unique matrice U triangulaire supérieure telles que*

$$A = LU$$

De plus,

$$det(A) = \prod_{i=1}^{n} u_{ii}$$

Résolution par la méthode de factorisation LU

Étape 1 : On part de

$$A^{(1)} = A = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & a_{23}^{(1)} & \dots & a_{2j}^{(1)} & \dots & a_{2n}^{(1)} \\ a_{31}^{(1)} & a_{32}^{(1)} & a_{33}^{(1)} & \dots & a_{3j}^{(1)} & \dots & a_{3n}^{(1)} \\ & & \dots\dots\dots\dots\dots & & & & \\ a_{i1}^{(1)} & a_{i2}^{(1)} & a_{i3}^{(1)} & \dots & a_{ij}^{(1)} & \dots & a_{in}^{(1)} \\ & & \dots\dots\dots\dots\dots & & & & \\ a_{n1}^{(1)} & a_{n2}^{(1)} & a_{n3}^{(1)} & \dots & a_{nj}^{(1)} & \dots & a_{nn}^{(1)} \end{pmatrix} \quad \text{et} \quad b^{(1)} = b = \begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} \\ b_3^{(1)} \\ \vdots \\ b_i^{(1)} \\ \vdots \\ b_n^{(1)} \end{pmatrix}$$

La première étape consiste à éliminer l'inconnue x_1. Pour cela, on commence la triangularisation par inspecter la première colonne (correspondant aux coefficients de l'inconnue x_1 dans le système linéaire) et en supposant le premier pivot $a_{11}^{(1)}$ non nul, on introduit la matrice $P^{(1)}$

$$P^{(1)} = \begin{pmatrix} \boxed{1} & 0 & \dots & \dots & \dots & 0 \\ -p_{21}^{(1)} & 1 & 0 & \dots & \dots & 0 \\ -p_{31}^{(1)} & 0 & 1 & 0 & \dots & 0 \\ & \dots\dots\dots & & & & \\ -p_{i1}^{(1)} & 0 & 0 & 1 & 0 & 0 \\ & \dots\dots\dots & & & & \\ -p_{n1}^{(1)} & 0 & \dots & \dots & \dots & 1 \end{pmatrix}$$

avec

$$\boxed{p_{i1}^{(1)} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad \forall 1 \leq i \leq n}$$

dont la matrice inverse est :

$$\left(P^{(1)}\right)^{-1} = \begin{pmatrix} \boxed{1} & 0 & \dots & \dots & \dots & 0 \\ p_{21}^{(1)} & 1 & 0 & \dots & \dots & 0 \\ p_{31}^{(1)} & 0 & 1 & 0 & \dots & 0 \\ & \dots\dots\dots & & & & \\ p_{i1}^{(1)} & 0 & 0 & 1 & 0 & 0 \\ & \dots\dots\dots & & & & \\ p_{n1}^{(1)} & 0 & \dots & \dots & \dots & 1 \end{pmatrix}$$

On applique donc les transformations élémentaires suivantes sur la matrice $A^{(1)}$

$$\boxed{L_i^{(2)} \leftarrow L_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)} = L_i^{(1)} - p_{i1}^{(1)} \cdot L_1^{(1)}, \quad i = 2,\ 3,\ \dots, n.}$$

On obtient alors un système linéaire sous la forme $A^{(2)}\mathbf{x} = \mathbf{b}^{(2)}$, équivalent à $A^{(1)}\mathbf{x} = \mathbf{b}^{(1)}$, avec

$$\boxed{A^{(2)} \;=\; P^{(1)}A^{(1)} \;\text{ et }\; \mathbf{b}^{(2)} \;=\; P^{(1)}\mathbf{b}^{(1)}}$$

sont les matrices de la forme :

$$A^{(2)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \dots & a_{1j}^{(1)} & \dots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} - p_{21}^{(1)}\, a_{12}^{(1)} & \dots & a_{2j}^{(1)} - p_{21}^{(1)}\, a_{1j}^{(1)} & \dots & a_{2n}^{(1)} - p_{21}^{(1)}\, a_{1n}^{(1)} \\ 0 & a_{32}^{(1)} - p_{31}^{(1)}\, a_{12}^{(1)} & \dots & a_{3j}^{(1)} - p_{31}^{(1)}\, a_{1j}^{(1)} & \dots & a_{3n}^{(1)} - p_{31}^{(1)}\, a_{1n}^{(1)} \\ \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & a_{i2}^{(1)} - p_{i1}^{(1)}\, a_{12}^{(1)} & \dots & a_{ij}^{(1)} - p_{i1}^{(1)}\, a_{1j}^{(1)} & \dots & a_{in}^{(1)} - p_{i1}^{(1)}\, a_{1n}^{(1)} \\ \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & a_{n2}^{(1)} - p_{n1}^{(1)}\, a_{12}^{(1)} & \dots & a_{nj}^{(1)} - p_{n1}^{(1)}\, a_{1n}^{(1)} & \dots & a_{nn}^{(1)} - p_{n1}^{(1)}\, a_{1n}^{(1)} \end{pmatrix}$$

et

$$\mathbf{b}^{(2)} = \begin{pmatrix} b_1^{(1)} \\ b_2^{(1)} - p_{21}^{(1)}\, b_1^{(1)} \\ b_3^{(1)} - p_{31}^{(1)}\, b_1^{(1)} \\ \vdots \\ b_i^{(1)} - p_{i1}^{(1)}\, b_1^{(1)} \\ \vdots \\ b_n^{(1)} - p_{n1}^{(1)}\, b_1^{(1)} \end{pmatrix}$$

Les coefficients $a_{ij}^{(2)}$ et $b_i^{(2)}$ sont données par les relations suivantes :

$$\boxed{\begin{cases} a_{1j}^{(1)} = a_{1j}, \;\to a_{1j}^{(2)} = a_{1j}^{(1)}, & \forall\, j = 2, 3, ..., n, \\ a_{ij}^{(1)} = a_{ij}, \;\to a_{ij}^{(2)} = a_{ij}^{(1)} - \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}\, .\, a_{1j}^{(1)}, & \forall\, i, j = 2, 3, ..., n \\ b_i^{(1)} = b_i, \;\to b_i^{(2)} = b_i^{(1)} - \dfrac{a_{i1}^{(1)}}{a_{11}^{(1)}}\, .\, b_1^{(1)}, & \forall\, i, j = 2,\ 3, ..., n\ . \end{cases}}$$

Après d'éventuelles permutations de lignes, on suppose que $a_{22}^{(2)} \neq 0$ et on applique le même procédé à la deuxième colonne de la matrice $A^{(2)}$. C'est ce processus que nous allons continuer :

Étape k : A l'étape k, on a le système $A^{(K)}\mathbf{x} = \mathbf{b}^{(K)}$ où la matrice $A^{(K)}$ et le vecteur $\mathbf{b}^{(K)}$ sont donnés par :

$$\begin{cases} A^{(k)} = P^{(k-1)}A^{(k-1)} = P^{(k-1)}P^{(k-2)}A^{(k-2)} = P^{(k-1)}P^{(k-2)}...P^{(1)}A^{(1)} \\ \mathbf{b}^{(k)} = P^{(k-1)}\mathbf{b}^{(k-1)} = P^{(k-1)}P^{(k-2)}\mathbf{b}^{(k-2)} = P^{(k-1)}P^{(k-2)}...P^{(1)}\mathbf{b}^{(1)} \end{cases}$$

La matrice $A^{(k)}$ est de la forme suivante :

$$A^{(k)} = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \dots & a_{1k-1}^{(k)} & a_{1k}^{(k)} & a_{1k+1}^{(k)} & \dots & a_{1n}^{(k)} \\ 0 & a_{22}^{(k)} & \dots & a_{2k-1}^{(k)} & a_{2k}^{(k)} & a_{2k+1}^{(k)} & \dots & a_{2n}^{(k)} \\ & \dots\dots\dots\dots & & & & & & \\ 0 & 0 & \dots & a_{k-1k-1}^{(k)} & a_{k-1k}^{(k)} & a_{k-1k+1}^{(k)} & \dots & a_{k-1n}^{(k)} \\ 0 & 0 & \dots & 0 & a_{kk}^{(k)} & a_{kk+1}^{(k)} & \dots & a_{kn}^{(k)} \\ \dots\dots\dots & & & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \dots & 0 & a_{nk}^{(k)} & a_{nk+1}^{(k)} & \dots & a_{nn}^{(k)} \end{pmatrix}$$

Supposons maintenant que le pivot $a_{kk}^{(k)}$ est non nul et introduisons la matrice $P^{(k)}$

$$P^{(k)} = \begin{pmatrix} 1 & 0 & \dots & 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 0 & \dots & 0 \\ & \dots\dots\dots\dots & & & & & & \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \boxed{1} & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & -p_{k+1k} & 1 & \dots & 0 \\ \dots\dots\dots & & & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \dots & 0 & -p_{nk} & 0 & \dots & 1 \end{pmatrix}$$

avec

$$\boxed{p_{ik}^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \quad \forall k \leq i \leq n}$$

On obtient alors un système linéaire sous la forme $A^{(k+1)}\mathbf{x} = \mathbf{b}^{(k+1)}$, équivalent à $A^{(k)}\mathbf{x} = \mathbf{b}^{(k)}$, avec

$$\begin{cases} A^{(k+1)} = P^{(k)}A^{(k)} = P^{(k)}P^{(k-1)}...P^{(1)}A^{(1)} \\ \mathbf{b}^{(k+1)} = P^{(k)}\mathbf{b}^{(k)} = P^{(k)}P^{(k-1)}...P^{(1)}\mathbf{b}^{(1)} \end{cases}$$

La matrice $A^{(k+1)}$ est de la forme suivante :

$$A^{(k+1)} = \begin{pmatrix} a_{11}^{(k+1)} & a_{12}^{(k+1)} & \dots & a_{1k-1}^{(k+1)} & a_{1k}^{(k+1)} & a_{1k+1}^{(k+1)} & \dots & a_{1n}^{(k+1)} \\ 0 & a_{22}^{(k+1)} & \dots & a_{2k-1}^{(k+1)} & a_{2k}^{(k+1)} & a_{2k+1}^{(k+1)} & \dots & a_{2n}^{(k+1)} \\ & \dots\dots\dots\dots & & & & & & \\ 0 & 0 & \dots & a_{k-1k-1}^{(k+1)} & a_{k-1k}^{(k+1)} & a_{k-1k+1}^{(k+1)} & \dots & a_{k-1n}^{(k+1)} \\ 0 & 0 & \dots & 0 & a_{kk}^{(k+1)} & a_{kk+1}^{(k+1)} & \dots & a_{kn}^{(k+1)} \\ \dots\dots\dots & & & \vdots & \vdots & \vdots & \vdots & \\ 0 & 0 & \dots & 0 & 0 & a_{nk+1}^{(k+1)} & \dots & a_{nn}^{(k+1)} \end{pmatrix}$$

Étape $n-1$ ***:*** A l'étape $(n-1)$, on obtient une matrice $A^{(n)}$ qui est triangulaire supérieure, et le système

$$A^{(n)}\mathbf{x} = \mathbf{b}^{(n)}$$

avec

$$\begin{cases} A^{(n)} = P^{(n-1)}A^{(n-1)} = P^{(n-1)}P^{(n-2)}...P^{(1)}A^{(1)} = P^{(n-1)}P^{(n-2)}...P^{(1)}A \\ \mathbf{b}^{(n)} = P^{(n-1)}\,\mathbf{b}^{(n-1)} = P^{(n-1)}P^{(n-2)}...P^{(1)}\,\mathbf{b}^{(1)} = P^{(n-1)}P^{(n-2)}...P^{(1)}\,\mathbf{b} \end{cases}$$

Posons maintenant

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

Alors,

$$A = \left(P^{(n-1)}P^{(n-2)}...P^{(1)}\right)^{-1}U = \left(P^{(1)}\right)^{-1}...\left(P^{(n-2)}\right)^{-1}\left(P^{(n-1)}\right)^{-1}U$$

Posons

$$L = \left(P^{(1)}\right)^{-1}...\left(P^{(n-2)}\right)^{-1}\left(P^{(n-1)}\right)^{-1}$$

Puisque tous les $\left(P^{(i)}\right)^{-1}$ sont des matrices triangulaires inférieures, alors le produit de ces dernières l'est aussi et, dans le cas ici, possède des 1 sur la diagonale. Nous obtenons pratiquement la décomposition LU :

$$A = LU$$

Exemple : On considère la matrice

$$A = \begin{pmatrix} 5 & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

1) Factoriser la matrice A par la méthode de Gauss en un produit LU

2) Résoudre le système suivant :

$$\begin{cases} 5 & x_1 & - & 4 & x_2 & + & & x_3 & = & 2, \\ -4 & x_1 & + & 4 & x_2 & & & & = & 0, \\ & x_1 & & & & + & 2 & x_3 & = & 3. \end{cases}$$

1) Pour la mise en œuvre de la méthode de Factorisation LU, nous procédons tout d'abord à la triangularisation de la matrice A en la décomposant en le produit de deux matrices triangulaires l'une inférieure L avec des 1 sur la diagonale et l'autre supérieure U telles que

$$A = LU$$

On pose

$$A^{(1)} = \begin{pmatrix} \boxed{5} & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

Étape 1 : On choisit $a_{11}^{(1)} = 5 \neq 0$ pivot de la première étape. La matrice $P^{(1)}$ s'écrit :

$$P^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{4}{5} & 1 & 0 \\ -\frac{1}{5} & 0 & 1 \end{pmatrix}$$

dont la matrice inverse est :

$$\left(P^{(1)}\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{1}{5} & 0 & 1 \end{pmatrix}$$

On obtient alors la matrice $A^{(2)}$ telle que :

$$A^{(2)} = P^{(1)}A^{(1)} = \begin{pmatrix} 5 & -4 & 1 \\ 0 & \boxed{\frac{4}{5}} & \frac{4}{5} \\ 0 & \frac{4}{5} & \frac{9}{5} \end{pmatrix}$$

Étape 2 : On choisit $a_{22}^{(2)} = \frac{4}{5} \neq 0$ pivot de la deuxième étape. La matrice $P^{(2)}$ s'écrit :

$$P^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

dont la matrice inverse est :

$$\left(P^{(2)}\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

On obtient alors la matrice $A^{(3)}$ telle que

$$A^{(3)} = P^{(2)}A^{(2)} = P^{(2)}P^{(1)}A^{(1)} = \begin{pmatrix} 5 & -4 & 1 \\ 0 & \frac{4}{5} & \frac{4}{5} \\ 0 & 0 & 1 \end{pmatrix}$$

Nous obtenons la décomposition de la matrice A en le produit des deux matrices L et U où

$$L = \left(P^{(1)}\right)^{-1}\left(P^{(2)}\right)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{1}{5} & 1 & 1 \end{pmatrix}$$

et

$$U = A^{(3)} = P^{(2)}P^{(1)}A = \begin{pmatrix} 5 & -4 & 1 \\ 0 & \frac{4}{5} & \frac{4}{5} \\ 0 & 0 & 1 \end{pmatrix}$$

2) On résout le système $A^{(3)}\mathbf{x} = \mathbf{b}^{(3)}$ par remontée avec

$$\mathbf{b}^{(3)} = P^{(2)}P^{(1)}\mathbf{b} = \begin{pmatrix} 1 \\ \frac{8}{5} \\ 2 \end{pmatrix}$$

$$\left\{ \begin{array}{rcrcrcl} & & & & x_3 & = & 1, \\ & & \frac{4}{5} & x_2 + & \frac{4}{5} & x_3 = & \frac{8}{5} \iff x_2 = 1, \\ 5 & x_1 - & 4 & x_2 + & & x_3 = & 2 \iff x_1 = 1. \end{array} \right.$$

$\underline{2^{me}}$ ***méthode :*** Le système

$$\left\{ \begin{array}{rcrcrcrcl} 5 & x_1 & - & 4 & x_2 & + & x_3 & = & 2, \\ -4 & x_1 & + & 4 & x_2 & & & = & 0, \\ & x_1 & & & & + & 2\ x_3 & = & 3. \end{array} \right.$$

s'écrit sous forme matricielle $A\mathbf{x} = \mathbf{b}$:

$$\begin{pmatrix} 5 & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}$$

équivalent au système $LU\mathbf{x} = b$:

$$\begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{1}{5} & 1 & 1 \end{pmatrix} \begin{pmatrix} 5 & -4 & 1 \\ 0 & \frac{4}{5} & \frac{4}{5} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix}$$

On résout le système $L\mathbf{y} = \mathbf{b}$ et ensuite le système $U\mathbf{x} = \mathbf{y}$:

$$\begin{pmatrix} 1 & 0 & 0 \\ -\frac{4}{5} & 1 & 0 \\ \frac{1}{5} & 1 & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix} \Longrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 8 \\ \frac{5}{1} \end{pmatrix}$$

et

$$\begin{pmatrix} 5 & -4 & 1 \\ 0 & \frac{4}{5} & \frac{4}{5} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

On trouve comme solution du système linéaire

$$S = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

2.9.1.1 Coût et intérêt de la méthode de Factorisation LU

La méthode de Factorisation LU requiert environ $2n^2$ opérations. Elle est utile lors de la résolution à plusieurs fois le même système linéaire $A\mathbf{x} = \mathbf{b}$ pour différents seconds membre b. Ainsi on effectue simultanément les même opérations élémentaires nécessaires sur tous les seconds membres.

Ceci nous permet donc de conclure que si n est grand, il est plus plus rapide d'utiliser la méthode de factorisation LU qui est d'ordre 2 ($\theta(n^2)$), que la méthode de Gauss, qui est d'ordre 3 ($\theta(n^3)$).

2.9.2 Méthode de Crout

La méthode de Factorisation de Crout, aussi appelée méthode d'***élimination de Gauss par colonnes*** ou ***active columm***, est une autre variante de la méthode de Gauss qui nécessite le même nombre d'opérations et consiste à procéder par substitution en privilégiant l'opération

de produit scalaire.

Supposons l'existence de la décomposition

$$A \;=\; LU$$

Développons l'équation précédente. Les termes a_{ij} sont déterminés en fonction des termes l_{ij} et u_{ij} des matrices L et Ude la manière suivante :

$$\boxed{\begin{cases} a_{ij} \;=\; u_{ij} \;+\; \sum_{k=1}^{i-1} l_{ik}.u_{kj}, & \forall \;\; i \leq j \leq n \quad (1) \\ \\ a_{ij} \;=\; \sum_{k=1}^{j} l_{ik}.u_{kj}, & \forall \;\; i > j \qquad (2) \end{cases}}$$

Résolution par la méthode de Crout En appliquant l'équation (2), on détermine les termes l_{ij} de la matrice L qui nous serviront à déterminer les termes u_{ij} de la matrice U en appliquant l'équation (1). Les valeurs des l_{ij} et u_{ij} sont déterminés en fonctions des a_{ij} de la manière suivante :

Ligne 1 **:** En appliquant l'équation (1) pour $i = 1$ et $1 \leq j \leq n$, l'équation (1) donne :

$$\boxed{u_{1j} \;=\; a_{1j}, \;\; \forall 1 \leq j \leq n}$$

Sachant que L est une matrice triangulaire avec des 1 sur la diagonale et U est une matrice triangulaire supérieure, on vient de déterminer les termes de la première ligne de ces deux matrices : $u_{1j}, \;\; \forall 1 \leq j \leq n$ et $l_{11} = 1$.

Ligne 2 **:** En appliquant l'équation (2) pour $i = 2$ et $j = 1$, on a :

$$a_{21} \;=\; l_{21}.u_{11} \;\;\Longrightarrow\;\; \boxed{l_{21} \;=\; \frac{a_{21}}{u_{11}}}$$

En appliquant l'équation (1) pour $i = 2$ et $2 \leq j \leq n$ et sachant que $u_{1j} \;=\; a_{1j}$, on a :

$$a_{2j} \;=\; u_{2j} \;+\; l_{21}.u_{1j}$$

ceci implique

$$\boxed{u_{2j} \;=\; a_{2j} \;-\; l_{21}.u_{1j}, \;\; \forall 2 \leq j \leq n}$$

On vient de déterminer les termes de la deuxième ligne de ces deux matrices : $u_{2j}, \;\; \forall 2 \leq j \leq n$ et l_{12} sachant que $u_{21} = 0$ et $l_{22} = 1$.

Ligne 3 **:** En appliquant l'équation (2) pour $i = 3$ et $j < i$, on a :

$$a_{3j} \;=\; \sum_{k=1}^{j} l_{3k}.u_{kj} \;\Longrightarrow\; \begin{cases} a_{31} \;=\; l_{31}.u_{11} & \Longrightarrow \;\; \boxed{l_{31} = \dfrac{a_{31}}{u_{11}}} \\ \\ a_{32} \;=\; l_{31}.u_{12} + l_{32}.u_{22} & \Longrightarrow \;\; \boxed{l_{32} = \dfrac{a_{32} - l_{31}.u_{12}}{u_{22}}} \end{cases}$$

En appliquant l'équation (1) pour $i = 3$ et $3 \leq j \leq n$, on a :

$$a_{3j} \;=\; u_{3j} \;+\; \sum_{k=1}^{2} l_{3k}.u_{kj}$$

ceci implique

$$\boxed{u_{3j} \;=\; a_{3j} \;-\; \sum_{k=1}^{2} l_{3k}.u_{kj} \;=\; a_{3j} \;-\; \big(l_{31}.u_{1j} \;+\; l_{32}.u_{2j}\big),\;\; \forall 3 \leq j \leq n}$$

On vient de déterminer les termes de la troisième ligne de ces deux matrices : $u_{3j},\;\; \forall 3 \leq j \leq n$, l_{31} et l_{32} sachant que $u_{31} = u_{32} = 0$ et $l_{33} = 1$.

Ligne i : Sachant qu'on a déterminé à l'étape précédente, les termes de la $(i-1)^{me}$ ligne de ces deux matrices : $u_{i-1j},\;\; \forall i-1 \leq j \leq n$ et $l_{i-11}, ..., l_{i-1i-2}$ sachant que $u_{i-11} = ... = u_{i-1i-2} = 0$ et $l_{i-1i-1} = 1$.

En appliquant l'équation (2) pour $j = 1, ..., i-1$, on a :

$$a_{ij} \;=\; \sum_{k=1}^{j} l_{ik}.u_{kj} \;\;\Longrightarrow\;\; \boxed{l_{ij} \;=\; \frac{a_{ij} \;-\; \sum_{k=1}^{j-1} l_{ik}.u_{kj}}{u_{jj}}}$$

En appliquant l'équation (1) pour $i \leq j$, il vient alors

$$a_{ij} \;=\; u_{ij} \;+\; \sum_{k=1}^{i-1} l_{ik}.u_{kj}$$

ceci implique

$$\boxed{u_{ij} \;=\; a_{ij} \;-\; \sum_{k=1}^{i-1} l_{ik}.u_{kj},\;\; \forall i \leq j \leq n}$$

Ainsi, l'***algorithme de Crout*** devient :

$$\boxed{\begin{cases} l_{ij} \;=\; \dfrac{a_{ij} \;-\; \sum_{k=1}^{j-1} l_{ik}.u_{kj}}{u_{jj}}, & \forall 1 \leq j \leq i-1 \\ \\ u_{ij} \;=\; a_{ij} \;-\; \sum_{k=1}^{i-1} l_{ik}.u_{kj}, & \forall i \leq j \leq n \end{cases}}$$

Exemple : On considère la matrice

$$A \;=\; \begin{pmatrix} 5 & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

Factoriser la matrice A par la méthode de Crout

Sachant que L est une matrice triangulaire avec des 1 sur la diagonale et U est une matrice triangulaire supérieure, cherchons les termes de ces deux matrices. En appliquant les deux formules suivantes :

$$\begin{cases} l_{ij} = \dfrac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}.u_{kj}}{u_{jj}}, \quad \forall \quad i > j \quad (1) \\ \\ u_{ij} = a_{ij} - \sum_{k=1}^{i-1} l_{ik}.u_{kj}, \quad \forall \quad i \leq j \leq n \quad (2) \end{cases}$$

***Ligne** 1 :* Les éléments l_{1j} et u_{1j}, $\forall j = 1, ..., 3$ deviennent :

$$\begin{cases} \bullet\ l_{11} = 1, \\ \bullet\ l_{12} = l_{13} = 0, \\ \bullet\ u_{11} = a_{11} = 5, \\ \bullet\ u_{12} = a_{12} = -4, \\ \bullet\ u_{13} = a_{13} = 1. \end{cases}$$

***Ligne** 2 :* Les éléments l_{2j} et u_{2j}, $\forall j = 1, ..., 3$ deviennent :

$$\begin{cases} \bullet\ l_{21} = \dfrac{a_{21}}{u_{11}} = -\dfrac{4}{5}, \\ \bullet\ l_{22} = 1, \\ \bullet\ l_{23} = 0, \\ \bullet\ u_{21} = 0, \\ \bullet\ u_{22} = a_{22} - l_{21}.u_{12} = \dfrac{4}{5}, \\ \bullet\ u_{23} = a_{23} - l_{21}.u_{13} = \dfrac{4}{5}. \end{cases}$$

***Ligne** 3 :* Les éléments l_{3j} et u_{3j}, $\forall j = 1, ..., 3$ deviennent :

$$\begin{cases} \bullet\ l_{31} = \dfrac{a_{31}}{u_{11}} = \dfrac{1}{5}, \\ \bullet\ l_{32} = \dfrac{a_{32} - l_{31}.u_{12}}{u_{22}} = 1, \\ \bullet\ l_{33} = 1, \\ \bullet\ u_{31} = 0, \\ \bullet\ u_{32} = 0, \\ \bullet\ u_{33} = a_{33} - \left(l_{31}.u_{13} + l_{32}.u_{23}\right) = 1. \end{cases}$$

Nous obtenons la décomposition de la matrice A en le produit des deux matrices L et U où

$$L = \begin{pmatrix} 1 & 0 & 0 \\ -\dfrac{4}{5} & 1 & 0 \\ \dfrac{1}{5} & 1 & 1 \end{pmatrix} \quad \text{et} \quad U = \begin{pmatrix} 5 & -4 & 1 \\ 0 & \dfrac{4}{5} & \dfrac{4}{5} \\ 0 & 0 & 1 \end{pmatrix}$$

2.9.3 Méthode de Cholesky

Théorème 2.2. *Soit A une matrice symétrique définie positive. Alors il existe une unique matrice L triangulaire inférieure à diagonale unité, et une unique matrice diagonale D à coefficients strictement positifs, telles que :*

$$A = LD^tL$$

Théorème 2.3. ***Théorème de Cholesky.*** *Soit A une matrice symétrique définie positive d'ordre n. Alors il existe une unique décomposition de Cholesky de A sous la forme*

$$A = L\,{}^tL,$$

où L est une matrice triangulaire inférieure à coefficients diagonaux strictement positifs.

Décomposition de la matrice A : Développons l'équation précédente, on a :

$$a_{ij} = \sum_{k=1}^{n} l_{ik}.l_{jk}$$

Sachant que $l_{ik} = 0$ si $k > i$ et $l_{jk} = 0$ si $k > j$, l'équation précédente s'écrit alors :

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik}.l_{jk}$$

Afin de déterminer les termes de la matrice L qui est une matrice triangulaire supérieure, la *r-ième* ligne de l'équation précédente implique

$$a_{rj} = \sum_{k=1}^{r} l_{rk}.l_{jk} = l_{rr}.l_{jr} + \sum_{k=1}^{r} l_{rk}.l_{jk}, \quad \forall r \leq j \leq n$$

Ainsi pour $\boxed{1 \leq r \leq n}$, il vient alors

$$\boxed{l_{rr} = \sqrt{a_{rr} - \sum_{k=1}^{r-1} l_{rk}^2}}$$

et

$$\boxed{l_{jr} = \frac{1}{l_{rr}} \left(a_{rj} - \sum_{k=1}^{r-1} l_{rk}.l_{jk} \right), \quad \forall r+1 \leq j \leq n}$$

Exemple : On considère la matrice

$$A = \begin{pmatrix} 5 & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix}$$

1) Montrer que la matrice A est une matrice définie positive

2) Factoriser la matrice A par la méthode de Cholesky

3) Résoudre le système suivant :

$$\left\{\begin{array}{rcrcrcrcl} 5 & x_1 & - & 4 & x_2 & + & x_3 & = & 2, \\ -4 & x_1 & + & 4 & x_2 & & & = & 0, \\ & x_1 & & & & + & 2\ x_3 & = & 3. \end{array}\right.$$

1) La matrice A est une matrice définie positive si et seulement si la forme quadratique qui lui est associée est définie positive, i.e.

$$^t\mathbf{x}\ A\ \mathbf{x} \ = \ \sum_{i=1}^{3}\sum_{j=1}^{3} x_i\ a_{ij}\ x_j > 0, \quad \forall \mathbf{x} = (x_1, x_2, x_3) > 0$$

Considérons donc un vecteur $\mathbf{x} = (x_1, x_2, x_3)$, on a

$$\begin{array}{rcl} (x_1, x_2, x_3) \begin{pmatrix} 5 & -4 & 1 \\ -4 & 4 & 0 \\ 1 & 0 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} & = & 5x_1^2 \ + \ 4x_2^2 \ + \ 2x_3^2 \ - \ 8x_1x_2 \ + \ 2x_1x_3 \\ & = & 4(x_1 - x_2)^2 \ + \ (x_1 + x_3)^2 \ + \ x_3^2 > 0. \end{array}$$

2) L'algorithme de Cholesky génère une matrice L triangulaire inférieure à coefficients diagonaux strictement positifs.

Colonne 1 **:** Les éléments l_{j1}, $\forall j = 1, ..., 3$ deviennent :

$$\left\{\begin{array}{lcll} \bullet\ l_{11} & = & \sqrt{a_{11}} & = \sqrt{5}, \\ \bullet\ l_{21} & = & \dfrac{a_{12}}{l_{11}} & = -\dfrac{4}{\sqrt{5}}, \\ \bullet\ l_{21} & = & \dfrac{a_{13}}{l_{11}} & = \dfrac{1}{\sqrt{5}}. \end{array}\right.$$

Colonne 2 **:** Les éléments l_{j2}, $\forall j = 1, ..., 3$ deviennent :

$$\left\{\begin{array}{lcll} \bullet\ l_{12} & = & 0, & \\ \bullet\ l_{22} & = & \sqrt{a_{12} - l_{21}^2} & = \dfrac{2}{\sqrt{5}}, \\ \bullet\ l_{32} & = & \dfrac{1}{l_{22}}\Big(a_{23} - l_{21}.l_{31}\Big) & = \dfrac{2\sqrt{5}}{5}. \end{array}\right.$$

Colonne 3 **:** Les éléments l_{j3}, $\forall j = 1, ..., 3$ deviennent :

$$\left\{\begin{array}{lcll} \bullet\ l_{13} & = & 0, & \\ \bullet\ l_{23} & = & 0, & \\ \bullet\ l_{33} & = & \sqrt{a_{33} - \Big(l_{31}^2 + l_{32}^2\Big)} & = 1. \end{array}\right.$$

Nous obtenons la décomposition de la matrice A en le produit des deux matrices L et ${}^tL$ où

$$L = \begin{pmatrix} \sqrt{5} & 0 & 0 \\ -\frac{4}{\sqrt{5}} & \frac{2}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{5}} & \frac{2\sqrt{5}}{5} & 1 \end{pmatrix}$$

3) Après décomposition de la matrice A, la résolution du système $A\mathbf{x} = \mathbf{b}$ peut se décomposer en la résolution successive des deux systèmes :

$$\begin{cases} L\mathbf{y} = \mathbf{b} & \text{d'inconnue } \mathbf{y} \\ {}^tL\mathbf{x} = \mathbf{y} & \text{d'inconnue } \mathbf{x} \end{cases}$$

où

$$\begin{pmatrix} \sqrt{5} & 0 & 0 \\ -\frac{4}{\sqrt{5}} & \frac{2}{\sqrt{5}} & 0 \\ \frac{1}{\sqrt{5}} & \frac{2\sqrt{5}}{5} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 3 \end{pmatrix} \Longrightarrow \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} \frac{2}{\sqrt{5}} \\ \frac{4\sqrt{5}}{5} \\ 1 \end{pmatrix}$$

et

$$\begin{pmatrix} \sqrt{5} & -\frac{4}{\sqrt{5}} & \frac{1}{\sqrt{5}} \\ 0 & \frac{2}{\sqrt{5}} & \frac{2\sqrt{5}}{5} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} \Longrightarrow \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

On trouve comme solution du système linéaire

$$S = \left\{ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

Il reste à noter que la matrice L est en quelque sorte une racine carrée de A. Dans le cas où il n'est pas nécessaire de déterminer la décomposition de la matrice A, il suffit d'adapter la stratégie suivante dite ***Factorisation de Cholesky sans racine carrée***.

Soit la matrice D telle que $d_{ii} = l_{ii}$, $\forall i$. Les factorisations $L'D'^tL'$ et L^tL sont liées :

$$A = L'D'^tL' = \left(LD^{-1}\right) D^2 \left(D^{-1}\,{}^tL\right)$$

où L' est une matrice triangulaire inférieure unitaire telle que $\boxed{l'_{ii} = 1, \quad \forall i.}$

La *ième* ligne de la matrice A donne :

$$a_{ij} = \sum_{k=1}^{i} l'_{ik}\, d'_{kk}\, l'_{jk} = \sum_{k=1}^{i-1} l'_{ik}\, d'_{kk}\, l'_{jk} + l'_{ii}\, d'_{ii}\, l'_{ji} = \sum_{k=1}^{i-1} l'_{ik}\, d'_{kk}\, l'_{jk} + d'_{ii}\, l'_{ji}, \quad \forall i+1 \leq j \leq n.$$

Ainsi pour $j = i$, on a :

$$a_{ii} \;=\; \sum_{k=1}^{i-1} l'_{ik}\; d'_{kk}\; l'_{ik} \;+\; l'_{ii}\; d'_{ii}\; l'_{ii} \;=\; \sum_{k=1}^{i-1} l'_{ik}\; d'_{kk}\; l'_{ik} \;+\; d'_{ii}$$

Ainsi, on obtient l'algorithme de Cholesky suivant :

$$\left\{\begin{array}{ll} l'_{ii} & = 1, \quad \forall 1 \leq i \leq n \\ d'_{ii} & = a_{ii} \;-\; \sum_{k=1}^{i-1} l'_{ik}\; d'_{kk}\; l'_{ik}, \quad \forall 1 \leq i \leq n \\ l'_{ji} & = \dfrac{1}{d'_{ii}}\left(a_{ij} \;-\; \sum_{k=1}^{i-1} l'_{ik}\; d'_{kk}\; l'_{jk}\right), \quad i+1 \leq j \leq n. \end{array}\right.$$

Chapitre 3

Résolution des systèmes linéaires-Méthodes itératives

3.1 Principe

Soit $A \in M_n(\mathbb{R})$ qu'on supposera inversible et $\mathbf{b} \in \mathbb{R}^n$. On cherche à calculer $\mathbf{x} \in \mathbb{R}^n$ tel que $A\mathbf{x} = \mathbf{b}$, autrement dit on cherche à résoudre le système linéaire

$$A\,\mathbf{x} \;=\; \mathbf{b}.$$

Dans ce chapitre, on présentera quelques méthodes itératives à un pas qui consistent à générer une suite de vecteurs $\left(\mathbf{x}^{(i)}\right)_i$ telle que $\mathbf{x}^{(i)} = {}^t\left(x_1^{(i)}, ..., x_n^{(i)}\right)$ convergente vers le vecteur solution $\mathbf{x}$ en partant d'un vecteur initial $\mathbf{x}^{(0)} = {}^t\left(x_1^{(0)}, ..., x_n^{(0)}\right)$. Les méthodes itératives sont avantageuses par rapport aux méthodes directes lorsque n est assez grand ou que le problème est mal conditionné, afin de minimiser la propagation des erreurs qui reste le problème majeur des calculs flottants.

Les méthodes itératives reposent sur la décomposition suivante de la matrice A qui est supposée être régulière et la matrice M inversible :

$$\boxed{A \;=\; M - N}$$

Ainsi le système linéaire $A\mathbf{x} = \mathbf{b}$ peut s'écrire :

$$(M - N)\,\mathbf{x} \;=\; \mathbf{b}$$

qui est équivalent à

$$M\mathbf{x} \;=\; N\mathbf{x} + \mathbf{b} \;\text{ ou }\; \mathbf{x} \;=\; M^{-1}N\mathbf{x} + M^{-1}\,\mathbf{b}$$

On se donne un vecteur initial $\mathbf{x}^{(0)}$ et on définit une suite de vecteurs $\left(\mathbf{x}^{(i)}\right)_i$ telle que elle peut être représentée par la relation itérative suivante :

$$\boxed{M\mathbf{x}^{(i+1)} \;=\; N\,\mathbf{x}^{(i)} + \mathbf{b}}$$

qui est équivalente à

$$\boxed{\mathbf{x}^{(i+1)} \;=\; P\,\mathbf{x}^{(i)} + Q}$$

où

$$\boxed{P \;=\; M^{-1}N \;\text{ et }\; Q \;=\; M^{-1}\,\mathbf{b}}$$

sont indépendants de i.

3.2 Convergence de la méthode itérative

Par contre, la question de la vitesse de convergence reste cruciale. Cette question fondamentale a fait l'objet d'une grand intérêt des mathématiciens au cours des années. Pour répondre à cette question, on cite les théorèmes de convergence du processus itératif suivant :

Théorème 3.1. ***Théorème de convergence.*** *On se donne un vecteur initial $\boldsymbol{x}^{(0)}$. La suite de vecteurs $\left(\boldsymbol{x}^{(i)}\right)_i$ telle que $\boldsymbol{x}^{(i+1)} \;=\; P\,\boldsymbol{x}^{(i)} + Q$ converge vers le vecteur solution $\boldsymbol{x} = (I-P)^{-1}Q$ quel que soit $\boldsymbol{x}^{(0)}$ si et seulement si*

$$\rho(P) \;<\; 1$$

Proposition 3.1. ***Condition de convergence.*** *Sachant que $\rho(P) \leq \|P\|$ et $\rho(P) < 1$. La suite de vecteurs $\left(\boldsymbol{x}^{(i)}\right)_i$ converge vers le vecteur solution $\boldsymbol{x}$ quel que soit $\boldsymbol{x}^{(0)}$ si et seulement si*

$$\|P\| <\; 1$$

La démonstration de ces théorème et proposition reposent sur le théorème suivant qu'on énonce :

Théorème 3.2. *Soit A une matrice carrée. Les conditions suivantes sont équivalentes :*

- $\lim_{i\to\infty} A^i = \mathbf{0}$,
- $\lim_{i\to\infty} A^i\mathbf{v} = \mathbf{0}$, pour tout vecteur $\mathbf{v}$,
- $\rho(A) \;<\; 1$,
- $\|A\| <\; 1$ pour au moins une norme matricielle subordonnée $\|.\|$.

De plus, on a :

Théorème 3.3. *Soit $\|.\|$ une norme matricielle subordonnée et A une matrice carrée. Alors, on a*

$$\lim_{i\to\infty} \|A^i\|^{1/i} = \;\rho(A)$$

Preuve 3.1. *En effet, définissons le vecteur associé à la i-ème itération*

$$\boldsymbol{e}^{(i)} \;=\; \boldsymbol{x}^{(i)} \;-\; \boldsymbol{x} \;=\; P\,\boldsymbol{x}^{(i-1)} \;=\; P^i\,\boldsymbol{x}^{(0)}$$

La convergence de la méthode itérative est assurée si et seulement si

$$\lim_{i\to\infty} \boldsymbol{e}^{(i)} = \boldsymbol{0}$$

où $\boldsymbol{0}$ dénote la matrice nulle d'ordre n. Autrement dit, pour tout vecteur initial $\boldsymbol{x}^{(0)}$, on a

$$\lim_{i\to\infty} P^i\,\boldsymbol{x}^{(0)} = \boldsymbol{0}$$

Ce qui revient à dire :

$$\lim_{i\to\infty} P^i = \boldsymbol{0}$$

Ainsi en appliquant le théorème précédent, on tire le théorème de convergence.

3.3 Décomposition de la matrice A

Les méthodes itératives sont déduites de la décomposition suivante de la matrice A

$$A = M - N$$

de telle façon que la matrice M soit inversible et vérifie la condition de convergence suivante $\rho(M^{-1}N) < 1$ ou d'une manière équivalente $\|M^{-1}N\| < 1$.

On définit les matrices suivantes D, L et U telles que :

- D est une matrice diagonale, i.e.

$$d_{ii} = a_{ii}, \quad \forall 1 \leq i \leq n$$

- L est une matrice triangulaire inférieure à diagonale nulle, i.e.

$$l_{ij} = -a_{ij}, \quad \forall 1 \leq j < i \quad \text{et} \quad l_{ij} = 0, \quad \forall i \leq j$$

- U est une matrice triangulaire supérieure à diagonale nulle, i.e.

$$u_{ij} = -a_{ij}, \quad \forall i < j \quad \text{et} \quad u_{ij} = 0, \quad \forall 1 \leq j \leq i$$

On a alors la relation suivante :

$$\boxed{A = D - L - U}$$

et on présente les types de décomposition suivants :

- ***Méthode de Jacobi***

$$\boxed{M = D \quad \text{et} \quad N = L + U}$$

et la matrice P_J sera

$$\boxed{P_J = D^{-1}(L + U)}$$

- ***Méthode de Gauss-Seidel***

$$\boxed{M = D - L \quad \text{et} \quad N = U}$$

et la matrice P_G sera

$$\boxed{P_G = (D - L)^{-1} U}$$

- ***Méthode de Relaxation***

$$\boxed{M = \frac{D}{w} - L \quad \text{et} \quad N = \frac{1 - w}{w} D + U}$$

et la matrice P_R sera

$$\boxed{P_R = \left(\frac{D}{w} - L\right)^{-1} \left(\frac{1 - w}{w} D + U\right)}$$

3.4 Méthode de Jacobi

Selon la méthode de Jacobi qui est due au mathématicien allemand ***Karl Jacobi***, on décompose la matrice A de la façon suivante :

$$M = D \quad \text{et} \quad N = L + U$$

Le système $A\mathbf{x} = \mathbf{b}$ s'écrit alors :

$$(D - (L+U))\mathbf{x} = \mathbf{b} \implies D\mathbf{x} = (L+U)\mathbf{x} + \mathbf{b}$$

Ainsi, la méthode itérative de Jacobi devient :

$$D\,\mathbf{x}^{(i+1)} \;=\; (L+U)\,\mathbf{x}^{(i)} + \mathbf{b}$$

équivalente à

$$\boxed{\mathbf{x}^{(i+1)} \;=\; D^{-1}(L+U)\,\mathbf{x}^{(i)} + D^{-1}\,\mathbf{b} \;=\; \left(I - D^{-1}A\right)\mathbf{x}^{(i)} + D^{-1}\,\mathbf{b}} \tag{3.1}$$

En supposant que les pivots $a_{ii} \neq 0$, (3.1) s'écrit sous forme matricielle :

$$\begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \\ \vdots \\ x_j^{(i+1)} \\ \vdots \\ x_n^{(i+1)} \end{pmatrix} = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \dots & -\frac{a_{1j}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} & \dots & -\frac{a_{2j}}{a_{22}} & \dots & -\frac{a_{2n}}{a_{22}} \\ -\frac{a_{31}}{a_{33}} & -\frac{a_{32}}{a_{33}} & 0 & \dots & -\frac{a_{3j}}{a_{33}} & \dots & -\frac{a_{3n}}{a_{33}} \\ & & & \dots\dots\dots\dots & & & \\ -\frac{a_{j1}}{a_{jj}} & -\frac{a_{j2}}{a_{jj}} & -\frac{a_{j3}}{a_{jj}} & \dots & 0 & \dots & -\frac{a_{jn}}{a_{jj}} \\ & & & \dots\dots\dots\dots & & & \\ -\frac{a_{n1}}{a_{nn}} & -\frac{a_{n2}}{a_{nn}} & -\frac{a_{n3}}{a_{nn}} & \dots & -\frac{a_{nj}}{a_{nn}} & \dots & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ \vdots \\ x_j^{(i)} \\ \vdots \\ x_n^{(i)} \end{pmatrix} + \begin{pmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \frac{b_3}{a_{33}} \\ \vdots \\ \frac{b_j}{a_{ii}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix}$$

équivalent à la forme développée suivante :

$$\begin{cases} x_1^{(i+1)} &= \Big(b_1 \;- a_{12}\,x_2^{(i)} \;- a_{13}\,x_3^{(i)} \;- \dots \;- a_{1n}\,x_n^{(i)}\Big)/a_{11}, \\ x_2^{(i+1)} &= \Big(b_2 \;- a_{21}\,x_1^{(i)} \;- a_{23}\,x_3^{(i)} \;- \dots \;- a_{2n}\,x_n^{(i)}\Big)/a_{22}, \\ & \dots\dots\dots\dots \\ x_n^{(i+1)} &= \Big(b_n \;- a_{n1}\,x_1^{(i)} \;- a_{n2}\,x_2^{(i)} \;- \dots \;- a_{nn-1}\,x_{n-1}^{(i)}\Big)/a_{nn}. \end{cases}$$

Les itérations précédentes dont du type $\mathbf{x}^{(i+1)} = F(\mathbf{x}^{(i)})$ où F est une fonction linéaire indépendante de i.

3.4.1 Convergence et critère d'arrêt de la méthode de Jacobi

Théorème 3.4. ***Condition suffisante.*** *La méthode de Jacobi converge quel que soit le vecteur initial $\boldsymbol{x}^{(0)}$ pour les systèmes linéaires $A\boldsymbol{x} = \boldsymbol{b}$ dont la matrice A est à diagonale fortement dominante. Ceci se traduit par la condition suivante :*

$$\boxed{\sum_{j=1, i\neq j} |a_{ij}| < |a_{ii}| \quad \forall i = 1, 2, ..., n}$$

Preuve 3.2. *En appliquant le théorème 3.1, la convergence de la méthode Jacobi est assurée si et seulement si $\rho(P_J) < 1$ ou la condition équivalente $\|P_J\|_1 < 1$. Ceci se traduit par*

$$\sum_{j=1}^{n} |p_{ij}| < 1$$

ou encore

$$\sum_{j=1}^{n} |l_{ij} + u_{ij}| < |d_{ii}|, \quad \forall i = 1, 2, ..., n$$

Théorème 3.5. ***Test d'arrêt.*** *Un critère est d'arrêter les itérations quand on utilise l'erreur relative sur le vecteur résidu $\boldsymbol{r}^{(i)} = b - A\boldsymbol{x}^{(i)}$, ce qui donne pour une précision donnée ε :*

$$\frac{\|\mathbf{r}^{(i)}\|}{\|\mathbf{b}\|} < \varepsilon$$

Ainsi, l'***algorithme de Jacobi*** devient :

$$\boxed{\left\{\begin{array}{l} \text{Etant donnés } A,\ b,\ x^{(0)},\ imax \text{ et } \varepsilon,\ \forall i = 1, 2, ..., imax \\ \\ r_j^{(i)} = b_j - \sum_{k=1}^{n} a_{jk}\, x_k^{(i)}, \quad \forall j = 1, 2, ..., n \\ \\ x_j^{(i+1)} = \left(b_j - \sum_{k=1}^{n} a_{jk}\, x_k^{(i)}\right)/a_{jj}, \quad \forall j = 1, 2, ..., n \\ \\ \text{arrêter si : } \dfrac{\|\mathbf{r}^{(i)}\|}{\|\mathbf{b}\|} < \varepsilon \end{array}\right.}$$

Exemple : On considère le système linéaire suivant :

$$\left\{\begin{array}{rrcrcrcl} 5 & x_1 & - & 4 & x_2 & + & x_3 & = 2, \\ -4 & x_1 & + & 4 & x_2 & & & = 0, \\ & x_1 & & & & + & 2\ x_3 & = 3. \end{array}\right.$$

1) Donner la matrice de la méthode de Jacobi et expliciter sous forme matricielle l'algorithme de cette méthode.

2) Donner la solution du système en partant du vecteur initial $x^0 = {}^t(0.9, 0.9, 0.9)$.

1) On décompose la matrice A de la façon suivante :

$$A = D - (L + U)$$

où

$$D = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{pmatrix} \quad L = \begin{pmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix} \quad U = \begin{pmatrix} 0 & 4 & -1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

La matrice de Jacobi est donnée par :

$$D^{-1}(L+U) \;=\; \left(I - D^{-1}A\right) \;=\; \begin{pmatrix} 0 & \frac{4}{5} & -\frac{1}{5} \\ 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix}$$

L'algorithme de Jacobi s'écrit sous la forme matricielle suivante :

$$\begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \end{pmatrix} = \begin{pmatrix} 0 & \frac{4}{5} & -\frac{1}{5} \\ 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \end{pmatrix} + \begin{pmatrix} \frac{2}{5} \\ 0 \\ \frac{3}{2} \end{pmatrix}$$

équivalent à la forme développée suivante :

$$\begin{cases} x_1^{(i+1)} &= \frac{4}{5}\, x_2^{(i)} \;-\; \frac{1}{5}\, x_3^{(i)} \;+\; \frac{2}{5} \\ x_2^{(i+1)} &= x_1^{(i)} \\ x_3^{(i+1)} &= -\frac{1}{2}\, x_1^{(i)} \;+\; \frac{3}{2} \end{cases}$$

2) La solution du système linéaire par la méthode de Jacobi est :

n	x_1^n	x_2^n	x_3^n	n	x_1^n	x_2^n	x_3^n
1	0.9400000	0.9000000	1.050000	2	0.9100000	0.9400000	1.030000
3	0.9460001	0.9100000	1.045000	4	0.9190000	0.9460001	1.027000
5	0.9514002	0.9190000	1.040500	6	0.9271001	0.9514002	1.024300
7	0.9562602	0.9271001	1.036450	8	0.9343901	0.9562602	1.021870
9	0.9606342	0.9343901	1.032805	10	0.9409511	0.9606342	1.019683
11	0.9645709	0.9409511	1.029524	12	0.9468560	0.9645709	1.017715
13	0.9681138	0.9468560	1.026572	14	0.9521704	0.9681138	1.015943
15	0.9713024	0.9521704	1.023915	16	0.9569534	0.9713024	1.014349
17	0.9741722	0.9569534	1.021523	18	0.9612581	0.9741722	1.012914
19	0.9767551	0.9612581	1.019371	20	0.9651323	0.9767551	1.011622
21	0.9790796	0.9651323	1.017434	22	0.9686191	0.9790796	1.010460
23	0.9811716	0.9686191	1.015690	24	0.9717572	0.9811716	1.009414
25	0.9830545	0.9717572	1.014121	26	0.9745814	0.9830545	1.008473
27	0.9847491	0.9745814	1.012709	28	0.9771234	0.9847491	1.007625
29	0.9862742	0.9771234	1.011438	30	0.9794110	0.9862742	1.006863
31	0.9876469	0.9794110	1.010294	32	0.9814699	0.9876469	1.006177
33	0.9888822	0.9814699	1.009265	34	0.9833229	0.9888822	1.005559
35	0.9899939	0.9833229	1.008339	36	0.9849905	0.9899939	1.005003
37	0.9909946	0.9849905	1.007505	38	0.9864915	0.9909946	1.004503
39	0.9918951	0.9864915	1.006754	40	0.9878424	0.9918951	1.004052
41	0.9927056	0.9878424	1.006079	42	0.9890581	0.9927056	1.003647
43	0.9934351	0.9890581	1.005471	44	0.9901523	0.9934351	1.003282
45	0.9940916	0.9901523	1.004924	46	0.9911371	0.9940916	1.002954
47	0.9946824	0.9911371	1.004431	48	0.9920234	0.9946824	1.002659
49	0.9952142	0.9920234	1.003988	50	0.9928211	0.9952142	1.002393
51	0.9956928	0.9928211	1.003589	52	0.9935390	0.9956928	1.002154
53	0.9961236	0.9935390	1.003230	54	0.9941851	0.9961236	1.001938
55	0.9965112	0.9941851	1.002907	56	0.9947667	0.9965112	1.001744
57	0.9968601	0.9947667	1.002617	58	0.9952900	0.9968601	1.001570
59	0.9971742	0.9952900	1.002355	60	0.9957610	0.9971742	1.001413
61	0.9974568	0.9957610	1.002120	62	0.9961849	0.9974568	1.001272
63	0.9977112	0.9961849	1.001908	64	0.9965664	0.9977112	1.001144
65	0.9979401	0.9965664	1.001717	66	0.9969097	0.9979401	1.001030
67	0.9981461	0.9969097	1.001545	68	0.9972187	0.9981461	1.000927
69	0.9983315	0.9972187	1.001391	70	0.9974968	0.9983315	1.000834
71	0.9984984	0.9974968	1.001252	72	0.9977471	0.9984984	1.000751
73	0.9986486	0.9977471	1.001126	74	0.9979724	0.9986486	1.000676
75	0.9987838	0.9979724	1.001014	76	0.9981752	0.9987838	1.000608
77	0.9989054	0.9981752	1.000912	78	0.9983577	0.9989054	1.000547
79	0.9990149	0.9983577	1.000821	80	0.9985220	0.9990149	1.000493
81	0.9991134	0.9985220	1.000739	82	0.9986698	0.9991134	1.000443
83	0.9992021	0.9986698	1.000665	84	0.9988028	0.9992021	1.000399
85	0.9992819	0.9988028	1.000599	86	0.9989226	0.9992819	1.000359
87	0.9993538	0.9989226	1.000539	88	0.9990303	0.9993538	1.000323

TABLE 3.1 – Vecteurs solutions obtenus par la méthode de Jacobi

n	x_1^n	x_2^n	x_3^n	n	x_1^n	x_2^n	x_3^n
89	0.9994184	0.9990303	1.000485	90	0.9991273	0.9994184	1.000291
91	0.9994766	0.9991273	1.000436	92	0.9992146	0.9994766	1.000262
93	0.9995289	0.9992146	1.000393	94	0.9992931	0.9995289	1.000236
95	0.9995761	0.9992931	1.000353	96	0.9993638	0.9995761	1.000212
97	0.9996185	0.9993638	1.000318	98	0.9994276	0.9996185	1.000191
99	0.9996567	0.9994276	1.000286	100	0.9994848	0.9996567	1.000172
101	0.9996911	0.9994848	1.000258	102	0.9995363	0.9996911	1.000154
103	0.9997220	0.9995363	1.000232	104	0.9995827	0.9997220	1.000139
105	0.9997498	0.9995827	1.000209	106	0.9996244	0.9997498	1.000125
107	0.9997749	0.9996244	1.000188	108	0.9996620	0.9997749	1.000113
109	0.9997974	0.9996620	1.000169	110	0.9996958	0.9997974	1.000101
111	0.9998177	0.9996958	1.000152	112	0.9997262	0.9998177	1.000091
113	0.9998359	0.9997262	1.000137	114	0.9997536	0.9998359	1.000082
115	0.9998524	0.9997536	1.000123	116	0.9997782	0.9998524	1.000074
117	0.9998671	0.9997782	1.000111	118	0.9998004	0.9998671	1.000066
119	0.9998804	0.9998004	1.000100	120	0.9998204	0.9998804	1.000060
121	0.9998924	0.9998204	1.000090	122	0.9998383	0.9998924	1.000054
123	0.9999031	0.9998383	1.000081	124	0.9998546	0.9999031	1.000048
125	0.9999127	0.9998546	1.000073	126	0.9998692	0.9999127	1.000044
127	0.9999214	0.9998692	1.000065	128	0.9998823	0.9999214	1.000039
131	0.9999365	0.9998941	1.000053	132	0.9999047	0.9999365	1.000032
133	0.9999428	0.9999047	1.000048	134	0.9999142	0.9999428	1.000029
135	0.9999486	0.9999142	1.000043	136	0.9999228	0.9999486	1.000026
137	0.9999537	0.9999228	1.000039	138	0.9999306	0.9999537	1.000023
139	0.9999583	0.9999306	1.000035	140	0.9999375	0.9999583	1.000021
141	0.9999625	0.9999375	1.000031	142	0.9999439	0.9999625	1.000019
143	0.9999663	0.9999439	1.000028	144	0.9999495	0.9999663	1.000017
145	0.9999696	0.9999495	1.000025	146	0.9999546	0.9999696	1.000015
147	0.9999727	0.9999546	1.000023	148	0.9999592	0.9999727	1.000014
149	0.9999755	0.9999592	1.000020	150	0.9999633	0.9999755	1.000012
151	0.9999779	0.9999633	1.000018	152	0.9999669	0.9999779	1.000011
153	0.9999802	0.9999669	1.000017	154	0.9999703	0.9999802	1.000010
155	0.9999822	0.9999703	1.000015	156	0.9999732	0.9999822	1.000009
157	0.9999841	0.9999732	1.000013	158	0.9999759	0.9999841	1.000008
159	0.9999857	0.9999759	1.000012	160	0.9999783	0.9999857	1.000007
161	0.9999872	0.9999783	1.000011	162	0.9999804	0.9999872	1.000006
163	0.9999885	0.9999804	1.000010	164	0.9999825	0.9999885	1.000006
165	0.9999897	0.9999825	1.000009	166	0.9999842	0.9999897	1.000005
167	0.9999908	0.9999842	1.000008	168	0.9999858	0.9999908	1.000005
169	0.9999918	0.9999858	1.000007	170	0.9999873	0.9999918	1.000004
171	0.9999927	0.9999873	1.000006	172	0.9999887	0.9999927	1.000004
173	0.9999934	0.9999887	1.000006	174	0.9999899	0.9999934	1.000003
175	0.9999942	0.9999899	1.000005	176	0.9999909	0.9999942	1.000003
177	***0.9999948***	***0.9999909***	***1.000005***				

TABLE 3.2 – Vecteurs solutions obtenus par la méthode de Jacobi

Par suite, on en conclut que le vecteur solution obtenu avec un critère d'arrêt d'ordre 10^{-6} est :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.9999948 \\ 0.9999909 \\ 1.0000050 \end{pmatrix}$$

Sachant qu'à partir de $n = 224$, l'algorithme de la méthode de Jacobi converge vers le vecteur solution stable :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.9999996 \\ 0.9999996 \\ 1.0000000 \end{pmatrix}$$

3.5 Méthode de Gauss-Seidel

On décompose la matrice A de la façon suivante :

$$M = D - L \quad \text{et} \quad N = U$$

Le système $A\mathbf{x} = \mathbf{b}$ s'écrit alors :

$$((D - L) - U)\ \mathbf{x} = \mathbf{b} \implies (D - L)\ \mathbf{x} = U\mathbf{x} + \mathbf{b}$$

Ainsi, la méthode itérative de Jacobi devient :

$$(D - L)\ \mathbf{x}^{(i+1)} = U\ \mathbf{x}^{(i)} + \mathbf{b}$$

équivalente à

$$\mathbf{x}^{(i+1)} = (D - L)^{-1} U\ \mathbf{x}^{(i)} + (D - L)^{-1}\ \mathbf{b} \tag{3.2}$$

Puisque l'inverse de la matrice $D - L$ peut avérer être compliquée à calculer, on écrit le système (3.2) de la manière suivante :

$$\boxed{\mathbf{x}^{(i+1)} = D^{-1}L\ \mathbf{x}^{(i+1)} + D^{-1}U\ \mathbf{x}^{(i)} + D^{-1}\ \mathbf{b}} \tag{3.3}$$

En supposant que les pivots $a_{ii} \neq 0$, (3.3) s'écrit sous forme matricielle :

$$\begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \\ \vdots \\ x_n^{(i+1)} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ -\dfrac{a_{21}}{a_{22}} & 0 & 0 & \dots & 0 \\ -\dfrac{a_{31}}{a_{33}} & -\dfrac{a_{32}}{a_{33}} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \ddots & \\ -\dfrac{a_{n1}}{a_{nn}} & -\dfrac{a_{n2}}{a_{nn}} & & \dots & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \\ \vdots \\ x_n^{(i+1)} \end{pmatrix}$$

$$+\begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & -\frac{a_{13}}{a_{11}} & \dots & -\frac{a_{1n}}{a_{11}} \\ 0 & 0 & -\frac{a_{23}}{a_{22}} & \dots & -\frac{a_{2n}}{a_{22}} \\ 0 & 0 & 0 & \dots & -\frac{a_{3n}}{a_{33}} \\ \vdots & \vdots & & \ddots & \ddots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \\ \vdots \\ x_n^{(i)} \end{pmatrix} + \begin{pmatrix} \frac{b_1}{a_{11}} \\ \frac{b_2}{a_{22}} \\ \frac{b_3}{a_{33}} \\ \vdots \\ \frac{b_n}{a_{nn}} \end{pmatrix}$$

équivalente à la forme développée de la récurrence vectorielle suivante :

$$\begin{cases} x_1^{(i+1)} &= \Big(b_1 - a_{12}\, x_2^{(i)} - a_{13}\, x_3^{(i)} - \dots - a_{1n}\, x_n^{(i)}\Big)/a_{11}, \\ x_2^{(i+1)} &= \Big(b_2 - a_{21}\, x_1^{(i+1)} - a_{23}\, x_3^{(i)} - \dots - a_{2n}\, x_n^{(i)}\Big)/a_{22}, \\ & \dots\dots\dots\dots \\ x_n^{(i+1)} &= \Big(b_n - a_{n1}\, x_1^{(i+1)} - a_{n2}\, x_2^{(i+1)} - \dots - a_{nn-1}\, x_{n-1}^{(i+1)}\Big)/a_{nn}. \end{cases}$$

3.5.1 Convergence de la méthode de Gauss-Seidel

Théorème 3.6. ***Condition suffisante.*** *La méthode de Gauss-Seidel converge quel que soit le vecteur initial $\boldsymbol{x}^{(0)}$ pour les systèmes linéaires $A\boldsymbol{x} = \boldsymbol{b}$ dont la matrice A est à diagonale fortement dominante. Ceci se traduit par la condition suivante :*

$$\boxed{\sum_{j=1, i\neq j} |a_{ij}| < |a_{ii}| \quad \forall i = 1, 2, ..., n}$$

Ainsi, l'***algorithme de Gauss-Seidel*** devient :

$$\begin{cases} \text{Etant donnés } A,\ b,\ x^{(0)},\ imax,\ \varepsilon_1 \text{ et } \varepsilon_2,\ \forall i = 1, 2, ..., imax \\ x_j^{(i+1)} = \Big(b_j - \sum_{k=1}^{j-1} a_{jk}\, x_k^{(i+1)} - \sum_{k=j+1}^{n} a_{jk}\, x_k^{(i+1)}\Big)/a_{jj}, \quad \forall j = 1, 2, ..., n \\ \text{arrêter si : } |x_j^{(i+1)} - x_j^{(i)}| < \varepsilon_1 \text{ ou bien si } \dfrac{|x_j^{(i+1)} - x_j^{(i)}|}{|x_j^{(i+1)}|} < \varepsilon_2 \end{cases}$$

Exemple : On considère le système linéaire suivant :

$$\begin{cases} 5\ x_1 - 4\ x_2 + x_3 = 2, \\ -4\ x_1 + 4\ x_2 = 0, \\ x_1 + 2\ x_3 = 3. \end{cases}$$

1) Donner la matrice de la méthode de Jacobi et expliciter sous forme matricielle l'algorithme de cette méthode.

2) Donner la solution du système en partant du vecteur initial $x^0 = {}^t(0.9, 0.9, 0.9)$.

1) La matrice de Gauss-Seidel est donnée par :

$$D^{-1}L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix}$$

L'algorithme de Jacobi s'écrit sous la forme matricielle suivante :

$$\begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i+1)} \\ x_2^{(i+1)} \\ x_3^{(i+1)} \end{pmatrix} + \begin{pmatrix} 0 & \frac{4}{5} & -\frac{1}{5} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(i)} \\ x_2^{(i)} \\ x_3^{(i)} \end{pmatrix} + \begin{pmatrix} \frac{2}{5} \\ 0 \\ \frac{3}{2} \end{pmatrix}$$

équivalent à la forme développée suivante :

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

2) La solution du système linéaire par la méthode de Gauss-Seidel est :

n	x_1^n	x_2^n	x_3^n	n	x_1^n	x_2^n	x_3^n
0	0.9	0.9	0.9	33	0.9979401	0.9979401	1.001030
1	0.9400000	0.9400000	1.030000	34	0.9981461	0.9981461	1.000927
2	0.9460001	0.9460001	1.027000	35	0.9983315	0.9983315	1.000834
3	0.9514002	0.9514002	1.024300	36	0.9984984	0.9984984	1.000751
4	0.9562602	0.9562602	1.021870	37	0.9986486	0.9986486	1.000676
5	0.9606342	0.9606342	1.019683	38	0.9987838	0.9987838	1.000608
6	0.9645709	0.9645709	1.017715	39	0.9989054	0.9989054	1.000547
7	0.9681138	0.9681138	1.015943	40	0.9990149	0.9990149	1.000493
8	0.9713024	0.9713024	1.014349	41	0.9991134	0.9991134	1.000443
9	0.9741722	0.9741722	1.012914	42	0.9992021	0.9992021	1.000399
10	0.9767551	0.9767551	1.011622	43	0.9992819	0.9992819	1.000359
11	0.9790796	0.9790796	1.010460	44	0.9993538	0.9993538	1.000323
12	0.9811716	0.9811716	1.009414	45	0.9994184	0.9994184	1.000291
13	0.9830545	0.9830545	1.008473	46	0.9994766	0.9994766	1.000262
14	0.9847491	0.9847491	1.007625	47	0.9995289	0.9995289	1.000236
15	0.9862742	0.9862742	1.006863	48	0.9995761	0.9995761	1.000212
16	0.9876469	0.9876469	1.006177	49	0.9996185	0.9996185	1.000191
17	0.9888822	0.9888822	1.005559	50	0.9996567	0.9996567	1.000172
18	0.9899939	0.9899939	1.005003	51	0.9996911	0.9996911	1.000154
19	0.9909946	0.9909946	1.004503	52	0.9997220	0.9997220	1.000139
20	0.9918951	0.9918951	1.004052	53	0.9997498	0.9997498	1.000125
21	0.9927056	0.9927056	1.003647	54	0.9997749	0.9997749	1.000113
22	0.9934351	0.9934351	1.003282	55	0.9997974	0.9997974	1.000101
23	0.9940916	0.9940916	1.002954	56	0.9998177	0.9998177	1.000091
24	0.9946824	0.9946824	1.002659	57	0.9998359	0.9998359	1.000082
25	0.9952142	0.9952142	1.002393	58	0.9998524	0.9998524	1.000074
26	0.9956928	0.9956928	1.002154	59	0.9998671	0.9998671	1.000066
27	0.9961236	0.9961236	1.001938	60	0.9998804	0.9998804	1.000060
28	0.9965112	0.9965112	1.001744	61	0.9998924	0.9998924	1.000054
29	0.9968601	0.9968601	1.001570	62	0.9999031	0.9999031	1.000048
30	0.9971742	0.9971742	1.001413	63	0.9999127	0.9999127	1.000044
31	0.9974568	0.9974568	1.001272	64	***0.9999214***	***0.9999214***	***1.000039***
32	0.9977112	0.9977112	1.001144	65	***0.9999294***	***0.9999294***	***1.000035***

TABLE 3.3 – Vecteurs solutions obtenus par la méthode de Gauss-Seidel

Par suite, on en conclut que le vecteur solution obtenu avec un critère d'arrêt d'ordre 10^{-6} est :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.9999294 \\ 0.9999294 \\ 1.0000350 \end{pmatrix}$$

Sachant qu'à partir de $n = 108$, l'algorithme de la méthode de Gauss-Seidel converge vers le vecteur solution stable :

$$\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 0.9999996 \\ 0.9999996 \\ 1.0000000 \end{pmatrix}$$

Chapitre 4

Interpolation polynomiale

Dans ce chapitre, on s'intéresse à approcher une fonction f, connue par ses valeurs en certains points, par une fonction plus simple le plus souvent par un polynôme. Nous verrons dans ce contexte, l'***interpolation polynomiale***, par la méthode de Lagrange ou la méthode de Newton, qui consiste à rechercher un polynôme qui passe exactement par les points donnés, et l'***approximation au sens des moindres carrés*** ou on cherche à approcher au mieux la fonction.

Considérons le relevé expérimental de la pression atmosphérique en fonction de l'altitude au-dessus du sol. On a mesuré l'altitude (en mètres) et la pression atmosphérique (en hPa) de 10 différentes altitudes et on reporte les résultats dans le tableau suivant en prenant l'altitude en abscisse et la pression en ordonnée, comme ci-dessous

Altitude en m	**Pression atmosphérique en hPa**	**Altitude en m**	**Pression atmosphérique en hPa**
0	1 013.25	**4 000**	616.45
500	954.61	**5 000**	540.25
1 000	898.76	**6 000**	471.87
1 500	854.58	**7 000**	410.66
2 000	794.98	**8 000**	356.06
2 500	746.86	**9 000**	307.48
3 000	701.12	**10 000**	264.42
3 500	657.68	**11 000**	226.37

TABLE 4.1 – Relevé expérimental de la "variation de la pression atmosphérique en fonction de l'altitude"

A des altitudes discrètes, notées t_i, on mesure les pressions P_i. Sachant que $P = f(t)$, il faut pouvoir comparer les P_i à $f(t_i)$. Ainsi l'approximation de f par un polynôme, qui reste la fonction la plus simples que l'on puisse construire, s'impose naturellement. Parmi les méthodes les plus courantes, on cite celle de Lagrange et celle de Newton. Plus tard, on présentera une autre approche d'approximation, celle de la méthode d'approximation au sens de moindres carrés

4.1 Principe de l'interpolation polynomiale

Étant donné un ensemble de $(n+1)$ points expérimentaux $(t_i; f(t_i))$, l'interpolation polynomiale consiste à approcher le nuage de points $(t_i; f(t_i))$ par un polynôme p_n de degré n. Le polynôme p_n a la particularité de passer par tous les points donnés et permet également d'estimer la valeur de la fonction f en une valeur de $t \notin (t_i)_{\{i=1,...,n+1\}}$.

Le polynôme p_n est le ***polynôme d'interpolation de la fonction*** f ***en*** $t_0, t_1, ...t_{n+1}$. Il est de la forme suivante :

$$\boxed{p_n(t) = a_0 + a_1\, t + a_2\, t^2 + ... + a_{n-1}\, t^{n-1} + a_n\, t^n}$$

Le polynôme p_n est exprimé dans la base canonique $\{1, t, t^2, ..., t^n\}$.

Dès lors, il est naturel de considérer la question suivante : ***se donner des critères d'existence et d'unicité pour le polynôme*** p_n.

Du fait que le polynôme p_n passe par les $(n+1$ points $(t_i; f(t_i))$ donnés, on a :

$$\boxed{p_n(t_i) \;=\; f(t_i), \quad \forall i \in \{1, ..., n+1\}}$$

Par suite, les coefficients a_0, a_1,...,a_n vérifient le système linéaire suivant à $(n+1)$ équations et $(n+1)$ inconnues :

$$f(t_i) \;=\; \sum_{j=0}^{n} a_j t_i^j, \quad \forall i \in \{1, ..., n+1\}.$$

ou sous forme développée :

$$\begin{cases} a_0 + a_1\, t_1 & + a_2\, t_1^2 & + \ldots & + a_n\, t_1^n & = f(t_1), \\ a_0 + a_1\, t_2 & + a_2\, t_2^2 & + \ldots & + a_n\, t_2^n & = f(t_2), \\ a_0 + a_1\, t_3 & + a_2\, t_3^2 & + \ldots & + a_n\, t_3^n & = f(t_3), \\ \quad \ldots\ldots\ldots\ldots & & & & \\ a_0 + a_1\, t_n & + a_2\, t_n^2 & + \ldots & + a_n\, t_n^n & = f(t_n), \\ a_0 + a_1\, t_{n+1} & + a_2\, t_{n+1}^2 & + \ldots & + a_n\, t_{n+1}^n & = f(t_{n+1}). \end{cases}$$

Ce système est dit de '***Cramer***'. En notation matricielle, on le note :

$$\begin{pmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^n \\ 1 & t_2 & t_2^2 & \ldots & t_2^n \\ 1 & t_3 & t_3^2 & \ldots & t_3^n \\ & \ldots\ldots & & & \\ 1 & t_n & t_n^2 & \ldots & t_n^n \\ 1 & t_{n+1} & t_{n+1}^2 & \ldots & t_{n+1}^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_{n-1} \\ a_n \end{pmatrix} = \begin{pmatrix} f(t_1) \\ f(t_2) \\ f(t_3) \\ \vdots \\ f(t_n) \\ f(t_{n+1}) \end{pmatrix}$$

C'est un système linéaire de déterminant :

$$\Delta \;=\; \begin{vmatrix} 1 & t_1 & t_1^2 & \ldots & t_1^n \\ 1 & t_2 & t_2^2 & \ldots & t_2^n \\ 1 & t_3 & t_3^2 & \ldots & t_3^n \\ \ldots\ldots & & & & \\ 1 & t_n & t_n^2 & \ldots & t_n^n \\ 1 & t_{n+1} & t_{n+1}^2 & \ldots & t_{n+1}^n \end{vmatrix} = \prod_{1 \le i < j \le n+1} (t_i - t_j).$$

Le déterminant de la matrice de cette équation est appelée ***déterminant de Vandermonde***.

On a :

Théorème 4.1. *Le système de Cramer admet une solution unique si et seulement si les réels* $(t_i)_{i=1}^{n+1}$ *sont distincts deux à deux* ($\Delta \neq 0$).

Théorème 4.2. *Il existe un seul polynôme* p_n *de degré* n *interpolant une série de* $(n+1)$ *points donnés tels que* $t_i \neq t_j$ *si* $i \neq j$.

Preuve 4.1. *L'unicité résulte aussi du fait que s'il existe deux polynômes* P *et* R *de degré* n *et coïncidant en* $(n+1)$ *points tels que :*

$$P(t_i) = f(t_i) \;\; et \;\; Q(t_i) = f(t_i), \;\; \forall i \in \{1, 2, ..., n+1\},$$

alors on a :

$$R(t_i) = P(t_i) \; - \; Q(t_i) = 0, \;\; \forall i \in \{1, 2, ..., n+1\}.$$

Le polynôme R_n *dont le degré est au plus* n, *a donc* $(n+1)$ *zéros distincts deux à deux. Il est donc identiquement nul. La matrice est une matrice du type matrice de Vandermonde. Son déterminant est non nul, ce qui prouve le théorème d'unisolvance.*

Remarque 4.1. *On peut en revanche trouver à l'aide de la méthode des moindres carrés une infinité de polynômes de degré inférieur à* n *approchant le nuage des* $(n+1)$ *points.*

4.2 Polynôme d'interpolation de Lagrange

La technique d'interpolation polynomiale de Lagrange, du nom de ***Joseph-Louis Lagrange***, a été découverte par ***Edward Waring*** en 1779 et plus tard par ***Leonhard Euler*** en 1783. Le polynôme p est exprimé dans la base de Lagrange.

Définition 4.1. *On appelle* ***base de Lagrange relative aux points*** t_i *les polynômes* L_i, $i \in \{1, 2, ..., n+1\}$ *:*

$$\boxed{L_i(t) \; = \; \prod_{j=1, j\neq i}^{n+1} \frac{t - t_j}{t_i - t_j} \; = \; \frac{(t-t_1) \; ... \; (t-t_{i-1}) \; (t-t_{i+1}) \; ... \; (t-t_{n+1})}{(t_i-t_1) \; ... \; (t_i-t_{i-1}) \; (t_i-t_{i+1}) \; ... \; (t_i-t_{n+1})}}$$

Le polynôme $L_i(t)$ *est le polynôme de Lagrange associé au point* $(t_i, f(t_i))$ *et est de degré* n. *De ce fait, on en déduit immédiatement le résultat suivant :*

$$\boxed{L_i(t_j) \; = \; \delta_{i,j}, \;\; 1 \leq i, j \leq n+1, \text{ c-à-d } \; L_i(t_i) = 1 \; \text{ et } \; L_i(t_j) = 0, \; i \neq j}$$

Il vient alors :

Proposition 4.1. *Le polynôme qui interpole les* $(n+1)$ *valeurs* $f(t_i)$ *aux points* t_i *s'écrit :*

$$\boxed{p_n(t) \; = \; \sum_{i=1}^{n+1} f(t_i) \; L_i(t)}$$

Le polynôme $p_n(t)$ s'appelle le polynôme d'interpolation de Lagrange qui s'exprime sous forme d'une combinaison linéaire des $(n+1)$ polynômes $L_i(t)$. En outre, $p_n(t)$ est l'unique polynôme de degré au plus n vérifiant

$$p_n(t_i) \;=\; f(t_i), \;\; \forall i \in \{1, 2, ..., n+1\}.$$

Exemple : Construire le polynôme d'interpolation de Lagrange associé aux points $(0, -2)$, $(1, -1)$, $(3, 3)$, $(4, 5)$. En déduire une approximation de $f(0.5)$, $f'(0.5)$ et $\int_0^4 f(t)\, dt$.

Dans ce cas, on a 4 points et on cherche par suite un polynôme de degré inférieur ou égal à 3 de la forme :

$$p_3(t) \;=\; \sum_{i=1}^{4} f(t_i)\; L_i(t)$$

avec

$$L_1(t) \;=\; \frac{(t-t_2)\,(t-t_3)\,(t-t_4)}{(t_1-t_2)\,(t_1-t_3)\,(t_1-t_4)} \;=\; \frac{(t-1)\,(t-3)\,(t-4)}{(0-1)\,(0-3)\,(0-4)} \;=\; \frac{t^3-8t^2+19t-12}{-12},$$

$$L_2(t) \;=\; \frac{(t-t_1)\,(t-t_3)\,(t-t_4)}{(t_2-t_1)\,(t_2-t_3)\,(t_2-t_4)} \;=\; \frac{(t-0)\,(t-3)\,(t-4)}{(1-0)\,(1-3)\,(1-4)} \;=\; \frac{t^3-7t^2+12t}{6},$$

$$L_3(t) \;=\; \frac{(t-t_1)\,(t-t_2)\,(t-t_4)}{(t_3-t_1)\,(t_3-t_2)\,(t_3-t_4)} \;=\; \frac{(t-0)\,(t-1)\,(t-4)}{(3-0)\,(3-1)\,(3-4)} \;=\; \frac{t^3-5t^2+4t}{-6}$$

$$L_4(t) \;=\; \frac{(t-t_1)\,(t-t_2)\,(t-t_3)}{(t_4-t_1)\,(t_4-t_2)\,(t_4-t_3)} \;=\; \frac{(t-0)\,(t-1)\,(t-3)}{(4-0)\,(4-1)\,(4-3)} \;=\; \frac{t^3-4t^2+3t}{12}.$$

Il s'en suit alors :

$$\begin{aligned} p_3(t) \;&=\; f(t_1)\,L_1(t) + f(t_2)\,L_2(t) + f(t_3)\,L_3(t) + f(t_4)\,L_4(t) \\ &=\; (-2)\times\left(\frac{t^3-8t^2+19t-12}{-12}\right) + (-1)\times\left(\frac{t^3-7t^2+12t}{6}\right) + (3)\times\left(\frac{t^3-5t^2+4t}{-6}\right) \\ &\quad +(5)\times\left(\frac{t^3-4t^2+3t}{12}\right) \;=\; \frac{-t^3+12t^2+5t-24}{12}. \end{aligned}$$

Sachant que $f(t) \simeq p_3(t)$ et $f'(t) \simeq p'_3(t) = \dfrac{-3t^2+24t+5}{12}$ sur l'intervalle $[0, 4]$, on a

$$\begin{aligned} f(0.5) \;&\simeq p_3(0.5) \;&=\; \frac{-(0.5)^3+12\times(0.5)^2+5\times(0.5)-24}{12} \;=\; \frac{-149}{96}, \\ f'(0.5) \;&\simeq p'_3(0.5) \;&=\; \frac{-3\times(0.5)^2+24\times(0.5)+5}{12} \;=\; \frac{65}{48}, \\ \int_0^4 f(t)\,dt \;&\simeq \int_0^4 p_3(t)\,dt \;&=\; \int_0^4 \frac{-t^3+12t^2+5t-24}{12}\,dt \;=\; \frac{104}{3}. \end{aligned}$$

Exemple : Construire le polynôme d'interpolation de Lagrange associé aux points $(-2, -1)$, $(-1, 0)$, $(0, 1)$, $(1, 2)$ et $(2, 3)$.

Dans ce cas, on a 5 points et on cherche par suite un polynôme de degré inférieur ou égal à 4 de la forme :

$$p_4(t) \;=\; \sum_{i=1}^{5} f(t_i)\; L_i(t)$$

avec

$$\begin{aligned}
L_1(t) &= \frac{(t-t_2)\,(t-t_3)\,(t-t_4)\,(t-t_5)}{(t_1-t_2)\,(t_1-t_3)\,(t_1-t_4)\,(t_1-t_5)} = \frac{(t+1)\,(t)\,(t-1)\,(t-2)}{(-2+1)\,(-2-0)\,(-2-1)\,(-2-2)} \\
&= \frac{t^4-2t^3-t^2+2t}{24}, \\
L_2(t) &= \frac{(t-t_1)\,(t-t_3)\,(t-t_4)\,(t-t_5)}{(t_2-t_1)\,(t_2-t_3)\,(t_2-t_4)\,(t_2-t_5)} = \frac{(t+2)\,(t)\,(t-1)\,(t-2)}{(-1+2)\,(-1-0)\,(-1-1)\,(-1-2)} \\
&= \frac{-t^4+t^3+4t^2-4t}{6}, \\
L_3(t) &= \frac{(t-t_1)\,(t-t_2)\,(t-t_4)\,(t-t_5)}{(t_3-t_1)\,(t_3-t_2)\,(t_3-t_4)\,(t_3-t_5)} = \frac{(t+2)\,(t+1)\,(t-1)\,(t-2)}{(0-2)\,(0+1)\,(0-1)\,(0-2)} \\
&= \frac{t^4-5t^2+4}{4}, \\
L_4(t) &= \frac{(t-t_1)\,(t-t_2)\,(t-t_3)\,(t-t_5)}{(t_4-t_1)\,(t_4-t_2)\,(t_4-t_3)\,(t_4-t_5)} = \frac{(t+2)\,(t+1)\,(t)\,(t-2)}{(1+2)\,(1+1)\,(1-0)\,(1-2)} \\
&= \frac{t^4+t^3-4t^2-4t}{-6}, \\
L_5(t) &= \frac{(t-t_1)\,(t-t_2)\,(t-t_3)\,(t-t_4)}{(t_5-t_1)\,(t_5-t_2)\,(t_5-t_3)\,(t_5-t_4)} = \frac{(t+2)\,(t+1)\,(t-0)\,(t-1)}{(2+2)\,(2+1)\,(2-)\,(2-1)} \\
&= \frac{t^4+2t^3-t^2-2t}{24}.
\end{aligned}$$

Il s'en suit alors :

$$\begin{aligned}
p_4(t) &= f(t_1)\,L_1(t) + f(t_2)\,L_2(t) + f(t_3)\,L_3(t) + f(t_4)\,L_4(t) + f(t_5)\,L_5(t) \\
&= (-1)\times\left(\frac{t^4-2t^3-t^2+2t}{24}\right) + (0)\times\left(\frac{-t^4+t^3+4t^2-4t}{6}\right) + (1)\times\left(\frac{t^4-5t^2+4}{4}\right) \\
&\quad +(2)\times\left(\frac{t^4+t^3-4t^2-4t}{-6}\right) + (3)\times\left(\frac{t^4+2t^3-t^2-2t}{24}\right) \;=\; t+1.
\end{aligned}$$

4.3 Méthode d'interpolation de Newton

Définition 4.2. ***Différences divisées*** *Étant donnés les points $(t_i, f(t_i))_{\{1\le i\le n+1\}}$, on définit les différences divisées de la fonction f aux points $(t_i)_{\{1\le i\le n+1\}}$ par les relations de récurrence*

suivantes :

$$
\begin{aligned}
\delta(t_1) &:= f(t_1),\\
\delta_1(t_1,t_2) &:= \frac{\delta(t_1)-\delta(t_2)}{t_1-t_2} = \frac{f(t_1)-f(t_2)}{t_1-t_2},\\
\delta_2(t_1,t_2,t_3) &:= \frac{\delta_1(t_1,t_2)-\delta_1(t_2,t_3)}{t_1-t_3},\\
\delta_3(t_1,t_2,t_3,t_4) &:= \frac{\delta_2(t_1,t_2,t_3)-\delta_2(t_2,t_3,t_4)}{t_1-t_4},\\
&\vdots\\
\delta_{n-1}(t_1,t_2,...,t_n) &:= \frac{\delta_{n-2}(t_1,t_2,...,t_{n-1})-\delta_{n-2}(t_2,t_3,...,t_n)}{t_1-t_n},\\
\delta_n(t_1,t_2,...,t_{n+1}) &:= \frac{\delta_{n-1}(t_1,t_2,...,t_n)-\delta_{n-1}(t_2,t_3,...,t_{n+1})}{t_1-t_{n+1}}.
\end{aligned}
$$

$\delta_n(t_1,t_2,...,t_{n+1})$ *est appelée* ***différence divisée d'ordre*** n ***de la fonction*** f ***aux points*** $t_1,t_2,...,t_{n+1}$.

Vu que le calcul de $\delta_n(t_1,t_2,...,t_{n+1})$ implique le calcul de plusieurs différences divisées d'ordre différents et afin de faciliter le calcul des valeurs, on dressera la table des différences divisées suivante :

t_i	$f(t_i)$	δ_1	δ_2	$\dots$	δ_{n-1}	δ_n
t_1	$f(t_1)=\boxed{\delta(t_1)}$					
		$\boxed{\delta_1(t_1,t_2)}$				
t_2	$f(t_2)=\delta(t_2)$		$\boxed{\delta_2(t_1,t_2,t_3)}$			
		$\delta_1(t_2,t_3)$				
t_3	$f(t_3)=\delta(t_3)$		$\delta_2(t_2,t_3,t_4)$			
			$\delta_1(t_3,t_4)$			
t_4	$f(t_4)=\delta(t_4)$					
					$\boxed{\delta_{n-1}(t_1,t_2,...,t_n)}$	
$\vdots$	$\vdots$					$\boxed{\delta_n(t_1,t_2,...,t_{n+1})}$
					$\delta_{n-1}(t_2,t_3,...,t_{n+1})$	
t_n	$f(t_n)=\delta(t_n)$		$\delta_1(t_{n-1},t_n,t_{n+1})$			
		$\delta_1(t_n,t_{n+1})$				
t_{n+1}	$f(t_{n+1})=\delta(t_{n+1})$					

Seules les valeurs encadrées sur la diagonale interviennent dans l'expression du polynôme d'interpolation

Théorème 4.3. ***Formule de Newton (1669)*** *Le polynôme d'interpolation de degré n qui passe par les $n+1$ points $(t_1, f(t_1))$, $(t_2, f(t_2))$,...,$(t_{n+1}, f(t_{n+1}))$, où les t_i sont distincts deux à deux, est unique et donné par*

$$\boxed{p_n(t) \;=\; \delta(t_1) + (t-t_1)\,\delta_1(t_1,t_2) + (t-t_1)\,(t-t_2)\,\delta_2(t_1,t_2,t_3) + ... + \prod_{i=1}^{n}(t-t_i)\,\delta_n(t_1,t_2,...,t_{n+1})}$$

Théorème 4.4. *Soit f une fonction continue sur l'intervalle $I = [\min(t_1,t_2,...,t_{n+1}), \max(t_1,t_2,...,t_{n+1})]$ et les t_i sont distincts deux à deux. On a alors :*

$$\begin{aligned} f(t) \;&=\; \delta(t_1) + (t-t_1)\,\delta_1(t_1,t_2) + (t-t_1)\,(t-t_2)\,\delta_2(t_1,t_2,t_3) + ... + \textstyle\prod_{i=1}^{n}(t-t_i)\,\delta_n(t_1,t_2,...,t_{n+1}) \\ &\quad + \textstyle\prod_{i=1}^{n}(t-t_i)\,\delta_n(t,t_1,t_2,...,t_{n+1}). \end{aligned}$$

Théorème 4.5. *Soit l'intervalle $I = [\min(t_1,t_2,...,t_{n+1}), \max(t_1,t_2,...,t_{n+1})]$, contenant les points t_i distincts deux à deux, et f une fonction $(n+1)$ fois différentiable sur I. Alors il existe $\xi \in I$ tel que :*

$$\boxed{\delta_n(t_1,t_2,...,t_{n+1}) \;=\; \frac{f^{(n)}(\xi)}{n!}}$$

Exemple : On considère la fonction $f(t)$ définie sur l'intervalle $[0,4]$ par la table des valeurs

t_i	0	1	3	4
$f(t_i)$	-2	-1	3	5

Écrire le polynôme d'interpolation de Newton de la fonction f associé aux points $t_1 = 0$, $t_2 = 1$, $t_3 = 3$, $t_4 = 4$.

Le polynôme d'interpolation de Newton, de degré inférieur ou égal à 3, est donné par :

$$\begin{aligned} p_3(t) \;&=\; \delta(t_1) + (t-t_1)\,\delta_1(t_1,t_2) + (t-t_1)\,(t-t_2)\,\delta_2(t_1,t_2,t_3) + (t-t_1)\,(t-t_2)\,(t-t_3)\,\delta_3(t_1,t_2,t_3,t_4) \\ &=\; \delta(0) + (t-0)\,\delta_1(0,1) + (t-0)\,(t-1)\,\delta_2(0,1,3) + (t-0)\,(t-1)\,(t-3)\,\delta_3(0,1,3,4). \end{aligned}$$

les différences divisées sont présentées dans le tableau suivant

t_i	$f(t_i)$		δ_1	δ_2	δ_3
0	-2	$=$ $\boxed{\delta(0)}$			
			$\boxed{\delta_1(0,1)}$		
1	-1	$=\delta(1)$		$\boxed{\delta_2(0,1,3)}$	
			$\delta_1(1,3)$		$\boxed{\delta_3(0,1,3,4)}$
3	3	$=\delta(3)$		$\delta_2(1,3,4)$	
			$\delta_1(3,4)$		
4	5	$=\delta(4)$			

avec

$$\begin{cases} \delta_1(0,1) = \dfrac{\delta(0)-\delta(1)}{0-1} = \dfrac{-2+1}{-1} = 1, \\ \delta_1(1,3) = \dfrac{\delta(1)-\delta(3)}{1-3} = \dfrac{-1-3}{1-3} = 2, \\ \delta_1(3,4) = \dfrac{\delta(3)-\delta(4)}{3-4} = \dfrac{3-5}{3-4} = 2, \\ \delta_2(0,1,3) = \dfrac{\delta_1(0,1)-\delta_1(1,3)}{0-3} = \dfrac{1-2}{0-3} = \dfrac{1}{3}, \\ \delta_2(1,3,4) = \dfrac{\delta_1(1,3)-\delta_1(3,4)}{1-4} = \dfrac{2-2}{1-4} = 0, \\ \delta_3(0,1,3,4) = \dfrac{\delta_2(-0,1,3)-\delta_2(1,3,4)}{0-4} = \dfrac{\frac{1}{3}-0}{0-4} = -\dfrac{1}{12}. \end{cases}$$

il vient alors

$$p_3(t) = \frac{-t^3+8t^2+5t-24}{12}.$$

Exemple : On considère la fonction $f(t)$ définie sur l'intervalle $[-2,2]$ par la table des valeurs

t_i	-2	-1	0	1	2
$f(t_i)$	-1	0	1	2	3

Écrire le polynôme d'interpolation de Newton de la fonction f associé aux points $t_1=-2$, $t_2=-1$, $t_3=0$, $t_4=1$ et $t_5=2$. En déduire une valeur approchée de $f(0,5)$.

Le polynôme d'interpolation de Newton, de degré inférieur ou égal à 4, est donné par :

$$\begin{aligned} p_4(t) &= \delta(t_1)+(t-t_1)\,\delta_1(t_1,t_2)+(t-t_1)(t-t_2)\,\delta_2(t_1,t_2,t_3)+(t-t_1)(t-t_2)(t-t_3)\,\delta_3(t_1,t_2,t_3,t_4) \\ &\quad +(t-t_1)(t-t_2)(t-t_3)(t-t_4)\,\delta_4(t_1,t_2,t_3,t_4,t_5) \\ &= \delta(-2)+(t+2)\,\delta_1(-2,-1)+(t+2)(t+1)\,\delta_2(-2,-1,0)+(t+2)(t+1)(t)\,\delta_3(-2,-1,0,1) \\ &\quad +(t+2)(t+1)(t)(t-1)\,\delta_4(-2,-1,0,1,2). \end{aligned}$$

les différences divisées sont présentées dans le tableau suivant

t_i	$f(t_i)$		δ_1	δ_2	δ_3	δ_4
-2	-1	$= \boxed{\delta(-2)}$				
			$\boxed{\delta_1(-2,-1)}$			
-1	0	$= \delta(-1)$		$\boxed{\delta_2(-2,-1,0)}$		
			$\delta_1(-1,0)$		$\boxed{\delta_3(-2,-1,0,1)}$	
0	1	$= \delta(0)$		$\delta_2(-1,0,1)$		
			$\delta_1(0,1)$		$\delta_3(-1,0,1,2)$	
1	2	$= \delta(1)$		$\delta_2(0,1,2)$		$\boxed{\delta_4(-2,-1,0,1,2)}$
			$\delta_1(1,2)$			
2	3	$= \delta(2)$				

avec

$$
\begin{cases}
\delta_1(-2,-1) = \dfrac{\delta(-2)-\delta(-1)}{-2-(-1)} = \dfrac{-1-0}{-2-(-1)} = 1, \\
\delta_1(-1,0) = \dfrac{\delta(-1)-\delta(0)}{-1-0} = \dfrac{0-1}{-1-0} = 1, \\
\delta_1(0,1) = \dfrac{\delta(0)-\delta(1)}{0-1} = \dfrac{1-2}{0-1} = 1, \\
\delta_1(1,2) = \dfrac{\delta(1)-\delta(2)}{1-2} = \dfrac{2-3}{1-2} = 1, \\
\delta_2(-2,-1,0) = \dfrac{\delta_1(-2,-1)-\delta_1(-1,0)}{-2-0} = \dfrac{1-1}{-2-0} = 0, \\
\delta_2(-1,0,1) = \dfrac{\delta_1(-1,0)-\delta_1(0,1)}{-1-1} = \dfrac{1-1}{-1-1} = 0, \\
\delta_2(0,1,2) = \dfrac{\delta_1(0,1)-\delta_1(1,2)}{0-2} = \dfrac{1-1}{0-2} = 0, \\
\delta_3(-2,-1,0,1) = \dfrac{\delta_2(-2,-1,0)-\delta_2(-1,0,1)}{-2-1} = \dfrac{0-0}{-2-1} = 0, \\
\delta_3(-1,0,1,2) = \dfrac{\delta_2(-1,0,1)-\delta_2(0,1,2)}{-1-2} = \dfrac{0-0}{-1-2} = 0, \\
\delta_4(-2,-1,0,1,2) = \dfrac{\delta_3(-2,-1,0,1)-\delta_3(-1,0,1,2)}{-2-2} = \dfrac{0-0}{-2-2} = 0.
\end{cases}
$$

il vient alors

$$p_4(t) \;=\; -1+(t+2) \;=\; t+1.$$

4.4 Étude de l'erreur d'interpolation

L'interpolation polynomiale est une opération mathématique permettant de construire un polynôme à partir de la donnée d'un nombre fini de points dits aussi ***noeuds***. La solution du problème d'interpolation passe par les points prescrits. On approche ainsi la fonction f par le polynôme de degré aussi grand que nécessaire afin d'estimer localement f. On parle alors d'une méthode itérative et éventuellement des erreurs de troncature commises.

Remarque 4.2. *Étant donné la valeur connue d'une certaine fonction f en un certain nombre de points, le polynôme d'interpolation construit permet d'évaluer la valeur que peut prendre cette fonction en un point quelconque situé dans l'intervalle des points donnés. Ainsi, si on désire déterminer la valeur de cette fonction en un point situé hors de l'intervalle des points connus, on emploiera la* ***méthode d'extrapolation*** *qui est très similaire à la méthode d'interpolation.*

Théorème 4.6. *Étant donné la valeur connue d'une certaine fonction f en un certain nombre de points $(t_i, f(t_i))_{1\leq i\leq n+1}$, soit l'intervalle $I = [\min(t_1, t_2, ..., t_{n+1})], \max(t_1, t_2, ..., t_{n+1})]$. On suppose que f est $(n+1)$ fois différentiable sur l'intervalle I. Alors pour tout $t \in I$, il existe $\xi \in I$, tel que :*

$$\boxed{f(t) - p_n(t) \;=\; \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=1}^{n+1}(t - t_i)}$$

La formule précédente ne permet pas le calcul d'une valeur exacte de l'erreur étant donné que ξ est inconnu. Cependant, on peut en déduire une majoration, d'où

Lemme 4.1. *Étant donné la valeur connue d'une certaine fonction f en un certain nombre de points $(t_i, f(t_i))_{1\leq i\leq n+1}$, soit l'intervalle $I = [\min(t_1, t_2, ..., t_{n+1})], \max(t_1, t_2, ..., t_{n+1})]$. On suppose que f est $(n+1)$ fois différentiable sur l'intervalle I. Alors pour tout $t \in I$, on a :*

$$\boxed{|f(t) - p_n(t)| \leq \frac{M^{(n+1)}}{(n+1)!} \left| \prod_{i=1}^{n+1}(t - t_i) \right| \quad \text{où} \quad M^{(n+1)} \;=\; \max_{t\in I} |f^{(n+1)}(t)|}$$

Soit l'intervalle $I = [a, b]$, remarquons que l'erreur d'interpolation est le produit de deux expressions : l'une dépendant de f et l'autre de la subdivision choisie. Il est donc naturel de chercher pour quels points t_i la quantité $\prod_{i=1}^{n+1}(t - t_i)$ est minimale. Les ***polynômes de Tchebytcheff*** permettent de résoudre ce problème. En fait, Le choix des points d'interpolation est donné par :

$$t_i \;=\; \frac{a+b}{2} \;+\; \frac{b-a}{2} \cos\left(\frac{(2i-1)\Pi}{2n+2}\right), \quad i \in \{1, 2, ..., n+1\}.$$

Exemple : Avec quelle précision peut-on calculer $\sqrt{15}$ à l'aide de la formule d'erreur d'interpolation de la méthode de Lagrange pour la fonction $f(t) = \sqrt{t}$ associée aux points d'interpolations suivants : $t_1 = 9$, $t_2 = 16$ et $t_3 = 25$.

Notons $p_2(t)$ le polynôme d'interpolation de Lagrange, de degré inférieur ou égal à 2, interpolant la fonction f aux points t_1, t_2 et t_3.

L'erreur au point $t = 15$ est donnée par la formule suivante avec $n = 2$:

$$|f(15) - p_2(15)| = \frac{M^{(3)}}{(3)!} \left| \prod_{i=1}^{3} (15 - t_i) \right| \text{ où } M^{(3)} = \max_{t \in [9,25]} \left| f^{(3)}(t) \right|.$$

D'une part, sachant que

$$f'(t) = \frac{1}{2\sqrt{t}}, \quad f''(t) = \frac{-1}{4t\sqrt{t}}, \text{ et } f^{(3)}(t) = \frac{3}{8t^2\sqrt{t}},$$

le maximum de la fonction $f^{(3)}(t)$ est atteint pour $t = 9$, d'où

$$M^{(3)} = \frac{3}{8\sqrt{9^5}} = \frac{1}{648}.$$

D'autre part, on trouve

$$\prod_{i=1}^{3} (15 - t_i) = (15 - t_1)(15 - t_2)(15 - t_3) = (15 - 9)(15 - 16)(15 - 25) = 60.$$

Il vient alors

$$|f(15) - p_2(15)| \leq \frac{60}{648} = 0.154320.10^{-1}.$$

Par suite, $p_2(15)$ est une valeur approchée de $f(15)$, calculée avec au moins une décimale exacte.

Exemple : On considère la fonction $f(t) = \dfrac{4}{1-t}$.

1 Écrire le polynôme d'interpolation de Lagrange, dont on précisera le degré, associé aux abscisses $t_1 = -1$, $t_2 = 0$, $t_3 = 2$ et $t_4 = 3$.

2 Évaluer l'erreur au point $t = \sqrt{5}$.

3 Tracer les courbes $f(t)$ et $p_3(t)$.

1 Dans ce cas, on a 4 points et on cherche par suite un polynôme de degré inférieur ou égal à 3 de la forme :

$$p_3(t) = \sum_{i=1}^{4} f(t_i)\, L_i(t)$$

avec

$$L_1(t) = \frac{(t-t_2)\,(t-t_3)\,(t-t_4)}{(t_1-t_2)\,(t_1-t_3)\,(t_1-t_4)} = \frac{(t)\,(t-2)\,(t-3)}{(-1)\,(-3)\,(-4)} = \frac{-t^3+5t^2-6t}{12},$$

$$L_2(t) = \frac{(t-t_1)\,(t-t_3)\,(t-t_4)}{(t_2-t_1)\,(t_2-t_3)\,(t_2-t_4)} = \frac{(t+1)\,(t-2)\,(t-3)}{(1)\,(-2)\,(-3)} = \frac{t^3-4t^2+t+6}{6},$$

$$L_3(t) = \frac{(t-t_1)\,(t-t_2)\,(t-t_4)}{(t_3-t_1)\,(t_3-t_2)\,(t_3-t_4)} = \frac{(t+1)\,(t-2)\,(t-3)}{(3)\,(2)\,(-1)} = \frac{-t^3+2t^2+3t}{6},$$

$$L_4(t) = \frac{(t-t_1)\,(t-t_2)\,(t-t_3)}{(t_4-t_1)\,(t_4-t_2)\,(t_4-t_3)} = \frac{(t)\,(t+1)\,(t-2)}{(4)\,(3)\,(1)} = \frac{t^3-t^2-2t}{12}.$$

Sachant que

$$f(t_1) = f(-1) = 2,\ f(t_2) = f(0) = 4,\ f(t_3) = f(2) = -4 \text{ et } f(t_4) = f(3) = -2,$$

il s'en suit alors :

$$\begin{aligned} p_3(t) &= f(t_1)\,L_1(t) + f(t_2)\,L_2(t) + f(t_3)\,L_3(t) + f(t_4)\,L_4(t) \\ &= (2) \times \left(\frac{-t^3+5t^2-6t}{12}\right) + (4) \times \left(\frac{t^3-4t^2+t+6}{6}\right) + (-4) \times \left(\frac{-t^3+2t^2+3t}{6}\right) \\ &\quad +(-2) \times \left(\frac{t^3-t^2-2t}{12}\right) \\ &= t^3 - 3t^2 - 2t + 4. \end{aligned}$$

2 L'erreur au point $t = \sqrt{5}$ est donnée par la formule suivante avec $n = 3$:

$$\left|f(\sqrt{5}) - p_3(\sqrt{5})\right| = \frac{M^{(4)}}{(4)!}\left|\prod_{i=1}^{4}(\sqrt{5}-t_i)\right| \text{ où } M^{(4)} = \max_{t\in[-1,3]}\left|f^{(4)}(t)\right|.$$

D'une part, sachant que

$$f'(t) = \frac{4}{(1-t)^2},\ f''(t) = \frac{8}{(1-t)^3},\ f^{(3)}(t) = \frac{24}{(1-t)^4} \text{ et } f^{(4)}(t) = \frac{96}{(1-t)^5},$$

le maximum de la fonction $f^{(4)}(t)$ est atteint pour $t = 0$, d'où

$$M^{(4)} = 96.$$

D'autre part, on trouve

$$\begin{aligned} \textstyle\prod_{i=1}^{4}(\sqrt{5}-t_i) &= (\sqrt{5}-t_1)(1+\sqrt{5}-t_2)(\sqrt{5}-t_3)(\sqrt{5}-t_4) \\ &= (\sqrt{5}+1)(\sqrt{5}-0)(\sqrt{5}-2)(\sqrt{5}-3) = 30-14\sqrt{5}. \end{aligned}$$

Il vient alors

$$\left|f(\sqrt{5}) - p_3(\sqrt{5})\right| \leq \frac{|30-14\sqrt{5}|}{4}.$$

Or $f(\sqrt{5}) = \dfrac{4}{1-\sqrt{5}}$ et $p_3(\sqrt{5}) = 3\sqrt{5}-11$, d'où $\left|f(\sqrt{5}) - p_3(\sqrt{5})\right| = \dfrac{4}{\sqrt{5}} \leq \dfrac{|30-14\sqrt{5}|}{4}$.

L'interpolation polynômiale d'un grand nombre de points peut présenter des oscillations. C'est le ***phénomène de Runge*** qui s'explique par le fait que le polynôme d'interpolation p_n ne converge pas toujours uniformément vers la fonction f quel que soit le choix des points d'interpolation (sauf lorsque les t_i sont les racines du polynôme de Tchebytcheff).

Pour l'éviter, il est préférable d'employer une ***approximation ou interpolation par des fonctions polynômiales par morceaux*** qui consiste à subdiviser l'intervalle $I = [\min(t_1, t_2, ..., t_{n+1}), \max(t_1, t_2, ..., t_{n+1})]$ en plusieurs sous-intervalles et d'appliquer les méthodes d'interpolations sur

chacun de ces sous-intervalles, ce qui permet de construire un polynôme d'interpolation globalement continu dont la restriction à chaque intervalle $[t_i, t_{i+1}]$ est polynômiale.

Parmi les approximations les plus courantes, on cite l'***approximation linéaire par morceaux***, les ***fonctions splines*** où la fonction d'approximation est non seulement à dérivée continue, mais à dérivée seconde continue ou encore l'***approximation cubique*** où les polynômes de base sont de degré 3, on parle alors de l'interpolation d'Hermite :

4.5 Interpolation d'Hermite

L'interpolation d'Hermite consiste à chercher un polynôme interpolateur et osculateur qui non seulement prend les valeurs fixées en les abscisses données, mais dont également la dérivée. L'interpolation de Hermite, nommée d'après le mathématicien ***Charles Hermite***, est une extension de la méthode d'interpolation de Lagrange.

On suppose que la fonction $f \in \mathcal{C}^1$, on se donne $(n+1)$ points $(t_i)_{i=1}^{n+1}$ deux à deux distincts. Ainsi, le polynôme d'interpolation d'Hermite vérifie les conditions suivantes :

$$\boxed{p_n(t_i) \;=\; f(t_i) \;\text{ et }\; p_n'(t_i) \;=\; f'(t_i), \;\; \forall i \in \{1, 2, ..., n+1\}}$$

Ces $(2n+2)$ degrés de liberté nous conduisent à construire un polynôme de degré inférieur ou égal à $(2n+1)$. L'unicité du polynôme interpolateur de Hermite se montre de façon similaire à celle du polynôme interpolateur de Lagrange

Théorème 4.7. *Étant donné la valeur connue d'une certaine fonction f en un certain nombre de points $(t_i, f(t_i))_{1\leq i\leq n+1}$, soit l'intervalle $I = [\min(t_1, t_2, ..., t_{n+1})], \max(t_1, t_2, ..., t_{n+1})]$. On suppose que f est $(n+1)$ fois différentiable sur l'intervalle I. Alors pour tout $t \in I$, il existe $\xi \in I$, tel que :*

$$\boxed{f(t) - p_{2n+1}(t) \;=\; \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\left(\prod_{i=1}^{n+1}(t-t_i)\right)^2}$$

La formule précédente ne permet pas le calcul d'une valeur exacte de l'erreur étant donné que ξ est inconnu. Cependant, on en déduit une majoration d'où

Lemme 4.2. *Étant donné la valeur connue d'une certaine fonction f en un certain nombre de points $(t_i, f(t_i))_{1\leq i\leq n+1}$, soit l'intervalle $I = [\min(t_1, t_2, ..., t_{n+1})], \max(t_1, t_2, ..., t_{n+1})]$. On suppose que f est $(n+1)$ fois différentiable sur l'intervalle I. Alors pour tout $t \in I$, on a :*

$$\boxed{|f(t) - p_{2n+1}(t)| \leq \frac{M^{(2n+2)}}{(2n+2)!}\left|\prod_{i=1}^{n+1}(t-t_i)\right| \;\text{ où }\; M^{(2n+2)} \;=\; \max_{t\in I}|f^{(2n+2)}(t)|}$$

4.6 Approximation au sens de moindres carrés

Reprenons l'exemple du relevé expérimental de la pression en fonction de l'altitude, le nuage de points est présenté dans la figure ci-dessous :

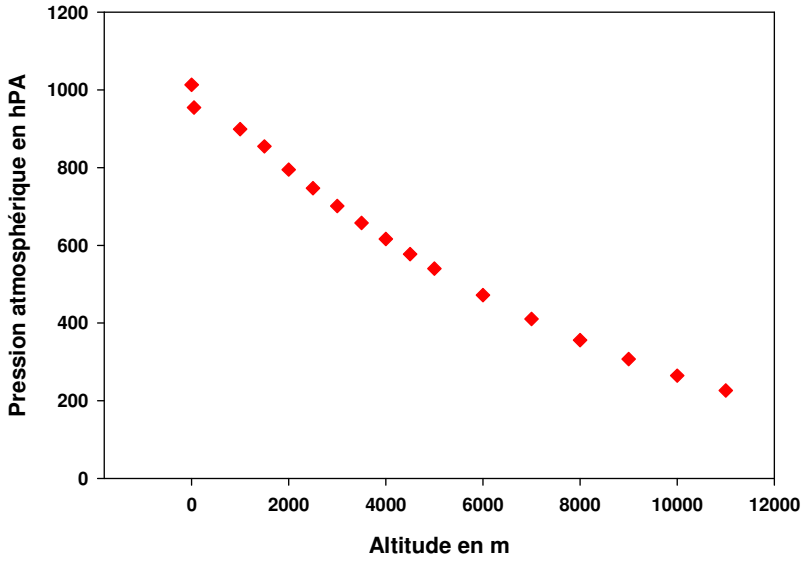


FIGURE 4.1 – Nuage de points (Altitude ; Pression)

Le problème qui s'impose maintenant est de construire le meilleur modèle mathématique qui approche au mieux les données à l'aide de certains types de fonctions (non seulement des polynômes) dépendant de plusieurs paramètres selon chaque modèle.

Ainsi la fonction construite p_f ne passe pas nécessairement par toutes les données $(t_i, f(t_i))_{\{1\leq i\leq n+1\}}$ mais les approche au plus près d'où la notion de

la minimisation de l'écart entre chaque donnée $f(t_i)$ et chaque point de la fonction associé $p(t_i)$.

On parle alors de la ***méthode des moindres carrés*** qui consiste à minimiser la quantité suivante :

$$Q_f \;=\; \sum_{i=1}^{n+1} \big(f(t_i) - p_f(t_i)\big)^2$$

p_f peut prendre plusieurs formes :

- une droite $p_f(t) = a_0 + a_1\, t$ avec des paramètres libres a_0 et a_1,
- un polynôme de degré n, $p_f(t) = \sum_{j=0}^{n} a_j\, t^i$ avec des paramètres libres $(a_j)_{\{1\leq j\leq n\}}$,
- une fonction exponentielle $p_f(t) = a_0 e^{a_1\, t}$ avec des paramètres libres a_0 et a_1,...
- des polynômes trigonométriques

$$p_f(t) \;=\; a_0 \;+\; \sum_{j=0}^{n} a_j\, \sin(j\, t) \;+\; \sum_{j=0}^{n} b_j\, \cos(j\, t)$$

où les coefficients a_i et b_j sont donnés par

$$\left\{\begin{array}{lcl} a_0 &=& \dfrac{1}{2\pi}\ \int_{-\pi}^{\pi} f(t)\ dt, \\ a_j &=& \dfrac{1}{\pi}\ \int_{-\pi}^{\pi} f(t)\ \sin(jt)dt, \\ b_j &=& \dfrac{1}{\pi}\ \int_{-\pi}^{\pi} f(t)\ \cos(jt)dt. \end{array}\right.$$

Ce type d'approximation est surtout utilisé pour les fonctions périodiques.

Afin de déterminer les paramètres libres, il suffit de minimiser la quantité Q, i.e :

$$\boxed{\frac{\partial Q}{\partial a_j}\ =\ 0,\ \ \forall j}$$

4.6.1 Approximation par des polynômes de degré n

Rappelons nous que

$$Q_f\ =\ \sum_{i=1}^{n+1}\Big(f(t_i)-p_f(t_i)\Big)^2\ =\ \sum_{i=1}^{n+1}\Big(f(t_i)-\sum_{j=0}^{n+1} a_j\ t_i^j\Big)^2.$$

Afin de déterminer les $(n+1)$ paramètres libres, il suffit de calculer

$$\frac{\partial Q}{\partial a_j}\ =\ 0,\ \ \forall 1\leq j\leq n.$$

Il vient alors

$$\frac{\partial Q}{\partial a_j}\ =\ \sum_{i=1}^{n+1} -2\ \frac{\partial p_f(t_i)}{\partial a_j}\Big(f(t_i)-p_f(t_i)\Big)\ =\ \sum_{i=1}^{n+1} -2\ t_i^j\ \Big(f(t_i)-\sum_{j=0}^{n} a_j\ t_i^j\Big)\ =\ 0,\ \ \forall 1\leq j\leq n.$$

Ceci implique,

$$\sum_{i=1}^{n+1} t_i^j\ f(t_i)\ =\ \sum_{i=1}^{n+1} t_i^j\ \sum_{j=0}^{n} a_j\ t_i^j.$$

Remarque 4.3. *Remarquons bien que dans l'expression* $\sum_{i=1}^{n+1} t_i^j\ \sum_{j=0}^{n} a_j\ t_i^j$, *la somme* $\sum_{i=1}^{n+1} t_i^j$ *correspond à une valeur* t_i^j *invariante par rapport à* j *tandis que* $\sum_{j=0}^{n} a_j\ t_i^j$ *correspond à une valeur* t_i^j *invariante par rapport à* i. *Par abus de simplification, on remplacera dans l'expression* $\sum_{j=0}^{n} a_j\ t_i^j$, j *par* k. *Ainsi, on écrit :*

$$\boxed{\sum_{i=1}^{n+1} t_i^j\ f(t_i)\ =\ \sum_{i=1}^{n+1} t_i^j\ \sum_{k=0}^{n} a_k\ t_i^k\ =\ \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\ t_i^{j+k}}$$

Par suite, les coefficients a_0, a_1,..., a_n vérifient les système linéaire suivant à $(n+1)$ inconnues et équations :

$$\left\{\begin{array}{rcl}
\dfrac{\partial Q}{\partial a_0} = 0 \implies \sum_{i=1}^{n+1} f(t_i) & = & \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\, t_i^j, \\
\dfrac{\partial Q}{\partial a_1} = 0 \implies \sum_{i=1}^{n+1} t_i\, f(t_i) & = & \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\, t_i^{j+1}, \\
\dfrac{\partial Q}{\partial a_2} = 0 \implies \sum_{i=1}^{n+1} t_i^2\, f(t_i) & = & \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\, t_i^{j+2}, \\
\dfrac{\partial Q}{\partial a_3} = 0 \implies \sum_{i=1}^{n+1} t_i^3\, f(t_i) & = & \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\, t_i^{j+3}, \\
\vdots & & \\
\dfrac{\partial Q}{\partial a_{n-1}} = 0 \implies \sum_{i=1}^{n+1} t_i^{n-1}\, f(t_i) & = & \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\, t_i^{j+n-1}, \\
\dfrac{\partial Q}{\partial a_n} = 0 \implies \sum_{i=1}^{n+1} t_i^{n}\, f(t_i) & = & \sum_{i=1}^{n+1}\sum_{k=0}^{n} a_k\, t_i^{j+n}.
\end{array}\right.$$

En notation matricielle, on écrit :

$$\begin{pmatrix}
\sum_{i=1}^{n+1} 1 & \sum_{i=1}^{n+1} t_i & \sum_{i=1}^{n+1} t_i^2 & \cdots & \sum_{i=1}^{n+1} t_i^n \\
\sum_{i=1}^{n+1} t_i & \sum_{i=1}^{n+1} t_i^2 & \sum_{i=1}^{n+1} t_i^3 & \cdots & \sum_{i=1}^{n+1} t_i^{n+1} \\
\sum_{i=1}^{n+1} t_i^2 & \sum_{i=1}^{n+1} t_i^3 & \sum_{i=1}^{n+1} t_i^4 & \cdots & \sum_{i=1}^{n+1} t_i^{n+2} \\
\sum_{i=1}^{n+1} t_i^3 & \sum_{i=1}^{n+1} t_i^4 & \sum_{i=1}^{n+1} t_i^5 & \cdots & \sum_{i=1}^{n+1} t_i^{n+3} \\
\multicolumn{5}{c}{\cdots\cdots\cdots\cdots\cdots} \\
\sum_{i=1}^{n+1} t_i^{n-1} & \sum_{i=1}^{n+1} t_i^{n} & \sum_{i=1}^{n+1} t_i^{n+1} & \cdots & \sum_{i=1}^{n+1} t_i^{2n-1} \\
\sum_{i=1}^{n+1} t_i^{n} & \sum_{i=1}^{n+1} t_i^{n+1} & \sum_{i=1}^{n+1} t_i^{n+2} & \cdots & \sum_{i=1}^{n+1} t_i^{2n}
\end{pmatrix}
\begin{pmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_{n-1} \\ a_n \end{pmatrix}
=
\begin{pmatrix}
\sum_{i=1}^{n+1} f(t_i) \\
\sum_{i=1}^{n+1} t_i\, f(t_i) \\
\sum_{i=1}^{n+1} t_i^2\, f(t_i) \\
\sum_{i=1}^{n+1} t_i^3\, f(t_i) \\
\vdots \\
\sum_{i=1}^{n+1} t_i^{n-1}\, f(t_i) \\
\sum_{i=1}^{n+1} t_i^{n}\, f(t_i)
\end{pmatrix}$$

On pose $(M)[A] = [S]$ tel que

$$M_{jk} \;=\; \sum_{i=1}^{n+1} t_i^{j+k}, \;\; A_j \;=\; a_j \;\text{ et }\; S_j \;=\; \sum_{i=1}^{n+1} t_i^{j}\, f(t_i), \;\; \forall j,k = 0,1,...,n.$$

On obtient :

$$\boxed{\sum_{k=0}^{n} M_{jk}\, a_j \;=\; S_j, \;\; \forall j,k = 0,1,...,n.}$$

Remarque 4.4. *La matrice M est une matrice symétrique, par suite il suffit de calculer $\frac{n^2}{2}$ termes et en déduire les autres par symétrie.*

Exemple : On considère la fonction $f(t)$ définie sur l'intervalle $[-2,2]$ par la table des valeurs

t_i	-2	-1	0	1	2
$f(t_i)$	-1	0	1	2	3

Déterminer le polynôme de meilleure approximation au sens de moindre carré de degré inférieur ou égal à 2, de la fonction $f(t)$ sur l'intervalle $[-2, 2]$.

Soit le polynôme p_2 de degré 2 :

$$p_2(t) \;=\; a_0 + a_1\, t + a_2\, t^2.$$

En appliquant les formules suivantes, calculons les termes de la matrice M et du vecteur S :

$$\begin{cases}
M_{00} & = \sum_{i=1}^{5} t_i^0 & = 1, \\
M_{01} = M_{10} & = \sum_{i=1}^{5} t_i & = (-2-1+0+1+2) \;=\; 0, \\
M_{02} = M_{20} = M_{11} & = \sum_{i=1}^{5} t_i^2 & = (-2)^2+(-1)^2+(0)^2+(1)^2+(2)^2 \;=\; 10, \\
M_{12} = M_{21} & = \sum_{i=1}^{5} t_i^3 & = (-2)^3+(-1)^3+(0)^3+(1)^3+(2)^3 \;=\; 0, \\
M_{22} & = \sum_{i=1}^{5} t_i^4 & = (-2)^4+(-1)^4+(0)^4+(1)^4+(2)^4 \;=\; 34, \\
S_0 & = \sum_{i=1}^{5} f(t_i) & = -1+0+1+2+3 \;=\; 5, \\
S_1 & = \sum_{i=1}^{5} t_i\, f(t_i) & = (-2)(-1)+(-1)(0)+(0)(1)+(1)(2)+(2)(3) \;=\; 10, \\
S_2 & = \sum_{i=1}^{5} t_i^2\, f(t_i) & = (-2)^2(-1)+(-1)^2(0)+(0)^2(1)+(1)^2(2)+(2)^2(3) \;=\; 10.
\end{cases}$$

Par suite, on obtient le système matriciel suivant :

$$\begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 10 & 0 & 34 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \;=\; \begin{pmatrix} 5 \\ 10 \\ 10 \end{pmatrix}$$

équivalent à

$$\begin{pmatrix} 5 & 0 & 10 \\ 0 & 10 & 0 \\ 0 & 0 & 14 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \;=\; \begin{pmatrix} 5 \\ 10 \\ 0 \end{pmatrix}$$

On obtient alors :

$$a_0 \;=\; 1, \;\; a_1 \;=\; 1 \;\text{ et }\; a_2 \;=\; 0,$$

ce qui donne

$$p_2(t) \;=\; 1 + t.$$

4.6.2 Droites de régression

Soit X et Y deux variables quantitatives dont l'une dépend de l'autre et on dispose donc d'un échantillon de N couples d'observations (x_i, y_i) que l'on peut représenter dans le plan $\mathbb{R}^2$, où chaque point i, d'abscisse x_i et d'ordonnée y_i, correspond à un couple d'observations.

L'objectif ici est de chercher à exprimer la relation entre les deux variables X et Y. La première est souvent appelée ***variable explicative*** et la seconde est appelée ***variable expliquée*** .

Si l'examen du nuage de points décèle une relation de dépendance linéaire de Y en X, il est naturel de représenter graphiquement cette relation particulière à l'aide d'une droite suivant une forme allongée traversant le nuage de points, appelée ***droite de régression de*** Y ***en***

X qui ajuste au mieux le nuage et qui consiste à se baser sur les propriétés de la covariance et de l'espérance. Le critère d'optimalité que nous allons considérer est celui des moindres carrés.

Comme toute droite, la droite de régression de Y en X peut être définie au moyen d'une équation du premier degré de la forme :

$$\boxed{y \;=\; a\,x \;+b}$$

Le principe de moindres carrés consiste à minimiser la fonctionnelle :

$$\boxed{F(a,b) \;=\; \sum_i [y_i \;-\; (a\,x_i \;+\; b)]^2}$$

où $r_i \;=\; y_i \;-\; (a\,x_i \;+\; b)$ est le ***résidu*** ou l'***erreur d'ajustement*** entre la valeur réellement observée y_i pour la variable dépendante et la valeur ajustée fournie par la droite de régression $a\,x_i \;+\; b$.

Nous pouvons dès lors considérer que le meilleur ajustement est fourni par la droite qui minimise globalement l'amplitude des erreurs d'ajustement r_i, $\forall i$. Cette méthode d'optimisation est nommée "***méthode des multiplicateurs de Lagrange***".

La détermination de a et b et est un problème classique de minimisation, dont la solution est :

$$\boxed{a \;=\; \frac{Cov(X,Y)}{\sigma(X)^2} \quad et \quad b \;=\; \overline{Y} \;-\; a\,\overline{X}}$$

Dans cette expression, $Cov(X,Y)$ est la covariance, σ_X l'écart-type de la distribution marginale en X et $\overline{X}$ et $\overline{Y}$ les moyennes marginales de X et Y respectivement définis comme suit :

$$\boxed{\begin{aligned} Cov(X,Y) \;&=\; \frac{1}{N}\,\textstyle\sum_{i=1}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) \\ \sigma_X^2 \;&=\; V(X) \;=\; \frac{1}{N}\,\textstyle\sum_i (x_i \;-\; \overline{X})^2 \\ \overline{X} \;&=\; \frac{1}{N}\,\textstyle\sum_i x_i \;\text{ et }\; \overline{Y} \;=\; \frac{1}{N}\,\textstyle\sum_i y_i \end{aligned}}$$

Lorsqu'une série statistique à deux variables quantitatives est représentée au moyen d'un graphique de dispersion ou un nuage de point, on s'intéresse toujours à déceler une structure d'association entre les deux variables. Le ***coefficient de corrélation*** permettra de quantifier l'intensité de la liaison qui pourra exister entre ces deux variables dans le cas où l'association est de nature linéaire.

La notion du coefficient de corrélation peut être attribuée au physicien français ***Auguste Bravais*** par le biais de ses travaux effectués dans l'étude des erreurs dans les tirs d'artillerie mais aussi à ***Francis Galton*** qui grâce à lui que la corrélation devient un concept statistique. C'est

ensuite ***karl Pearson*** qui propose en 1986 la formulation mathématique actuelle.

Le ***coefficient de corrélation de Bravais-Pearson***, désigné par r ou r_{XY}, est défini comme suit :

$$\boxed{r_{XY} = \frac{Cov(X,Y)}{\sigma_X \, \sigma_Y}}$$

Dans cette expression, $Cov(X,Y)$ est la covariance et σ_X et σ_Y sont les écarts-types des distributions marginales en X et en Y.

Nous constatons ainsi que la droite de régression de Y en X passe par le barycentre ou centre de gravité du nuage de points de coordonnées $(\overline{X}, \overline{Y})$ et que l'orientation de la droite de régression indique la nature de liaison entre les deux variables de telle manière que :

• si la covariance est positive, par suite le coefficient de corrélation et a le sont, ce qui signifie que la droite est ascendante,

• si la covariance est négative, par suite le coefficient de corrélation et a le sont, ce qui signifie que la droite est descendante.

Exemple : Sur une année glissante, on a mesuré pour un échantillon de 10 jeunes âgés de 11 à 16 ans, l'âge (variable X) et la durée journalière moyenne durée d'écoute de leur MP3 (variable Y exprimée en heure). La série statistique observée ainsi que sa représentation graphique au moyen d'un nuage de points sont présentées ci-après :

x_i	11	11	12	12	13	13	14	15	15	16
y_i	2	2.1	2.7	3	4	4.5	5	6.5	6.8	7.6

TABLE 4.2 – Répartition d'une population de 10 jeunes suivant l'âge et la durée journalière moyenne durée d'écoute de leur MP3

Le calcul de la covariance peut s'effectuer au moyen du tableau suivant :

x_i	y_i	$(x_i - \overline{X})$ avec $\overline{X} = 13.2$	$(y_i - \overline{Y})$ avec $\overline{Y} = 4.42$	$(x_i - \overline{X})^2$	$(y_i - \overline{Y})^2$	$(x_i - \overline{X})\,(y_i - \overline{Y})$
11	2	-2.2	-2.42	4.84	5.8564	5.324
11	2.1	-2.2	-2.32	4.84	5.3824	5.104
12	2.7	-1.2	-1.72	1.44	2.9584	2.064
12	3	-1.2	-1.42	1.44	2.0164	1.704
13	4	-0.2	-0.42	0.04	0.1764	0.084
13	4.5	-0.2	0.08	0.04	0.0064	-0.016
14	5	0.8	0.58	0.64	0.3364	0.464
15	6.5	1.8	2.08	3.24	4.3264	3.744
15	6.8	1.8	2.38	3.24	5.6644	4.284
16	7.6	2.8	3.18	7.84	10.1124	8.904
132	44.2	0	0	27.6	36.836	31.66

Par suite, on obtient :

$$Cov(X,Y) \;=\; \frac{1}{10}\sum_{i=1}^{10}\left(x_i - \overline{X}\right)\left(y_i - \overline{Y}\right) \;=\; \frac{31.66}{10} \;=\; 3.166.$$

Sachant que

$$\sigma_X^2 \;=\; V(X) \;=\; \frac{1}{N}\sum_{i=1}^{10}(x_i \;-\; \overline{X})^2 \;=\; 2.76$$

et

$$\sigma_Y^2 \;=\; V(Y) \;=\; \frac{1}{N}\sum_{i=1}^{10}(y_i \;-\; \overline{Y})^2 \;=\; 3.68,$$

on obtient :

$$a \;=\; \frac{3.166}{2.76} \;=\; 1.147 \quad et \quad b \;=\; 4.42 \;-\; 1.147 \times 13.2 \;=\; -10.7204.$$

Par suite, la droite de régression de Y en X s'écrit ainsi :

$$y \;=\; 1.147\,x \;-\; 10.7204.$$

Le coefficient de corrélation est égale à :

$$r_{XY} \;=\; \frac{Cov(X,Y)}{\sigma_X\;\sigma_Y} \;=\; \frac{31.66)}{\sqrt{2.76}\,\sqrt{3.68}} \;=\; 0.9929.$$

Le coefficient de corrélation a nécessairement le même signe que la covariance, puisque les écarts-types sont des nombres positifs. Ceci s'interprète de la même manière suivante :

• s'il existe une association linéaire et positive entre les deux variables, la covariance et le coefficient de corrélation sont tous deux positifs,

• s'il existe une association linéaire et négative entre les deux variables, la covariance et le coefficient de corrélation sont tous deux négatifs,

• s'il n'existe pas d'association entre les deux variables, la covariance et le coefficient de corrélation ont tous deux des valeurs proches de zéro.

Le coefficient de corrélation est un nombre sans dimension et est compris entre -1 et 1. Le coefficient de corrélation peut être positif ou négatif selon la position des points par rapport au centre de gravité du nuage de points, dont les coordonnées sont $(\overline{X}, \overline{Y})$. Considérons à présent les trois sortes nuages de points :

• On remarque bien que si dans un nuage A, faisant apparaître une association linéaire, les points du nuage ont tendance à se concentrer autour d'une droite (c-à-d une structure positive, car lorsque la valeur de x_i augmente, celle de y_i a également tendance à augmenter), ainsi la plupart des points du nuage se trouvent dans les quadrants I et III et donnent donc lieu à des produits $\left(x_i - \overline{X}\right)\ \left(y_i - \overline{Y}\right)$ presque positifs, aussi bien que la covariance et le coefficient de corrélation.

• Contrairement au nuage A, si un nuage B met en évidence une association linéaire et négative (puisque lorsque la valeur de x_i augmente, celle de y_i a au contraire tendance à diminuer), on remarque alors que presque tous les points du nuage occupent les quadrants II et IV, d'où tous les produits $\left(x_i - \overline{X}\right)\ \left(y_i - \overline{Y}\right)$ sont négatifs et la covariance ainsi que le coefficient de corrélation sont alors négatifs.

• Si un nuage C ne montre aucune structure particulière puisqu'il semble que les deux variables ne sont pas liées entre elles, alors les points du nuage se répartissent d'une manière équitable dans les quatre quadrants du plan et la somme des produits positifs compense la somme des produits négatifs, donnant ainsi lieu à une covariance et un coefficient de corrélation pratiquement nuls.

Ce coefficient est une mesure de la dispersion du nuage. On considère que l'approximation d'un nuage par sa droite des moindres carrés est de bonne qualité lorsque $|r|$ est proche de 1 et de qualité médiocre lorsque $|r|$ est proche de 0.

Chapitre 5

Méthodes d'intégration numérique

Le but de ce chapitre consiste à déterminer l'intégrale d'une fonction f sur un domaine fini délimité par des bornes finies a et b :

$$I \;=\; \int_a^b f(t)\; w(t)\; dt$$

Les méthodes d'intégration numérique employées dans ce chapitre, consistent à remplacer l'intégrale I par une expression de la forme :

$$\sum_{i=0}^{n} w_i\; f(t_i)$$

où les w_i sont des coefficients réel et $t_0, t_1, ..., t_n$ ses points distincts équirépartis entre a et b. Nous distinguons trois différentes méthodes :

- Les méthodes de Newton-Cotes simples
- Les méthodes de Newton-Cotes composites
- Les méthodes de Gauss-Legendre

Les ***méthodes de Newton-Cotes*** Dans ce cas, $w = 1$ et le nombre

$$I \;=\; \int_a^b f(t)\; dt$$

désigne l'aire de la région délimitée par l'horizontale $y = 0$ et les deux verticales $t = a$ et $t = b$ dans le graphe de la fonction f.

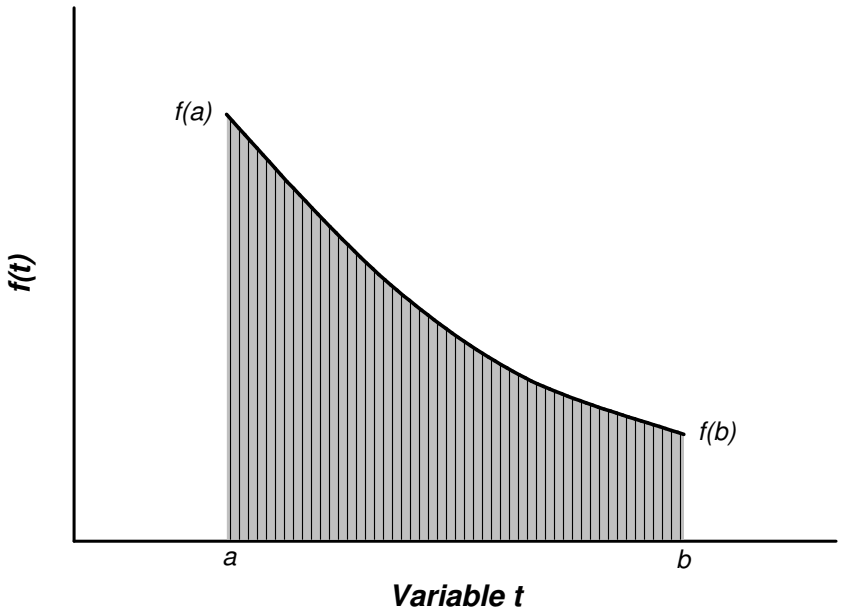


FIGURE 5.1 – L'aire correspondant à la valeur de l'intégrale I

Deux cas se présentent pour le calcul exact de cette intégrale. Soit une telle intégrale peut être calculée analytiquement, soit le calcul d'une aire se ramène à un calcul de primitives F, de la fonction f qu'on suppose continue, où

$$\int_a^b f(t)\,dx \;=\; F(b) - F(a),$$

sauf que F peut aborder être une fonction compliquée à calculer.

Les méthodes de Newton-Cotes consistent ainsi à approcher l'intégrale I par

$$I \;\simeq\; \int_a^b P_f(t)\,dt$$

où P_f est le polynôme qui interpole la fonction f aux points distincts $t_0, t_1, ..., t_n$ équirépartis entre a et b. P_f s'exprime en fonction des $n+1$ polynômes de Lagrange L_i, $i = 0, ..., n$:

$$P(t) \;=\; \sum_{i=0}^{n} f(t_i)\,L_i(t) \;=\; \sum_{i=0}^{n} f(t_i) \prod_{j=0, j\neq i} \frac{t - t_j}{t_i - t_j}.$$

Ainsi l'intégrale I est approchée par

$$I \;\simeq\; \int_a^b P_f(t)\,dt \;=\; \sum_{i=0}^{n} f(t_i) \int_a^b L_i(t) \;=\; (b-a) \sum_{i=0}^{n} f(t_i)\,\frac{\int_a^b L_i(t)}{b-a} \;=\; (b-a) \sum_{i=0}^{n} w_i\, f(t_i)$$

Par suite, le calcul de l'intégrale d'une fonction est approchée par le calcul exact de l'intégrale d'un polynôme interpolant cette fonction en certains points t_i pour $i = 0, ..., n$. Plus précisément

on aura recours à des combinaisons linéaires de valeurs de la fonction$f$ à intégrer en des points t_i pour $i = 0, ..., n$ de l'intervalle $[a, b]$: $f_i = f(t_i)$, pour $i = 0, ..., n$. On obtient la ***formule d'intégration numérique*** dite aussi ***formule de quadrature***.

$$I = \int_a^b f(t)\, dt \simeq \sum_{i=0}^{n} w_i\, f(t_i) = \sum_{i=0}^{n} w_i\, f_i.$$

Les points t_i sont appelés les ***noeuds*** de la formule et les w_i sont les ***poids***.

5.1 Les méthodes de Newton-Cotes simples

Soit l'intervalle $[a, b]$, on notera $h=b-a$. Nous présentons à présent quelques une des méthodes d'ordres les plus bas.

5.1.1 Méthode des rectangles

Cette méthode consiste à interpoler la fonction f par des polynômes constants (correspondant à $n = 1$). $P_f(t)$ est l'interpolé linéaire de la fonction f au point a, soit b, soit $\frac{a+b}{2}$. Ainsi, on approche l'intégrale I par

$$\boxed{I \simeq (b-a)\, f(\alpha)}$$

On distingue trois choix courants :

- La ***Méthode des rectangles à gauche*** consistant à approcher l'intégrale I par l'aire du rectangle de base $[a, b]$ et de hauteur $f(a)$:

$$\boxed{I \simeq (b-a)\, f(a)}$$

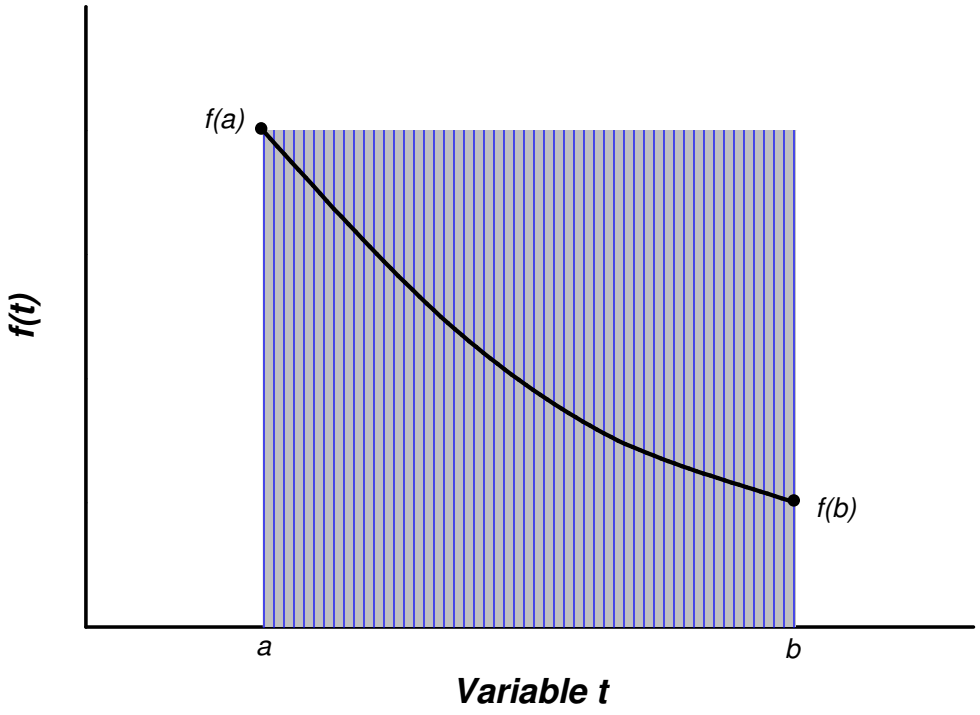


FIGURE 5.2 – Méthode des rectangles à gauche appliquée à l'intervalle $[a, b]$

- La ***Méthode des rectangles à droite*** consistant à approcher l'intégrale I par l'aire du rectangle de base $[a, b]$ et de hauteur $f(b)$:

$$\boxed{I \simeq (b-a)\, f(b)}$$

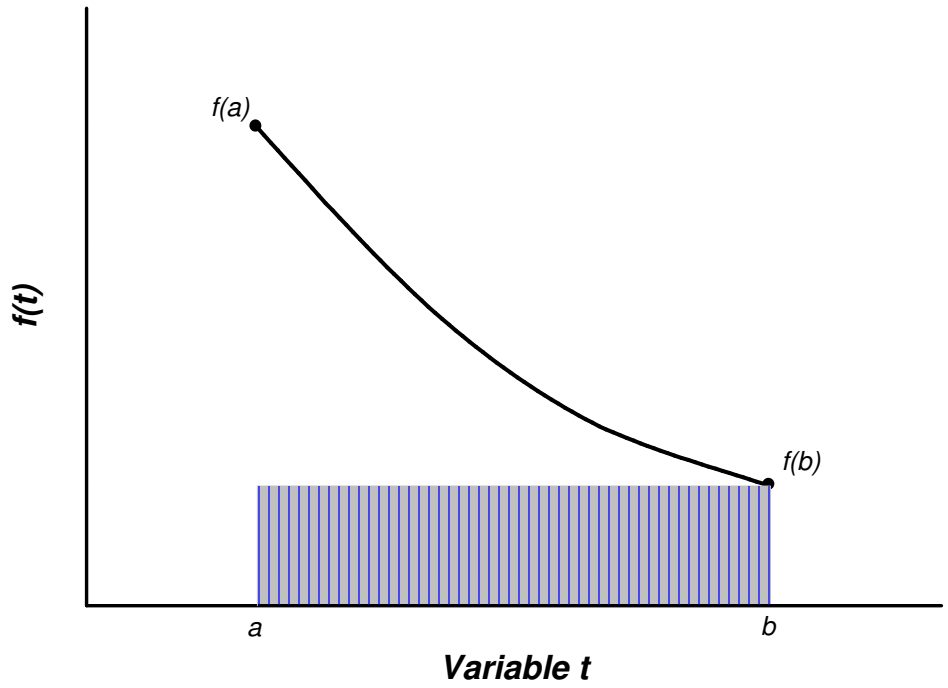


FIGURE 5.3 – Méthode des rectangles à droite appliquée à l'intervalle $[a, b]$

- La méthode des points milieux correspondant à $\alpha = \frac{a+b}{2}$. Cette méthode utilise également le polynôme constant pour approcher la fonction f. On approche l'intégrale I par l'aire du rectangle de base $[a, b]$ et de hauteur $f\left(\frac{a+b}{2}\right)$:

$$\boxed{I \simeq (b-a)\, f\left(\frac{a+b}{2}\right)}$$

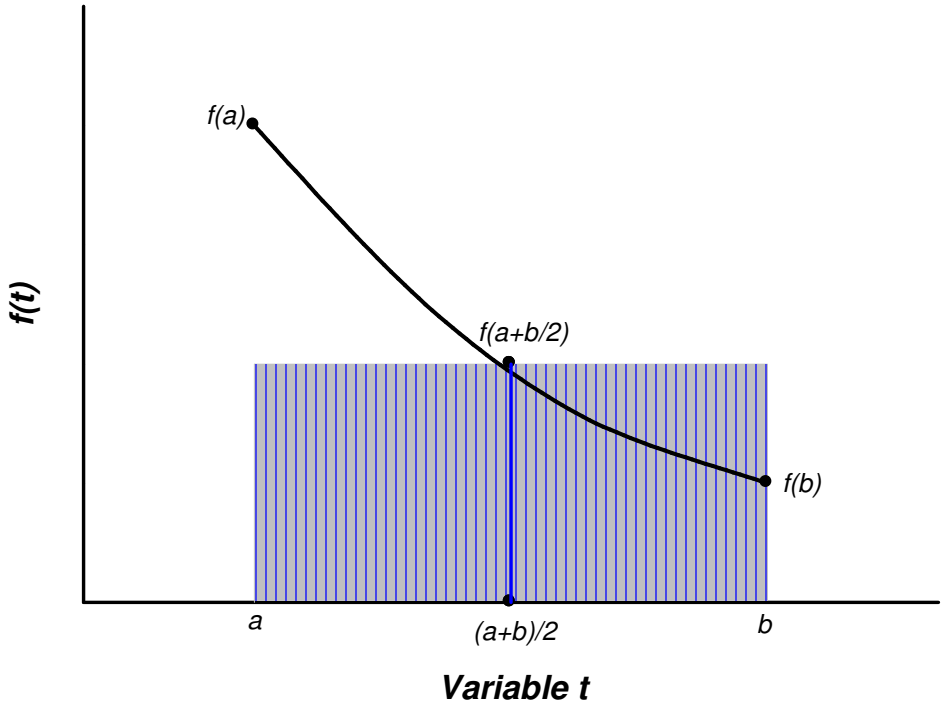


FIGURE 5.4 – Méthode des points milieux appliquée à l'intervalle $[a, b]$

5.1.2 Méthode des Trapèzes

Cette méthode consiste à interpoler la fonction f par des polynômes de degré 1 (correspondant à $n = 2$). $P_f(t)$ est l'interpolé linéaire de la fonction f aux points a et b. Ainsi, on approche l'intégrale I par l'aire du trapèze délimité par les 4 points $(a, 0)$, $(b, 0)$, $(a, f(a))$ et $(b, f(b)$. On écrit :

$$\boxed{I \simeq \frac{(b-a)}{2}\big(f(a)+f(b)\big)}$$

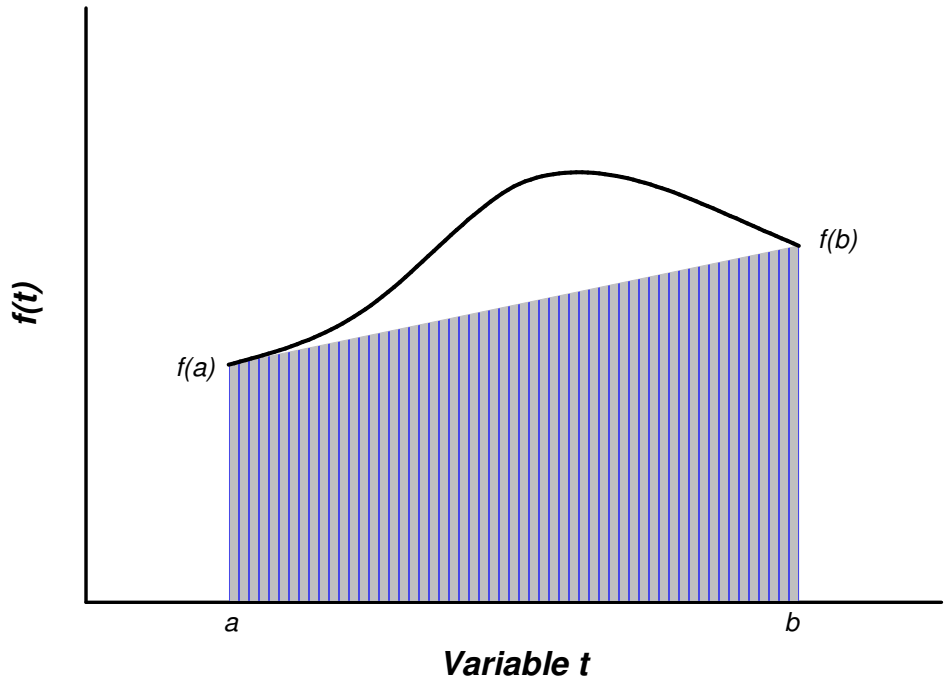


FIGURE 5.5 – Méthode des Trapèzes appliquée à l'intervalle $[a, b]$

5.1.3 Méthode de Simpson

Cette méthode consiste à interpoler la fonction f par des polynômes de degré 2 passant par les trois points $(a, f(a))$, $(\frac{a+b}{2}, f\big(\frac{a+b}{2}\big))$ et $(b, f(b)$. On remplace le segment joignant (a, f(a)) et (b, f(b)) de la méthode du trapèze par la portion de parabole passant par ces deux points et le troisième point $(\frac{a+b}{2}, f\big(\frac{a+b}{2}\big)$. On obtient :

$$\boxed{I \simeq \frac{(b-a)}{6}\left(f(a) + 4\, f\left(\frac{a+b}{2}\right) + f(b)\right)}$$

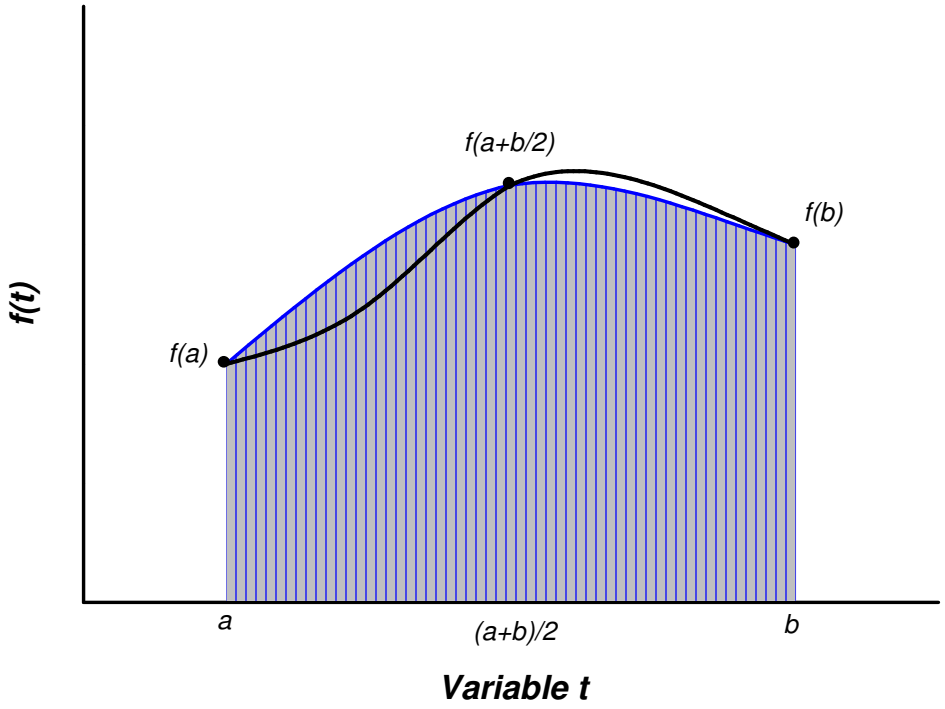


FIGURE 5.6 – Méthode de Simpson appliquée à l'intervalle $[a, b]$

5.1.3.1 Méthodes d'ordre élevé

Pour des ordres supérieurs, on remarque que plus les méthodes de Newton-Cotes simples sont basées sur des polynômes de degré élevé, plus elles sont lentes mais plus elles sont précises. On cite parmi lesquelles :

- ***Méthode de Simpson 3/8*** Sa formule relative de degré 3 s'écrit :

$$\boxed{I \simeq \frac{(b-a)}{8}\left(f(t_0) + 3\left(f(t_1)+f(t_2)\right) + f(t_3)\right)}$$

- ***Méthode de Boole-Villarceau*** Sa formule relative de degré 4 s'écrit :

$$\boxed{I \simeq \frac{(b-a)}{90}\left(7\,f(t_0) + 32\,f(t_1) + 12\,f(t_2) + 32\,f(t_3) + 7\,f(t_4)\right)}$$

- ***Méthode à 8 points*** Sa formule relative s'écrit :

$$\boxed{I \simeq \frac{7\,(b-a)}{138240}\left(751\left(f(t_0)+f(t_7)\right) + 3577\left(f(t_1)+f(t_6)\right) + 1323\left(f(t_2)+f(t_5)\right) + 2989\left(f(t_3)+f(t_4)\right)\right)}$$

- ***Méthode à 9 points*** Sa formule relative s'écrit :

$$I \simeq \frac{4\,(b-a)}{127575}\left(989\ \ (f(t_0)+f(t_8))\ +\ 5888\,(f(t_1)+f(t_7))\ -\ 928\,(f(t_2)+f(t_6)) + 10496\,f(t_5)\right)$$

- ***Méthode à 10 points*** Sa formule relative s'écrit :

$$I \simeq \frac{9\,(b-a)}{89600}\left(2857\,(f(t_0)+f(t_9))\ +\ 15741\,(f(t_1)+f(t_8))\ +\ 1080\,(f(t_2)+f(t_7)) + 19344\,(f(t_3)+f(t_6))\ +\ 5778\,(f(t_4)+f(t_5))\right)$$

- ***Méthode à 11 points*** Sa formule relative s'écrit :

$$I \simeq \frac{5\,(b-a)}{3293136}\left(160067\,(f(t_0)+f(t_{10}))\ +\ 106300\,(f(t_1)+f(t_9))\ -\ 48525\,(f(t_2)+f(t_8)) + 272400\,(f(t_3)+f(t_7))\ -\ 260550\,(f(t_4)+f(t_6))\ +\ 427368\,f(t_5)\right)$$

5.2 Les méthodes de Newton-Cotes composites

On découpe l'intervalle $[a,b]$ en n segments de même longueur $h=\dfrac{b-a}{n}$. On suppose que $t_0=a$ et $t_n=b$ et que les points t_i sont espacés régulièrement entre les bornes a et b : $\{t_i=a+i\,h\}$ pour $i=0,...,n$ avec $t_{i+1}-t_i=h$. On note $I_j=[t_j,t_{j-1}]$ pour $j=0,...,n-1$ tel que

$$I\ =\ \sum_{j=0}^{n-1} I_j.$$

Ainsi, on approche l'intégrale I par une simple somme des approximations des segments I_j, obtenue en appliquant sur chacun desquels la méthode de Newton-Cotes simple. Sur chaque intervalle, une méthode de degré $m+1$ évalue la fonction à intégrer en $m+1$ points. Ainsi la précision de toutes les méthodes de Newton-Cote augmente avec le nombre de points utilisés et constitue par suite une alternative au problème des polynômes de grand degré.

5.2.1 Méthode des rectangles

La méthode des rectangles composite consiste à appliquer la méthode des rectangles simples sur chacun des n intervalles I_j en interpolant la fonction f par les n polynômes constants $P_f^j(t)$ au point t_j, t_{j+1} ou $\frac{t_j+t_{j+1}}{2}$, pour $j=0,...,n-1$. Ainsi, on approche l'intégrale I par

$$I\ \simeq\ h\ \sum_{j=0}^{n-1} f(\alpha_i)$$

On distingue trois choix courants :

- La ***Méthode des rectangles à gauche*** consistant à approcher chacune des intégrales I_j par l'aire du rectangle de base $[t_j, t_{j+1}]$ et de hauteur $f(t_j)$:

$$I \simeq h \sum_{j=0}^{n-1} f(t_j)$$

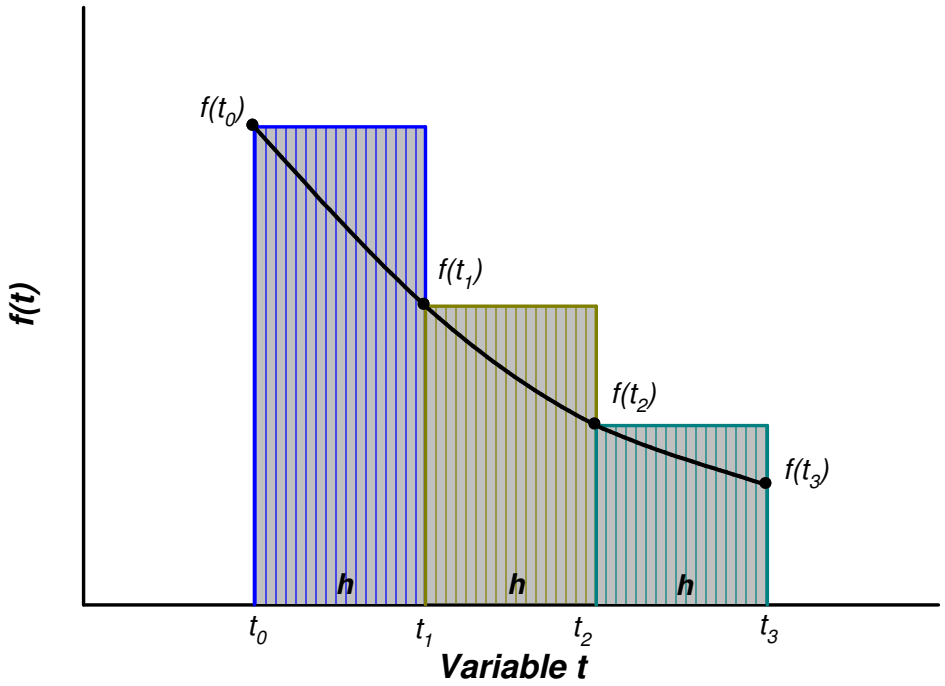


FIGURE 5.7 – Méthode composite des rectangles à gauche correspondant à $n = 3$

Dans cette figure, la méthode composite des rectangles à gauche est appliquée à l'intervalle $[a, b] = [t_0, t_3]$, qu'on a subdivisé en 3 sous-intervalles $[t_0, t_1]$, $[t_1, t_2]$ et $[t_2, t_3]$ sur chacun desquels on applique la méthode des rectangles à gauche.

- La ***Méthode des rectangles à droite*** consistant à approcher chacune des intégrales I_j par l'aire du rectangle de base $[t_j, t_{j+1}]$ et de hauteur $f(t_{j+1})$:

$$I \simeq h \sum_{j=0}^{n-1} f(t_{j+1})$$

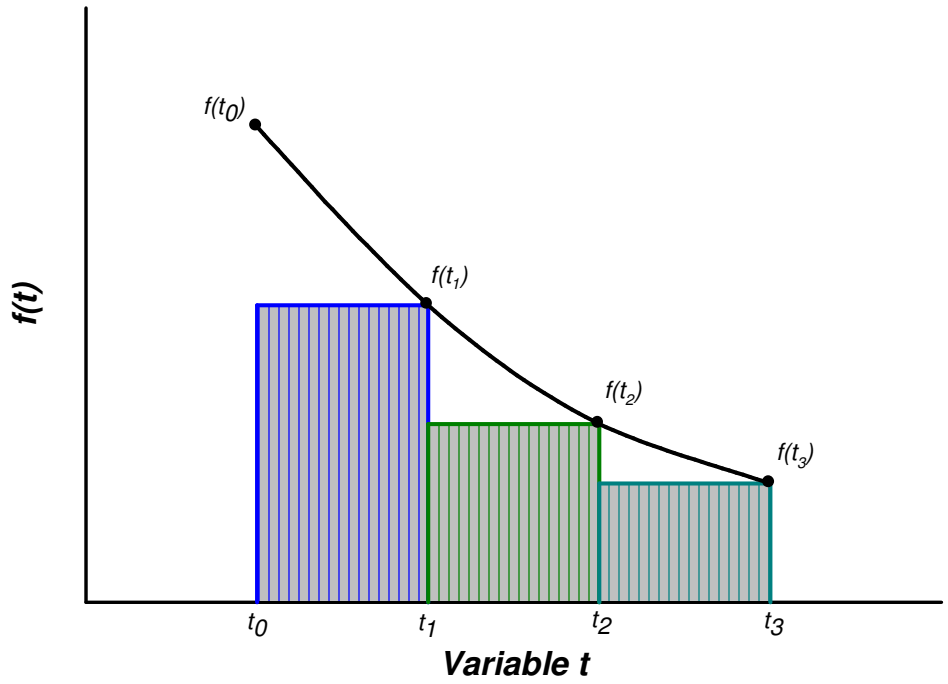


FIGURE 5.8 – Méthode composite des rectangles à droite correspondant à $n = 3$

- La méthode des points milieux correspondant à $\alpha = \frac{a+b}{2}$. Cette méthode consiste à approcher chacune des intégrales I_j par l'aire du rectangle de base $[t_j, t_{j+1}]$ et de hauteur $f\left(\frac{t_j+t_{j+1}}{2}\right)$:

$$\boxed{I \simeq h \sum_{j=0}^{n-1} f\left(\frac{t_j + t_{j+1}}{2}\right)}$$

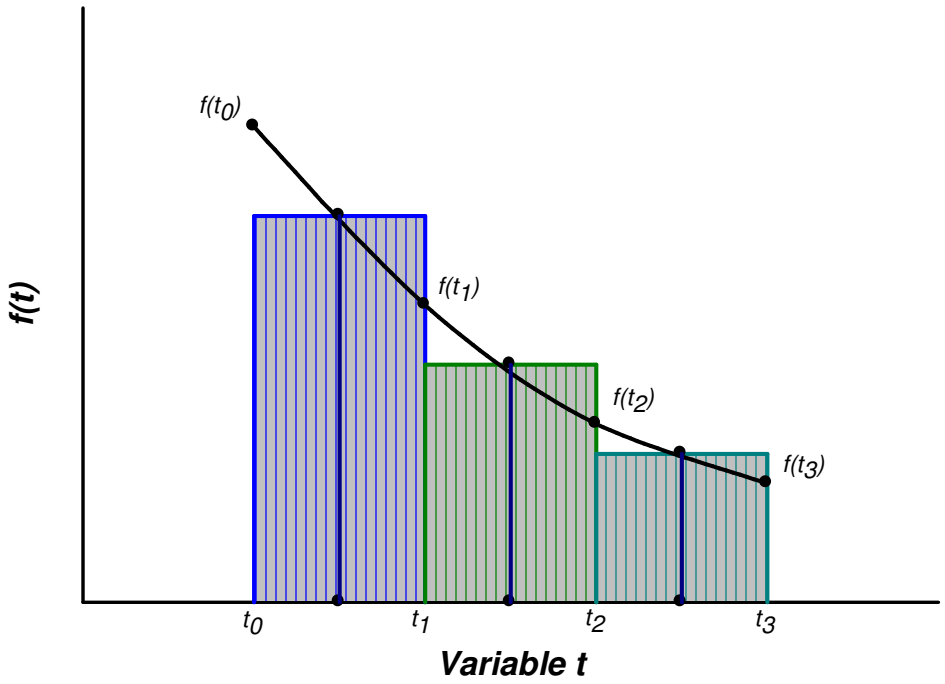


FIGURE 5.9 – Méthode composite des points milieux correspondant à $n = 3$

5.2.2 Méthode des Trapèzes

On approche l'intégrale I par la somme des n aires des trapèzes délimités par les 4 points $(t_i, 0)$, $(t_{i+1}, 0)$, $t_i, f(t_i))$ et $(t_{i+1}, f(t_{i+1}))$, pour $j = 0, ..., n-1$. Chacune de ces aires vaut $h\ \frac{f(t_i)+f(t_{i+1})}{2}$. On écrit :

$$\boxed{I \simeq h\left(\frac{f(t_0)+f(t_n)}{2} + \sum_{j=1}^{n-1} f(t_j)\right)}$$

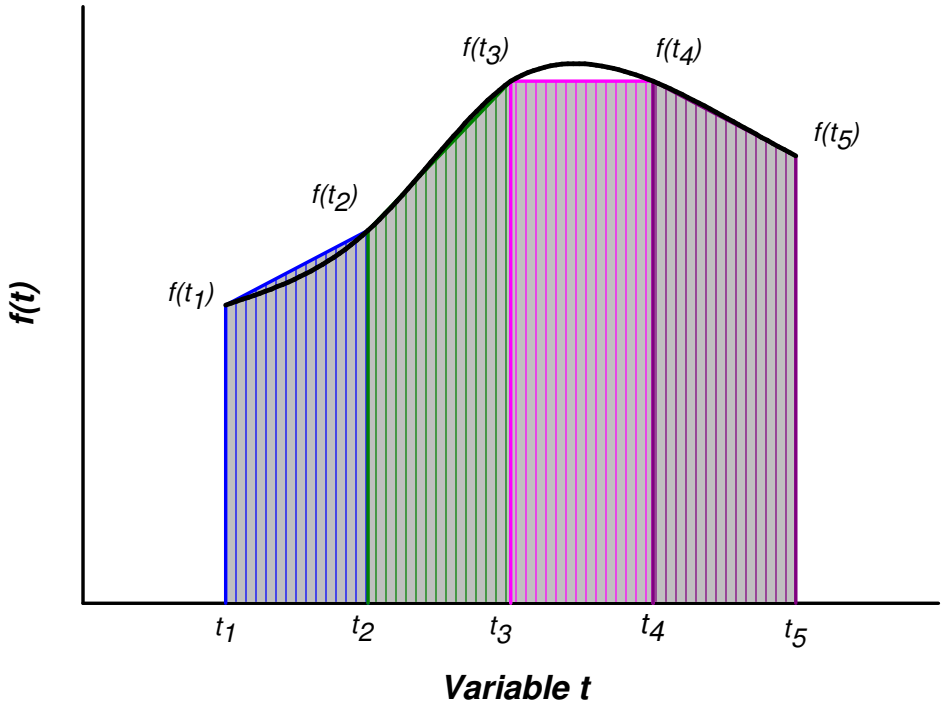


FIGURE 5.10 – Méthode composite des Trapèzes correspondant à $n = 4$

5.2.3 Méthode de Simpson

Cette méthode consiste à interpoler la fonction f sur chacun des n intervalles par des polynômes de degré 2 reposant sur une interpolation à trois points $(t_j, f(t_j))$, $\left(\frac{t_j+t_{j+1}}{2}, f\left(\frac{t_j+t_{j+1}}{2}\right)\right)$ et $(t_{j+1}, f(t_{j+1})$. On obtient :

$$\boxed{I \simeq \frac{h}{3}\left(f(t_0) + 4\sum_{j=1}^{n/2} f(t_{2j-1}) + 2\sum_{j=1}^{n/2-1} f(t_{2j}) + f(t_n)\right)}$$

Remarque 5.1. *On ne peut pas appliquer la méthode de Simpson pour des valeurs de n impaires car cette dernière n'est appliquée que sur chacun des n intervalles $[t_{2i}; t_{2i+2}]$, tirées d'une subdivision en $2n$ intervalles par les points :*

$$t_i = a + ih \text{ pour } i = 0, ..., 2n-1 \text{ et } h = \frac{b-a}{2n}.$$

Exemple : Déterminer l'intégrale

$$I = \int_{0.1}^{1.1} \ln(t)\, dt$$

pour $n = 10$.

1. Par la méthode composite des Trapèzes généralisée

2. Par la méthode composite de Simpson

3. Par la méthode composite des rectangles à droite et à gauche

4. Par la méthode composite des points milieux

5. Comparer les même pour les différents pas d'espace : $h = 0.2$ et $h = 0.5$. Déduire une conclusion.

On définit la fonction f telle que $f(t) = \ln(t)$ et on a que

$$I \;=\; \int_{0.1}^{1.1} \ln(t)\, dt \;=\; \big(t\, \ln(t) - t\big)_{(1.1)} - \big(t\, \ln(t) - t\big)_{(0.1)} = \textbf{\textit{- 0.664900292}}.$$

Sachant que $n = 10$ correspond à $h = \dfrac{1}{10} = 0.1$, on découpe l'intervalle $[0.1; 1.1]$ en 10 segments dont les bornes équidistants sont telles que

t_i	$f(t_i)$
$t_0 = 0.1$	$f(t_0) = f(0.1) = \ln(0.1) = -2.30259$
$t_1 = 0.2$	$f(t_1) = f(0.2) = \ln(0.2) = -1.60944$
$t_2 = 0.3$	$f(t_2) = f(0.3) = \ln(0.3) = -1.20397$
$t_3 = 0.4$	$f(t_3) = f(0.4) = \ln(0.4) = -0.91629$
$t_4 = 0.5$	$f(t_4) = f(0.5) = \ln(0.5) = -0.69315$
$t_5 = 0.6$	$f(t_5) = f(0.6) = \ln(0.6) = -0.51083$
$t_6 = 0.7$	$f(t_6) = f(0.7) = \ln(0.7) = -0.35667$
$t_7 = 0.8$	$f(t_7) = f(0.8) = \ln(0.8) = -0.22314$
$t_8 = 0.9$	$f(t_8) = f(0.9) = \ln(0.9) = -0.10536$
$t_9 = 1$	$f(t_9) = f(1) = \ln(1) = 0$
$t_{10} = 1.1$	$f(t_{10}) = f(1.1) = \ln(1.1) = 0.09531$

1. La formule de la méthode composite des Trapèzes appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$I_T \;\simeq\; h\left(\frac{f(t_0) + f(t_{10})}{2} + \sum_{i=1}^{9} f(t_i)\right) \;=\; h\left(\frac{\ln(t_0) + \ln(t_{10})}{2} + \sum_{i=1}^{9} \ln(t_i)\right) \simeq \textbf{\textit{- 0.67225}}.$$

2. La formule de la méthode composite de Simpson appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$\begin{aligned} I_S \;&\simeq\; \frac{h}{3}\left(f(t_0) + f(t_{10}) + 4\, \textstyle\sum_{i=0}^{4} f(t_{2i+1}) + 2\, \sum_{i=0}^{4} f(t_{2i+2})\right) \\ &= \frac{h}{3}\left(\ln(t_0) + \ln(t_{10}) + 4\, \textstyle\sum_{i=0}^{4} \ln(t_{2i+1}) + 2\, \sum_{i=0}^{4} \ln(t_{2i+2})\right) \;\simeq\; \textbf{\textit{- 0.66548}}. \end{aligned}$$

3. La formule de la méthode composite des rectangles à droite appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$I_{Rd} \;\simeq\; h\, \sum_{i=1}^{10} f(t_i) \;=\; h\, \sum_{i=1}^{10} \ln(t_i) \simeq \textbf{\textit{- 0.55235}}$$

et celles des rectangles à gauche

$$I_{Rg} \simeq h \sum_{i=0}^{9} f(t_i) = h \sum_{i=0}^{9} \ln(t_i) \simeq \textbf{\textit{- 0.79214}}.$$

4. La formule de la méthode composite du point milieu appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$I_{pm} \simeq h \sum_{i=0}^{9} f\left(\frac{t_i + t_{i+1}}{2}\right) = h \sum_{i=0}^{9} \ln\left(\frac{t_i + t_{i+1}}{2}\right) \simeq \textbf{\textit{- 0.661306}}.$$

On remarque bien d'après le tableau ci-dessous, rassemblant les valeurs des intégrales approchées de l'intégrale I, que la méthode de Simpson approche le mieux la valeur de cette intégrale.

Méthode	**Exacte**	**Trapèzes**	**Simpson**	**Rectangle à droite**	**Rectangle à gauche**	**Point milieu**
Valeur	***-0.664900292***	-0.67225	-0.66548	-0.55235	-0.79214	-0.661306

5. Nous regroupons les différents résultats dans le tableau ci-dessous

Méthode	$h = 0.1$	$h = 0.2$	$h = 0.5$
Exacte	***-0.664900292***	***-0.664900292***	***-0.664900292***
Trapèzes	-0.67225	-0.692558	-0.807231
Simpson	-0.66548	-0.6649686	-0.69637816
Rectangle à gauche	-0.55235	-0.32348	-1.406705
Rectangle à droite	-0.79214	-0.452769	-0.20757
Point milieu	-0.89157	-0.651939	-0.606170

Exemple : Déterminer l'intégrale

$$I = \int_0^{\pi} \cos(t)\, dt$$

pour $n = 4$.

1. Par la méthode composite des Trapèzes généralisée

2. Par la méthode composite de Simpson

3. Par la méthode composite des rectangles à droite et à gauche

4. Par la méthode composite des points milieux

On définit la fonction f telle que $f(t) = \cos(t)$ et on a que

$$I = \int_0^{\pi} \cos(t)\, dt = \big(\sin(t)\big)_{(\pi)} - \big(\sin(t)\big)_{(0)} = \textbf{\textit{0}}.$$

Sachant que $n = 4$ correspond à $h = \dfrac{\pi}{4}$, on découpe l'intervalle $[0;\pi]$ en 4 segments dont les bornes équidistants sont telles que

t_i	$f(t_i)$
$t_0 = 0$	$f(t_0) = \cos(0) = 1$
$t_1 = \dfrac{\pi}{4}$	$f(t_1) = \cos\left(\dfrac{\pi}{4}\right) = 0.707106$
$t_2 = \dfrac{\pi}{2}$	$f(t_2) = \cos\left(\dfrac{\pi}{2}\right) = 0$
$t_3 = \dfrac{3\pi}{4}$	$f(t_3) = \cos\left(\dfrac{3\pi}{4}\right) = -0.707106$
$t_4 = \pi$	$f(t_4) = \cos(\pi) = -1$

1. La formule de la méthode composite des Trapèzes appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$I_T \simeq h\left(\frac{f(t_0)+f(t_4)}{2} + \sum_{i=1}^{3} f(t_i)\right) = h\left(\frac{\cos(t_0)+\cos(t_4)}{2} + \sum_{i=1}^{3}\cos(t_i)\right) \simeq \boldsymbol{0}.$$

2. La formule de la méthode composite de Simpson appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$\begin{aligned} I_S &\simeq \frac{h}{3}\left(f(t_0)+f(t_4)+4\ \textstyle\sum_{i=0}^{1} f(t_{2i+1}) + 2\ f(t_2)\right) \\ &= \frac{h}{3}\left(\cos(t_0)+\cos(t_4)+4\ \textstyle\sum_{i=0}^{1}\cos(t_{2i+1}) + 2\ \cos(t_2)\right) \simeq \boldsymbol{0}. \end{aligned}$$

3. La formule de la méthode composite des rectangles à droite appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$I_{Rd} \simeq h\sum_{i=1}^{4} f(t_i) = h\sum_{i=1}^{4}\cos(t_i) \simeq \frac{\pi}{4}$$

et celles des rectangles à gauche

$$I_{Rg} \simeq h\sum_{i=0}^{3} f(t_i) = h\sum_{i=0}^{3}\cos(t_i) \simeq \frac{\pi}{4}.$$

4. La formule de la méthode composite du point milieu appliquée à l'intégrale I pour $n = 10$ et $h = 0.1$ s'écrit :

$$I_{pm} \simeq h \sum_{i=0}^{3} f\left(\frac{t_i + t_{i+1}}{2}\right) = h \sum_{i=0}^{3} \cos\left(\frac{t_i + t_{i+1}}{2}\right) \simeq \boldsymbol{0}.$$

On remarque bien d'après le tableau ci-dessous, rassemblant les valeurs des intégrales approchées de l'intégrale I, que la méthode de Simpson approche le mieux la valeur de cette intégrale.

Méthode	**Exacte**	**Trapèzes**	**Simpson**	**Rectangle à droite**	**Rectangle à gauche**	**Point milieu**
Valeur	***0***	0	0	$\frac{\pi}{4}$	$-\frac{\pi}{4}$	0

5.3 Analyse de l'erreur dans les méthodes d'intégration

On définit

$$E_f = \int_a^b f(t)\, dt - \int_a^b P_f(t)\, dt$$

comme étant l'erreur faite lorsqu'on remplace $\int_a^b f(t)\, dt$ par $\int_a^b P_f(t)\, dt$. On obtient les majoration des erreurs suivantes :

Théorème 5.1. *Si la fonction f est de classe $\mathcal{C}$ (c-à-d continu et différentiable sur l'intervalle $[a, b]$), L'erreur de la méthode des rectangles peut être estimée en utilisant les développements en série de Taylor ou le théorème des accroissements finis on trouve alors pour $h = b - a$:*

$$\left|E_f\right| \leq \frac{(b-a)^2}{2} M'$$

où $M' = \sup_{[a,b]} \left(|f'|\right)$ est un majorant de f' sur l'intervalle $[a, b]$.

On déduit que la méthode des rectangles est une méthode d'ordre 1 car $E_f = \theta\left(\frac{1}{n}\right) = \theta(h)$.

L'erreur de la méthode des rectangles composée est simplement la somme de toutes les erreurs :

$$\left|E_f\right| \leq \frac{(b-a)^2}{2\,n} M'.$$

- *Si la fonction f est de classe $\mathcal{C}^2$ (c-à-d 2 fois continument différentiable sur l'intervalle $[a, b]$), la méthode des trapèzes donnera une erreur majorée par :*

$$\left|E_f\right| \leq \frac{(b-a)^3}{12} M''$$

où $M'' = \sup_{[a,b]} \left(|f''|\right)$ est un majorant de f'' sur l'intervalle $[a, b]$.

On déduit que la méthode des Trapèzes est une méthode d'ordre 2 car $E_f = \theta\left(\frac{1}{n^2}\right) = \theta(h^2)$.

Ainsi la méthode des trapèzes composée donnera une erreur majorée par :

$$|E_f| \leq \frac{(b-a)^3}{12\,n^2}\,M''.$$

- *Si la fonction f est de classe $\mathcal{C}^4$ (c-à-d 4 fois continument différentiable sur l'intervalle $[a,b]$), la méthode de Simpson donnera une erreur majorée par :*

$$|E_f| \leq \frac{(b-a)^5}{90}\,M^{(4)}$$

où $M^{(4)} = \sup_{[a,b]}\left(|f^{(4)}|\right)$ est un majorant de $f^{(4)}$ sur l'intervalle $[a,b]$.
la méthode de Simpson composée donnera une erreur majorée par :

$$|E_f| \leq \frac{(b-a)^5}{180\,n^4}\,M^{(4)}.$$

On déduit que la méthode de Simpson est une méthode d'ordre 4 car $E_f = \theta\left(\frac{1}{n^4}\right) = \theta(h^4)$.

Cette expression du terme d'erreur signifie que la méthode de Simpson est exacte (c'est-à-dire que le terme d'erreur s'annule) pour tout polynôme de degré inférieur ou égal à 3 car elles vérifient $f^{(4)} = 0$.

5.4 Les méthodes de Gauss

La méthodes de Gauss, dites aussi ***méthodes de quadratures de Gauss*** sont des méthodes exactes pour des polynômes de degré inférieur ou égal à $(2n+1)$ points et permettent également de calculer des intégrales singulières où la fonction f à intégrer devient infinie en certains points.

Les méthodes de Gauss consiste à approcher numériquement l'intégrale par la somme pondérée suivante :

$$\int_a^b f(t)\tilde{w}(t)\,dt \simeq \sum_{i=1}^{n} w_i\,f(t_i)$$

où

- w_i sont appelés les ***coefficients de quadrature*** ou ***poids***,

- les points t_i ou ***nœuds*** sont des réels distincts racines des polynômes de Legendre calculés sur $[-1,1]$

- l'intervalle $[a,b]$ peut prendre les formes suivantes $[a,b]$ ou $[a,+\infty]$ ou encore $\mathbb{R}$.

Il existe plusieurs types de la quadrature de Gauss qui varient en fonction du domaine d'intégration $[a,b]$ et de la fonction de pondération $\tilde{w}$, on cite parmi les plus courantes :

- ***Méthode de Gauss-Legendre*** définie pour $[a,b]=[-1,1]$ et $\tilde{w}=1$

$$\int_{-1}^{1} f(t)\,dt \;\simeq\; \sum_{i=1}^{n} w_i\, f(t_i)$$

où les t_i sont les zéros des polynômes de Legendre.

- ***Méthode de Gauss-Tchebychev*** définie pour $[a,b]=[-1,1]$ et $\tilde{w}=\sqrt{1-t^2}$ ou $\tilde{w}=\dfrac{1}{\sqrt{1-t^2}}$

$$\int_{-1}^{1} \frac{1}{\sqrt{1-t^2}} f(t)\,dt \;\simeq\; \sum_{i=1}^{n} w_i\, f(t_i)$$

où les t_i sont les zéros des polynômes de Tchebychev.

- ***Méthode de Gauss-Hermite*** définie pour $[a,b]=\mathbb{R}$ et $\tilde{w}=e^{-t^2}$

$$\int_{-1}^{1} e^{-t^2} f(t)\,dt \;\simeq\; \sum_{i=1}^{n} w_i\, f(t_i)$$

où les t_i sont les zéros des polynômes de Hermite.

5.4.1 Méthode de Gauss-Legendre

La formule de Legendre-Gauss s'écrit sous la forme suivante :

$$\int_{-1}^{1} f(t)\,dt \;\simeq\; \sum_{i=1}^{n} w_i\, f(t_i)$$

où les racines ζ_i sont données dans le tableau ci-dessous :

n	t_i	w_i
2	± 0.5773502691	1
3	0	0.8888888888
	± 0.7745966692	0.5555555555
4	± 0.3399810435	0.6521451548
	± 0.8611363115	0.3478548451
5	0	0.5688888888
	± 0.5384693101	0.4786286704
	± 0.9061798459	0.2369268850
6	± 0.2386191860	0.4679139345
	± 0.6612093864	0.3607615730
	± 0.9324695142	0.1713244923

Dans le cas où l'intégrale n'est pas calculée sur l'intervalle $[-1,1]$, mais sur un domaine $[a,b]$, on procède avec un changement de variable afin d'appliquer la méthode de quadrature de

Gauss-Legendre. Ainsi l'intégrale I est approchée par

$$I \simeq \frac{b-a}{2} \sum_{i=1}^{n} w_i \, f\left(\frac{b-a}{2}\, t_i + \frac{a+b}{2}\right)$$

car

$$\int_a^b f(t)\, dt = \frac{b-a}{2} \int_{-1}^{1} f\left(\frac{b-a}{2}\, t + \frac{a+b}{2}\right) dt.$$

Exemple : Calculer l'intégrale $I = \int_0^1 \sqrt{1+2t}\, dt$ en employant la formule de Gauss-Legendre appliquée à $n = 3$.

On pose $f(t) = \sqrt{1+2t}$. L'intégrale I est approchée par

$$I \simeq \frac{b-a}{2} \sum_{i=1}^{3} w_i \, f\left(\frac{b-a}{2}\, t_i + \frac{a+b}{2}\right) \simeq \frac{1}{2} \sum_{i=1}^{3} w_i \, f\left(\frac{t_i}{2} + \frac{1}{2}\right)$$

avec

t_i	$\frac{t_i}{2} + \frac{1}{2}$	$f\left(\frac{t_i}{2} + \frac{1}{2}\right)$
0	0.5	1.414213562
-0.7745966692	0.1127016654	1.106979372
0.7745966692	0.8872983346	1.665712061

Par suite,

$$I \simeq \frac{1}{2}\Big((0.8888888888) \times (1.414213562) + (0.5555555555) \times (1.106979372) + (0.5555555555) \times (1.665712061)\Big).$$

D'où

$$I = \int_0^1 \sqrt{1+2t}\, dt \simeq 1.398731424.$$

Exemple : Calculer l'intégrale $I = \int_0^{\frac{\pi}{4}} t^2 \cos t \, dt$ en employant la formule de Gauss-Legendre appliquée à $n = 4$.

On pose $f(t) = t^2 \cos t$. L'intégrale I est approchée par

$$I \simeq \frac{b-a}{2} \sum_{i=1}^{4} w_i \, f\left(\frac{b-a}{2}\, t_i + \frac{a+b}{2}\right) \simeq \frac{\pi}{4} \sum_{i=1}^{4} w_i \, f\left(\frac{\pi}{4}(t_i+1)\right)$$

avec

t_i	$\frac{\pi}{4}(t_i+1)$	$f\left(\frac{\pi}{4}(t_i+1)\right)$
-0.3399810435	0.518377676	0.233412695
0.3399810435	1.052418651	0.548776921
-0.8611363115	0.109063285	0.011894778
0.8611363115	1.461733041	0.232569837

Par suite,

$$I \simeq \frac{\pi}{4}\left((0.6521451548)\times(0.233412695+0.548776921)+(0.3478548451)\times(0.011894778+0.232569837)\right).$$

D'où

$$I = \int_0^{\frac{\pi}{4}} t^2 \cos t \, dt \simeq 0.467421367.$$

Exercice :

1. Déterminer l'intégrale

$$I = \int_0^1 e^{-t^2} \, dt$$

pour $n = 8$

1.1 Par la méthode composite des Trapèzes généralisée

1.2 Par la méthode composite de Simpson

1.3 Par la méthode composite des rectangles à droite et à gauche

1.4 Par la méthode composite des points milieux

2. Déterminer le nombre de points n nécessaire afin d'approcher l'intégrale I

2.1 à l'aide de la méthode composite des Trapèzes généralisée avec un ordre de précision de 10^{-2}

2.2 à l'aide de la méthode composite de Simpson avec un ordre de précision de 10^{-4}

2.3 à l'aide de la méthode composite des rectangles avec un ordre de précision de 10^{-1}

3. Évaluer les erreurs faites pour chacune des méthodes précédentes et comparer les avec les résultats de la question 2

On définit la fonction f telle que $f(t) = e^{-t^2}$. Sachant que $n = 8$ correspond à $h = \frac{1}{8} = 0.125$, on découpe l'intervalle $[0; 1]$ en 8 segments dont les bornes équidistants sont telles que

t_i	$f(t_i)$
$t_0 = 0$	$f(t_0) = f(0) = 1$
$t_1 = 0.125$	$f(t_1) = f(0.125) = 0.984496437$
$t_2 = 0.250$	$f(t_2) = f(0.250) = 0.939413062$
$t_3 = 0.375$	$f(t_3) = f(0.375) = 0.868815056$
$t_4 = 0.5$	$f(t_4) = f(0.5) = 0.77880783$
$t_5 = 0.625$	$f(t_5) = f(0.625) = 0.676633846$
$t_6 = 0.750$	$f(t_6) = f(0.750) = 0.569782824$
$t_7 = 0.875$	$f(t_7) = f(0.875) = 0.465043188$
$t_8 = 1$	$f(t_8) = f(1) = 0.367879441$

1. La valeur exacte de l'intégrale I est égale à $0,74682$.

1.1 La formule de la méthode composite des Trapèzes appliquée à l'intégrale I pour $n = 8$ et $h = 0.125$ s'écrit :

$$I_T \simeq h\left(\frac{f(t_0)+f(t_8)}{2}+\sum_{i=1}^{7} f(t_i)\right) = h\left(\frac{e^{-t_0^2}+e^{-t_{10}^2}}{2}+\sum_{i=1}^{7} e^{-t_i^2}\right) \simeq \boldsymbol{0.745865614}.$$

1.2 La formule de la méthode composite de Simpson appliquée à l'intégrale I pour $n = 8$ et $h = 0.125$ s'écrit :

$$\begin{aligned} I_S &\simeq \frac{h}{3}\left(f(t_0)+f(t_8)+4\ \textstyle\sum_{i=0}^{3} f(t_{2i+1})+2\ \sum_{i=0}^{3} f(t_{2i})\right) \\ &= \frac{h}{3}\left(e^{-t_0^2}+e^{-t_{10}^2}+4\ \textstyle\sum_{i=0}^{3} e^{-t_{2i+1}^2}+2\ \sum_{i=0}^{3} e^{-t_{2i}^2}\right) \simeq \boldsymbol{0.74682612}. \end{aligned}$$

1.3 La formule de la méthode composite des rectangles à droite appliquée à l'intégrale I pour $n = 8$ et $h = 0.125$ s'écrit :

$$I_{Rd} \simeq h\sum_{i=1}^{8} f(t_i) = h\sum_{i=1}^{8} e^{-t_i^2} \simeq \boldsymbol{0.785373149}$$

et celles des rectangles à gauche

$$I_{Rg} \simeq h\sum_{i=0}^{7} f(t_i) = h\sum_{i=0}^{7} e^{-t_i^2} \simeq \boldsymbol{0.706358079}.$$

1.4 La formule de la méthode composite du point milieu appliquée à l'intégrale I pour $n = 8$ et $h = 0.125$ s'écrit :

$$I_{pm} \simeq h\sum_{i=0}^{7} f\left(\frac{t_i+t_{i+1}}{2}\right) = h\sum_{i=0}^{7} e^{-\frac{(t_i+t_{i+1})^2}{2}} \simeq \boldsymbol{0.747303578}.$$

2. Déterminons les valeurs de n selon :

2.1 la méthode des trapèzes composée donnera une erreur majorée par :

$$\big|E_f\big| \ \leq \ \frac{1}{12\,n^2} \ \sup_{[0,1]} \big(|f''(t)|\big)$$

Afin de déterminer la valeur de $\sup_{[0,1]} \big(|f''|\big)$, construisons tout d'abord le tableau de variation de la fonction f''. Sachant que $f'(t) = -2te^{-t^2}$, $f''(t) = -2e^{-t^2} + 4t^2e^{-t^2}$ et $f^{(3)}(t) = 4te^{-t^2}(3-2t^2)$, on a :

t	$-\frac{\sqrt{3}}{2}$		0		1		$\frac{\sqrt{3}}{2}$
t		$-$	$\vert$	$+$	$\vert$	$+$	
e^{-t^2}		$+$	$\vert$	$+$	$\vert$	$+$	
$3-2t^2$		$+$	$\vert$	$+$	$\vert$	$+$	
$f^{(3)}(t)$		$-$	$\vert$	$+$	$\vert$	$+$	

Sur l'intervalle $[0,1]$, la fonction $f^{(3)}(t)$ est croissante et on a $f''(0) = -2$ et $f''(1) = \frac{2}{e}$. Ainsi la fonction $|f''(t)|$ atteint son sup pour $|f''(0)| = 2$. D'où

$$\big|E_f\big| \ \leq \ \frac{1}{6\,n^2}.$$

Par suite, le nombre de points n nécessaire afin d'approcher l'intégrale I à l'aide de la méthode composite des Trapèzes généralisée avec un ordre de précision de 10^{-2} est égal à :

$$\frac{1}{6\,n^2} \ \leq 10^{-2} \implies n \geq \frac{10}{\sqrt{6}} \implies \boldsymbol{n \ \geq \ 4.0824.}$$

2.2 la méthode de Simpson composée donnera une erreur majorée par :

$$\big|E_f\big| \ \leq \ \frac{1}{180\,n^4} \ \sup_{[0,1]} \big(|f^{(4)}(t)|\big)$$

Afin de déterminer la valeur de $\sup_{[0,1]} \big(|f^{(4)}(t)|\big)$, construisons tout d'abord le tableau de variation de la fonction $f^{(3)}(t)$. Sachant que $f^{(4)}(t) = 4e^{-t^2}(3 - 12t^2 + 4t^4)$ et $f^{(5)}(t) = 8te^{-t^2}(t^2-1)(-4t^2+15)$, sur l'intervalle $[0,1]$, la fonction $f^{(3)}(t)$ est croissante et on a $f^{(4)}(0) = 12$ et $f^{(4)}(1) = \frac{20}{e}$. Ainsi la fonction $|f^{(4)}(t)|$ atteint son sup pour $|f^{(4)}(0)| = 12$. D'où

$$\big|E_f\big| \ \leq \ \frac{12}{180\,n^4}.$$

Par suite, le nombre de points n nécessaire afin d'approcher l'intégrale I à l'aide de la méthode composite des Trapèzes généralisée avec un ordre de précision de 10^{-2} est égal à :

$$\frac{1}{15\,n^4} \ \leq 10^{-4} \implies \boldsymbol{n} \geq \frac{10}{\sqrt[4]{15}}.$$

3. Déterminons les erreurs relatives à $n = 8$ selon les méthodes suivantes :

3.1 la méthode des trapèzes composée donnera une erreur majorée par :

$$\left|E_f\right| \leq \frac{1}{6\,n^2} = \frac{1}{6\times 8^2} = 0.002604166.$$

3.2 la méthode de Simpson composée donnera une erreur majorée par :

$$\left|E_f\right| \leq \frac{12}{180\,n^4} = \frac{1}{15\times 8^4} = 0.00001627.$$

Ces résultats sont tout à fait en accord avec les considérations théoriques précédentes. Noter comment la méthode des Trapèzes donne une approximation à $2.10^{-3} \leq 10^{-2}$ et la méthode de Simpson donne une approximation à $2.10^{-5} \leq 10^{-4}$ avec seulement 8 valeurs de la fonction f.

Chapitre 6

Résolution des équations différentielles

6.1 Solution générale d'une équation différentielle

Soit $\mathcal{O}$ un ouvert de $\mathbb{R} \times \mathbb{R}^m$ et $f : \mathcal{O} \to \mathbb{R}^m$ une application continue. On considère l'équation différentielle :

$$y' \ = \ f(t, y), \quad (t, y) \in \mathcal{O}. \tag{6.1}$$

On appelle ***solution sur*** $I \in \mathbb{R}$ ***de l'équation différentielle*** :

$$\left\{ \begin{array}{l} \textit{toute application} \quad y : U \to \mathbb{R}^m, \ \textit{dérivable sur} I \ \textit{telle que} \\ (t, y(t)) \in \mathcal{O}, \ \forall t \in U \\ y'(t) \ = \ f(t, y(t)), \ \forall t \in U \end{array} \right.$$

avec $y(t) \ = \ (y_1(t), \ y_2(t), ..., \ y_m(t))$ et $f \ = \ (f_1(t), \ f_2(t), ..., \ f_m(t))$.

Ainsi l'équation 6.1 s'écrit sous forme d'un système différentiel du premier ordre à m fonctions inconnues $(y_1, \ y_2, ..., \ y_m)$:

$$\left\{ \begin{array}{rcl} y_1'(t) & = & f_1(t, y_1(t), \ y_2(t), ..., \ y_m(t)), \\ y_2'(t) & = & f_2(t, y_1(t), \ y_2(t), ..., \ y_m(t)), \\ & \vdots & \\ y_m'(t) & = & f_m(t, y_1(t), \ y_2(t), ..., \ y_m(t)). \end{array} \right.$$

6.2 Problème de Cauchy

Le problème de Cauchy consiste à déterminer une solution en partant d'une valeur donnée t_0, appelée aussi *donnée initiale*.

Définition 6.1. ***Problème de Cauchy.*** *Étant donné un point $(t_0, y_0) \in \mathcal{O}$, le problème de Cauchy consiste à déterminer une solution $y : U \to \mathbb{R}^m$ de l'équation différentielle 6.1 sur un intervalle $U = [0, T]$ contenant t_0, avec T un réel strictement positif et telle que $y(t_0) = y_0$.*

Le problème de Cauchy s'écrit :

$$\boxed{\begin{cases} y'(t) &= f(t, y(t)), \\ y(t_0) &= y_0 \end{cases}}$$

Résoudre le problème de Cauchy revient à prévoir l'évolution du problème suivant le paramètre t, sachant qu'en $t = t_0$; le système est décrit par les paramètres $y_0 = (y_{01},\ y_{02}, ...,\ y_{0m})$. (t_0, y_0) sont les données initiales du problème de Cauchy.

Lemme 6.1. *Une fonction $y : U \to \mathbb{R}^m$ est une solution du problème de Cauchy de données initiales (t_0, y_0) si et seulement si*

(1) *La fonction y est continue*

(2) $(t, y(t)) \in \mathcal{O},\ \forall t \in U$

(3) $\forall t \in U,\ y(t) \ = \ y_0 \ + \ \int_{t_0}^{t} f(x, y(x))\, dx$

Ce lemme montre que la résolution d'une équation différentielle est équivalente à la résolution d'une équation intégrale et donc la solution à l'instant t dépend uniquement de la solution aux instants $\tilde{t}$ tels que $t_0 \leq \tilde{t} \leq t$.

Théorème 6.1. ***(Théorème de Cauchy-Lipschitz)*** *Si la fonction f est continue par rapport aux deux variables t et y, et que f est uniformément Lipschitzienne par rapport à y dans le sens :*

$$\exists L > 0,\ \forall t \in [t_0, T],\ \forall y_1, y_2 \in \mathbb{R},\ |f(t, y_1) - f(t, y_2)| \leq \ L\, |y_1 - y_2|$$

alors par tout point (t_0, y_0) passe une solution unique et toute suite de solutions approchées converge vers la solution exacte $y \in \mathcal{C}^1([0, T])$.

6.3 Résolution explicite des équations différentielles

Ce paragraphe est consacré à la résolution d'équations de type suivant :

- Équations différentielles linéaires sans second membre : $y'(t) \ = \ \lambda(t)\, y(t)$,
- Équations différentielles linéaires avec second membre : $y'(t) \ = \ \lambda(t)\, y(t) \ + \ \beta(t)$.

6.3.1 Résolution des équations différentielles linéaires sans second membre de la forme $y'(t) = \lambda(t)\, y(t)$

Théorème 6.2. *Soit $\alpha : U \to \mathbb{R}$ une primitive de la fonction λ sur U. L'ensemble des solutions sur l'intervalle U de l'équation différentielle $y'(t) \ = \ \lambda(t)\ y(t)$ est l'ensemble des fonctions $y : U \to \mathbb{R}$, telles que :*

$$y(t) = C\, e^{\alpha(t)},\ \ \forall\, t \in U$$

où C est une constante réelle quelconque.

Recherche d'une solution des équations différentielles linéaires sans second membre :

L'équation différentielle s'écrit sous la forme :

$$\frac{y'(t)}{y(t)} = \lambda(t)$$

où λ est une fonction continue. Si on prend une primitive de chaque membre, les deux primitives doivent être égales à une constante près :

$$\begin{aligned} \ln|y(t)| &= \alpha(t) + C,\ C \in \mathbb{R} \\ |y(t)| &= e^C\, e^{\alpha(t)} \\ y(t) &= \eta\, e^C\, e^{\alpha(t)},\ \text{ avec } \eta = \pm 1. \end{aligned}$$

Exemple : Résoudre et déterminer une solution sur $\mathbb{R}$ de l'équation différentielle :

$$y'(t) - \frac{2}{t}\, y(t) = 0.$$

Observons que dans ce cas la fonction $\lambda(t) = \dfrac{2}{t}$ n'est pas définie en $t = 0$. On pourra donc résoudre l'équation différentielle sur les deux intervalles $]-\infty; 0[$ et $]0; +\infty[$ séparément. La méthode de résolution se détaille comme suit : La fonction $t \to \dfrac{2}{t}$ est continue sur chacun des intervalles et on a :

$$\frac{y'(t)}{y(t)} = \frac{2}{t}.$$

Si on prend une primitive de chaque membre, les deux primitives doivent être égales à une constante près :

$$\ln|y(t)| = \ln|t^2| + \ln C,\ C \in \mathbb{R}.$$

L'exponentielle des deux membres donne

$$y(t) = C\, t^2.$$

La solution de l'équation différentielle est donc

$$y(t) = \begin{cases} C_1\, t^2 & \text{sur }]-\infty, 0[, \\ C_2\, t^2 & \text{sur }]0, +\infty[. \end{cases}$$

Un cas particulier correspond à $C_1 = -4$ et $C_2 = 1$. La fonction définie par :

$$y(t) = \begin{cases} -4\, t^2 & \text{sur }]-\infty; 0[, \\ t^2 & \text{sur }]0; +\infty[\end{cases}$$

est solution de l'équation $y'(t) - \dfrac{2}{t}\, y(t) = 0$ sur $]-\infty; 0[\times]0; +\infty[$.

On peut aussi chercher une solution vérifiant une condition initiale.

Proposition 6.1. *Soit t_0 un point de l'intervalle U et y_0 un réel quelconque. Il existe une fonction $y(t)$ vérifiant $y'(t) = \lambda(t)\, y(t)$ et telle que $y(t_0) = y_0$. Sa solution est définie par :*

$$y(t) = y_0 \left(\int_{t_0}^{t} \lambda(x)\, dx \right).$$

Remarque 6.1. *On applique les mêmes techniques de résolution des équations différentielles linéaires sans second membre pour les équations différentielles du type :*

$$y'(t) = \lambda(t)\ f(y(t)).$$

Théorème 6.3. ***Existence et unicité de la solution du problème de Cauchy*** *Pour tout couple $(t_0, y_0) \in U \times \mathbb{R}$, il existe une solution et une seule y de l'équation différentielle $y'(t) = \lambda(t)\ y(t)$ sur l'intervalle U telle que $y(t_0) = y_0$.*

Définition 6.2. *L'équation différentielle $y'(t) = \lambda(t)\ y(t)$ admet une infinité de solutions. Elles sont toutes du type $y(t) = C\ e^{\alpha(t)}$, avec $C \in \mathbb{R}$. L'ensemble de ces fonctions solutions est appelé* ***famille de fonctions*** *et l'ensemble des courbes représentatives de ces fonctions est appelé* ***famille de courbes****.*

Exemple : Résoudre et déterminer une solution sur $\mathbb{R}$ de l'équation différentielle :

$$y'(t) - \frac{t^3}{1+t^4}\ y(t) = 0, \quad \text{avec } y(0) = 1.$$

La fonction $\lambda(t) = \dfrac{t^3}{1+t^4}$ est continue sur $\mathbb{R}$ et admet comme primitive la fonction $t \to \dfrac{1}{4}\ \ln(1+t^4) = (\ln(1+t^4))^{\frac{1}{4}}$, d'où

$$\begin{aligned} \frac{y'(t)}{y(t)} &= \frac{t^3}{1+t^4} \\ \ln|y(t)| &= \frac{1}{4}\ \ln(1+t^4)\ +\ \ln C,\ C \in \mathbb{R}. \end{aligned}$$

La solution générale de l'équation différentielle est donc la fonction définie par

$$y(t) = C\ e^{\ln \sqrt[4]{1+t^4}}\ =\ C\ \sqrt[4]{1+t^4}.$$

Puisque la donnée d'une condition détermine une solution unique, le calcul de la solution de l'équation différentielle vérifiant $y(0) = 1$ est :

$$y(0)\ =\ C\ \sqrt[4]{1+0^4}\ =\ 1 \iff C\ =\ 1.$$

Par suite, la solution de l'équation $y'(t) - \dfrac{t^3}{1+t^4}\ y(t) = 0$ vérifiant $y(0) = 1$ est :

$$y(t) = \sqrt[4]{1+t^4}.$$

Exemple : Résoudre et déterminer une solution sur l'intervalle $]0;+\infty[$ de l'équation différentielle :

$$\sqrt{t}\ y'(t)\ -\ y(t) = 0.$$

La fonction $\lambda(t) = \dfrac{1}{\sqrt{t}}$ est continue sur $]0;+\infty[$ et admet comme primitive la fonction $t \to 2\sqrt{t}$ à une constante près. Par suite, la solution de l'équation $\sqrt{t}\ y'(t) - y(t) = 0$ est :

$$y(t) = C\ e^{2\sqrt{t}}.$$

Exemple : Résoudre et déterminer une solution sur l'intervalle $]0;\frac{\pi}{2}[$ de l'équation différentielle :

$$y'(t) - (1 + \tan^2 t)\, y(t) = 0.$$

La fonction $\lambda(t) = 1 + \tan^2 t = \dfrac{1}{\cos^2 t}$ est continue sur $]0;\frac{\pi}{2}[$ et admet comme primitive la fonction $t \to \tan t$ à une constante près. Par suite, la solution de l'équation $y'(t) - (1 + \tan^2 t)\, y(t) = 0$ est :

$$y(t) = C\, e^{\tan t}.$$

Proposition 6.2. *Dans le cas particulier où la fonction λ est la fonction constante, la solution de l'équation différentielle $y'(t) = \lambda y(t)$ est :*

$$y(t) = C\, e^{\lambda\, t}$$

Exemple : Résoudre et déterminer une solution sur l'intervalle $\mathbb{R}$ de l'équation différentielle :

$$y'(t) + 4\, y(t) = 0, \quad \text{avec } y(0) = 1.$$

La fonction $\lambda(t) = 4$ est continue sur $\mathbb{R}$ et admet comme primitive la fonction $t \to 4\, t$ à une constante près. Par suite, la solution de l'équation $y'(t) + 4\, y(t) = 0$ est :

$$y(t) = C\, e^{-4\, t}.$$

Puisque la donnée d'une condition détermine une solution unique, le calcul de la solution de l'équation différentielle vérifiant $y(0) = 1$ est :

$$y(0) \;=\; C\, e^{-4} \;=\; 1 \iff C \;=\; 1.$$

Par suite, la solution de l'équation $y'(t) + 4\, y(t) = 0$ vérifiant $y(0) = 1$ est :

$$y(t) = e^{-4\, t}.$$

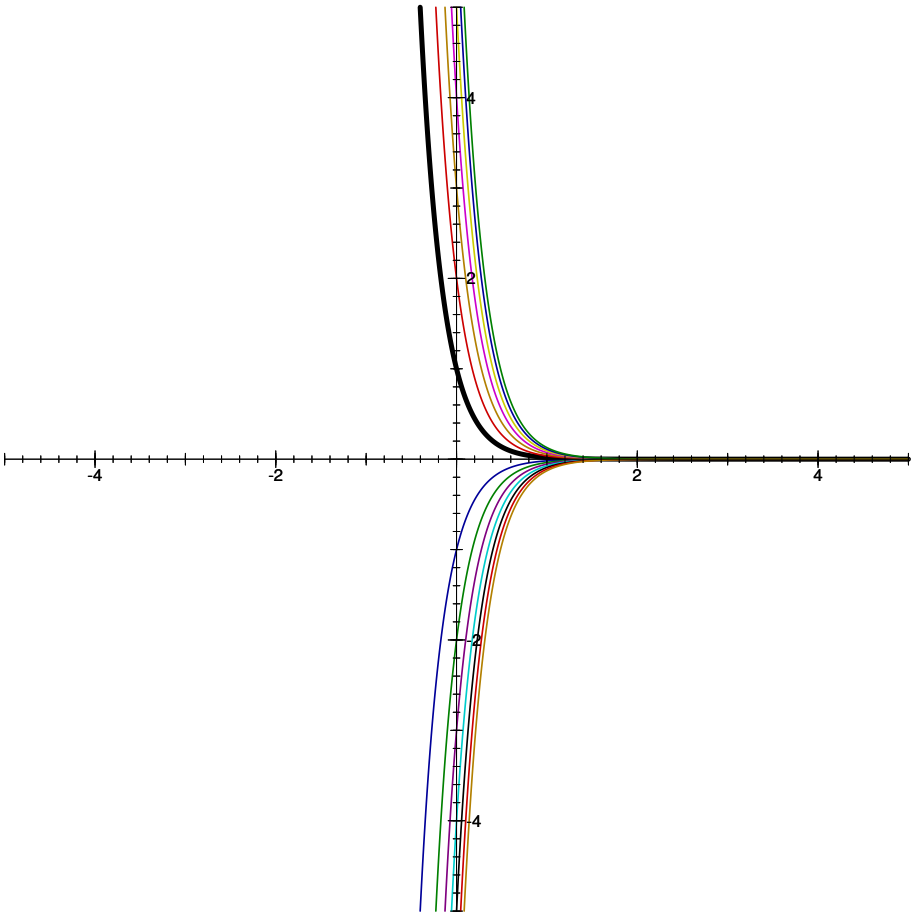


FIGURE 6.1 – Partie de la famille de fonctions solutions de l'équation différentielle $y'(t)+4\ y(t) = 0$ (La fonction solution avec condition initiale est représentée en noir).

6.3.2 Résolution des équations différentielles linéaires sans second membre de la forme $y'(t) + \lambda(t)\ y(t) = \gamma(t)$

On cherche à résoudre des équations du type :

$$y'(t) = \lambda(t)\ y(t) + \beta(t)$$

où λ et β, le second membre, sont deux fonctions définies et continues sur un intervalle U de $\mathbb{R}$.

Théorème 6.4. *Soit $\alpha : U \to \mathbb{R}$ une primitive de la fonction λ sur U et $\tilde{y}$ une solution particulière de l'équation différentielle $y'(t) = \lambda(t)\ y(t) + \beta(t)$ définie sur l'intervalle U.*
L'ensemble des solutions sur l'intervalle U de l'équation différentielle $y'(t) = \lambda(t)\ y(t) + \beta(t)$ est l'ensemble des fonctions $y : U \to \mathbb{R}$, telles que :

$$y(t) = \tilde{y}(t) + C\ e^{\alpha(t)}, \quad \forall\ t \in U$$

où C est une constante réelle quelconque.

On retiendra que la solution générale de l'équation différentielle $y'(t) = \lambda(t)\ y(t) + \beta(t)$ est la somme d'une solution particulière et de la solution de l'équation homogène.

$$\underbrace{y}_{\text{solution générale de l'équation avec second membre}} = \underbrace{\tilde{y}}_{\text{solution particulière}} + \underbrace{C\ e^{\alpha(t)}}_{\text{solution générale de l'équation sans second membre}}$$

Recherche d'une solution des équations différentielles linéaires avec second membre : Méthode de variation de la constante

Pour résoudre une équation différentielle du premier ordre $y'(t) = \lambda(t)\ y(t) + \beta(t)$, on procède en deux étapes.

Étape 1 : Déterminer les solutions de l'équation différentielle linéaire sans second membre associée. Ces solutions sont de la forme : $C\ e^{\alpha(t)}$ où C est une constante réelle et α une primitive de la fonction λ sur l'intervalle U.

Étape 2 : Trouver une solution particulière $\tilde{y}$ de l'équation avec second membre $y'(t) = \lambda(t)\ y(t) + \beta(t)$, à l'aide de la méthode de variation de la constante qui n'est valable que pour les équations différentielles linéaires et qui consiste à chercher la solution particulière $\tilde{y}$ sous la forme :

$$\tilde{y} = C(t)\ e^{\alpha(t)},$$

où $C(t)$ est une fonction différentiable. En dérivant la fonction $\tilde{y}$, il vient alors :

$$\tilde{y}'(t) = C'(t)\ e^{\alpha(t)} + \lambda(t)\ C(t)\ e^{\alpha(t)} = \lambda(t)\ \tilde{y}(t) + C'(t)\ e^{\alpha(t)}.$$

En supposant que $\tilde{y}$ est solution de l'équation différentielle avec second membre, on obtient :

$$\tilde{y}'(t) = \lambda(t)\ \tilde{y}(t) + \beta(t) = \lambda(t)\ \tilde{y}(t) + C'(t)\ e^{\alpha(t)}.$$

Soit après simplification on en déduit $C'(t) = \beta(t)\ e^{-\alpha(t)}$ et on obtient la fonction C par primitivation et ensuite la solution particulière :

$$\tilde{y}(t) = e^{\alpha(t)} \int_{t_0}^{t} \beta(s)\ e^{-\alpha(s)}\ ds = exp\left(\int_{t_0}^{t} \lambda(s)\ ds\right)\left(\int_{t_0}^{t} \beta(s)\ exp\left(-\int_{t_0}^{s} \lambda(u)\ du\right)\right).$$

Si de plus une condition initiale est imposée, alors on ajuste la constante C en conséquence.

Exemple : Résoudre et déterminer une solution sur $\mathbb{R}$ de l'équation différentielle :

$$y'(t) = \frac{t^3}{1+t^4}\ y(t) + \sqrt[4]{1+t^4}, \quad \text{avec } y(0) = 1.$$

La solution de l'équation différentielle sans second membre $y'(t) = \dfrac{t^3}{1+t^4}\ y(t)$ est la fonction définie par

$$y(t) = C\ \sqrt[4]{1+t^4}.$$

Remplaçons maintenant la constante C par une fonction $C(t)$, dérivable sur $\mathbb{R}$, dans la solution $y(t)$ et en dérivant la fonction y, il vient alors :

$$y'(t) = C'(t)\sqrt[4]{1+t^4} + C(t)\frac{t^3}{\sqrt[4]{(1+t^4)^3}}.$$

En supposant que $C(t)\sqrt[4]{1+t^4}$ est solution de l'équation différentielle linéaire avec second membre, cela donne :

$$y'(t) = C(t)\frac{t^3}{\sqrt[4]{(1+t^4)^3}} + \sqrt[4]{1+t^4} = C(t)\frac{t^3}{\sqrt[4]{(1+t^4)^3}} + C'(t)\sqrt[4]{1+t^4}.$$

Il reste alors après simplification $C'(t) = 1$ d'où $C(t) = t$.

Une solution particulière est donc la fonction définie par $\tilde{y}(t) = t\sqrt[4]{1+t^4}$.
La solution générale obtenue par superposition de solutions est ainsi $y(t) = t\sqrt[4]{1+t^4} + C\sqrt[4]{1+t^4}$.
En tenant compte de la condition initiale $y(0) = 1$, on obtient $C = 1$.

Ainsi, la solution générale de l'équation différentielle avec second membre $y'(t) = \dfrac{t^3}{1+t^4}\, y(t) + \sqrt[4]{1+t^4}$ vérifiant $y(0) = 1$, est :

$$y(t) = (t+1)\sqrt[4]{1+t^4}.$$

Exemple : Résoudre et déterminer une solution sur $\mathbb{R}$ de l'équation différentielle :

$$t\, y'(t) = 2\, y(t) + t^4, \quad \text{avec } y(1) = \frac{1}{2} \text{ et } y(-1) = -\frac{1}{2}.$$

La solution de l'équation différentielle sans second membre $y'(t) = \dfrac{2}{t}\, y(t)$ est la fonction définie par

$$y(t) \;=\; \begin{cases} C_1\, t^2 & \text{sur }]-\infty, 0[, \\ C_2\, t^2 & \text{sur }]0, +\infty[. \end{cases}$$

Remplaçons maintenant les constantes C_1 et C_2 par deux fonctions $C_1(t)$ et $C_2(t)$, dérivables respectivement sur $\left]-\infty; 0\right[$ et $\left]0; \infty\right[$, dans la solution $y(t)$ et en dérivant la fonction y par rapport à t, il vient alors :

$$y'(t) \;=\; \begin{cases} C_1'(t)t^2 + 2\, t\, C_1(t) & \text{sur }]-\infty, 0[, \\ C_2'(t)t^2 + 2\, t\, C_2(t) & \text{sur }]0, +\infty[. \end{cases}$$

En supposant que $C(t)t^2$ est solution de l'équation différentielle linéaire avec second membre, cela donne :

$$\begin{cases} t^3 + 2\, t\, C_1(t) = C_1'(t)\, t^2 + 2\, t\, C_1(t) & \text{sur }]-\infty, 0[, \\ t^3 + 2\, t\, C_2(t) = C_2'(t)\, t^2 + 2\, t\, C_2(t) & \text{sur }]0, +\infty[. \end{cases}$$

Il reste alors après simplification $C_1'(t) = C_2'(t) = t$ d'où $C_1(t) = C_2(t) = \dfrac{t^2}{2}$.

Une solution particulière est donc la fonction définie par $\tilde{y}(t) = \dfrac{t^4}{2}$.

La solution générale obtenue par superposition de solutions est ainsi

$$y(t) \;=\; \begin{cases} C_1\, t^2 + \dfrac{t^4}{2} & \text{sur }]-\infty, 0[, \\ C_2\, t^2 + \dfrac{t^4}{2} & \text{sur }]0, +\infty[. \end{cases}$$

En tenant compte de la condition initiale $y(1) = \dfrac{1}{2}$ et $y(-1) = \dfrac{-1}{2}$, on obtient respectivement $C_1 = 0$ et $C_2 = -1$ sur les intervalles $\left]-\infty; 0\right[$ et $\left]0; \infty\right[$.

Ainsi, la solution générale de l'équation différentielle avec second membre est :

$$y(t) \;=\; \begin{cases} \dfrac{t^4}{2} & \text{sur }]-\infty, 0[, \\ \dfrac{t^4}{2} - t^2 & \text{sur }]0, +\infty[. \end{cases}$$

Exemple : Résoudre et déterminer une solution sur $\mathbb{R}$ de l'équation différentielle :

$$t\,(t-1)\,y'(t) + 2\,y(t) + t^2, \;\text{ avec } y\left(-\frac{1}{2}\right) = \frac{\ln 2}{9}, \;\; y\left(\frac{1}{2}\right) = \ln 2 \;\text{ et }\; y\left(\frac{3}{2}\right) = 9\ln\frac{2}{3}.$$

Observons que dans ce cas la fonction $\lambda(t) = \dfrac{-2}{t\,(t-1)}$ n'est pas définie en $t = 0$ et en $t = 1$. On pourra donc résoudre l'équation différentielle sur les trois intervalles $]-\infty; 0[$, $]0; 1[$ et $]1; +\infty[$ séparément. La méthode de résolution se détaille comme suit : La fonction $t \to \dfrac{-2}{t\,(t-1)}$ est continue sur chacun des intervalles et on a :

$$\frac{y'(t)}{y(t)} = \frac{-2}{t\,(t-1)}.$$

Si on prend une primitive de chaque membre, les deux primitives doivent être égales à une constante près :

$$\ln|y(t)| = -\ln\left|\frac{(t-1)^2}{t^2}\right| \;+\; \ln C,\; C \in \mathbb{R}.$$

L'exponentielle des deux membres donne

$$y(t) = C\,\frac{t^2}{(t-1)^2}.$$

La solution de l'équation différentielle est donc

$$y(t) \;=\; \begin{cases} C_1\,\dfrac{t^2}{(t-1)^2} & \text{sur }]-\infty, 0[, \\ C_2\,\dfrac{t^2}{(t-1)^2} & \text{sur }]0, 1[, \\ C_3\,\dfrac{t^2}{(t-1)^2} & \text{sur }]1, +\infty[. \end{cases}$$

Remplaçons maintenant les constantes C_1, C_2 et C_3 par deux fonctions $C_1(t)$, $C_2(t)$ et $C_3(t)$, dérivables respectivement sur $]-\infty;0[$, $]0;1[$ et $]1;\infty[$, dans la solution $y(t)$ et en dérivant la fonction y par rapport à t, il vient alors :

$$y'(t) \;=\; \begin{cases} C_1'(t)\,\dfrac{t^2}{(t-1)^2} - C_1(t)\,\dfrac{2\,t}{(t-1)^3} & \text{sur }]-\infty,0[, \\ C_2'(t)\,\dfrac{t^2}{(t-1)^2} - C_2(t)\,\dfrac{2\,t}{(t-1)^3} & \text{sur }]0,1[, \\ C_3'(t)\,\dfrac{t^2}{(t-1)^2} - C_3(t)\,\dfrac{2\,t}{(t-1)^3} & \text{sur }]1,+\infty[. \end{cases}$$

En reportant dans l'équation différentielle linéaire avec second membre, cela donne :

$$\begin{cases} C_1'(t)\,\dfrac{t^3}{t-1} - C_1(t)\,\dfrac{2\,t^2}{(t-1)^2} + C_1(t)\,\dfrac{2\,t^2}{(t-1)^2} = t^2 & \text{sur }]-\infty,0[, \\ C_2'(t)\,\dfrac{t^3}{t-1} - C_2(t)\,\dfrac{2\,t^2}{(t-1)^2} + C_2(t)\,\dfrac{2\,t^2}{(t-1)^2} = t^2 & \text{sur }]0,1[, \\ C_3'(t)\,\dfrac{t^3}{t-1} - C_3(t)\,\dfrac{2\,t^2}{(t-1)^2} + C_3(t)\,\dfrac{2\,t^2}{(t-1)^2} = t^2 & \text{sur }]1,+\infty[. \end{cases}$$

Il reste alors après simplification $C_1'(t) = C_2'(t) = C_3'(t) = \dfrac{t-1}{t}$ d'où $C_1(t) = C_2(t) = C_3(t) = t - \ln t$.

Une solution particulière est donc la fonction définie par $\tilde{y}(t) = \dfrac{t^2\,(t-\ln t)}{(t-1)^2}$.

La solution générale obtenue par superposition de solutions est ainsi

$$y(t) \;=\; \begin{cases} C_1\,\dfrac{t^2}{(t-1)^2} + \dfrac{t^2\,(t-\ln t)}{(t-1)^2} & \text{sur }]-\infty,0[, \\ C_2\,\dfrac{t^2}{(t-1)^2} + \dfrac{t^2\,(t-\ln t)}{(t-1)^2} & \text{sur }]0,1[, \\ C_3\,\dfrac{t^2}{(t-1)^2} + \dfrac{t^2\,(t-\ln t)}{(t-1)^2} & \text{sur }]1,+\infty[. \end{cases}$$

En tenant compte de la condition initiale $y\left(-\dfrac{1}{2}\right) = \dfrac{\ln 2}{9}$, $y\left(\dfrac{1}{2}\right) = \ln 2$ et $y\left(\dfrac{3}{2}\right) = \ln\dfrac{2}{3}$, on obtient respectivement $C_1 = \dfrac{1}{2}$, $C_2 = -\dfrac{1}{2}$ et $C_3 = -\dfrac{3}{2}$ sur les intervalles $]-\infty;0[$, $]0;1[$ et $]1;\infty[$.

Ainsi, la solution générale de l'équation différentielle avec second membre est :

$$y(t) = \begin{cases} \dfrac{t^2\left(t-\ln t+\dfrac{1}{2}\right)}{(t-1)^2} & \text{sur }]-\infty,0[, \\ \dfrac{t^2\left(t-\ln t-\dfrac{1}{2}\right)}{(t-1)^2} & \text{sur }]0,1[, \\ \dfrac{t^2\left(t-\ln t-\dfrac{3}{2}\right)}{(t-1)^2} & \text{sur }]1,+\infty[. \end{cases}$$

6.4 Résolution numérique des équations différentielles

Les techniques de résolution des équations différentielles sont basées sur les formulations d'intégration numérique pour les second membre $f(t, y(t))$ (méthode de trapèze, méthode de Simpson, méthode des rectangles,...) ou les développements de Taylor au voisinage de y_n. Introduisons tout d'abord les schémas explicites et schémas implicites :

Schéma explicite et schéma implicite : Si une approximation y_{n+1} de la valeur de la fonction $y(t)$ cherchée en un point t_{n+1} s'exprime en fonction de y_n, t_n et le pas d'espace h, on dit que le ***schéma*** est ***explicite*** et on écrit :

$$y_{n+1} = y_n + \varphi(t_n, y_n, h).$$

Si de plus, on fait intervenir y_{n+1}, on dit que le ***schéma*** est ***implicite*** et on écrit :

$$y_{n+1} = y_n + \varphi(t_n, y_n, y_{n+1}, h).$$

On distingue deux grandes familles de schémas de résolution numérique des problèmes aux conditions initiales pour les équations différentielles :

Les schémas à un pas : Ce type de schéma consiste à calculer une approximation y_{n+1} de la valeur de la fonction $y(t)$ cherchée en un point t_{n+1}, en fonction de y_n, t_n et le pas d'espace h en utilisant des formules d'intégration numérique approchée. A titre d'exemple, on cite les *schémas d'Euler explicite, amélioré ou schéma de Lax-Wendroff, implicite, modifié* et le *schéma de Cranck Nicolson.*

Les schémas à multi-pas ou pas multiples : Dans le but de construire des méthodes d'ordre de précision élevé, on augmente le nombre de pas. Dans ce cas, la solution au point t_{n+1} sera calculée en fonction des solutions aux pas précédents t_n, t_{n-1},..., i.e, on utilise les solutions calculées aux pas antérieurs y_n, y_{n-1}, y_{n-2}, ... On cite à titre d'exemple les *schémas d'Adams-Bashforth*, les *schémas de Gear*, les *schémas de de Nystrom ou saute-mouton* ou aussi les *schémas implicites d'Adams-Moulton.*

6.4.1 Méthodes numériques à un pas

Étant donné une subdivision $t_0 < t_1 < ... < t_N$ de $[t_0, t_N]$ avec $t_N = t_0 + T$, on cherche à calculer les valeurs approchées $y(t_0),\ y_1, ...,\ y(t_N)$ des valeurs $y(t_n)$ prises par la solution exacte y. Les pas successifs sont notés

$$h_n = t_{n+1} - t_n, \quad 0 \leq n \leq N-1,$$

et on pose

$$h_{max} = \max(h_1,\ h_2, ...,\ h_{N-1}).$$

6.4.1.1 Schéma d'Euler explicite et implicite

Sachant que la solution exacte $y(t)$ d'une équation différentielle vérifie $y'(t) = f(t, y(t))$, on écrit :

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

La ***méthode d'Euler*** (ou ***méthode de la tangente***) consiste à construire une solution approchée y affine par morceaux en approchant l'intégrale par une formule utilisant des valeurs de $f(t, y(t))$ sur l'intervalle $[t_n, t_{n+1}]$ bien que $y(t)$ ne soit pas connue sur cet intervalle.

Ainsi, en appliquant la *formule de Taylor* appliquée à $y(t)$ en $t = t_n$ (appelée aussi la *méthode des rectangles "à gauche"*) et en supposant que que le pas $h_n = t_{n+1} - t_n$ est constant, $h = \frac{T}{N}$ où N est un entier, on a au premier ordre l'approximation

$$y(t_{n+1}) = y(t_n + h) = y(t_n) + h\ y'(t_n) = y(t_n) + h\ f(t_n, y(t_n)).$$

On propose donc le schéma suivant :

Schéma d'Euler explicite : Partant de la donnée initiale y_0, on calcule donc y_{n+1} par récurrence en posant

$$\left\{ \begin{array}{lcl} t_{n+1} & = & t_n + h, \quad 0 \leq n \leq N-1, \\ y_{n+1} & = & y_n + hf(t_n, y_n). \end{array} \right.$$

La solution approchée y s'obtient graphiquement en traçant pour chaque n les segments joignant les points $(t_n, y(t_n))$ et $(t_{n+1}, y(t_{n+1}))$ puisqu'on confond la courbe intégrale sur $[t_n, t_{n+1}]$ avec sa tangente au point $(t_n, y(t_n))$.

On peut aussi approcher l'intégrale en utilisant la *formule de Taylor* appliquée à $y(t)$ en $t = t_n$ (appelée aussi la *méthode des rectangles "à droite"*) :

$$y(t_n) = y(t_{n+1}) - h\ f(t_{n+1}, y(t_{n+1})) + \frac{h^2}{2}\ y''(\zeta), \quad \zeta \in [t_n, t_{n+1}].$$

On propose donc le schéma suivant :

Schéma d'Euler implicite ou rétrograde : Partant de la donnée initiale y_0, on calcule donc y_{n+1} par récurrence en posant

$$\left\{ \begin{array}{lcl} t_{n+1} & = & t_n + h, \quad 0 \leq n \leq N-1, \\ y_{n+1} & = & y_n + hf(t_{n+1}, y_{n+1}). \end{array} \right.$$

On dit que ce schéma est implicite car y_{n+1} est défini implicitement comme solution de l'équation non-linéaire

$$x = y_n + h\ f(t_{n+1}, x).$$

Pour la résoudre, on fait alors appel aux méthodes de point fixe ou de Newton.

6.4.1.2 Schéma du point milieu ou schéma de Lax-Wendroff

Le schéma du point milieu, appelé aussi ***schéma d'Euler amélioré*** consiste à remplacer la pente de la tangente $(t_n, y(t_n))$ dans le schéma d'Euler par la valeur corrigée au milieu de l'intervalle $[t_n, t_{n+1}]$.

En appliquant la *formule d'intégration par la méthode du point milieu*, on a au premier ordre l'approximation

$$y(t_{n+1}) = y(t_n + h) \simeq y(t_n) + h\ y'\left(t_n + \frac{h}{2}\right) = y(t_n) + h\ f\left(t_n + \frac{h}{2}, y\left(t_n + \frac{h}{2}\right)\right).$$

Par ailleurs, on approche la solution au point $t_n + \frac{h}{2}$ en utilisant le schéma d'Euler explicite :

$$y\left(t_n + \frac{h}{2}\right) \simeq y(t_n) + \frac{h}{2}\ f(t_n, y(t_n)).$$

D'où en définitive

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

L'algorithme du point milieu donne lieu au schéma numérique suivant dit aussi Schéma de Lax-Wendroff :

Schéma du point milieu ou Lax-Wendroff : Partant de la donnée initiale y_0, on calcule donc y_{n+1} par récurrence en posant

$$\boxed{\begin{cases} t_{n+1} &= t_n + h, \quad 0 \leq n \leq N-1, \\ L_n &= hf(t_n, y_n), \\ y_{n+1} &= y_n + hf\left(t_n + \frac{h}{2}, y_n + \frac{L_n}{2}\right). \end{cases}}$$

Remarque 6.2. *On admet souvent la notation suivante :*

$$y_{n+\frac{1}{2}} = y_n + \frac{h}{2} f(t_n, y_n) \quad d'où \quad y_{n+1} = y_n + hf\left(t_n + \frac{h}{2}, y_{n+\frac{1}{2}}\right).$$

6.4.1.3 Schéma d'Euler modifié

Le schéma d'Euler modifié appelé aussi ***schéma prédicteur-correcteur d'Euler-Cauchy*** est inspiré du schéma implicite de Cranck-Nicolson où :

$$y_{n+1} \;=\; y_n + \frac{h}{2}\Big(f\big(t_n, y_n\big) \;+\; f\big(t_{n+1}, E_{n+1}\big)\Big)$$

avec

$$E_{n+1} \;=\; y_n \;+\; hf(t_n, y_n).$$

Le schéma d'Euler modifié consiste à remplacer dans le schéma d'Euler la pente de la tangente $(t_n, y(t_n))$ par la moyenne de cette pente avec la On propose donc le procédé itératif d'ordre 2 suivant :

Schéma d'Euler modifié : Partant de la donnée initiale y_0, on calcule donc y_{n+1} par récurrence en posant

$$\boxed{\begin{cases} t_{n+1} &= t_n + h, \\ E_{n+1} &= y_n \;+\; hf(t_n, y_n), \\ y_{n+1} &= y_n + \dfrac{h}{2}\Big(f\big(t_n, y_n\big) \;+\; f\big(t_{n+1}, E_{n+1}\big)\Big). \end{cases}}$$

6.4.1.4 Schéma de Cranck-Nicolson

En appliquant la *formule d'intégration par la méthode du trapèzes*, on a :

$$y(t_{n+1}) = y(t_n) + \frac{h}{2}\bigg(f\big(t_n, y(t_n)\big) + f\big(t_{n+1}, y(t_{n+1})\big)\bigg).$$

On propose donc le procédé itératif d'ordre 2 suivant :

Schéma de Cranck-Nicolson : Partant de la donnée initiale y_0, on calcule donc y_{n+1} par récurrence en posant

$$\boxed{\begin{cases} t_{n+1} &= t_n + h, \quad 0 \le n \le N-1, \\ y_{n+1}) &= y_n + \dfrac{h}{2}\bigg(f\big(t_n, y_n\big) + f\big(t_{n+1}, y_{n+1}\big)\bigg). \end{cases}}$$

6.4.2 Étude des méthodes à un pas : Ordre, convergence, stabilité et consistance

Lors de la résolution des équations différentielles, il convient d'introduire plusieurs propriétés mathématiques telles que la convergence, la stabilité et la consistance, on ajoute à ceci l'ordre de précision d'une méthode.

Dans les méthode à un pas, le calcul de y_{n+1} fait intervenir t_n, y_n et h. Ainsi les schémas à un pas explicites peuvent se mettre sous la forme générique suivante :

$$\begin{cases} y_0 \text{ donné,} \\ y_{n+1} = y_n + \varphi(t_n, y_n, h), \quad 0 \leq n \leq N-1. \end{cases}$$

et les schémas implicites sous la forme générique suivante :

$$\begin{cases} y_0 \text{ donné,} \\ y_{n+1} = y_n + \varphi(t_n, y_n, y_{n+1}, h), \quad 0 \leq n \leq N-1. \end{cases}$$

Étant donnés f , t_0 et $y(t_0)$, soit $y(t)$ la solution exacte de

$$\begin{cases} y(t_0) = y_0, \\ y'(t) = f(t, y(t)). \end{cases}$$

6.4.2.1 Ordre de précision d'un schéma

Définition 6.3. *L'**erreur locale de consistance** est la suite (e_n) :*

Schéma explicite $e_n = y(t_{n+1}) - y_{n+1} = y(t_{n+1}) - (y_n + \varphi(t_n, y_n, h)), \quad 0 \leq n \leq N-1,$

Schéma implicite $e_n = y(t_{n+1}) - y_{n+1} = y(t_{n+1}) - (y_n + \varphi(t_n, y_n, y_{n+1}, h), \quad 0 \leq n \leq N-1.$

Définition 6.4. *L'**erreur de troncature** est la suite (ξ_n) :*

$$\xi_n = |y(t_{n+1}) - y_{n+1}| = \theta(h^{p+1}), \quad 0 \leq n \leq N-1.$$

Définition 6.5. *Un schéma numérique est **d'ordre p** si et seulement si*

$$e_n = \theta(h^{p+1}) \text{ quand h tend vers 0.}$$

6.4.2.2 Convergence d'un schéma

Un schéma numérique est dite convergent si la solution y construite selon le schéma numérique considéré tend vers la solution $\tilde{y}$ exacte de l'équation continue lorsque le pas de discrétisation h tend vers 0.

Définition 6.6. *Le schéma numérique*

$$y_{n+1} = y_n + \varphi(t_n, y_n, h), \quad 0 \leq n \leq N-1, \quad y_0 \text{ donné}$$

*est dit **convergent** par rapport à l'équation différentielle*

$$y'(t) = f(t, y(t)), \quad y(t_0) = y_0$$

si pour toute solution exacte $y(t)$ définie sur un intervalle $[t_0, t_0 + T]$ et toute suite $(y_n)_n$ construite à partir de y_0 et d'une subdivision de $[t_0, t_0 + T]$, on a la relation suivante :

$$\lim_{h \to 0} \max_{0 \leq n \leq N-1} |y(t_n) - y_n| = \max_{0 \leq n \leq N-1} |e_n| = 0. \tag{6.2}$$

*(6.2) s'appelle **relation de convergence uniforme**.*

6.4.2.3 Stabilité et consistance d'un schéma

Définition 6.7. *Un schéma numérique est dit* ***stable*** *s'il existe un δ indépendant de h, telle que pour tout y_0 et $\tilde{y}_0$, les suites $(y_n)_n$ et $(\tilde{y}_n)_n$ construites par récurrence selon les formules*

$$y_{n+1} = y_n + \varphi(t_n, y_n, h), \quad 0 \leq n \leq N-1,$$

$$\tilde{y}_{n+1} = \tilde{y}_n + \varphi(t_n, \tilde{y}_n, h) + \varepsilon_n, \quad 0 \leq n \leq N-1,$$

on a

$$\max_{0 \leq n \leq N-1} |y_n - \tilde{y}_n| \leq \delta \left(|y_0 - \tilde{y}_0| + \sum_{n=0}^{N-1} |\varepsilon_n| \right).$$

Théorème 6.5. ***(Condition suffisante de stabilité)***

- *Si φ est lipschitzienne par rapport à la deuxième variable y, le schéma est stable.*

- *De plus, si L est la constante de Lipschitz pour φ, alors la constante de stabilité est $\delta = e^{LT}$.*

Théorème 6.6. *Un schéma à un pas est consistant si et seulement si*

$$\forall (t, y) \in \mathcal{O}, \ \varphi(t, y, 0) = f(t, y).$$

6.4.3 Méthodes numériques multi-pas

6.4.3.1 Schémas d'Adams-Bashforth

Les schémas d'Adams-Bashforths sont basés sur une approximation de l'intégrale $\int_{t_n}^{t_{n+1}} f(t, y(t))\, dt$ par la quantité

$$\int_{t_n}^{t_{n+1}} p(t)\, dt \ = \ h \ \sum_{i=0}^{r} b_i \ f(t_{n-i}, y(t_{n-i}))$$

avec les b_i donnés sont d'ordre r.

Soit donc le schéma suivant, appelé schéma d'Adams-Bashforth à $(r+1)$ pas

$$y_{n+1} \ = \ y_n \ + \ h \ \sum_{i=0}^{r} b_i \ f(t_{n-i}, y_{n-i}), \quad n \geq r.$$

Ce schéma fait intervenir $(r+1)$ points (t_{n-i}, y_{n-i}) pour $i = 0, 1, ..., r$ et n'est valable qu'à partir de $n = r$. Dans cet ordre, on approche les r premières valeurs $y_1,\ y_2, ...,\ y_r$ par un schéma à un pas.

Schéma explicite d'Adams-Bashforth à 2 pas et d'ordre 2 :

Étant donné y_0 et y_1 calculés avec une méthode à un pas, on a

$$\left\{ \begin{array}{lcl} t_{n+1} & = & t_n + h, \quad 1 \leq n \leq N-1, \\ y_{n+1} & = & y_n + \dfrac{h}{2} \Big(3\, f(t_n, y_n) - f(t_{n-1}, y_{n-1}) \Big). \end{array} \right.$$

Schéma explicite d'Adams-Bashforth à 3 pas et d'ordre 3 :

Étant donné y_0, y_1 et y_2 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{lcl} t_{n+1} &=& t_n + h, \quad 2 \leq n \leq N-1, \\ y_{n+1} &=& y_n + \dfrac{h}{12}\Big(23\ f(t_n, y_n) - 16\ f(t_{n-1}, y_{n-1}) + 5\ f(t_{n-2}, y_{n-2})\Big). \end{array}\right.$$

Schéma explicite d'Adams-Bashforth à 4 pas et d'ordre 4 :

Étant donné y_0, y_1, y_2 et y_3 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{lcl} t_{n+1} &=& t_n + h, \quad 3 \leq n \leq N-1, \\ y_{n+1} &=& y_n + \dfrac{h}{24}\Big(55\ f(t_n, y_n) - 59\ f(t_{n-1}, y_{n-1}) + 37\ f(t_{n-2}, y_{n-2}) - 9\ f(t_{n-3}, y_{n-3})\Big). \end{array}\right.$$

6.4.3.2 Schémas d'Adams-Moulton

Les schémas d'Adams-Bashforths sont basés sur une approximation de l'intégrale $\int_{t_n}^{t_{n+1}} f(t, y(t))\ dt$ par la quantité $\int_{t_n}^{t_{n+1}} p(t)\ dt$ où $p(t)$ est l'unique polynôme de degré $r+1$ vérifiant

$$p(t_i) \ = \ f(t_i, y(t_i)), \ \ i = n+1,\ n,\ n-1, ...,\ n-r.$$

Soit donc la famille des schémas multi-pas implicites appelés schémas d'Adams-Moulton à $r+1$ pas

$$y_{n+1} \ = \ y_n \ + \ h \ \sum_{i=-1}^{r} b_i\ f(t_{n-i}, y_{n-i}), \ \ n \geq r.$$

Ce schéma fait intervenir $(r+1)$ points (t_{n-i}, y_{n-i}) pour $i = -1, 0, ..., r$.

Schéma implicite d'Adams-Moulton à 1 pas et d'ordre 2 :

Étant donné y_0 calculé avec une méthode à un pas, on a

$$\left\{\begin{array}{lcl} t_{n+1} &=& t_n + h, \quad 0 \leq n \leq N-1, \\ y_{n+1} &=& y_n + \dfrac{h}{2}\Big(f(t_{n+1}, y_{n+1}) + f(t_n, y_n)\Big). \end{array}\right.$$

Schéma implicite d'Adams-Moulton à 2 pas et d'ordre 3 :

Étant donné y_0 et y_1 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{lcl} t_{n+1} &=& t_n + h, \quad 1 \leq n \leq N-1, \\ y_{n+1} &=& y_n + \dfrac{h}{12}\Big(5\ f(t_{n+1}, y_{n+1}) + 8\ f(t_n, y_n) - f(t_{n-1}, y_{n-1})\Big). \end{array}\right.$$

Schéma implicite d'Adams-Moulton à 3 pas et d'ordre 4 :

Étant donné y_0, y_1 et y_2 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{rcl} t_{n+1} &=& t_n + h, \quad 2 \leq n \leq N-1, \\ y_{n+1} &=& y(t_n) + \dfrac{h}{24}\Big(9\, f(t_{n+1}, y_{n+1}) + 19\, f(t_n, y_n) - 5\, f(t_{n-1}, y_{n-1}) + f(t_{n-2}, y_{n-2})\Big). \end{array}\right.$$

Schéma implicite d'Adams-Moulton à 4 pas et d'ordre 5 :

Étant donné y_0, y_1, y_2 et y_3 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{rcl} t_{n+1} &=& t_n + h, \quad 3 \leq n \leq N-1, \\ y_{n+1} &=& y_n + \dfrac{h}{720}\Big(251\, f(t_{n+1}, y_{n+1}) + 646\, f(t_n, y_n) - 264\, f(t_{n-1}, y_{n-1}) \\ & & \qquad +106\, f(t_{n-2}, y_{n-2}) - 19\, f(t_{n-3}, y_{n-3})\Big). \end{array}\right.$$

6.4.3.3 Schémas prédicteur-correcteur

Pour la résolution des schémas implicites d'Adams-Moulton, il suffit d'utiliser conjointement un schéma d'Adams-Bashforth et un schéma d'Adams-Moulton de même ordre en remplaçant la valeur de y_{n+1} par une estimation prédite par le schéma d'Adams-Bashforth. Les schémas ainsi construits sont appelés prédicteur-correcteur.

Schéma prédicteur-correcteur d'ordre 2 :

En se servant du schéma explicite d'Adams-Bashforth à 2 pas et d'ordre 2, le schéma implicite d'Adams-Moulton à un pas et d'ordre 2 s'écrit :

Étant donné y_0 et y_1 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{rcl} t_{n+1} &=& t_n + h, \quad 1 \leq n \leq N-1, \\ y^*_{n+1} &=& y_n + \dfrac{h}{2}\Big(3\, f(t_n, y_n) - f(t_{n-1}, y_{n-1})\Big) \leftarrow \textbf{\textit{Prédicteur}} \\ y_{n+1} &=& y_n + \dfrac{h}{2}\Big(f(t_{n+1}, y^*_{n+1}) + f(t_n, y_n)\Big) \leftarrow \textbf{\textit{Correcteur}} \end{array}\right.$$

Schéma prédicteur-correcteur d'ordre 4 :

En se servant du schéma explicite d'Adams-Bashforth à 4 pas et d'ordre 4, le schéma implicite d'Adams-Moulton à 3 pas et d'ordre 4 s'écrit :

Étant donné y_0, y_1, y_2 et y_3 calculés avec une méthode à un pas, on a

$$\left\{\begin{array}{rcl} t_{n+1} &=& t_n + h, \quad 3 \leq n \leq N-1, \\ y^*_{n+1} &=& y(t_n) + \dfrac{h}{24}\Big(55\, f(t_n, y_n) - 59\, f(t_{n-1}, y_{n-1}) \\ & & \qquad +37\, f(t_{n-2}, y_{n-2}) - 9\, f(t_{n-3}, y_{n-3})\Big) \leftarrow \textbf{\textit{Prédicteur}} \\ y_{n+1} &=& y_n + \dfrac{h}{24}\Big(9\, f(t_{n+1}, y^*_{n+1}) + 19\, f(t_n, y_n) \\ & & \qquad -5\, f(t_{n-1}, y_{n-1}) + f(t_{n-2}, y_{n-2})\Big) \quad \leftarrow \textbf{\textit{Correcteur}} \end{array}\right.$$

6.5 Application 1 à un système d'équations différentielles

Soit à résoudre sur l'intervalle $[0, 1]$, l'équation différentielle

$$y'(t) = y(t) + t - 1$$

vérifiant la condition initiale $y(0) = 1$.

1) Déterminer la solution exacte.

2) Déterminer une solution approchée de $y(1)$ en prenant comme pas d'espace $h = 0.2$ et en utilisant les différents schémas ci-dessous :

- **2.1)** Schéma d'Euler explicite,
- **2.2)** Schéma de Lax-Wendroff ou point milieu,
- **2.3)** Schéma d'Euler modifié,
- **2.4)** Schéma d'Euler implicite,
- **2.5)** Schéma de Cranck Nicolson.

1) ***Détermination de la solution exacte de l'équation différentielle :***

Détermination d'une solution de l'équation différentielle sans second membre : La fonction $\lambda(t) = 1$ est continue sur $\mathbb{R}$ et admet comme primitive la fonction t, ainsi la solution générale de l'équation différentielle sans second membre est donc la fonction définie par

$$y(t) = C\, e^t.$$

Détermination d'une solution de l'équation différentielle avec second membre : Remplaçons maintenant la constante C par une fonction $C(t)$, dérivable sur $\mathbb{R}$, dans la solution $y(t)$ et en dérivant la fonction y, il vient alors :

$$y'(t) = C'(t)\, e^t + C(t)\, e^t.$$

En supposant que $C(t)e^t$ est solution de l'équation différentielle linéaire avec second membre, cela donne :

$$y'(t) = C(t)\, e^t + t - 1 = C(t)\, e^t + C'(t)\, e^t.$$

Il reste alors après simplification $C'(t) = \dfrac{t-1}{e^t}$, d'où $C(t) = \dfrac{-t}{e^t}$.

Une solution particulière est donc la fonction définie par $\tilde{y}(t) = -t$.
La solution générale obtenue par superposition de solutions est ainsi $y(t) = C\ e^t - t$.
En tenant compte de la condition initiale $y(0) = 1$, on obtient $C = 1$.
Ainsi, la solution générale de l'équation différentielle avec second membre $y'(t) = y(t) + t - 1$ vérifiant $y(0) = 1$, est :

$$y(t) = e^t - t.$$

2) *Détermination d'une solution approchée :*

2.1) ***Schéma d'Euler explicite :*** L'algorithme d'Euler explicite s'écrit :

$$\begin{cases} y_0 & = 1, \\ t_{n+1} & = t_n + h, \\ y_{n+1} & = y_n + hf(t_n, y_n) \ = \ y_n + 0.2\ (y_n + t_n - 1). \end{cases}$$

Les valeurs de y_n pour $t \in [0, 1]$ avec $h = 0.2$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = y_0 = 1$
1	$t_1 = t_0 + h = 0.2$	$y(0.2) = y_1 = y_0 + h\left(y_0 + t_0 - 1\right) = 1$
2	$t_2 = t_1 + h = 0.4$	$y(0.4) = y_2 = y_1 + h\left(y_1 + t_1 - 1\right) = 1.04$
3	$t_3 = t_2 + h = 0.6$	$y(0.6) = y_3 = y_2 + h\left(y_2 + t_2 - 1\right) = 1.128$
4	$t_4 = t_3 + h = 0.8$	$y(0.8) = y_4 = y_3 + h\left(y_3 + t_3 - 1\right) = 1.2736$
5	$t_5 = t_4 + h = 1$	$y(1) = y_5 = y_4 + h\left(y_4 + t_4 - 1\right) = 1.48832$

Ainsi, la solution de l'équation différentielle en utilisant le schéma d'Euler explicite est $y(1) = 1.488$.

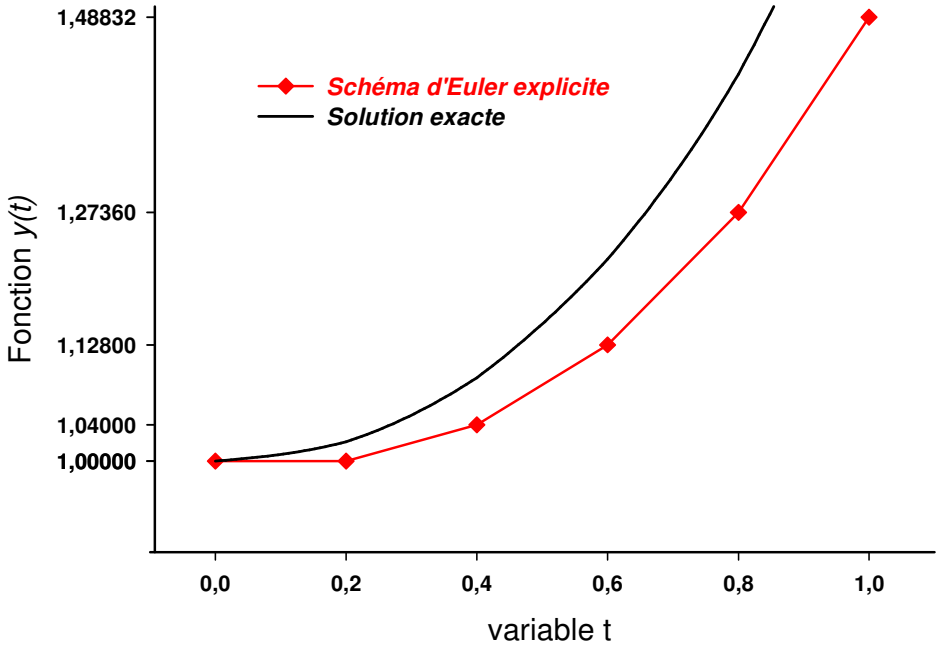


FIGURE 6.2 – Comparaison entre solution exacte et solution approchée obtenue en utilisant le schéma d'Euler explicite

Remarque 6.3. *Afin de vérifier la convergence de la solution approchée en appliquant le schéma d'Euler explicite vers la solution exacte, nous appliquons le schéma pour différentes valeurs de h, pour $h = 0.2,\ 0.1,\ 0.05$ et $h = 0.025$. Nous remarquons que plus le pas d'espace h est fin, plus la convergence est meilleure comme le montre ce tableau récapitulatif des valeurs.*

Ci-dessous le tableau des valeurs de y_n correspondant aux valeurs de h en appliquant le schéma d'Euler explicite. Sachant que la valeur de la solution exacte en $t = 1$ vaut $e^1 - 1 = 1,718281828$, nous remarquons que la solution la mieux approchée est celle correspondante au pas d'espace le plus fin $h = 0.025$.

$h=0.2$	$h=0.1$	$h=0.05$	$h=0.025$	
1	1	1	1	1.154581852
1	1.01	1	1	1.171571398
1.04	1.031	1.0025	1.000625	1.189610683
1.128	1.0641	1.007625	1.0018900625	1.208725595
1.2736	1.11051	1.01550625	1.003812891	1.228943735
1.48832	1.171561	1.026281563	1.006408213	1.250292328
	1.2487171	1.040095641	1.009693418	1.272799636
	1.34358881	1.057100423	1.013685753	1.296494627
	1.457947691	1.077455444	1.018402897	1.321406993
	1.59374246	1.103828216	1.023862969	1.347567168
		1.131519627	1.030084543	1.375006347
		1.163095608	1.037086657	1.403756506
		1.198750388	1.044888823	1.433850419
		1.238687907	1.053511044	1.465321679
		1.283122302	1.06297382	1.498204721
		1.332278417	1.073298166	1.532534839
		1.386392338	1.08450562	1.56834821
		1.445711955	1.096618261	1.605681915
		1.510497553	1.109658718	1.644573963
		1.581022431	1.123650186	*1.685063312*
		1.657573553	1.138616441	

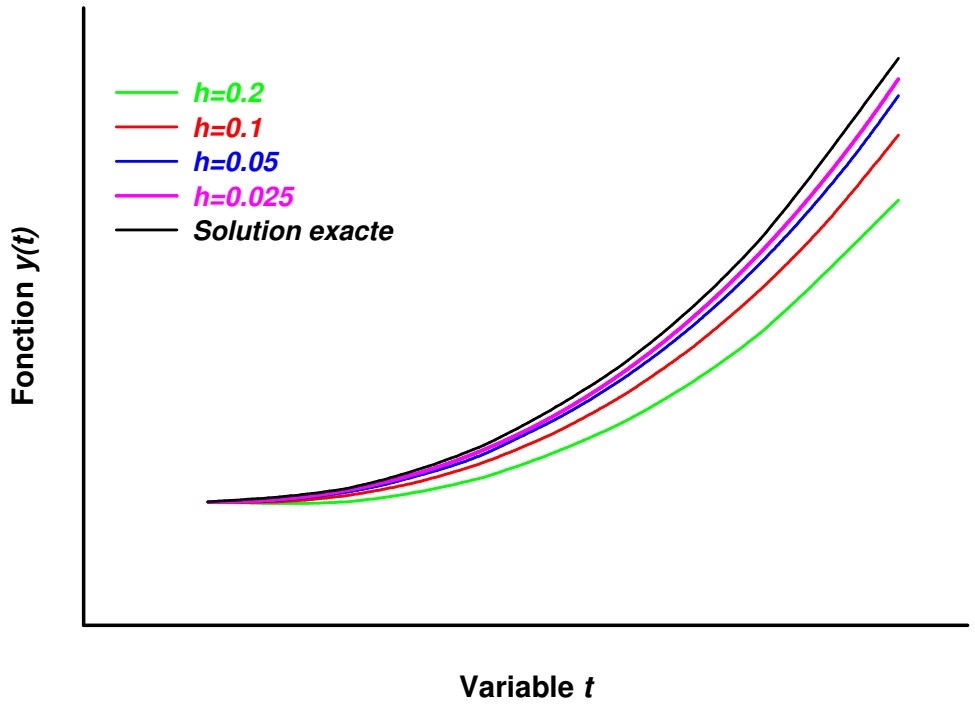


FIGURE 6.3 – Comparaison entre solution exacte et solution approchée correspondant aux valeurs de $h = 0.2,\ 0.1,\ 0.05,\ 0.025$ en utilisant le schéma d'Euler explicite

2.2) ***Schéma de Lax-Wendroff ou point milieu :*** L'algorithme de Lax-Wendroff ou point milieu s'écrit :

$$\begin{cases} y_0 & = 1, \\ t_{n+1} & = t_n + h, \\ L_n & = hf(t_n, y_n) \ = \ 0.2\ \big(y_n + t_n - 1\big), \\ y_{n+1} & = y_n + hf\left(t_n + \dfrac{h}{2}, y_n + \dfrac{L_n}{2}\right) \ = \ y_n + h\left(\left(y_n + \dfrac{L_n}{2}\right) + \left(t_n + \dfrac{h}{2}\right) - 1\right). \end{cases}$$

Les valeurs de y_n pour $t \in [0, 1]$ avec $h = 0.2$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$\boldsymbol{y(0)} = y_0 = \boldsymbol{1}$
1	$t_1 = 0.2$	$L_0 = h\left(y_0 + t_0 - 1\right) = 0$ $\boldsymbol{y(0.2)} = y_1 = y_0 + h\left(y_0 + \frac{L_0}{2} + t_0 + \frac{h}{2} - 1\right) = \boldsymbol{1.02}$
2	$t_2 = 0.4$	$L_1 = h\left(y_1 + t_1 - 1\right) = 0.044$ $\boldsymbol{y(0.4)} = y_2 = y_1 + h\left(y_1 + \frac{L_1}{2} + t_1 + \frac{h}{2} - 1\right) = \boldsymbol{1.0884}$
3	$t_3 = 0.6$	$L_2 = h\left(y_2 + t_2 - 1\right) = 0.09768$ $\boldsymbol{y(0.6)} = y_3 = y_2 + h\left(y_2 + \frac{L_2}{2} + t_2 + \frac{h}{2} - 1\right) = \boldsymbol{1.215848}$
4	$t_4 = 0.8$	$L_3 = h\left(y_3 + t_3 - 1\right) = 0.163169$ $\boldsymbol{y(0.8)} = y_4 = y_3 + h\left(y_3 + \frac{L_3}{2} + t_3 + \frac{h}{2} - 1\right) = \boldsymbol{1.415334}$
5	$t_5 = 1$	$L_4 = h\left(y_4 + t_4 - 1\right) = 0.243066$ $\boldsymbol{y(1)} = y_5 = y_4 + h\left(y_4 + \frac{L_4}{2} + t_4 + \frac{h}{2} - 1\right) = \boldsymbol{1.702708}$

Ainsi, la solution de l'équation différentielle en utilisant le schéma de Lax-Wendroff ou point milieu est $y(1) = 1.702708$.

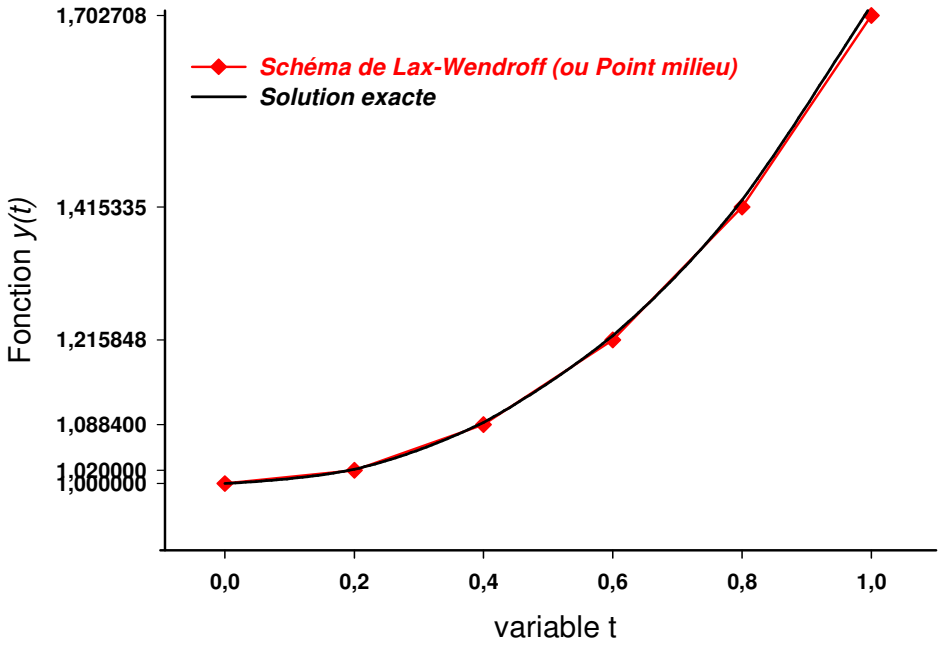


FIGURE 6.4 – Comparaison entre solution exacte et solution approchée obtenue en utilisant le schéma de Lax-Wendroff ou point milieu

2.3) ***Schéma d'Euler modifié :*** L'algorithme d'Euler modifié s'écrit :

$$\begin{cases} y_0 & = 1, \\ t_{n+1} & = t_n + h, \\ E_{n+1} & = y_n + hf(t_n, y_n) = y_n + h\left(y_n + t_n - 1\right), \\ y_{n+1} & = y_n + \dfrac{h}{2}\Big(f\left(t_n, y_n\right) + f\left(t_{n+1}, E_{n+1}\right)\Big) \\ & = y_n + \dfrac{h}{2}\Big(\left(y_n + t_n - 1\right) + \left(E_{n+1} + t_{n+1} - 1\right)\Big). \end{cases}$$

Les valeurs de y_n pour $t \in [0,1]$ avec $h = 0.2$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = y_0 = 1$
1	$t_1 = 0.2$	$E_1 = y_0 + h\,(y_0 + t_0 - 1) = 1$ $y(0.2) = y_1 = y_0 + \frac{h}{2}\Big((y_0 + t_0 - 1) + (E_1 + t_1 - 1)\Big) = 1.02$
2	$t_2 = 0.4$	$E_2 = y_1 + h\,(y_1 + t_1 - 1) = 1.064$ $y(0.4) = y_2 = y_1 + \frac{h}{2}\Big((y_1 + t_1 - 1) + (E_2 + t_2 - 1)\Big) = 1.0884$
3	$t_3 = 0.6$	$E_3 = y_2 + h\,(y_2 + t_2 - 1) = 1.18608$ $y(0.6) = y_3 = y_2 + \frac{h}{2}\Big((y_2 + t_2 - 1) + (E_3 + t_3 - 1)\Big) = 1.215848$
4	$t_4 = 0.8$	$E_4 = y_3 + h\,(y_3 + t_3 - 1) = 1.379017$ $y(0.8) = y_4 = y_3 + \frac{h}{2}\Big((y_3 + t_3 - 1) + (E_4 + t_4 - 1)\Big) = 1.415334$
5	$t_5 = 1$	$E_5 = y_4 + h\,(y_4 + t_4 - 1) = 1.658401$ $y(1) = y_5 = y_4 + \frac{h}{2}\Big((y_4 + t_4 - 1) + (E_5 + t_5 - 1)\Big) = 1.702708$

Ainsi, la solution de l'équation différentielle en utilisant le schéma d'Euler modifié est $y(1) = 1.702708$.

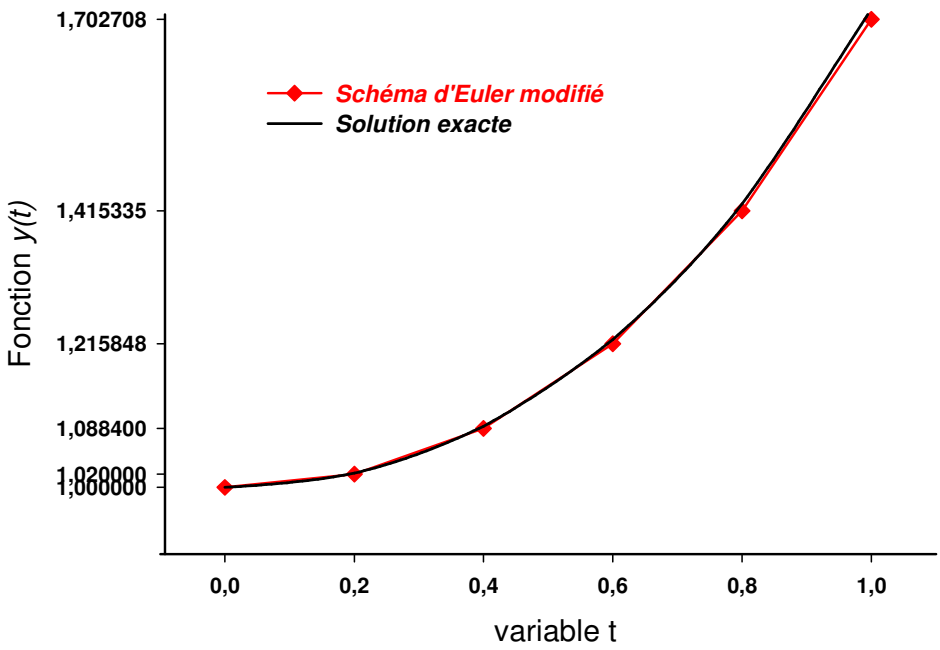


FIGURE 6.5 – Comparaison entre solution exacte et solution approchée obtenue en utilisant le schéma d'Euler modifié

2.4) ***Schéma d'Euler implicite :*** L'algorithme d'Euler implicite s'écrit :

$$\begin{cases} y_0 & = 1, \\ t_{n+1} & = t_n + h, \\ y_{n+1} & = y_n + hf\big(t_{n+1}, y_{n+1}\big) \\ & = y_n + h\,\big(y_{n+1} + t_{n+1} - 1\big) \\ & = \dfrac{1}{1-h}\,\big(y_n + h\,\big(t_{n+1} - 1\big)\big). \end{cases}$$

Les valeurs de y_n pour $t \in [0,1]$ avec $h = 0.2$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = y_0 = 1$
1	$t_1 = 0.2$	$y(0.2) = y_1 = (1/1-h)\,\big(y_0 + h\,\big(t_1 - 1\big)\big) = 1.05$
2	$t_2 = 0.4$	$y(0.4) = y_2 = (1/1-h)\,\big(y_1 + h\,\big(t_2 - 1\big)\big) = 1.1625$
3	$t_3 = 0.6$	$y(0.6) = y_3 = (1/1-h)\,\big(y_2 + h\,\big(t_3 - 1\big)\big) = 1.353125$
4	$t_4 = 0.8$	$y(0.8) = y_4 = (1/1-h)\,\big(y_3 + h\,\big(t_4 - 1\big)\big) = 1.641406$
5	$t_5 = 1$	$y(1) = y_5 = (1/1-h)\,\big(y_4 + h\,\big(t_5 - 1\big)\big) = 2.051757$

Ainsi, la solution de l'équation différentielle en utilisant le schéma d'Euler implicite est $y(1) = 1.727412$.

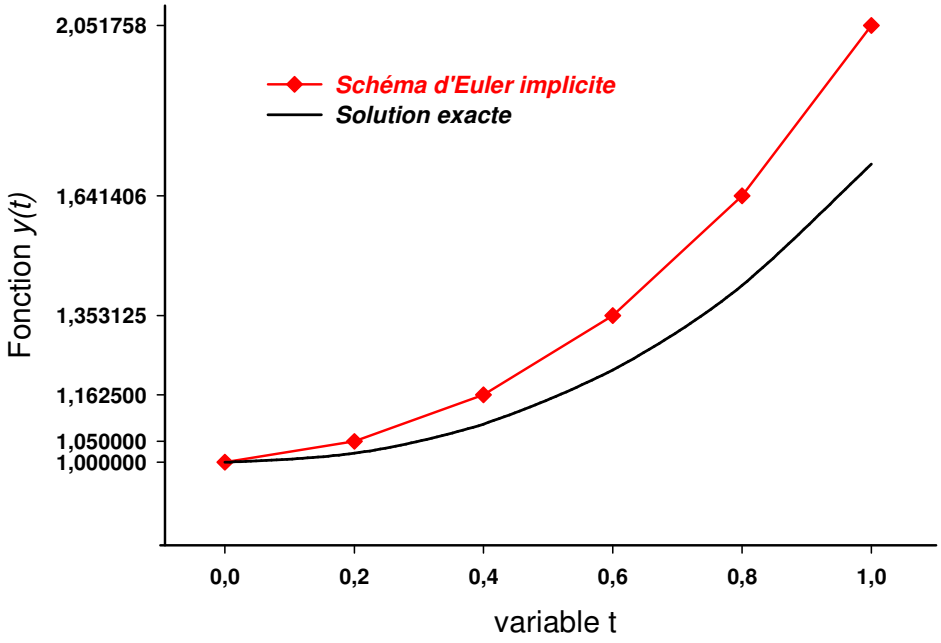


FIGURE 6.6 – Comparaison entre solution exacte et solution approchée obtenue en utilisant le schéma d'Euler implicite

2.5) ***Schéma de Cranck-Nicolson :*** L'algorithme de Cranck-Nicolson s'écrit :

$$\begin{cases} y_0 &= 1, \\ t_{n+1} &= t_n + h, \\ y_{n+1} &= y_n + \dfrac{h}{2}\Big(f\big(t_n, y_n\big) + f\big(t_{n+1}, y_{n+1}\big)\Big) \\ &= y_n + \dfrac{h}{2}\Big(\big(y_n + t_n - 1\big) + \big(y_{n+1} + t_{n+1} - 1\big)\Big) \\ &= \dfrac{1}{\Big(1 - \dfrac{h}{2}\Big)}\left(y_n + \dfrac{h}{2}\Big(y_n + t_n - 1\Big) + \dfrac{h}{2}\Big(t_{n+1} - 1\Big)\right). \end{cases}$$

Les valeurs de y_n pour $t \in [0,1]$ avec $h = 0.2$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = y_0 = 1$
1	$t_1 = 0.2$	$y(0.2) = \frac{1}{\left(1-\frac{h}{2}\right)}\left(y_0 + \frac{h}{2}\left(y_0 + t_0 - 1\right) + \frac{h}{2}\left(t_1 - 1\right)\right) = 1.022$
2	$t_2 = 0.4$	$y(0.4) = \frac{1}{\left(1-\frac{h}{2}\right)}\left(y_1 + \frac{h}{2}\left(y_1 + t_1 - 1\right) + \frac{h}{2}\left(t_2 - 1\right)\right) = 1.093827$
3	$t_3 = 0.6$	$y(0.6) = \frac{1}{\left(1-\frac{h}{2}\right)}\left(y_2 + \frac{h}{2}\left(y_2 + t_2 - 1\right) + \frac{h}{2}\left(t_3 - 1\right)\right) = 1.225788$
4	$t_4 = 0.8$	$y(0.8) = \frac{1}{\left(1-\frac{h}{2}\right)}\left(y_3 + \frac{h}{2}\left(y_3 + t_3 - 1\right) + \frac{h}{2}\left(t_4 - 1\right)\right) = 1.413519$
5	$t_5 = 1$	$y(1) = \frac{1}{\left(1-\frac{h}{2}\right)}\left(y_4 + \frac{h}{2}\left(y_4 + t_4 - 1\right) + \frac{h}{2}\left(t_5 - 1\right)\right) = 1.727412$

Ainsi, la solution de l'équation différentielle en utilisant le schéma de Cranck-Nicholson est $y(1) = 1.727412$.

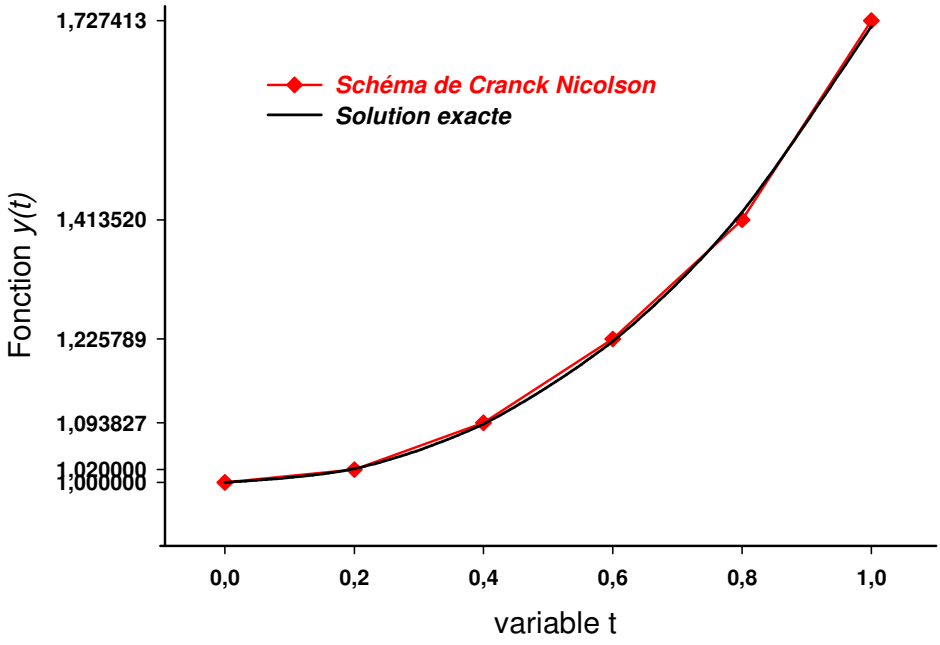


FIGURE 6.7 – Comparaison entre solution exacte et solution approchée obtenue en utilisant le schéma de Cranck Nicolson

Récapitulons dans cette figure les différents graphes des schémas numériques par rapport au graphe de la solution exacte de l'équation différentielle :

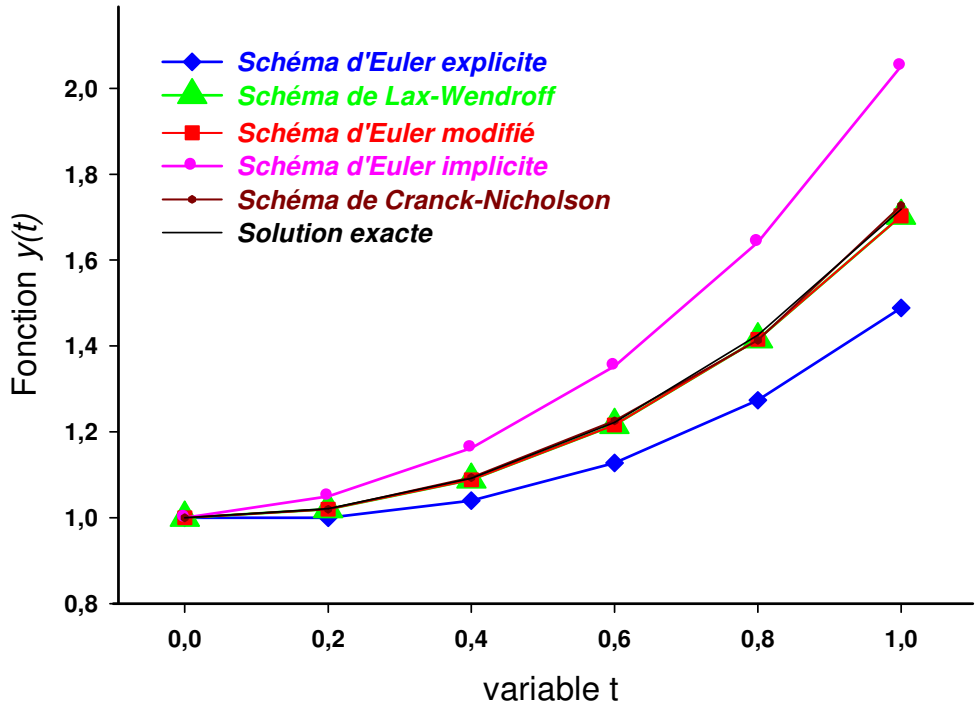


FIGURE 6.8 – Comparaison entre solution exacte et différentes solutions approchées

6.6 Application 2 à un système d'équations différentielles

Soit à résoudre sur l'intervalle $[0, \pi]$, l'équation différentielle

$$y'(t) = t^2 \cos\left(y(t) + 1\right)$$

vérifiant la condition initiale $y(0) = 1$.

1) Déterminer une solution approchée de $y(\pi)$ en prenant comme pas d'espace $h = \dfrac{\pi}{10}$ et en utilisant les différents schémas ci-dessous :

1.1) Schéma d'Euler explicite,

1.2) Schéma de Lax-Wendroff ou point milieu,

1.3) Schéma d'Euler modifié,

1.4) Schéma d'Euler implicite,

1.5) Schéma de Cranck Nicolson.

1) Détermination d'une solution approchée :

1.1) ***Schéma d'Euler explicite :*** L'algorithme d'Euler explicite s'écrit :

$$\begin{cases} y(0) &= 1, \\ t_{n+1} &= t_n + h, \\ y_{n+1} &= y_n + hf(t_n, y_n) \;=\; y_n + \dfrac{\pi}{10}\, t_n^2 \cos\,(y_n + 1). \end{cases}$$

Les valeurs de y_n pour $t \in [0, \pi]$ avec $h = \dfrac{\pi}{10}$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = y_0 = 1$
1	$t_1 = \dfrac{\pi}{10}$	$y\left(\dfrac{\pi}{10}\right) = y_1 = y_0 + h\, t_0^2 \cos\,(y_0 + 1) = 1$
2	$t_2 = \dfrac{2\pi}{10}$	$y\left(\dfrac{2\pi}{10}\right) = y_2 = y_1 + h\, t_1^2 \cos\,(y_1 + 1) = 0.987\ 096\ 836$
3	$t_3 = \dfrac{3\pi}{10}$	$y\left(\dfrac{3\pi}{10}\right) = y_3 = y_2 + h\, t_2^2 \cos\,(y_2 + 1) = 0.936\ 943\ 599$
4	$t_4 = \dfrac{4\pi}{10}$	$y\left(\dfrac{4\pi}{10}\right) = y_4 = y_3 + h\, t_3^2 \cos\,(y_3 + 1) = 0.837\ 035\ 584$
5	$t_5 = \dfrac{5\pi}{10}$	$y\left(\dfrac{5\pi}{10}\right) = y_5 = y_4 + h\, t_4^2 \cos\,(y_4 + 1) = 0.723\ 788\ 474$

6	$t_6 = \dfrac{6\pi}{10}$	$\boldsymbol{y}\left(\dfrac{6\pi}{10}\right) = y_6 = y_5 + h\, t_5^2 \cos\,(y_5 + 1) \quad = \boldsymbol{0.605\ 657\ 654}$
7	$t_7 = \dfrac{7\pi}{10}$	$\boldsymbol{y}\left(\dfrac{7\pi}{10}\right) = y_7 = y_6 + h\, t_6^2 \cos\,(y_6 + 1) \quad = \boldsymbol{0.566\ 752\ 417}$
8	$t_8 = \dfrac{8\pi}{10}$	$\boldsymbol{y}\left(\dfrac{8\pi}{10}\right) = y_8 = y_7 + h\, t_7^2 \cos\,(y_6 + 1) \quad = \boldsymbol{0.\ 572\ 896\ 343}$
9	$t_9 = \dfrac{9\pi}{10}$	$\boldsymbol{y}\left(\dfrac{9\pi}{10}\right) = y_9 = y_8 + h\, t_8^2 \cos\,(y_8 + 1) \quad = \boldsymbol{0.568\ 729\ 070}$
10	$t_{10} = \pi$	$\boldsymbol{y}(\pi) = y_{10} = y_9 + h\, t_9^2 \cos\,(y_9 + 1) \quad = \boldsymbol{0.573\ 920\ 999}$

Ainsi, la solution de l'équation différentielle en utilisant le schéma d'Euler explicite est $y(\pi) = 0.573920999$.

1.2) ***Schéma de Lax-Wendroff ou point milieu :*** L'algorithme de Lax-Wendroff ou point milieu s'écrit :

$$\begin{cases} y(0) &= 1, \\ t_{n+1} &= t_n + h, \\ L_n &= hf(t_n, y_n) \;=\; h\, t_n^2 \cos\,(y_n + 1), \\ y_{n+1} &= y_n + h\, f\left(t_n + \dfrac{h}{2}, y_n + \dfrac{L_n}{2}\right) \;=\; y_n + h\left(t_n^2 + \dfrac{h}{2}\right) \cos\,\left(y_n + \dfrac{L_n}{2} + 1\right). \end{cases}$$

Les valeurs de y_n pour $t \in [0, \pi]$ avec $h = \dfrac{\pi}{10}$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = y_0 = 1$
1	$t_1 = \frac{\pi}{10}$	$L_0 = h\, t_0^2 \cos(y_0 + 1) = 0$ $y\left(\frac{\pi}{10}\right) = y_1 = y_0 + h\left(t_0^2 + \frac{h}{2}\right)\cos\left(y_0 + \frac{L_0}{2} + 1\right) = 0.996\ 774\ 209$
2	$t_2 = \frac{2\pi}{10}$	$L_1 = h\, t_1^2 \cos(y_1 + 1) = -0.012\ 812\ 149$ $y\left(\frac{2\pi}{10}\right) = y_2 = y_1 + h\left(t_1^2 + \frac{h}{2}\right)\cos\left(y_1 + \frac{L_1}{2} + 1\right) = 0.968\ 354\ 437$
3	$t_3 = \frac{3\pi}{10}$	$L_2 = h\, t_2^2 \cos(y_2 + 1) = -0.048\ 018\ 559$ $y\left(\frac{3\pi}{10}\right) = y_3 = y_2 + h\left(t_2^2 + \frac{h}{2}\right)\cos\left(y_2 + \frac{L_2}{2} + 1\right) = 0.896\ 636\ 518$
4	$t_4 = \frac{4\pi}{10}$	$L_3 = h\, t_3^2 \cos(y_3 + 1) = -0.089\ 591\ 672$ $y\left(\frac{4\pi}{10}\right) = y_4 = y_3 + h\left(t_3^2 + \frac{h}{2}\right)\cos\left(y_3 + \frac{L_3}{2} + 1\right) = 0.791\ 923\ 170$
5	$t_5 = \frac{5\pi}{10}$	$L_4 = h\, t_4^2 \cos(y_4 + 1) = -0.108\ 809\ 293$ $y\left(\frac{5\pi}{10}\right) = y_5 = y_4 + h\left(t_4^2 + \frac{h}{2}\right)\cos\left(y_4 + \frac{L_4}{2} + 1\right) = 0.687\ 726\ 403$
6	$t_6 = \frac{6\pi}{10}$	$L_5 = h\, t_5^2 \cos(y_5 + 1) = -0.090\ 432\ 752$ $y\left(\frac{6\pi}{10}\right) = y_6 = y_5 + h\left(t_5^2 + \frac{h}{2}\right)\cos\left(y_5 + \frac{L_5}{2} + 1\right) = 0.620\ 520\ 903$
7	$t_7 = \frac{7\pi}{10}$	$L_6 = h\, t_6^2 \cos(y_6 + 1) = -0.055\ 480\ 993$ $y\left(\frac{7\pi}{10}\right) = y_7 = y_6 + h\left(t_6^2 + \frac{h}{2}\right)\cos\left(y_6 + \frac{L_6}{2} + 1\right) = 0.591\ 723\ 743$
8	$t_8 = \frac{8\pi}{10}$	$L_7 = h\, t_7^2 \cos(y_7 + 1) = -0.031\ 792\ 860$ $y\left(\frac{8\pi}{10}\right) = y_8 = y_7 + h\left(t_7^2 + \frac{h}{2}\right)\cos\left(y_7 + \frac{L_7}{2} + 1\right) = 0.582\ 949\ 221$
9	$t_9 = \frac{9\pi}{10}$	$L_8 = h\, t_8^2 \cos(y_8 + 1) = -0.024\ 115\ 630$ $y\left(\frac{9\pi}{10}\right) = y_9 = y_8 + h\left(t_8^2 + \frac{h}{2}\right)\cos\left(y_8 + \frac{L_8}{2} + 1\right) = 0.582\ 736\ 224$
10	$t_{10} = \pi$	$L_9 = h\, t_9^2 \cos(y_9 + 1) = -0.029\ 986\ 439$ $y(ß) = y_{10} = y_9 + h\left(t_9^2 + \frac{h}{2}\right)\cos\left(y_9 + \frac{L_9}{2} + 1\right) = 0.591\ 280\ 371$

Ainsi, la solution de l'équation différentielle en utilisant le schéma de Lax-Wendroff ou point milieu est $y(\pi) = 0.591280371$.

1.3) ***Schéma d'Euler modifié :*** L'algorithme d'Euler modifié s'écrit :

$$\begin{cases} y(0) &= 1, \\ t_{n+1} &= t_n + h, \\ E_{n+1} &= y_n \ + \ hf(t_n, y_n) \ = \ y_n \ + \ h\, t_n^2 \ \cos(y_n + 1), \\ y_{n+1} &= y_n + \dfrac{h}{2}\Big(f\big(t_n, y_n\big) \ + \ f\big(t_{n+1}, E_{n+1}\big)\Big) \\ &= y_n \ + \ \dfrac{h}{2}\ \big(t_n^2 \ \cos\,(y_n + 1) \ + \ t_{n+1}^2 \ \cos\,(E_{n+1} + 1)\big). \end{cases}$$

Les valeurs de y_n pour $t \in [0, \pi]$ avec $h = \dfrac{\pi}{10}$, sont données dans le tableau suivant :

n	t_n	y_n
0	$t_0 = 0$	$y(0) = 1$
1	$t_1 = \dfrac{\pi}{10}$	$E_1 = y_0 \ + \ h\, t_0^2 \cos\,(y_0 + 1) = 1$ $y\left(\dfrac{\pi}{10}\right) = y_0 \ + \ \dfrac{h}{2}\ \big(t_0^2 \ \cos\,(y_0 + 1) \ + \ t_1^2 \ \cos\,(E_1 + 1)\big) = 0.996\ 774\ 209$
2	$t_2 = \dfrac{2\pi}{10}$	$E_2 = y_1 \ + \ h\, t_1^2 \cos\,(y_1 + 1) = 0.980\ 827\ 416$ $y\left(\dfrac{2\pi}{10}\right) = y_1 \ + \ \dfrac{h}{2}\ \big(t_1^2 \ \cos\,(y_1 + 1) \ + \ t_2^2 \ \cos\,(E_2 + 1)\big) = 0.962\ 467\ 366$
3	$t_3 = \dfrac{3\pi}{10}$	$E_3 = y_2 \ + \ h\, t_2^2 \cos\,(y_2 + 1) = 0.915\ 122\ 835$ $y\left(\dfrac{3\pi}{10}\right) = y_2 \ + \ \dfrac{h}{2}\ \big(t_2^2 \ \cos\,(y_2 + 1) \ + \ t_3^2 \ \cos\,(E_3 + 1)\big) = 0.891\ 695\ 557$
4	$t_4 = \dfrac{4\pi}{10}$	$E_4 = y_3 \ + \ h\, t_3^2 \cos\,(y_3 + 1) = 0.803\ 675\ 554$ $y\left(\dfrac{4\pi}{10}\right) = y_3 \ + \ \dfrac{h}{2}\ \big(t_3^2 \ \cos\,(y_3 + 1) \ + \ t_4^2 \ \cos\,(E_4 + 1)\big) = 0.790\ 440\ 533$
5	$t_5 = \dfrac{5\pi}{10}$	$E_5 = y_4 \ + \ h\, t_4^2 \cos\,(y_4 + 1) = 0.790\ 440\ 533$ $y\left(\dfrac{5\pi}{10}\right) = y_4 \ + \ \dfrac{h}{2}\ \big(t_4^2 \ \cos\,(y_4 + 1) \ + \ t_5^2 \ \cos\,(E_5 + 1)\big) = 0.693\ 248\ 968$
6	$t_6 = \dfrac{6\pi}{10}$	$E_6 = y_5 \ + \ h\, t_5^2 \cos\,(y_5 + 1) = 0.598\ 565\ 993$ $y\left(\dfrac{6\pi}{10}\right) = y_5 \ + \ \dfrac{h}{2}\ \big(t_5^2 \ \cos\,(y_5 + 1) \ + \ t_6^2 \ \cos\,(E_6 + 1)\big) = 0.630\ 410\ 862$

7	$t_7 = \dfrac{7\pi}{10}$	$E_7 = y_6 \ + \ h\, t_6^2 \cos\,(y_6+1) = 0.563\ 906\ 977$ $\boldsymbol{y}\left(\dfrac{7\pi}{10}\right) = y_6 \ + \ \dfrac{h}{2}\left(t_6^2\ \cos\,(y_6+1) \ + \ t_7^2\ \cos\,(E_7+1)\right) \quad = \boldsymbol{0.602\ 392\ 399}$
8	$t_8 = \dfrac{8\pi}{10}$	$E_8 = y_7 \ + \ h\, t_7^2 \cos\,(y_7+1) = 0.554\ 396\ 234$ $\boldsymbol{y}\left(\dfrac{8\pi}{10}\right) = y_7 \ + \ \dfrac{h}{2}\left(t_7^2\ \cos\,(y_7+1) \ + \ t_8^2\ \cos\,(E_8+1)\right) \quad = \boldsymbol{0.594\ 665\ 773}$
9	$t_9 = \dfrac{9\pi}{10}$	$E_9 = y_8 \ + \ h\, t_8^2 \cos\,(y_8+1) = .547\ 303\ 700$ $\boldsymbol{y}\left(\frac{9\pi}{10}\right) = y_8 \ + \ \dfrac{h}{2}\left(t_8^2\ \cos\,(y_8+1) \ + \ t_9^2\ \cos\,(E_9+1)\right) \quad = \boldsymbol{0.600\ 482\ 988}$
10	$t_{10} = \pi$	$E_{10} = y_9 \ + \ h\, t_9^2 \cos\,(y_9+1) = 0.525\ 935\ 639$ $\boldsymbol{y}(ß) = y_9 \ + \ \dfrac{h}{2}\left(t_9^2\ \cos\,(y_9+1) \ + \ t_{10}^2\ \cos\,(E_{10}+1)\right) \quad = \boldsymbol{0.632\ 734\ 133}$

Ainsi, la solution de l'équation différentielle en utilisant le schéma d'Euler modifié est $y(1) = 0.632734133$.

1.4) ***Schéma d'Euler implicite :*** L'algorithme d'Euler implicite s'écrit :

$$\begin{cases} y(0) &= 1, \\ t_{n+1} &= t_n + h, \\ y_{n+1} &= y_n + hf\big(t_{n+1}, y_{n+1}\big) \\ &= y_n + h\, t_{n+1}^2 \cos\,(y_{n+1}+1). \end{cases}$$

Si on pose $z = y_{n+1}$ et $Y = y(t_n)$, on écrit :

$$z \ = \ Y + h\, t_{n+1}^2 \cos\,(z+1).$$

Soit la fonction $f(z)$ définie comme suit :

$$f(z) \ = \ -z + Y + h\, t_{n+1}^2 \cos\,(z+1).$$

En appliquant la méthode de Newton-Raphson à l'équation $f(z) = 0$, il vient alors :

$$z(t_{n+1}) \ = \ z_n - \frac{f(z_n)}{f'(z_n)} \ = \ z(t_n) - \frac{-z_n + Y + h\, t_{n+1}^2 \cos\,(z_n+1)}{-1 - h\, t_{n+1}^2 \sin(z_n+1)}.$$

L'algorithme d'Euler implicite s'écrit :

$$\begin{cases} z_0 &= 1, \\ t_{n+1} &= t_n + h, \\ z_{n+1} &= z_n - \dfrac{f(z_n)}{f'(z_n)} \ = \ z_n + \dfrac{-z_n + y_n + h\, t_{n+1}^2 \cos\,(z_n+1)}{1 + h\, t_{n+1}^2 \sin(z_n+1)}. \end{cases}$$

Les valeurs de z_n pour $t \in [0, \pi]$ avec $h = \dfrac{\pi}{10}$, sont données dans le tableau suivant :

n	t_n	z_n	
0	$t_0 = 0$	$z(0) = z_0$	$= 1$
1	$t_1 = \frac{\pi}{10}$	$z\left(\frac{\pi}{10}\right) = z(t_0) + \frac{-z_0 + y_0 + h\, t_1^2 \cos(z_0 + 1)}{1 + h\, t_1^2 \sin(z_0 + 1)}$	$= 0.987\ 450\ 652$
2	$t_2 = \frac{2\pi}{10}$	$z\left(\frac{2\pi}{10}\right) = z(t_1) + \frac{-z_1 + y_1 + h\, t_2^2 \cos(z_1 + 1)}{1 + h\, t_2^2 \sin(z_1 + 1)}$	$= 0.942\ 370\ 077$
3	$t_3 = \frac{3\pi}{10}$	$z\left(\frac{3\pi}{10}\right) = z(t_2) + \frac{-z_2 + y_2 + h\, t_3^2 \cos(z_2 + 1)}{1 + h\, t_3^2 \sin(z_2 + 1)}$	$= 0.861\ 957\ 841$
4	$t_4 = \frac{4\pi}{10}$	$z\left(\frac{4\pi}{10}\right) = z_3 + \frac{-z_3 + y_3 + h\, t_4^2 \cos(z_3 + 1)}{1 + h\, t_4^2 \sin(z_3 + 1)}$	$= 0.765\ 420\ 999$
5	$t_5 = \frac{5\pi}{10}$	$z\left(\frac{5\pi}{10}\right) = z(t_0) + \frac{-z_4 + y_4 + h\, t_5^2 \cos(z_4 + 1)}{1 + h\, t_5^2 \sin(z_4 + 1)}$	$= 0.680\ 267\ 836$
6	$t_6 = \frac{6\pi}{10}$	$z\left(\frac{6\pi}{10}\right) = z_5 + \frac{-z_5 + y_5 + h\, t_6^2 \cos(z_5 + 1)}{1 + h\, t_6^2 \sin(z_5 + 1)}$	$= 0.622\ 458\ 656$
7	$t_7 = \frac{7\pi}{10}$	$z\left(\frac{7\pi}{10}\right) = z_6 + \frac{-z_6 + y_6 + h\, t_7^2 \cos(z_6 + 1)}{1 + h\, t_7^2 \sin(z_6 + 1)}$	$= 0.591\ 291\ 667$
8	$t_8 = \frac{8\pi}{10}$	$z\left(\frac{8\pi}{10}\right) = z_7 + \frac{-z_7 + y_7 + h\, t_8^2 \cos(z_7 + 1)}{1 + h\, t_8^2 \sin(z_7 + 1)}$	$= 0.\ 577\ 662\ 865$
9	$t_9 = \frac{9\pi}{10}$	$z\left(\frac{9\pi}{10}\right) = y_9 = z_8 + \frac{-z_8 + y_8 + h\, t_9^2 \cos(z_8 + 1)}{1 + h\, t_9^2 \sin(z_8 + 1)}$	$= 0.572\ 751\ 722$
10	$t_{10} = \pi$	$z(\pi) = z_9 + \frac{-z_9 + y_9 + h\, t_{10}^2 \cos(z_9 + 1)}{1 + h\, t_{10}^2 \sin(z_9 + 1)}$	$= 0.572\ 692\ 917$

Ainsi, la solution de l'équation différentielle en utilisant le schéma d'Euler implicite est $y(1) = 0.572692917$.

Bibliographie

[1] D.Ameur, J. Dib. (2015). *Modélisation et simulation des écoulements de fluides compressibles et incompressibles.* Polycopié expertisé et validé, Département de Physique, Faculté des Sciences, Université de Tlemcen.

[2] G.Allaire, S.M.Kaber. (2002). *Algèbre linéaire numérique.* Ellipses.

[3] R.L.Burden, J.D.Faires. (2001). *Numerical Analysis.* Brooks/Cole : 7e edition.

[4] A-L.Cholesky. (1910). *Sur la résolution numérique des systèmes d'équations linéaires.* publié en 2005 dans le Bulletin de la société des amis de la bibliothèque de l'École polytechnique (SABIX), n 39, pp 81-95.

[5] P.Ciarlet. (1985). *Introduction à l'analyse numérique matricielle et à l'optimisation.* Collection Mathématiques Appliquées pour la Maîtrise. (Collection of Applied Mathematics for the Master's Degree). Masson, Paris, (rééd. 2001).

[6] P.D.Crout. (1941). *A short method for evaluating determinants and solving systems of linear equations with real or complex coefficients.* AIEE Trans Vol 60, pp 1235-1240.

[7] J.-P.Demailly. (1996). *Analyse numérique et équations différentielles.* PUG, Grenoble .

[8] J. Dib, D.Ameur. (2015). *Statistiques Descriptives.* Polycopié expertisé et validé, Département de Mathématiques, Faculté des Sciences, Université de Tlemcen.

[9] P.Lascaux, R.Theodor. (2001). *Analyse numérique matricielle appliquée à l'art de l'ingénieur.* Masson.

[10] C.Rose. (1993). *Une méthode multifrontale pour la résolution directe de systèmes linéaires.* Note EDF - DER HI-76/93/008.

[11] M.Schatzmann. (1991). *Analyse numérique.* Cours et exercices pour la licence. (Course and exercises for the bachelor's degree). InterEditions, Paris.

[12] M.Schatzmann. (2002). *Numerical Analysis, A Mathematical Introduction.* Oxford University Press.