Université Arbi Tébessi Tébessa

Faculté des Sciences Exactes et des sciences de la nature et de la vie

Département : Mathématiques et Informatique

Cours d'Analyse Numérique 2 Pour le deuxième Année Mathématiques

Dirigé par : Dr. Bouali Tahar

Année Universitaire : 2018-2019

Table des matières

0.1 Introduction Générale

D'après les historiens, le calcul numérique remonte au moins au troisième millénaire avant notre ère. Il est à l'origine favorisé par le besoin d'effectuer des mesures dans différents domaines de la vie courante, notamment en agriculture, commerce, architecture, géographie et navigation ainsi qu'en astronomie. Il semble que les Babyloniens (qui peuplaient l'actuelle Syrie/Iraq) sont parmi les premiers à réaliser des calculs algébriques et géométriques alliant complexité et haute précision. Surtout, ils donnent une importance et un sens au placement relatif des chiffres constituant un nombre, c'est-à-dire à introduire la notion de base de dénombrement, en l'occurrence, la base sexagésimale que nous avons fini par adopter dans certains domaines. Ils se distinguent ainsi d'autres civilisations, même bien plus récentes, qui développent des méthodes plus lourdes, en introduisant une pléthore de symboles. Il y a environ 3500 ans, les populations de la vallée de l'Indus (régions de l'Inde et du Pakistan) introduisent les notions de zéro et emploient les nombres négatifs. Il adapte également le système de comptage Babylonien au système décimal qui est le nôtre aujourd'hui. Ces premiers outils de calcul sont largement développés par la suite par les Grecs, puis transmis en Europe par l'intermédiaire des civilisations musulmanes peuplant le bassin méditerranéen.

Le calcul numérique tel que nous le concevons pratiquement aujourd'hui connaît son premier véritable essor à partir du XVIIème siècle avec les progrès fulgurants des Mathématiques et de la Physique, plus ou moins liés aux observations et aux calculs astronomiques. Plusieurs machines de calcul sont en effet construites, comme la \Pascaline" inventée par B. Pascal en 1643. Babbage en 1834 mais qui fonctionnait mal, ou encore le tabulateur de H. Hollerith spécialement conçu pour recenser la population américaine, vers 1890. Il s'agit bien-entendu de machines mécaniques imposantes et d'utilisation assez limitée. Le manque de moyens de calcul performants limite en fait l'expansion et la validation de certaines théories du début du XXème siècle. Ce fut le cas en particulier de la théorie de la Relativité Générale due à A. Einstein.

La Seconde Guerre Mondiale et les progrès technologiques qu'elle engendre va permettre au calcul numérique d'amorcer un second envol. Les anglais mettent au point le premier ordinateur en 1939, Colossus, dont la mission est de décrypter les messages codes envoyées par l'émetteur ENIGMA de l'Allemagne nazie. Cette machine introduit les concepts révolutionnaires émis par A. Turing dans les années 1936 concernant l'automatisation des calculs. Les calculateurs sont désormais entièrement électroniques. Autre machine qui fait date dans l'histoire, le ENIAC (\Electronic Numerical Integrator And Computer) construit en 1946. Malheureusement, ce type de machine ne dispose pas de mémoire interne et doit être en permanence reprogrammée.

A la fin des années 1940, un certain J. von Neumann repense l'architecture des ordinateurs et introduit, entre autres, les mémoires permettant de sauvegarder les programmes, et les concepts de hardware (matériel) et de software (logiciel). La première machine de calcul incluant les concepts de von Neumann (et ceux de Turing) est ainsi produite par la firme américaine IBM ; elle s'appelle MARK I et pèse 5 tonnes. Les premières applications concernent tous les domaines sciatiques et techniques. Le FORTRAN I, un langage de programmation destine aux scientifiques, est conçu dès 1954. . . mais il lui manque un vrai compilateur.

Vers la fin des années 1960, l'apparition progressive des transistors et de leur assemblage massif sur des surfaces de plus en plus réduites augmente considérablement les performances

des machines et permet des simulations numériques de réalisme croissant. Cet effort de miniaturisation est d'ailleurs imposé par la course à la conquête de l'espace. Apparaissent ainsi en 1970 les fameux microprocesseurs mis au point par les firmes Intel et Motorola qui équipent la majeure partie des sondes spatiales de l'époque. Le calcul numérique devient rapidement une science à part entière. Les années 70 marquent aussi le tournant pour les langages de programmation : certains sont définitivement produits à des fins scientifiques, alors que d'autres seront pensés pour la gestion, comme le Cobol. Au début des années 1980, l'ordinateur le plus puissant du monde s'appelle CRAY I. Sa forme est spécialement choisie pour optimiser la rapidité des calculs. C'est aussi le début de l'informatique familiale avec la mise sur le marché des PERSONAL COMPUTERS D'IBM

En une quinzaine d'années, la rapidité des calculateurs a été multipliée par plus de 10000. La vitesse d'exécution des opérations élémentaires se compte maintenant en dizaines de millions de millions d'opérations à la seconde (ou dizaines de Téra-flops, à comparer à la centaine de méga -flops du CRAY I). Les capacités de stockage ont gagné 7 ordres de grandeur au moins. Aujourd'hui, toutes ces performances doublent tous les ans. Pour le monde scientifique, celui de la Recherche Fondamentale et de l'Industrie, les calculateurs et le développement de techniques de programmation spécifiques (comme la programmation parallèle) sont devenus des outils incontournables à la connaissance et ouvrent de nouveaux horizons pour la modélisation et la compréhension des phénomènes complexes et la mise au point de nouvelles technologies.

On regroupe sous le terme générique de "méthodes numériques", toutes les techniques de calcul qui permettent de résoudre de manière exacte ou, le plus souvent, de manière approchée un problème donné. Le concept de calcul est assez vaste et doit être pris au sens large. Il peut s'agir de déterminer l'inconnue d'une équation, de calculer la valeur d'une fonction en un point ou sur un intervalle, d'intégrer une fonction, d'inverser une matrice, etc. Bien que la mise en équation d'un problème et sa résolution passent naturellement par les Mathématiques, les problématiques sous-jacentes concernent des disciplines aussi variées que la Physique, l'Astrophysique, la Biologie, la Médecine, l'Economie, etc. Il existe ainsi une grande variété de problèmes possibles avec pour chacun d'eux, des méthodes très spécifiques. De fait, le nombre total de méthodes numériques dont nous disposons à l'heure actuelle est vraisemblablement gigantesque.

Une méthode numérique met en œuvre une certaine procédure, une suite d'opérations, généralement en tries grand nombre, que l'on transcrira ensuite dans un langage de programmation. Bien qu'une méthode numérique puisse s'effectuer mentalement (du moins avec un crayon et un papier) comme inverser une matrice $2x2$, résoudre $\tan x - 1 = 0$, ou calculer $\sqrt{2}$, elle nécessite dans la majorité des cas un ordinateur qui a l'avantage de la rapidité (mais pas de la précision). Il convient à ce niveau de bien différencier la partie méthode numérique, souvent indépendante du calculateur et du langage, et la partie programmation qui met en œuvre d'une part l'algorithme et d'autre part une suite d'instructions écrites dans un langage de programmation. Bien sûr, une méthode numérique pourra dépendre de l'architecture d'un ordinateur et du langage utilisé. Toutefois, l'un des soucis majeurs de l'utilisateur et du programmeur est d'assurer à son programme une certaine portabilité, c'est-à-dire de pouvoir l'exécuter sur des machines différentes sans avoir besoin d'adaptations (trop) spécifiques.

Les méthodes numériques sont indispensables à la réalisation de programmes de calculs ou

codes de calcul. En particulier, pour les astrophysiciens qui ne bénéficient pas d'un laboratoire permettant de valider leurs théories à partir d'expériences renouvelables à loisir et contrôlables, ces outils sont le seul moyen de simuler ou de modéliser les phénomènes que nous observons, de les interpréter et de les comprendre. Rappelons que les méthodes numériques sont en effet présentent dans toutes les disciplines de l'Astrophysique moderne : la cosmologie, l'instrumentation, le traitement de données, la planétologie, la physique solaire, la physique des galaxies, la physique extragalactique, etc. S'il est vrai qu'il existe une très grande diversité de méthodes numériques avec lesquelles ont peut, en pratique, quasiment tout faire, certains problèmes (par exemple en traitement du signal, en mécanique céleste ou en mécanique des fluides) ont nécessité la mise au point de méthodes très spécifiques.

L'objet de l'analyse numérique est de concevoir et d'étudier des méthodes de résolution de certains problèmes mathématiques, en général issus de la modélisation de problèmes "réels", et dont on cherche à calculer la solution à l'aide d'un ordinateur.

Le cours est structuré en quatre grands chapitres :

– Résolution des systèmes linéaires
– Calcul des valeurs et vecteurs propres
– Systèmes non linéaires
– Equations différentielles.

Chapitre 1

Résolution des systèmes linéaires

On note $M_N(\mathbb{R})$ l'ensemble des matrices carrées d'ordre N. Soit $A \in M_N(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^n$, on a comme objectif de résoudre le système linéaire $Ax = b$, c'est à dire de trouver x solution de :

$$\begin{cases} x \in \mathbb{R}^N \\ Ax = b \end{cases} \tag{P}$$

Comme A est inversible, il existe un unique vecteur $x \in \mathbb{R}^N$ solution de (P). Nous allons étudier dans les deux chapitres suivants des méthodes de calcul de ce vecteur x : la première partie de ce chapitre sera consacrée aux méthodes "directes" et la deuxième aux méthodes "itératives". Nous aborderons ensuite en troisième partie les méthodes de résolution de problèmes aux valeurs propres.

Un des points essentiels dans l'efficacité des méthodes envisagées concerne la taille des systèmes à résoudre. Entre 1980 et 2000, la taille de la mémoire des ordinateurs a augmenté de façon drastique. La taille des systèmes qu'on peut résoudre sur ordinateur a donc également augmenté.

Le développement des méthodes de résolution de systèmes linéaires est liée à l'évolution des machines informatiques. Un grand nombre de recherches sont d'ailleurs en cours pour profiter au mieux de l'architecture des machines (méthodes de décomposition en sous domaines pour profiter des architectures parallèles, par exemple).

Dans la suite de ce chapitre, nous verrons deux types de méthodes pour résoudre les systèmes linéaires : les méthodes directes et les méthodes itératives. Pour faciliter la compréhension de leur étude, nous commençons par quelques rappels d'algèbre linéaire.

1.1 Quelques rappels d'algèbre linéaire

1.1.1 Norme induite

Définition 1.1.1 (Norme matricielle, norme induite) *On note $M_N(\mathbb{R})$ l'espace vectoriel (sur $\mathbb{R}$) des matrices carrées d'ordre N .*

1. On appelle norme matricielle sur $M_N(\mathbb{R})$ une norme $\|.\|$sur $M_N(\mathbb{R})$ t.q.

$$\|AB\| \leq \|A\| \|B\|, \forall A, B \in M_N(\mathbb{R})$$

2. On considère $M_N(\mathbb{R})$ muni d'une norme $\|.\|$. On appelle norme matricielle induite (ou norme induite) sur $M_N(\mathbb{R})$ par la norme $\|.\|$, encore notée $\|.\|$, la norme sur $M_N(\mathbb{R})$ définie par :

$$\|A\| = \sup \{\|Ax\|;\ x \in \mathbb{R}^n,\ \|x\| = 1\}, \forall A \in M_N(\mathbb{R})$$

Proposition 1 *Soit $M_N(\mathbb{R})$ muni d'une norme induite $\|.\|$. Alors pour toute matrice $A \in M_N(\mathbb{R})$, on a :*

1. $\|Ax\| \leq \|A\| \|x\|, \forall x \in \mathbb{R}^n$,

2. $\|A\| = \max \{\|Ax\|;\ \|x\| = 1,\ x \in \mathbb{R}^n\}$,

3. $\|A\| = \max \left\{ \frac{\|Ax\|}{\|x\|};\ x \in \mathbb{R}^n \backslash \{0\} \right\}$

4. $\|.\|$ *est une norme matricielle.*

Proposition 2 *Soit* $A = (a_{i,\,j})_{i,\,j \in \{1,\ldots,N\}} \in M_N(\mathbb{R})$

1. On munit $\mathbb{R}^n$ de la norme $\|.\|_\infty$ et $M_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|.\|_\infty$. Alors

$$\|A\|_\infty = \max_{i \in \{1,\ldots,N\}} \sum_{j=1}^{N} |a_{i,\,j}|$$

2. On munit $\mathbb{R}^n$ de la norme $\|.\|_1$ et $M_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|.\|_1$. Alors

$$\|A\|_1 = \max_{j \in \{1,\ldots,N\}} \sum_{i=1}^{N} |a_{i,\,j}|$$

3. On munit $\mathbb{R}^n$ de la norme $\|.\|_2$ et $M_N(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|.\|_2$. Alors

$$\|A\|_2 = \left(\rho\left(A^t A\right)\right)^{\frac{1}{2}}$$

1.1.2 Rayon spectral

Définition 1.1.2 (Valeurs propres et rayon spectral) *Soit $A \in M_N(\mathbb{R})$ une matrice inversible. On appelle valeur propre de A tout $\lambda \in \mathbb{C}$ tel qu'il existe $x \in \mathbb{C}^N$,$x \neq 0$ tel que $Ax = \lambda x$. L'élément x est appelé vecteur propre de A associé à λ. On appelle rayon spectral de A la quantité $\rho(A) = \max \{\lambda;\ \lambda \in \mathbb{C},\ \lambda$ valeur propre de $A\}$.*

Lemme 1.1.1 (Convergence et rayon spectral) *On munit $M_N(\mathbb{R})$ d'une norme, notée $\|.\|$. Soit $A \in M_N(\mathbb{R})$.*

Alors :

1. $\rho(A) < 1$ *si et seulement si* $A^k \to 0$ *quand* $k \to \infty$.

2. $\rho(A) < 1 \Rightarrow \limsup_{k \to \infty} \left\|A^k\right\|^{\frac{1}{k}} \leq 1$.

3. $\liminf_{k \to \infty} \left\|A^k\right\|^{\frac{1}{k}} < 1$.

4. $\rho(A) = \lim_{k\to\infty} \left\|A^k\right\|^{\frac{1}{k}}$.

5. On suppose de plus que $\|.\|$ *une norme matricielle (induite ou non). Alors*

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

Remarque 1.1.1 (Convergence des suites) *Une conséquence immédiate du lemme est que si* $x^{(0)}$ *est donné et* $x^{(k)}$ *défini par* $x^{(k+1)} = Ax^{(k)}$*, alors la suite* $\left(x^{(k)}\right)_{k\in\mathbb{N}}$ *converge vers* 0 *si et seulement si* $\rho(A) < 1$*.*

Proposition 3 (Rayon spectral et norme induite) *Soient* $A \in M_N(\mathbb{R})$ *et* $\varepsilon > 0$*. Il existe une norme sur* $\mathbb{R}^n$ *(qui dépend de* A *et* ε*) telle que la norme induite sur* $M_N(\mathbb{R})$*, notée* $\|.\|_{A,\,\varepsilon}$ *vérifie* $\|A\|_{A,\,\varepsilon} \leq \rho(A) + \varepsilon$*.*

Lemme 1.1.2 (Triangularisation d'une matrice) *Soit* $A \in M_N(\mathbb{R})$ *une matrice carrée quelconque, alors il existe une base* $(f_1, ..., f_N)$ *de* $\mathbb{C}$ *et une famille de complexes* $(\lambda_{i,\,j})_{i=1,...,N,j=1,...,N,j<i}$ *telles que* $Af_i = \lambda_{i,i}f_i + \sum_{j<i} \lambda_{i,j}f_j$ *De plus* $\lambda_{i,j}$ *est valeur propre de* A *pour tout* $i \in \{1, ..., N\}$*.*

On admettra ce lemme.

Nous donnons maintenant un théorème qui nous sera utile dans l'étude du conditionnement, ainsi que plus tard dans l'étude des méthodes itératives.

Théorème 1.1.1 (Matrices de la forme $Id + A$**)** *1. Soit une norme matricielle induite,* Id *la matrice identité de* $M_N(\mathbb{R})$ *et* $A \in M_N(\mathbb{R})$ *telle que* $\|A\| < 1$*.*

Alors la matrice $Id + A$ *est inversible et*

$$\left\|(Id + A)^{-1}\right\| \leq \frac{1}{1 - \|A\|}.$$

2. Si une matrice de la forme $Id + A \in M_N(\mathbb{R})$ *est singulière, alors* $\|A\| \geq 1$ *pour toute norme matricielle* $\|.\|$*.*

1.1.3 Matrices diagonalisables

Définition 1.1.3 (Matrice diagonalisable) *Soit* A *une matrice réelle carrée d'ordre* n*. On dit que* A *est diagonalisable dans* $\mathbb{R}$ *si il existe une base* $(\Phi_1, ..., \Phi_n)$ *et des réels* $\lambda_1, ..., \lambda_n$ *(pas forcément distincts) tels que* $A\Phi_i = \lambda\Phi_i$ *pour* $i = 1, ..., n$*. Les réels* $\lambda_1, ..., \lambda_n$ *sont les valeurs propres de* A*, et les vecteurs* $\Phi_1, ..., \Phi_n$ *sont les vecteurs propres associés.*

Lemme 1.1.3 *Soit* A *une matrice réelle carrée d'ordre* n*, diagonalisable dans* $\mathbb{R}$*. Alors*

$$A = Pdiag\,(\lambda_1, ..., \lambda_n)\,P^{-1},$$

où P *est la matrice dont les vecteurs colonnes sont égaux aux vecteurs* $\Phi_1, ..., \Phi_n$*.*

Lemme 1.1.4 *Soit E un espace vectoriel sur $\mathbb{R}$ de dimension finie : $dimE = n$, $n \in \mathbb{N}^*$, muni d'un produit scalaire i.e. d'une application*

$$\begin{aligned} E \times E &\rightarrow \mathbb{R}, \\ (x, y) &\rightarrow < x, y >_E \end{aligned}$$

qui vérifie :

$$\forall x \in E, < x, x >_E \geq 0 \text{ et } < x, x >_E = 0 \Leftrightarrow x = 0,$$

$$\forall (x, y) \in E^2, < x, y >_E = < y, x >_E,$$

$$\forall y \in E, \text{l'application de } E \text{ dans } \mathbb{R}, \text{définie par } x \rightarrow < x, y >_E \text{ est linéaire.}$$

Ce produit scalaire induit une norme sur E, $\|x\| = \sqrt{< x, x >_E}$.

Soit T une application linéaire de E dans E. On suppose que T est symétrique, c.à.d. que $< T(x), y >_E = < x, T(y) >_E$, $\forall (x, y) \in E^2$. Alors il existe une base orthonormée $(f_1 ... f_n)$ de E (c.à.d. telle que $(f_i, f_j)_E = \delta_{i,j}$) et $(\lambda_1, ..., \lambda_n) \in \mathbb{R}^n$ tels que $T(f_i) = \lambda_i f_i$ pour tout $i \in \{1...n\}$.

Conséquence immédiate : Dans le cas où $E = \mathbb{R}^n$, le produit scalaire canonique de $x = (x_1, ..., x_N)^t$ et $y = (y_1, ..., y_N)^t$ est défini par $< x, y >_E = x \cdot y = \sum_{i=1}^{N} x_i y_i$ Si $A \in M_N(\mathbb{R})$ est une matrice symétrique, alors l'application T définie de E dans E par : $T(x) = Ax$ est linéaire, et : $< T(x), y >= Ax \cdot y = x \cdot A^t y = x \cdot Ay = < x, T(y) >$. Donc T est linéaire symétrique. Par le lemme précédent, il existe $(f_1, ..., f_N)$ et $(\lambda_1, ..., \lambda_N) \in \mathbb{R}^N$ tels que $Tf_i = Af_i = \lambda_i f_i$ $\forall i \in \{1, ..., N\}$ et $f_i . f_j = \delta_{i,j}$, $\forall (i, j) \in \{1, ..., N\}^2$.

Interprétation algébrique : Il existe une matrice de passage P de $(e_1, ..., e_N)$ base canonique dans $(f_1, ..., f_N)$ dont la première colonne de P est constituée des coordonnées de f_i dans $(e_1, ..., e_N)$. On a : $Pe_i = f_i$. On a alors $P^{-1}APe_i = P^{-1}Af_i = P^{-1}(\lambda_i f_i) = \lambda_i e_i = diag\,(\lambda_1, ..., \lambda_N)\, e_i$, où $diag\,(\lambda_1, ..., \lambda_N)$ désigne la matrice diagonale de coefficients diagonaux $\lambda_1, ..., \lambda_N$ On a donc :

$P^{-1}AP = \begin{bmatrix} \lambda_i & 0 \\ 0 & \lambda_N \end{bmatrix} = D.$

De plus P est orthogonale, i.e. $P^{-1} = P^t$. En effet,

$$P^t Pe_i \cdot e_j = Pe_i \cdot Pe_j = < f_i, f_j >= \delta_{i,j} \forall i, j \in \{1...N\},$$

et donc $(P^t Pe_i - e_i) \cdot e_j = 0$ $\forall j \in \{1...N\}$ $\forall i \in \{1, ...N\}$. On en déduit $P^t Pe_i = e_i$ pour tout $i = 1, ...N$,

i.e. $P^t P = PP^t = Id$.

1.2 Méthodes directes

On appelle méthode directe de résolution de (P) une méthode qui donne exactement x (A et b étant connus) solution de (P) après un nombre fini d'opérations élémentaires $(+, -, \times, /)$.

Parmi les méthodes de résolution du système (P) on citera :

- Méthode de Gauss(méthode du pivot)
- Méthode de Gauss-Jordan
- La méthode $L.U$
- Méthode de Cholesky

1.2.1 Méthode de Gauss

Soit $Ax = b$ où A est une matrice $(n \times n)$, non singulière ($\det(A) \neq 0 \Leftrightarrow A$ non singulière).

Principe :

$\star$ Transformation de la matrice A en une matrice triangulaire supérieure .

on construit $\left[A \vdots b\right]$ et :

$\left[A \vdots b\right] -$ transformation $\longrightarrow \left[A' \vdots b'\right]$ une matrice triangulaire supérieure .*ie* :

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{2n} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} & b_n \end{pmatrix} \longrightarrow \begin{pmatrix} a_{11}^{'} & a_{12}^{'} & \cdots & a_{1n}^{'} & b_1^{'} \\ 0 & a_{22}^{'} & \cdots & a_{2n}^{'} & b_2^{'} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{'} & b_n^{'} \end{pmatrix}$$

$$\left[A \vdots b\right] \longrightarrow \left[A' \vdots b'\right]$$

$\star$ Puis, on résout le système $A^{'}x = b^{'}$ (dont la solution est exactement la solution du système $Ax = b$)

ETAPES : On pose $A = A^{(1)}$ et $b = b^{(1)}$.

1^{ere} étape :

$\star$ Si $a_{11}^{(1)} \neq 0$, on fait les affectations suivantes

- La ligne L_1 est maintenue ie : $L_1^{(2)} \longleftarrow L_1^{(1)}$
- pour $i = \overline{2, n}$; $\quad L_i^{(2)} \longleftarrow L_i^{(2)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot L_1^{(1)}$

On obtient alors :

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{pmatrix} \longrightarrow \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} & b_1^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{pmatrix}$$

$$\left[A^{(1)} \vdots b^{(1)}\right] \longrightarrow \left[A^{(2)} \vdots b^{(2)}\right]$$

où :

$$\begin{cases} a_{1j}^{(2)} = a_{1j}^{(1)} \qquad j = \overline{1, n} \\ a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot a_{1j}^{(1)} \quad , \quad i = \overline{2, n} \quad , \quad j = \overline{1, n} \end{cases}$$

et

$$\begin{cases} b_1^{(2)} = b_1^{(1)} \\ b_1^{(2)} = b_i^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} \cdot b_1^{(1)} \quad , \quad i = \overline{2, n} \end{cases}$$

$\star$ Si $a_{11}^{(1)} = 0$, on cherche une ligne $L_p^{(1)}$ avec $2 \leq p \leq n$ telle que $a_{p1}^{(1)} \neq 0$. Puis on permute les lignes $L_1^{(1)}$et $L_p^{(1)}$ pour obtenir

$Ax = b \iff A^{(1)}x = b^{(1)} \iff P^{(1)}A^{(1)}x = P^{(1)}b^{(1)}$

où $P^{(1)}$ est la matrice de permutation des lignes $L_1^{(1)}$et $L_p^{(1)}$.$P^{(1)}$ est elle meme la matrice identité dans laquelle on permute la premiére ligne et la P^{eme} ligne.

Dans ce cas au lieu de la matrice $A^{(1)}$ on considére la matrice : $\tilde{A}^{(1)} = P^{(1)}A^{(1)}$ dont on notera encore les éléments par $a_{ij}^{(1)}$ et on lui applique des transformations analogues à celles correspondantes au cas $a_{11}^{(1)} \neq 0$,étudié plus haut.

$\underline{k^{eme} \text{ étape :}}$

$\star$ $a_{kk}^{(k)} = 0$: On permute les lignes est une ligne d'indice p avec $k+1 \leq p \leq n$, telle que : $a_{pk}^{(k)} \neq 0$. Et dela

$$A^{(k)}x = b^{(k)} \iff P^{(k)}A^{(k)}x = P^{(k)}b^{(k)}$$

où P^k est la matrice de permutation des lignes $L_k^{(k)}$ et $L_p^{(k)}$. P^k est la matrice identité où on permute la k^{eme} et la P^{eme} ligne.

On considére alors : $\tilde{A}^{(k)} = P^{(k)}A^{(k)}$ et $\tilde{b}^{(k)} = P^{(k)}A^{(k)}$. Aprés transformation on obtient $A^{(k+1)}$ et $b^{(k+1)}$ avec :

$$\begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)}, \ \ i = \overline{1,k}; \ j = \overline{1,n} \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \cdot a_{kj}^{(k)}, \ i = \overline{k+1,n}; \ \ j = \overline{1,n} \end{cases}$$

et

$$\begin{cases} b_i^{(k+1)} = b_i^{(k)} \quad i = \overline{1,k} \\ b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \cdot b_k^{(k)}, i = \overline{k+1,n} \end{cases}$$

et ceci en faisant les affectations suivantes :

$$\begin{cases} L_1^{(k+1)} \longrightarrow L_1^{(k)} \\ L_2^{(k+1)} \longrightarrow L_2^{(k)} \\ \vdots \\ L_k^{(k+1)} \longrightarrow L_k^{(k)} \\ L_i^{(k+1)} \longrightarrow L_i^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} \cdot L_k^{(k)}, ,i = \overline{k+1,n} \end{cases}$$

$\underline{\text{Résolution de } A'x = b' :}$

On pose

$$\left[A' \vdots b'\right] = \left[A^{(n)} \vdots b^{(n)}\right] = \begin{pmatrix} a_{11}^{(n)} & a_{11}^{(n)} & \cdots & a_{11}^{(n)} & b_1^{(n)} \\ 0 & a_{11}^{(n)} & \cdots & a_{11}^{(n)} & b_2^{(n)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \vdots & \cdots & a_{11}^{(n)} & b_n^{(n)} \end{pmatrix}$$

Et delà

$$\begin{aligned} Ax = b &\iff A^{(1)}x = b^{(1)} \\ &\iff A^{(2)}x = b^{(2)} \\ &\vdots \\ &\iff A^{(n)}x = b^{(n)} \\ &\iff A'x = b' \end{aligned}$$

d'où (**Résolution par remontée)**

$$\begin{cases} x_1 = \frac{1}{a'_{11}} \cdot (b'_1 - a'_{1,2}x_2 - \cdots - a'_{1,n}x_n) \\ \quad\vdots \\ x_{n-1} = \frac{1}{a'_{n-1,n-1}} \cdot (b'_{n-1} - a'_{n-1,2}x_n) \\ x_n = \frac{1}{a'_{n,n}} \cdot b'_n \end{cases}$$

(On détermine x_n , puis x_{n-1} , etc. jusqu'à obtention de x_1.)

Remarque 1.2.1 *1. La méthode de Gauss nécessite $\frac{2}{3}n^3$ opérations pour un systéme d'ordre n.*

2. Elle permet de calculer $det\,(A)$ puisque $det\,(A) = (-1)^j \prod_{k=1}^{n} a_{kk}^{(k)}$ où j est le nombre de permulations.

Exemples 1 Soit à résoudre le systéme :

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

Le systéme (1) s'écrit encore : $Ax = b$ avec

$$A = \begin{pmatrix} 2 & 3 & -1 \\ 4 & 4 & -3 \\ -2 & 3 & -1 \end{pmatrix} \quad ; b = \begin{pmatrix} 5 \\ 3 \\ 1 \end{pmatrix} \qquad et \qquad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

$\underline{1^{ere}\text{ étape :}}$
$\star\ a_{11}^{(1)} = 2 \neq 0$

$$\begin{pmatrix} \boxed{2} & 3 & -1 & 5 \\ 4 & 4 & -3 & 3 \\ -2 & 3 & -1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 3 & -1 & 5 \\ 0 & -2 & -1 & -7 \\ 0 & 6 & -2 & 6 \end{pmatrix}$$

$$\left[A^{(1)} : b^{(1)}\right] \longrightarrow \left[A^{(2)} : b^{(2)}\right]$$

$\underline{2^{eme} \text{ étape :}}$

$\star\ a_{22}^{(2)} = -2 \neq 0$

$$\begin{pmatrix} 2 & 3 & -1 & 5 \\ 0 & \boxed{-2} & -1 & -7 \\ 0 & 6 & -2 & 6 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

$$\left[A^{(2)} \vdots b^{(2)}\right] \longrightarrow \left[A^{(3)} \vdots b^{(3)}\right]$$

$\underline{\text{Résolution de } A'x = b' :}$

Posons $\left[A' \vdots b'\right]$. On a alors :

$$A'x = b' \iff \begin{pmatrix} 2 & 3 & -1 \\ 0 & -2 & -1 \\ 0 & 0 & -5 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ -7 \\ -15 \end{pmatrix}$$

$$\iff \quad \begin{cases} 2x_1 + 3x_2 - x_3 = 5 \\ -2x_2 + x_3 = -7 \\ -5x_3 = 12 \end{cases} \iff \begin{cases} x_1 = 1 \\ x_2 = 2 \\ x_3 = 3 \end{cases}$$

1.2.2 Méthode de Gauss-Jordan

Soit le systéme linéaire $Ax = b$ où A est une matrice $(n \times n)$, réguliére.

$\underline{\text{Principe :}}$

$\star$ Transformation de la matrice A en matrice identité

ie : $\left[A \vdots b\right]$ _ transformation $\longrightarrow$ $\left[I \vdots b'\right]$ où I est la matrice identité.

$\star$ D'où : $Ax = b \iff Ix = b' \iff x = b'$

$\underline{\text{ETAPES :}}$ On pose $A = A^{(1)}$ et $b = b^{(1)}$.

$\underline{1^{ere} \text{ étape :}}$

$\star$ Si $a_{11}^{(1)} \neq 0$, on fait les affectations suivantes

• $L_1^{(2)} \longleftarrow \frac{1}{a_{11}^{(1)}} \cdot L_1^{(1)}$

• $L_i^{(2)} \longleftarrow L_i^{(2)} - a_{i1}^{(1)} \cdot L_i^{(2)} \quad ; \quad i = \overline{2, n}$

On a alors :

$$\begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & b_2^{(1)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & b_n^{(1)} \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} & b_1^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{pmatrix}$$

$$\left[A^{(1)} \vdots b^{(1)}\right] \longrightarrow \left[A^{(2)} \vdots b^{(2)}\right]$$

avec :

$$\begin{cases} a_{1j}^{(2)} = \frac{a_{1j}^{(1)}}{a_{11}^{(1)}} \ ; \quad j = \overline{1,n} \\ a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i1}^{(1)} \cdot a_{1j}^{(2)} \ ; \quad i = \overline{2,n} \ ; \quad j = \overline{2,n} \\ a_{ij}^{(2)} = 0 \ , \quad i = \overline{2,n} \ ; \quad j = 1 \end{cases}$$

et

$$\begin{cases} b_1^{(2)} = \frac{b_1^{(1)}}{a_{11}^{(1)}} \\ b_i^{(2)} = b_i^{(1)} - a_{i1}^{(1)} \cdot b_i^{(2)}, \quad i = \overline{2,n} \end{cases}$$

$\underline{1^{ere} \text{ étape :}}$

$\star$ Si $a_{22}^{(2)} \neq 0$, on fait les affectations suivantes

- $L_2^{(3)} \longleftarrow \frac{1}{a_{22}^{(2)}} \cdot L_2^{(2)}$
- $L_i^{(3)} \longleftarrow L_i^{(2)} - a_{i2}^{(2)} \cdot L_2^{(3)} \quad ; \quad i = \overline{1,n} \quad \text{avec} \quad i \neq 2$

d'où :

$$\begin{pmatrix} 1 & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} & b_1^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & b_2^{(2)} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & b_n^{(2)} \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & a_{13}^{(3)} & \cdots & a_{1n}^{(3)} & a_1^{(3)} \\ 0 & 1 & a_{23}^{(3)} & \cdots & a_{2n}^{(3)} & a_2^{(3)} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(3)} & \cdots & a_{nn}^{(3)} & a_n^{(3)} \end{pmatrix}$$

$$\left[A^{(2)} \vdots b^{(2)}\right] \longrightarrow \left[A^{(3)} \vdots b^{(3)}\right]$$

avec :

$$\begin{cases} a_{1j}^{(2)} = \frac{a_{1j}^{(1)}}{a_{11}^{(1)}} \ ; \quad j = \overline{1,n} \\ a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i1}^{(1)} \cdot a_{1j}^{(2)} \ ; \quad i = \overline{2,n} \ ; \quad j = \overline{2,n} \\ a_{ij}^{(2)} = 0; \quad i = \overline{2,n} \ ; \quad j = 1 \end{cases}$$

et

$$\begin{cases} b_1^{(2)} = \frac{b_1^{(1)}}{a_{11}^{(1)}} \\ b_i^{(3)} = b_i^{(2)} - a_{i2}^{(\acute{e})} \cdot b_2^{(")}; \quad i = \overline{1,n}, \quad i \neq 0 \end{cases}$$

$\underline{k^{eme} \text{ étape :}}$

$\star$ $a_{kk}^{(k)} \neq 0$:

- $L_k^{(1)} \longleftarrow \frac{1}{a_{kk}^{(k)}} \cdot L_k^{(k)}$
- $L_i^{(k+1)} \longleftarrow L_i^{(k)} - a_{ik}^{(k)} \cdot L_k^{(k+1)} \quad ; \quad i = \overline{1,n} \quad \text{avec} \quad i \neq k$

On obtient :

$$\left(\begin{array}{ccccccc} 1 & \cdots & 0 & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} & b_1^{(k)} \\ 0 & \cdots & 0 & a_{2k}^{(k)} & \cdots & a_{2n}^{(k)} & b_2^{(k)} \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 1 & a_{k-1,k}^{(k)} & \cdots & a_{k-1,n}^{(k)} & b_{k-1}^{(k)} \\ 0 & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{1k}^{(k)} & b_k^{(k)} \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{array}\right) \longrightarrow \left(\begin{array}{ccccccc} 1 & \cdots & 0 & a_{1,k+1}^{(k+1)} & \cdots & a_{1n}^{(k+1)} & b_1^{(k+1)} \\ 0 & \cdots & 0 & a_{2,k+1}^{(k+1)} & \cdots & a_{2n}^{(k+1)} & b_2^{(k+1)} \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 1 & a_{k,k+1}^{(k+1)} & \cdots & a_{k,n}^{(k+1)} & b_k^{(k+1)} \\ 0 & \cdots & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} & b_{k+1}^{(k+1)} \\ \vdots & & & & & & \vdots \\ 0 & \cdots & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{nn}^{(k+1)} & b_n^{(k+1)} \end{array}\right)$$

$$\left[A^{(k)} \vdots b^{(k)}\right] \longrightarrow \left[A^{(k+1)} \vdots b^{(k+1)}\right]$$

avec

$$\left\{\begin{array}{l} a_{kj}^{(k+1)} = \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}} \quad ; \qquad j = \overline{k,n} \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} \; - \; a_{ik}^{(k)} \cdot a_{kj}^{(k+1)} \; ; \quad i = \overline{1,n},\ i \neq k; \quad j = \overline{k+1,n} \end{array}\right.$$

et

$$\left\{\begin{array}{l} b_k^{(k+1)} = \frac{b_k^{(k)}}{a_{kk}^{(k)}} \\ b_i^{(k+1)} = b_i^{(k)} - a_{ik}^{(k)} \cdot b_k^{(k+1)}, \qquad i = \overline{1,n} \ , \quad i \neq k \end{array}\right.$$

Remarque 1.2.2 *1. La méthode de Gauss-Jordan nécessite n^3 opérations élémentaires (moins rapide que celle de Gauss et que celle de Cholesky que l'on verra par la suite).*

2. Elle est conseillée pour inverser une matrice, car elle évite la "remontée" (ie :la résolution par retour arrière) qu'on rencontre dans la méthode de Gauss.

Exemples 3 Soit à résoudre le systéme :

$$(1) \qquad \left\{\begin{array}{r} 2x_1 + 3x_2 \; - x_3 = 5 \\ 4x_1 + 4x_2 - 3x_3 = 3 \\ -\,2x_1 + 3x_2 - x_3 = 1 \end{array}\right.$$

Le systéme (1) s'écrit encore :$Ax = b$ avec

$$A = \left(\begin{array}{ccc} 2 & 3 & -1 \\ 4 & 4 & -3 \\ -2 & 3 & -1 \end{array}\right) \quad ; \quad b = \left(\begin{array}{c} 5 \\ 3 \\ 1 \end{array}\right) \qquad et \qquad x = \left(\begin{array}{c} x_1 \\ x_2 \\ x_3 \end{array}\right)$$

$\underline{1^{ere}\text{ étape :}}$

$\star\ a_{11}^{(1)} = 2 \neq 0$

$$\left(\begin{array}{cccc} \boxed{2} & 3 & -1 & 5 \\ 4 & 4 & -3 & 3 \\ -2 & 3 & -1 & 1 \end{array}\right) \longrightarrow \left(\begin{array}{cccc} 1 & 3/2 & -1/2 & 5/2 \\ 0 & -2 & -1 & -7 \\ 0 & 6 & -2 & 6 \end{array}\right)$$

$$\left[A^{(1)}:b^{(1)}\right] \longrightarrow A^{(2)}:b^{(2)}$$

$\underline{2^{eme} \text{ étape :}}$
$\star\ a_{22}^{(2)} = -2 \neq 0$

$$\begin{pmatrix} 1 & 3/2 & -1/2 & 5/2 \\ 0 & \boxed{-2} & -1 & -7 \\ 0 & 6 & -2 & 6 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & -5/4 & -11/4 \\ 0 & 1 & 1/2 & 7/2 \\ 0 & 0 & -5 & -15 \end{pmatrix}$$

$$\left[A^{(2)}:b^{(2)}\right] \longrightarrow \left[A^{(3)}:b^{(3)}\right]$$

$\underline{3^{eme} \text{ étape :}}$
$\star\ a_{33}^{(3)} = -5 \neq 0$

$$\begin{pmatrix} 1 & 0 & -5/4 & -11/4 \\ 0 & 1 & 1/2 & 7/2 \\ 0 & 0 & \boxed{-5} & -15 \end{pmatrix} \longrightarrow \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 1 & 3 \end{pmatrix}$$

$$\left[A^{(3)}:b^{(3)}\right] \longrightarrow \left[A^{(4)}:b^{(4)}\right]$$

$\underline{\text{Solution de } A'x = b' :}$
Posons $b' = b^{(4)}$.On a alors : $x = b' = (1.2.3)$.

1.2.3 Stratégie du choix du pivot

Exemples 4 Sachant que la solution exacte du systéme ci-aprés est $(x_1, x_2) = (1/3, 2/3)$, retrouvons la par la méthode de Gauss. Le systéme est

$$\begin{cases} 0.0003x_1 + 3x_2 = 2.0001 \\ x_1 + x_2 = 1 \end{cases}$$

$\star$ Posons

$$A = \begin{pmatrix} 0.0003 & 3 \\ 1 & 1 \end{pmatrix} \quad et \quad b = \begin{pmatrix} 2.0001 \\ 1 \end{pmatrix}$$

Nous avons : $a_{11} = 0.0003 \neq 0$, d'où :

$$\begin{pmatrix} 0.0003 & 3 & 2.0001 \\ 1 & 1 & 1 \end{pmatrix} \longrightarrow \begin{pmatrix} 0.0003 & 3 & 2.0001 \\ 0 & -9999 & -6666 \end{pmatrix}$$

$$\left[A \vdots b\right] \longrightarrow \left[A' \vdots b'\right]$$

et delà :

$$x_2 = \frac{6666}{9999} = \frac{2222}{3333} = ?$$
$$x_1 = \frac{1}{0.0003}(2.0001 - 3x_2) = ?$$

Question 1 : Effectuer les calculs avec quatre(04) c.s

Réponse 1 : $x_2 = 0.6667$ et $x_1 = -0.3333$

Question 2 : même question avec cinq (05) c.s

Réponse 2 : $x_2 = 0.66667$ et $x_1 = 0.3$

Question 3 : avec trois (03) c.s

Réponse 3 : $x_2 = 0.667$ et $x_1 = -3$

Remarque 1.2.3 *1. Les chiffres de la valeur x_2 restent stables. Par contre ceux de x_1 mûtent à chaque nouvelle précision.*

2. Les solutions auxquelles on aboutit sont, à des échelles différents, éloignées de la solution exacte.

Commentaire 1 Cette perte dans la précision est due au pivot $a_{11} = 0.0003$ qui est très petit.

Cas général

Soit le système

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ a_{22}x_2 + \cdots + a_{2n}x_n = b_2 \\ \ddots \\ a_{nn}x_n = b_n \end{cases}$$

avec $a_{11} \simeq 0$ et $a_{11} \neq 0$.

On suppose que les solutions $x_1, x_2, \ldots, x_n$ soient connues.Posons :

$x_1 = x_1^* \pm \triangle x_1$
$x_2 = x_2^* \pm \triangle x_2$ où $x_1^*, x_2^*, \ldots, x_n^*$ sont des valeurs
$\vdots$ approchées de $x_1, x_2, \ldots, x_n$
$x_n = x_n^* \pm \triangle x_n$

Déterminons $\triangle x_1 = ?$

Nous avons :

$$\begin{aligned} x_1 &= \frac{1}{a_{11}}(b_1 - a_{12}x_2 - \cdots - a_{1n}x_n) \\ &= \frac{1}{a_{11}}(b_1 - a_{12}x_2^* - \cdots - a_{1n}x_n^* \mp a_{12}\triangle x_2 \mp \cdots \mp a_{1n}\triangle x_n) \\ &= \frac{1}{a_{11}}b_1 - a_{12}x_2^* - \cdots - a_{1n}x_n^*) \mp \frac{a_{12}}{a_{11}}\triangle x_2 \mp \cdots \mp \frac{a_{1n}}{a_{11}}\triangle x_n \\ &= x_1^* \mp \frac{a_{12}}{a_{11}}\triangle x_2 \mp \cdots \mp \frac{a_{1n}}{a_{11}}\triangle x_n \end{aligned}$$

et delà : $\triangle x_1 = \mp\frac{a_{12}}{a_{11}}\triangle x_2 \mp \cdots \mp \frac{a_{1n}}{a_{11}}\triangle x_n$.Mais puisque $a_{11} \simeq 0$ alors $\frac{a_{12}}{a_{11}} \gg 0$.

De même pour les autres $\frac{a_{1j}}{a_{11}}$, $j = \overline{3, n}$. Et donc l'erreur $\triangle x_1$ sera importante.

Stratégie du pivot partiel

Supposons que l'on soit à la k^{eme} étape de la méthode Gauss :

$$\left[A^{(k)}\vdots b^{(k)}\right] \longrightarrow \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & & a_{1k}^{(k)} & \cdots & a_{1n}^{(k)} & b_1^{(k)} \\ 0 & a_{22}^{(k)} & \cdots & & a_{2k}^{(k)} & \cdots & a_{2n}^{(k)} & b_2^{(k)} \\ \vdots & & & & & & & \vdots \\ 0 & & \cdots & 0 & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} & b_k^{(k)} \\ \vdots & \vdots & & & & & & \vdots \\ 0 & & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} & b_n^{(k)} \end{pmatrix} \quad 1 \leq k \leq n$$

Parmi les coefficients $a_{kk}^{(k)}, a_{k+1,k}^{(k)}, \ldots, a_{nk}^{(k)}$, on choisit celui dont le module est le plus grand $\left(ie : \max_{i=\overline{k,n}} \left(\left|a_{ik}^{(k)}\right|\right)\right)$. Le pivot sera ce coefficient, et on permute alors la k^{eme}ligne et la ligne du pivot ainsi obtenu.

1.2.4 Stratégie du pivot total

A la k^{eme} étape le pivot est choisit parmi les coefficients $a_{ij}^{(k)}, i = \overline{k,n} \quad j = \overline{k,n}$,tel que son module soit le plus grand $\left(ie : \max_{i=\overline{k,n}} \max_{j=\overline{k,n}} \left(\left|a_{ij}^{(k)}\right|\right)\right)$.

Attention ! : Achaque permutation de colonnes les inconnues changent de places .

Exemples 7 Soit le systéme

$$\begin{cases} x_1 + 3x_2 + 3x_3 = -2 \\ 2\ x_1 + 2x_2 + 5x_3 = 7 \\ 3x_1 + 2x_2 + 6x_3 = 12 \end{cases}$$

Posons :

$$\left[A \vdots b\right] = \begin{pmatrix} 1 & 3 & 3 & -2 \\ 2 & 2 & 5 & 7 \\ 3 & 2 & 6 & 12 \end{pmatrix}$$

$\boxed{k=1}$: $\max\left(\left|a_{ij}^{(1)}\right|, i = 1,2,3\ j = 1,2,3\right) = 6$. La ligne du pivot total sera alors L_3. En permutant les lignes 1 et 3 on obtient :

$$\begin{pmatrix} 1 & 3 & 3 & -2 \\ 2 & 2 & 5 & 7 \\ 3 & 2 & 6 & 12 \end{pmatrix}$$

La colonne du pivot total est Col_3, et on permute les colonnes 1 et 3 :

$$\begin{pmatrix} \boxed{6} & 2 & 3 & 12 \\ 5 & 2 & 2 & 7 \\ 3 & 3 & 1 & -2 \end{pmatrix} \longrightarrow \begin{pmatrix} 6 & 2 & 3 & 12 \\ 0 & 1/3 & -1/2 & -3 \\ 0 & 2 & -1/2 & -8 \end{pmatrix}$$

$\boxed{k=2}$: $\max\left(\left|a_{ij}^{(2)}\right|, i=2,3\ j=2,3\right)=2$ La ligne du pivot total est alor L_3. Et donc on permute les lignes 2 et 3 :

$$\begin{pmatrix} 6 & 2 & 3 & 12 \\ 0 & \boxed{2} & -1/2 & -8 \\ 0 & 1/3 & -1/2 & -2 \end{pmatrix} \longrightarrow \begin{pmatrix} 6 & 2 & 3 & 12 \\ 0 & 2 & -1/2 & -8 \\ 0 & 0 & -5/12 & -5/3 \end{pmatrix}$$

Delà, du fait de la permutation des colonnes 1 et 3. On obtient le systéme équivalent suivant :

$$\begin{cases} 6x_3+2x_2+3x_1=1 \\ 2x_2-\frac{1}{2}x_1=-8 \\ -\frac{5}{12}x_1=-\frac{5}{3} \end{cases} \iff \begin{cases} x_1=1 \\ x_2=-3 \\ x_3=4 \end{cases}$$

1.2.5 La méthode $L.U$

Soit le systéme linéaire $Ax=b$ (1)

Principe : Décomposition de la matrice A de façon à la mettre sous la forme : $A=L.U$ où L est une matrice triangulaire unitaire inférieure et U une matrice triangulaire supérieure.

Résolution : Le systéme (1) devient :

$$Ax=b \iff L\underbrace{Ux}_{y}=b \iff \begin{cases} Ly=b \\ Ux=y \end{cases}$$

Donc la résolution du systéme $Ax=b$ revient à la résolutions des deux systémes $Ly=b$ et $Ux=y$, et la résolution de ces derniers est immédiate, puisque les matrices L et U sont triangulaires.

Méthode :

Par la méthode de Gauss on obtient

$$\left[A^{(k)}:b^{(k)}\right] = \left[A^{(k)}:b^{(k)}\right] = \begin{pmatrix} a_{11}^{(n)} & a_{12}^{(n)} & a_{13}^{(n)} & \cdots & a_{1n}^{(n)} & b_1^{(n)} \\ 0 & a_{22}^{(n)} & a_{23}^{(n)} & \cdots & a_{1n}^{(n)} & b_2^{(n)} \\ 0 & 0 & a_{33}^{(n)} & \cdots & a_{1n}^{(n)} & b_3^{(n)} \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & a_{nn}^{(n)} & b_n^{(n)} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{1n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{1n}^{(3)} & b_3^{(3)} \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & a_{nn}^{(n)} & b_n^{(n)} \end{pmatrix}$$

avec : $Ax=b \iff A'x=b'$

On pose alors $U=A'$

$$ie: \quad U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1n}^{(1)} & b_1^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{1n}^{(2)} & b_2^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & a_{1n}^{(3)} & b_3^{(3)} \\ \vdots & & & \ddots & & \vdots \\ 0 & \cdots & \cdots & \cdots & a_{nn}^{(n)} & b_n^{(n)} \end{pmatrix}$$

et on montre que : $A = L.U$ où

$$L = \begin{pmatrix} 1 & 0 & & \cdots & & 0 \\ l_{21} & 1 & & \cdots & & 0 \\ l_{31} & l_{32} & 1 & \cdots & & 0 \\ \vdots & & & \ddots & & \vdots \\ l_{n1} & l_{n2} & \cdots & \cdots & l_{n,n-1} & 1 \end{pmatrix} \qquad \text{avec } l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(n)}}$$

Vérification : Pour $n = 4$
On a :

$$U = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & a_{14}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} \end{pmatrix} \quad ; L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{pmatrix} \quad l_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(n)}}$$

d'où

$$l_{21} = \frac{a_{21}^{(1)}}{a_{11}^{(1)}} = \frac{a_{21}}{a_{11}} \qquad l_{31} = \frac{a_{31}^{(1)}}{a_{11}^{(1)}} = \frac{a_{31}}{a_{11}} \qquad l_{41} = \frac{a_{41}^{(1)}}{a_{11}^{(1)}} = \frac{a_{41}}{a_{11}}$$

$$l_{32} = \frac{a_{32}^{(2)}}{a_{22}^{(2)}} \qquad l_{42} = \frac{a_{42}^{(2)}}{a_{22}^{(2)}} \qquad l_{43} = \frac{a_{43}^{(3)}}{a_{33}^{(3)}}$$

Ainsi

$$L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(2)}}{a_{22}^{(2)}} & 1 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(2)}}{a_{22}^{(2)}} & \frac{a_{43}^{(3)}}{a_{33}^{(3)}} & 1 \end{pmatrix}$$

$$L.U = \begin{pmatrix} 1 & 0 & 0 & 0 \\ \frac{a_{21}}{a_{11}} & 1 & 0 & 0 \\ \frac{a_{31}}{a_{11}} & \frac{a_{32}^{(2)}}{a_{22}^{(2)}} & 1 & 0 \\ \frac{a_{41}}{a_{11}} & \frac{a_{42}^{(2)}}{a_{22}^{(2)}} & \frac{a_{43}^{(3)}}{a_{33}^{(3)}} & 1 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & a_{24}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & a_{34}^{(3)} \\ 0 & 0 & 0 & a_{44}^{(4)} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{21}\frac{a_{21}}{a_{11}} + a_{22}^{(2)} & a_{21}\frac{a_{13}}{a_{11}} + a_{23}^{(2)} & a_{21}\frac{a_{14}}{a_{11}} + a_{24}^{(2)} \\ a_{31} & a_{31}\frac{a_{21}}{a_{11}} + a_{32}^{(2)} & a_{31}\frac{a_{13}}{a_{11}} + a_{32}^{(2)}\frac{a_{23}^{(2)}}{a_{22}^{(2)}} + a_{33}^{(3)} & a_{31}\frac{a_{14}}{a_{11}} + a_{32}^{(2)}\frac{a_{24}^{(2)}}{a_{22}^{(2)}} + a_{34}^{(3)} \\ a_{41} & a_{41}\frac{a_{12}}{a_{11}} + a_{42}^{(2)} & a_{41}\frac{a_{13}}{a_{11}} + a_{42}^{(2)}\frac{a_{23}^{(2)}}{a_{22}^{(2)}} + a_{43}^{(3)} & a_{41}\frac{a_{14}}{a_{11}} + a_{42}^{(2)}\frac{a_{24}^{(2)}}{a_{22}^{(2)}} + a_{43}^{(3)}\frac{a_{34}^{(3)}}{a_{33}^{(3)}} + a_{44}^{(4)} \end{pmatrix}$$

$$= \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} = A \text{ car } a_{ij}^{(k+1)} = a_{ij}^{(k)} + a_{ik}^{(k)} \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}}$$

1.2.6 Méthode de Cholesky

Soit A une matrice carrée d'ordre n; $A = (a_{ij})$ $i, j = \overline{1, n}$

Définition 1.2.1 •*Soit $A \in M_n(\mathbb{R})$ est dite symétrique si elle coincide avec sa transposée ie : $A = A^T$ ou encore : $a_{ij} = a_{ji}$, $i, j = \overline{1, n}$*

Théorème 1.2.1 (Sylvester) *Soit $A \in M_n(\mathbb{R})$ une matrice symétrique. Alors A est définie positive si et seulement si*

$$\forall x \in \mathbb{R}^n, x \neq 0, \langle Ax, x \rangle = x^T.Ax \succ 0.$$

Théorème 1.2.2 *A est définie positive si et seulement si tous ses mineurs :*

$$\triangle_1 = a_{11}; \triangle_2 = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}; \quad \triangle_2 = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}; \ldots \quad ; \triangle_n = \det(A)$$

sont strictement positifs.

Théorème 1.2.3 (Cholesky) *Soit A une matrice carrée d'ordre n, non singulière et symétrique. Pour qu'il existe une matrice triangulaire inférieure L, de meme dimension que A, telle que : $A = LL^T$, il faut et il suffit que A soit définie positive .*

Remarque 1.2.4 • *L n'est pas unique*

• *La décomposition devient unique si l'on fixe à l'avance les élèments diagonaux l_{ii} avec $l_{ii} \succ 0$.*

Algorithme de décomposition

Afin d'obtenir les élèments l_{ij} de la matrice L on multiplie les matrices L et L^T, puis on identifie les coefficients respectifs dans l'égalité : $A = L\,L^T$ pour obtenir les équations

$$\begin{aligned} l_{11}^2 &= a_{11} \\ l_{i1}^2 + l_{i2}^2 + l_{i3}^2 + \cdots + l_{i,i-1}^2 + l_{ii}^2 &= a_{ii} \\ l_{i1}l_{ji} + l_{i2}l_{j2} + l_{i3}l_{j3} + + l_{ij}l_{jj} &= a_{ij} \ , \qquad i \succ j \\ l_{ij} &= 0 \qquad i \prec j \end{aligned}$$

D'où, on a successivement (en choissant systématiquement le signe +) :

$$\begin{cases} l_{11} = \sqrt{a_{11}} \\ l_{ij} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}\ , \quad i = \overline{2,n} \\ l_{ij} = \frac{1}{l_{jj}} \left(a_{ij} - \sum_{k=1}^{j-1} l_{ik}\ l_{jk} \right),\ i \succ j \\ l_{ij} = 0\ i, \prec j \end{cases}$$

Résolution du $Ax = b$

Résoudre le systéme $Ax = b$ revient alors à résoudre :

$$L \underbrace{L^T x}_{y} = b \quad \Longleftrightarrow \quad \begin{cases} Ly = b \\ L^T x = y \end{cases}$$

d'où l'on déduit :

$$\begin{cases} y_1 = \frac{b_1}{l_{11}} \\ y_i = \frac{1}{l_{ii}} \left(b_i - \sum_{k=1}^{i-1} l_{ik}\ y_k \right),\ i = \overline{2,n} \end{cases}$$

et

$$\begin{cases} x_n = \frac{y_n}{l_{nn}} \\ x_i = \frac{1}{l_{ii}} \left(y_i - \sum_{k=i+1}^{n} l_{ki}\ x_k \right),\ i = \overline{1,n-1} \end{cases}$$

Remarque 1.2.5 • *La méthode de Cholesky nécessite* $\frac{n^3}{3}$ *opirations élèmentaires (meilleur que celle de Gauss).*

• *La méthode de Cholesky permet le calcul du déterminant de A.*

En effet : $A = L\ L^T \implies \det(A) = \det\left(LL^T\right) = \det(L) \det\left(L^T\right)$

$$= (\det(L))^2 = \left(\prod_{i=1}^{n} l_{ii} \right)^2$$

1.3 Conditionnement

Dans ce paragraphe, nous allons définir la notion de conditionnement d'une matrice, qui peut servir à établir une majoration des erreurs d'arrondi dues aux erreurs sur les données. Malheureusement, nous verrons également que cette majoration n'est pas forcément très utile dans des cas pratiques, et nous nous efforcerons d'y remédier. la notion de conditionnement est également utilisée dans l'étude des méthodes itératives que nous verrons plus loin.

1.3.1 Le problème des erreurs d'arrondis

Soient $A \in M_n(\mathbb{R})$ inversible et $b \in \mathbb{R}^n$; supposons que les données A et b ne soient connues qu'à une erreur près. Ceci est souvent le cas dans les applications pratiques. Considérons par exemple le problème de la conduction thermique dans une tige métallique de longueur 1, modélisée par l'intervalle $[0,1]$. Supposons que la température u de la tige soit imposée aux extrémités, $u(0) = u_0$ et $u(1) = u_1$. On suppose que la température dans la tige satisfait à l'équation de conduction de la chaleur, qui s'écrit $(k(x)u'(x))' = 0$, où k est la conductivité thermique. Cette équation différentielle du second ordre peut se discrétiser par exemple par différences finies, et donne lieu à un système linéaire de matrice A. Si la conductivité k n'est connue qu'avec une certaine précision, alors la matrice A sera également connue à une erreur près, notée (A. On aimerait que l'erreur commise sur les données du modèle (ici la conductivité thermique k) n'ait pas une conséquence catastrophique sur le calcul de la solution du modèle (ici la température u). Si par exemple 1% d'erreur sur k entraîne 100% d'erreur sur u, le modèle ne sera pas d'une utilité redoutable. . .

L'objectif est donc d'estimer les erreurs commises sur x solution de (P) à partir des erreurs commises sur b et A. Notons ($\delta_b \in \mathbb{R}^n$ l'erreur commise sur b et δ_A l'erreur commise sur A. On cherche alors à évaluer δ_x est solution (si elle existe) du système :

$$\begin{cases} x + \delta_x \in \mathbb{R}^n \\ (A + \delta_A)(x + \delta_x) = b + \delta_b \end{cases} \tag{1.1}$$

On va montrer que si δ_A "n'est pas trop grand", alors la matrice $A + \delta_A$ est inversible, et qu'on peut estimer δ_x enfonction de δ_A et δ_A.

1.3.2 Conditionnement et majoration de l'erreur d'arrondi

Définition 1.3.1 (Conditionnement) *Soit $\mathbb{R}^n$ muni d'une norme $\|.\|$ et $M_n(\mathbb{R})$ muni de la norme induite. Soit $A \in M_n(\mathbb{R})$ une matrice inversible. On appelle conditionnement de A par rapport à la norme $\|.\|$ le nombre réel positif $cond(A)$ défini par :*

$$cond(A) = \|A\| \left\|A^{-1}\right\|$$

Proposition 4 (Propriétés générales du conditionnement) *Soit $\mathbb{R}^n$ muni d'une norme $\|.\|$ et $M_n(\mathbb{R})$ muni de la norme induite.*

1. Soit $A \in M_n(\mathbb{R})$ une matrice inversible, alors $cond(A) \geq 1$.

2. Soit $A \in M_n(\mathbb{R})$ et $\alpha \in \mathbb{R}^*$, alors $cond(\alpha A) = cond(A)$.

3. Soient A et $B \in M_n(\mathbb{R})$ des matrices inversibles, alors $cond(AB) = cond(A)cond(B)$.

Proposition 5 (Propriétés du conditionnement pour la norme 2) *Soit $\mathbb{R}^n$ muni d'une norme $\|.\|_2$ et $M_n(\mathbb{R})$ muni de la norme induite. Soit $A \in M_n(\mathbb{R})$ une matrice inversible. On note $cond_2(A)$ le conditionnement associé à la norme induite par la norme euclidienne sur $\mathbb{R}^n$.*

1. Soit $A \in M_n(\mathbb{R})$ une matrice inversible. On note σ_n[resp. σ_1] la plus grande [resp. petite] valeur propre de A^tA (noter que A^tA est une matrice symétrique définie positive). Alors $cond_2(A) = \sqrt{\sigma_n/\sigma_1}$.

2. Si de plus A une matrice symétrique définie positive, alors $cond_2(A) = \frac{\lambda_n}{\lambda_1}$, où λ_n [resp. λ_1] est la plus grande [resp. petite] valeur propre de A.

3. Si A et B sont deuxmatrices symétriques définies positives, alors $cond_2(A+B) \leq \max(cond_2(A),$

4. Soit $A \in M_n(\mathbb{R})$ une matrice inversible. Alors $cond_2(A) = 1$ si et seulement si $A = \alpha Q$ où $\alpha \in \mathbb{R}^*$ et Q est une matrice orthogonale (c'est-à-dire $Q^t = Q^{-1}$).

5. Soit $A \in M_n(\mathbb{R})$ une matrice inversible. On suppose que $A = QR$ où Q est une matrice orthogonale. Alors $cond_2(A) = cond_2(R)$.

6. Soient $A, B \in M_n(\mathbb{R})$ deux matrices symétriques définies positives. Montrer que $cond_2(A+B) \leq max\{cond_2(A), cond_2(B)\}$.

Théorème 1.3.1 *Soit $A \in M_n(\mathbb{R})$ une matrice inversible, et $b \in \mathbb{R}^n$, $b \neq 0$. On munit $\mathbb{R}^n$ d'une norme $\|.\|$, et $M_n(\mathbb{R})$ de la norme induite. Soient $\delta_A \in M_n(\mathbb{R})$ et $\delta_b \in \mathbb{R}^n$. On suppose que $\|\delta_A\| < \frac{1}{\|A^{-1}\|}$. Alors la matrice $(A + \delta_A)$ est inversible et si x est solution de (P) et $x + \delta_x$ est solution de (1.1), alors*

$$\frac{\|\delta_x\|}{\|x\|} \leq \frac{cond(A)}{1 - \|A^{-1}\| \, \|\delta_A\|} \left(\frac{\delta_b}{b} + \frac{\delta_A}{A} \right) \tag{1.2}$$

1.4 Méthodes itératives

Les méthodes directes que nous avons étudiées dans le paragraphe précédent sont très efficaces : elles donnent la solution exacte (aux erreurs d'arrondi près) du système linéaire considéré. Elles ont l'inconvénient de nécessiter une assez grande place mémoire car elles nécessitent le stockage de toute la matrice en mémoire vive. Si la matrice est pleine, c.à.d. si la plupart des coefficients de la matrice sont non nuls et qu'elle est trop grosse pour la mémoire vive de l'ordinateur dont on dispose, il ne reste plus qu'à gérer habilement le "swapping" c'est–à–dire l'échange de données entre mémoire disque et mémoire vive pour pouvoir résoudre le système.

Cependant, si le système a été obtenu à partir de la discrétisation d'équations aux dérivés partielles, il est en général "creux", c.à.d. qu'un grand nombre des coefficients de la matrice du système sont nuls ; de plus la matrice a souvent une structure "bande", i.e. les éléments non nuls de la matrice sont localisés sur certaines diagonales.

1.4.1 Définition et propriétés

Soit $A \in M_n(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^n$, on cherche toujours ici à résoudre le système linéaire (P) c'est–à–dire à trouver $x \in \mathbb{R}^n$ tel que $Ax = b$.

Définition 1.4.1 *On appelle méthode itérative de résolution du système linéaire (P) une méthode qui construit une suite $(x^{(k)})_{k\in\mathbb{N}}$ (où "l'itéré" $x^{(k)}$ est calculé à partir des itérés $x^{(0)}...x^{(k-1)}$) censée converger vers x solution de (P)).*

Définition 1.4.2 *On dit qu'une méthode itérative est convergente si pour tout choix initial $x^{(0)} \in \mathbb{R}^n$, on a :*

$$x^{(k)} \to x \text{ quand } n \to +\infty$$

Puisqu'il s'agit de résoudre un sytème linéaire, il est naturel d'essayer de construire la suite des itérés sous la forme $x^{(k+1)} = Bx^{(k)} + c$, *où* $B \in M_n(\mathbb{R})$ *et* $c \in \mathbb{R}^n$ *seront choisis de manière à ce que la méthode itérative ainsi définie soit convergente.On appellera ce type de méthode* ***Méthode I****, et on verra par la suite un choix plus restrictif qu'on appellera* ***Méthode II****.*

Définition 1.4.3 (Méthode I) *On appelle méthode itérative de type* ***I*** *pour la résolution du système linéaire (P) une méthode itérative où la suite des itérés* $(x^{(k)})_{k\in\mathbb{N}}$ *est donnée par :*

$$\begin{cases} \text{Initialisation} \quad x^{(0)} \in \mathbb{R}^n \\ \text{Itération } n \quad x^{(k+1)} = Bx^{(k)} + c \end{cases}$$

où $B \in M_n(\mathbb{R})$ *et* $c \in \mathbb{R}^n$.

Remarque 1.4.1 (Condition nécessaire de convergence) *Une condition nécessaire pour que la méthode* ***I*** *converge est que* $c = (Id - B)A^{-1}b$. *En effet, supposons que la méthode converge. En passant à la limite lorsque n tend vers l'infini sur l'itération n de l'algorithme, on obtient* $x = Bx + c$ *et comme* $x = A^{-1}b$, *ceci entraîne* $c = (Id - B)A^{-1}b$.

Remarque 1.4.2 (Intérêt pratique) *La "****méthode I****" est assez peu intéressante en pratique, car il faut calculer* $A^{-1}b$, *sauf si* $(Id - B)A^{-1} = \alpha Id$, *avec* $\alpha \in \mathbb{R}$. *On obtient dans ce cas :*

$$B = -\alpha A + Id$$

$$et \qquad c = \alpha b$$

c'est–à–dire

$$x^{n+1} = x^n + \alpha(b - Ax^n).$$

Le terme $b - Ax^n$ *est appelé résidu et la méthode s'appelle dans ce cas la méthode d'extrapolation de Richardson.*

Théorème 1.4.1 (Convergence de la méthode de type I) *Soit* $A \in M_n(\mathbb{R})$ *inversible,* $b \in \mathbb{R}^n$. *On considère la* ***méthode I*** *avec* $B \in M_n(\mathbb{R})$ *et*

$$c = (Id - B)A^{-1}b. \tag{1.3}$$

Alors la ***méthode I*** *converge si et seulement si le rayon spectral* $\rho(B)$ *de la matrice B vérifie* $\rho(B) < 1$.

Définition 1.4.4 (Méthode II) *Soit* $A \in M_n(\mathbb{R})$ *une matrice inversible,* $b \in \mathbb{R}^n$. *Soient* $\tilde{M}$ *et* $\tilde{N} \in M_n(\mathbb{R})$ *des matrices telles que* $A = \tilde{M} - \tilde{N}$ *et* $\tilde{M}$ *est inversible (et facile à inverser).*

On appelle méthode de type ***II*** *pour la résolution du système linéaire (P) une méthode itérative où la suite des itérés* $(x^{(k)})_{k\in\mathbb{N}}$ *est donnée par :*

$$\begin{cases} \text{Initialisation} \quad x^{(0)} \in \mathbb{R}^n \\ \text{Itération } n \quad \tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b \end{cases} \tag{1.4}$$

Remarque 1.4.3 *Si $\tilde{M}x^{(k+1)} = \tilde{N}x^{(k)} + b$ pour tout $k \in \mathbb{N}$ et $x(k) \to y$ quand $n \to +\infty$ alors $\tilde{M}y = \tilde{N}y + b$, c.à.d. $(\tilde{M} - \tilde{N})y = b$ et donc $Ay = b$. En conclusion, si la méthode de type **II** converge, alors elle converge bien vers la solution du système linéaire.*

Théorème 1.4.2 (Convergence de la méthode II) *Soit $A \in M_n(\mathbb{R})$ une matrice inversible, $b \in \mathbb{R}^n$. Soient $\tilde{M}$ et $\tilde{N} \in M_n(\mathbb{R})$ des matrices telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible. Alors :*

1. La méthode définie par (1.2) converge si et seulement si $\rho\left(\tilde{M}^{-1}\tilde{N}\right) < 1$.

2. La méthode itérative définie par (1.2) converge si et seulement si il existe une norme induite notée $\|.\|$ telle que $\left\|\tilde{M}^{-1}\tilde{N}\right\| < 1$.

Théorème 1.4.3 (Condition suffisante de convergence, méthode II) *Soit $A \in M_n(\mathbb{R})$ une matrice symétrique définie positive, et soient $\tilde{M}$ et $\tilde{N} \in M_n(\mathbb{R})$ telles que $A = \tilde{M} - \tilde{N}$ et $\tilde{M}$ est inversible. Si la matrice $\tilde{M}^t + \tilde{N}$ est symétrique définie positive alors $\rho\left(\tilde{M}^{-1}\tilde{N}\right) < 1$, et donc la méthode II converge*

1.4.2 Méthodes de Jacobi, Gauss-Seidel et SOR/SSOR

Décomposition par blocs de A :

Dans de nombreux cas pratiques, les matrices des systèmes linéaires à résoudre ont une structure “par blocs", et on se sert de cette structure lors de la résolution par une méthode itérative.

Définition 1.4.5 *Soit $A \in M_n(\mathbb{R})$ une matrice inversible. Une décomposition par blocs de A est définie par un entier $S \leq N$, des entiers $(n_i)_{i=1,...,S}$ tels que $\sum_{i=1}^{S} n_i = N$, et S^2 matrices $A_{i,j} \in M_{n_i,n_j}(\mathbb{R})$ (ensemble des matrices rectangulaires à n_i lignes et n_j colonnes, telles que les matrices $A_{i,j}$ soient inversibles pour $i = 1, ..., S$ et*

$$A = \begin{bmatrix} A_{1,1} & A_{1,2} & \cdots & & \cdots & A_{1,S} \\ A_{2,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & A_{S-1,S} \\ A_{S,1} & \cdots & & \cdots & A_{S,S-1} & A_{S,S} \end{bmatrix} \tag{1.5}$$

Remarque 1.4.4 *1. Si $S = N$ et $n_i = 1$ $\forall i \in \{1, ..., S\}$, chaque bloc est constitué d'un seul coefficient.*

2. Si A est symétrique définie positive, la condition $A_{i,j}$ inversible dans la définition 1.3.5 est inutile car $A_{i,j}$ est nécessairement symétrique définie positive donc inversible. Prenons par exemple $i = 1$; soit $y \in \mathbb{R}^{n_1}$, $y \neq 0$ et $x = (y, 0..., 0)^t \in \mathbb{R}^n$. Alors $A_{1,1}y \cdot y = Ax \cdot x > 0$ donc $A_{1,1}$ est symétrique définie positive.

3. Si A est une matrice triangulaire par blocs, c.à.d. de la forme (1.3) avec $A_{i,j} = 0$ si $j > i$, alors

$$\det(A) = \prod_{i=1}^{S} \det(A_{i,i}).$$

Par contre si A est décomposée en 2×2 blocs carrés (i.e. tels que $n_i = m_j, \forall (i,j) \in \{1,2\}$), on a en général : $\det(A) \neq \det(A_{1,1}) \det(A_{2,2}) - \det(A_{1,2}) \det(A_{2,1})$.

Méthode de Jacobi

On peut remarquer que le choix le plus simple pour le système $\tilde{M}x = d$ soit facile à résoudre (on rappelle que c'est un objectif de la mise sous forme méthode de type **II**) est de prendre pour $\tilde{M}$ une matrice diagonale. La méthode de Jacobi consiste à prendre pour $\tilde{M}$ la matrice diagonale D formée par les blocs diagonaux de A :

$$D = \begin{bmatrix} A_{1,1} & 0 & \cdots & & \cdots & 0 \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ 0 & \cdots & & \cdots & 0 & A_{S,S} \end{bmatrix}$$

Dans la matrice ci-dessus, 0 désigne un bloc nul.

On a alors $\tilde{N} = E + F$, où E et F sont constitués des blocs triangulaires inférieurs et supérieurs de la matrice A :

$$E = \begin{bmatrix} 0 & 0 & \cdots & & \cdots & 0 \\ -A_{1,1} & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & 0 \\ -A_{S,1} & \cdots & & \cdots & -A_{S,S-1} & 0 \end{bmatrix}$$

et

$$F = \begin{bmatrix} 0 & -A_{1,2} & \cdots & & \cdots & -A_{1,S} \\ 0 & \ddots & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & \ddots & -A_{S-1,S} \\ 0 & \cdots & & \cdots & 0 & 0 \end{bmatrix}.$$

On a bien $A = \tilde{M} - \tilde{N}$ et avec D, E et F définies comme ci-dessus, la méthode de Jacobi s'écrit :

$$\begin{cases} x^{(0)} \in \mathbb{R}^n \\ Dx^{(k+1)} = (E+F)\, x^{(k)} + b \end{cases} \tag{1.6}$$

Lorsqu'on écrit la méthode de Jacobi comme une méthode **I**, on a $B = D^{-1}(E+F)$; on notera J cette matrice.

En introduisant la décomposition par blocs de x, solution recherchée de (P), c.à.d. : $x = [x_1, ..., x_S]$, où $x_i \in \mathbb{R}^{n_i}$, on peut aussi écrire la méthode de Jacobi sous la forme :

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k)} - \sum_{j>i} A_{i,j}x_j^{(k)} + b_i \ i = 1, ..., S \end{array} \right. \tag{1.7}$$

Si $S = N$ et $n_i = 1 \ \forall i \in \{1, ..., S\}$, chaque bloc est constitué d'un seul coefficient, et on obtient la méthode de Jacobi par points (aussi appelée méthode de Jacobi), qui s'écrit donc :

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \quad i = 1, ..., n \end{array} \right. \tag{1.8}$$

Méthode de Gauss-Seidel

L'idée de la méthode de Gauss-Seidel est d'utiliser le calcul des composantes de l'itéré $(k+1)$ dès qu'il est effectué. Par exemple, pour calculer la deuxième composante $x_2^{(k+1)}$ du vecteur $x^{(k+1)}$, on pourrait employer la "nouvelle" valeur $x_1^{(k+1)}$ qu'on vient de calculer plutôt que la valeur $x_1^{(k)}$ comme dans (1.5) ; de même, dans le calcul de $x_3^{(k+1)}$, on pourrait employer les "nouvelles" valeurs $x_1^{(k+1)}$ et $x_2^{(k+1)}$ plutôt que les valeurs $x_1^{(k)}$ et $x_2^{(k)}$.

Cette idée nous suggère de remplacer dans (1.5) $x_j^{(k)}$ par $x_j^{(k+1)}$ si $j < i$. On obtient donc l'algorithme suivant :

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \ i = 1, ..., S \end{array} \right. \tag{1.9}$$

Notons que l'algorithme de Gauss–Seidel par points (cas ou $S = N$ et $n_i = 1$) s'écrit donc :

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ a_{i,i}x_i^{(k+1)} = -\sum_{j<i} a_{i,j}x_j^{(k)} - \sum_{j>i} a_{i,j}x_j^{(k)} + b_i \quad i = 1, ..., n \end{array} \right. \tag{1.10}$$

La méthode de Gauss–Seidel s'écrit donc sous forme de méthode **II** avec $M = D - E$ et $N = F$:

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ (D-E)\, x^{(k+1)} = Fx^{(k)} + b \end{array} \right. \tag{1.11}$$

Lorsqu'on écrit la méthode de Gauss–Seidel comme une méthode **I**, on a $B = (D-E)^{-1}F$; on notera $\mathcal{L}_1$ cette matrice, dite matrice de Gauss-Seidel.

Méthodes SOR et SSOR

L'idée de la méthode de sur-relaxation (**SOR** = Successive Over Relaxation) est d'utiliser la méthode de Gauss-Seidel pour calculer un itéré intermédiaire $\tilde{x}^{(k+1)}$ qu'on "relaxe" ensuite

pour améliorer la vitesse de convergence de la méthode. On se donne $0 < \omega < 2$, et on modifie l'algorithme de Gauss–Seidel de la manière suivante :

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}\tilde{x}_i^{(k+1)} = -\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i \\ x_i^{(k+1)} = \omega\tilde{x}_i^{(k+1)} + (1-\omega)\, x_i^{(k)} \quad i = 1, ..., S. \end{array} \right. \tag{1.12}$$

(Pour $\omega = 1$ on retrouve la méthode de Gauss–Seidel.)

L'algorithme ci-dessus peut aussi s'écrire (en multipliant par $A_{i,i}$ la ligne 3 de l'algorithme (1.10)) :

$$\left\{ \begin{array}{c} x^{(0)} \in \mathbb{R}^n \\ A_{i,i}x_i^{(k+1)} = \omega\left[-\sum_{j<i} A_{i,j}x_j^{(k+1)} - \sum_{i<j} A_{i,j}x_j^{(k)} + b_i\right] + (1-\omega)\, A_{i,i}x_i^{(k)} \end{array} \right. \tag{1.13}$$

On obtient donc

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

L'algorithme **SOR** s'écrit donc comme une méthode **II** avec

$$\tilde{M} = \frac{D}{\omega} - E \text{ et } \tilde{N} = F + \left(\frac{1-\omega}{\omega}\right) D.$$

Il est facile de vérifier que $A = \tilde{M} - \tilde{N}$.

L'algorithme **SOR** s'écrit aussi comme une méthode **I** avec

$$B = \left(\frac{D}{\omega} - E\right)^{-1} (F + \left(\frac{1-\omega}{\omega}\right) D).$$

On notera $\mathcal{L}_\omega$ cette matrice.

En "symétrisant" le procédé de la méthode **SOR**, c.à.d. en effectuant les calculs **SOR** sur les blocs dans l'ordre 1 à N puis dans l'ordre N à 1, on obtient la méthode de sur-relaxation symétrisée (**SSOR** = Symmetric Successive Over Relaxation) qui s'écrit dans le formalisme de la méthode **I** avec

$$B = \underbrace{\left(\frac{D}{\omega} - F\right)^{-1} (E + \frac{1-\omega}{\omega} D)}_{\text{calcul dans l'ordre } S...1} \underbrace{\left(\frac{D}{\omega} - E\right)^{-1} (F + \frac{1-\omega}{\omega} D)}_{\text{calcul dans l'ordre } 1...S}.$$

Etude théorique de convergence

On aimerait pouvoir répondre aux questions suivantes :

1. Les méthodes sont–elles convergentes ?
2. Peut-on estimer leur vitesse de convergence ?

3. Peut-on estimer le coefficient de relaxation ω optimal dans la méthode **SOR**, c.à.d. celui qui donnera la plus grande vitesse de convergence ?

On va maintenant donner des réponses, partielles dans certains cas, faute de mieux, à ces questions.

Convergence On rappelle qu'une méthode itérative de type **I**, i.e. écrite sous la forme $x^{(n+1)} = Bx^{(n)} + C$ converge si et seulement si $\rho(B)\backslash 1$.

Théorème 1.4.4 (Sur la convergence de la méthode SOR) *Soit $A \in M_n(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.3 ; soient D la matrice constituée par les blocs diagonaux, $-E$ (resp. $-F$) la matrice constituée par les blocs triangulaires inférieurs (resp. supérieurs) ; on a donc : $A = D - E - F$. Soit $\mathcal{L}_\omega$ la matrice d'itération de la méthode **SOR** (et de la méthode de Gauss–Seidel pour $\omega = 1$) définie par :*

$$\mathcal{L}_\omega = \left(\frac{D}{\omega} - E\right)^{-1} (F + \frac{1-\omega}{\omega} D), \omega \neq 0.$$

Alors :

1. *Si $\rho(\mathcal{L}_\omega) < 1$ alors $0 < \omega < 2$.*
2. *Si on suppose de plus que A symétrique définie positive, alors :*

$$\rho(\mathcal{L}_\omega) < 1 \text{ si et seulement si } 0 < \omega < 2.$$

En particulier , si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge.

Remarque 1.4.5 *On a vu (théorème 1.35) que si A est une matrice symétrique définie positive, la méthode de Gauss–Seidel converge. Par contre, même dans le cas où A est symétrique définie positive, il existe des cas où la méthode de Jacobi ne converge pas.*

Remarquons que le résultat de convergence des méthodes itératives donné par le théorème précédent n'est que partiel, puisqu'il ne concerne que les matrices symétriques définies positives et que les méthodes Gauss-Seidel et **SOR**. On a aussi un résultat de convergence de la méthode de Jacobi pour les matrices à diagonale dominante stricte, et un résultat de comparaison des méthodes pour les matrices tridiagonales par blocs, voir le théorème 1.3.5 donné ci-après. Dans la pratique, il faudra souvent compter sur sa bonne étoile. . .

Estimation du coefficient de relaxation optimal de SOR

La question est ici d'estimer le coefficient de relaxation, optimal dans la méthode **SOR**, c.à.d. le coefficient, $\omega_0 \in]0, 2[$ (condition nécessaire pour que la méthode **SOR** converge, voir théorème 1.3.4) tel que $\rho(\mathcal{L}_{\omega_0}) < \rho(\mathcal{L}_\omega)\ \forall \omega \in]0, 2[$.

D'après le paragraphe précédent ce ω_0 donnera lameilleure convergence possible pour **SOR**. On sait le faire dans le cas assez restrictif des matrices tridiagonales par blocs.

Théorème 1.4.5 (Coefficient optimal, matrice tridiagonale) *On considère une matrice $A \in M_n(\mathbb{R})$ qui admet une décomposition par blocs définie dans la définition 1.3 ; on suppose*

que la matrice A est tridiagonale par blocs, c.à.d. $A_{i,j} = 0$ si $|i-j| > 1$; soient $\mathcal{L}_1$ et J les matrices d'itération respectives des méthodes de Gauss-Seidel et Jacobi, alors :

1. $\rho(\mathcal{L}_1) < (\rho(J))^2$: la méthode de Gauss–Seidel converge (ou diverge) donc plus vite que celle de Jacobi.

2. On suppose de plus que toutes les valeurs propres de la matrice d'itération J de la méthode de Jacobi sont réelles. alors le paramètre de relaxation optimal, c.à.d. le paramètre ω_0 tel que $\rho(\mathcal{L}_{\omega_0}) = \min\{\rho(\mathcal{L}_\omega),\ \omega \in]0,2[\}$, s'exprime en fonction du rayon spectral $\rho(J)$ de la matrice J par la formule :

$$\omega_0 = \frac{2}{1+\sqrt{1-\rho(J)^2}} > 1,$$

et on a : $\rho(\mathcal{L}_{\omega_0}) = \omega_0 - 1$.

Travaux dirigés 1

Exercice 01 :
Soit le système linéaire suivant :

$$(1)\ \begin{cases} 3x_1 - 2x_2 + x_3 &= 2 \\ 2x_1 + x_2 + x_3 &= 7 \\ 4x_1 - 3x_2 + 2x_3 &= 4 \end{cases}$$

a) Résoudre ce système par la méthode de Gauss.
b) Factoriser la matrice A du système en produit LU où L est une matrice triangulaire inférieure (avec des 1 sur la diagonale principale) et U triangulaire supérieure, puis résoudre ce système.

Exercice 02 :

Soit le système linéaire $AX = B$ où : $A = \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 4 & -3 & 0 \end{pmatrix}$, $X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$, et $B = \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix}$. Factoriser la matrice A en produit LU puis résoudre le système.

Exercice 03

Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique dé nie positive.

1. On suppose ici que A est tridiagonale. Estimer le nombre d'opérations de la fact orisation LL^t dans ce cas.
2. Même question si A est une matrice bande (c'est-à-dire p diagonales non nulles).
3. En déduire une estimation du nombre d'opérations nécessa ires pour la discrétisation de l'équation $-u'' =$ f Même question pour la discrétisation de l'équation $-\Delta u = f$.

Exercice 04

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

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

Montrer que A est symétrique définie positive si et seulement si $-1/2 < a < 1$

et que la méthode de Jacobi

converge si et seulement si $-1/2 < a < 1/2$.

Exercice 05

1. Soit $A = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}$.
Calculer la décomposition LDL^t de A. Existe-t-il une décomposition LL^t de A ?

2. Montrer que toute matrice de $\mathcal{M}_N(\mathbb{R})$ symétrique définie positive admet une décomposition LDL^t.

3. Ecrire l'algorithme de décomposition LDL^t. La matrice $A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ admet-elle une décomposition LDL^t ?

Exercice 06

Soit $N \geq 1$. Soit $A = (a_{i,j})_{i,j=1,\ldots,N} \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. On note D la partie diagonale de A, $-E$ la partie triangulaire inférieure de A et $-F$ la partie triangulaire supérieure de A, c'est-à-dire :

$$\begin{array}{l} D = (d_{i,j})_{i,j=1,\ldots,N},\ d_{i,j} = 0 \text{ si } i \neq j,\ d_{i,i} = a_{i,i}, \\ E = (e_{i,j})_{i,j=1,\ldots,N},\ e_{i,j} = 0 \text{ si } i \leq j,\ e_{i,j} = -a_{i,j} \text{ si } i > j, \\ F = (f_{i,j})_{i,j=1,\ldots,N},\ f_{i,j} = 0 \text{ si } i \geq j,\ f_{i,j} = -a_{i,j} \text{ si } i < j. \end{array}$$

Noter que $A = D - E - F$. Soit $b \in \mathbb{R}^N$. On cherche à calculer $x \in \mathbb{R}^N$ t.q. $Ax = b$. On suppose que D est définie positive (noter que A n'est pas forcément inversible). On s'intéresse ici à la méthode de Jacobi (par points),

Initialisation. $x^{(0)} \in \mathbb{R}^N$
Itérations. Pour $n \in \mathbb{N}$, $Dx^{(n+1)} = (E + F)x^{(n)} + b$.

On pose $J = D^{-1}(E + F)$.

1. Montrer, en donnant un exemple avec $N = 2$, que J peut ne pas être symétrique.
2. Montrer que J est diagonalisable dans $\mathbb{R}$ et, plus précisement, qu'il existe une base de $\mathbb{R}^N$, notée $\{f_1, \ldots, f_N\}$, et il existe $\{\mu_1, \ldots, \mu_N\} \subset \mathbb{R}$ t.q. $Jf_i = \mu_i f_i$ pour tout $i \in \{1, \ldots, N\}$ et t.q. $Df_i \cdot f_j = \delta_{i,j}$ pour tout $i, j \in \{1, \ldots, N\}$.

   En ordonnant les valeurs propres de J, on a donc $\mu_1 \leq \ldots \leq \mu_N$, on conserve cette notation dans la suite.
3. Montrer que la trace de J est nulle et en déduire que $\mu_1 \leq 0$ et $\mu_N \geq 0$.

   On suppose maintenant que A et $2D - A$ sont symétriques définies positives et on pose $x = A^{-1}b$.
4. Montrer que la méthode de Jacobi (par points) converge (c'est-à-dire $x^{(n)} \to x$ quand $n \to \infty$). [Utiliser un théorème du cours.]

   On se propose maintenant d'améliorer la convergence de la méthode par une technique de relaxation. Soit $\omega > 0$, on considère la méthode suivante :
   Initialisation. $x^{(0)} \in \mathbb{R}^N$
   Itérations. Pour $n \in \mathbb{N}$, $D\tilde{x}^{(n+1)} = (E + F)x^{(n)} + b$, $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1 - \omega)x^{(n)}$.
5. Calculer les matrices M_ω (inversible) et N_ω telles que $M_\omega x^{(n+1)} = N_\omega x^{(n)} + b$ pour tout $n \in \mathbb{N}$, en fonction de ω, D et A. On note, dans la suite $J_\omega = (M_\omega)^{-1}N_\omega$.

6. On suppose dans cette question que $(2/\omega)D - A$ est symétrique définie positive. Montrer que la méthode converge (c'est-à-dire que $x^{(n)} \to x$ quand $n \to \infty$.)
7. Montrer que $(2/\omega)D - A$ est symétrique définie positive si et seulement si $\omega < 2/(1-\mu_1)$.
8. Calculer les valeurs propres de J_ω en fonction de celles de J. En déduire, en fonction des μ_i, la valeur "optimale" de ω, c'est-à-dire la valeur de ω minimisant le rayon spectral de J_ω.

Exercice 07

Soit $A = (a_{i,j})_{i,j\in\{1,\dots,N\}} \in M_n(\mathbb{R})$.
1. On munit $\mathbb{R}^n$ de la norme $\|\ \|_\infty$ et $M_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|.\|_\infty$.Montrer que $\|A\|_\infty = \max_{i\in\{1,\dots,N\}} \sum_{j=1}^{N} |a_{i,j}|$.
2. On munit $\mathbb{R}^n$ de la norme $\|.\|_1$ et $M_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|.\|_1$. Montrer que $\|A\|_1 = \max_{j\in\{1,\dots,N\}} \sum_{i=1}^{N} |a_{i,j}|$.
3. On munit $\mathbb{R}^n$ de la norme $\|.\|_2$ et $M_n(\mathbb{R})$ de la norme induite correspondante, notée aussi $\|.\|_2$. Montrer que $\|A\|_2 = (\rho(A^tA))^{\frac{1}{2}}$.

Exercice 08

Soient $A \in M_n(\mathbb{R})$ et $\|.\|$ une norme matricielle.
1. Montrer que si $\rho(A) < 1$, les matrices $Id - A$ et $Id + A$ sont inversibles.
2. Montrer que la série de terme général A^k converge (vers $(Id-A)^{-1}$) si et seulement si $\rho(A) < 1$.

Exercice 09

Résoudre le système par la méthode de Gauss

$$(1) \qquad \begin{cases} 2x_1 + x_2 \quad + x_4 = 2 \\ -4x_1 - 2x_2 + 3x_3 - 7x_4 = -9 \\ 4x_1 + x_2 - 2x_3 + 8x_4 = 2 \\ -3x_2 - 12x_3 - x_4 = 2 \end{cases}$$

Suggestions et Corrigés

Exercice 01

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

Ce système s'écrit sous la forme $AX = B$, où

$$A = \begin{pmatrix} 3 & -2 & 1 \\ 2 & 1 & 1 \\ 4 & -3 & 2 \end{pmatrix} \text{ et } B = \begin{pmatrix} 2 \\ 7 \\ 4 \end{pmatrix}$$

Posons $A^{(1)} = A$, on calcule $A^{(2)} = M^{(1)}A^{(1)}$, où

$$M^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ -\frac{2}{3} & 1 & 0 \\ -\frac{4}{3} & 0 & 1 \end{pmatrix},$$

d'où :

$$A^{(2)} = \begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{7}{3} & \frac{1}{3} \\ 0 & -\frac{1}{3} & \frac{2}{3} \end{pmatrix}$$

on calcule $A^{(3)} = M^{(2)}A^{(2)}$, où

$$M^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \frac{1}{7} & 1 \end{pmatrix}.$$

Donc,

$$A^{(3)} = \begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{7}{3} & \frac{1}{3} \\ 0 & 0 & \frac{5}{7} \end{pmatrix}.$$

La matrice $A^{(3)}$ est ainsi triangulaire supérieure, c'est la matrice U recherchée.

D'autre part, on a $A^{(3)} = M^{(2)}A^{(2)} = M^{(2)}M^{(1)}A^{(1)}$, on en déduit donc que

$$A^{(1)} = \underbrace{(M^{(1)})^{-1}(M^{(2)})^{-1}}_{L}\underbrace{A^{(3)}}_{U}.$$

Ainsi, $A = A^{(1)} = LU$, avec

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{4}{3} & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -\frac{1}{7} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{4}{3} & -\frac{1}{7} & 1 \end{pmatrix}.$$

On a ainsi factorisé A sous la forme :

$$A = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{4}{3} & -\frac{1}{7} & 1 \end{pmatrix} \begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{7}{3} & \frac{1}{3} \\ 0 & 0 & \frac{5}{7} \end{pmatrix}.$$

Présantation de la méthode d'identification

Résoudre $AX = B$ revient à résoudre $LUX = B$. On pose alors $Y = UX$, la résolution du système initial revient à résoudre successivement les deux systèmes triangulaires :

$$\begin{cases} LY &= B \\ UX &= Y \end{cases}$$

$$LY = B \iff \begin{pmatrix} 1 & 0 & 0 \\ \frac{2}{3} & 1 & 0 \\ \frac{4}{3} & -\frac{1}{7} & 1 \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 7 \\ 4 \end{pmatrix} \Rightarrow Y = \begin{pmatrix} 2 \\ \frac{17}{3} \\ \frac{15}{7} \end{pmatrix}.$$

Finalement, on résout :

$$UX = Y \iff \begin{pmatrix} 3 & -2 & 1 \\ 0 & \frac{7}{3} & \frac{1}{3} \\ 0 & 0 & \frac{3}{7} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ \frac{17}{3} \\ \frac{15}{7} \end{pmatrix} \Rightarrow X = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}.$$

Exercice 02

Soit le système linéaire $AX = B$ où :

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

Factorisons la matrice A en produit LU.

Posons $A^{(1)} = A$, on calcule $A^{(2)} = M^{(1)}A^{(1)}$, où

$$M^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ -0 & 1 & 0 \\ -4 & 0 & 1 \end{pmatrix},$$

d'où :

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

on calcule $A^{(3)} = M^{(2)}A^{(2)}$, où

$$M^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}.$$

Donc,

$$A^{(3)} = \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 0 & 0 & 18 \end{pmatrix}.$$

La matrice $A^{(3)}$ est ainsi triangulaire supérieure, c'est la matrice U recherchée. D'autre part, on a $A^{(3)} = M^{(2)}A^{(2)} = M^{(2)}M^{(1)}A^{(1)}$. On en déduit donc

$$A^{(1)} = \underbrace{(M^{(1)})^{-1}(M^{(2)})^{-1}}_{L}\underbrace{A^{(3)}}_{U}.$$

Ainsi, $A = A^{(1)} = LU$, avec

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

A se factorise donc sous la forme :

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & -3 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 0 & 0 & 18 \end{pmatrix}.$$

Résolvons le système $AX = B$. Cela revient à résoudre $LUX = B$, c'est à dire à résoudre successivement les systèmes $LY = B$ puis $UX = Y$.

$$LY = B \iff \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 4 & -3 & 1 \end{pmatrix}\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ -1 \end{pmatrix} \Rightarrow Y = \begin{pmatrix} 2 \\ 5 \\ 6 \end{pmatrix}.$$

Finalement, on résout :

$$UX = Y \iff \begin{pmatrix} 1 & 0 & -3 \\ 0 & 1 & 2 \\ 0 & 0 & 18 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 5 \\ 6 \end{pmatrix} \Rightarrow X = \begin{pmatrix} 3 \\ \frac{13}{3} \\ \frac{1}{3} \end{pmatrix}.$$

Exercice 03

On utilise le résultat de conservation du profil de la matrice énoncé dans le cours. Comme A est symétrique, le nombre p de diagonales de la matrice A est forcément impair si A ; notons $q = \frac{p-1}{2}$ le nombre de sous- et sur-diagonales non nulles de la matrice A, alors la matrice L aura également q sous-diagonales non nulles.

1. Cas d'une matrice tridiagonale. Si on reprend l'algorithme de construction de la matrice L vu en cours, on remarque que pour le calcul de la colonne $n+1$, avec $1 \leq n < n-1$, on a le nombre d'opérations suivant :

— Calcul de $\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}\ell_{n+1,k})^{1/2} > 0$:
une multiplication, une soustraction, une extraction de racine, soit 3 opérations élémentaires.

— Calcul de $\ell_{n+2,n+1} = \left(a_{n+2,n+1} - \sum_{k=1}^{n} \ell_{n+2,k}\ell_{n+1,k}\right) \frac{1}{\ell_{n+1,n+1}}$:
une division seulement car $\ell_{n+2,k} = 0$.

On en déduit que le nombre d'opérations élémentaires pour le calcul de la colonne $n+1$, avec $1 \leq n < n-1$, est de 4.
Or le nombre d'opérations pour la première et dernière colonnes est inférieur à 4 (2 opérations pour la première colonne, une seule pour la dernière). Le nombre $Z_1(n)$ d'opérations élémentaires pour la décomposition LL^t de A peut donc être estimé par : $4(n-2) \leq Z_1(n) \leq 4n$, ce qui donne que $Z_1(n)$ est de l'ordre de $4n$ (le calcul exact du nombre d'opérations, inutile ici car on demande une estimation, est $4n-3$.)

2. Cas d'une matrice à p diagonales.
On cherche une estimation du nombre d'opérations $Z_p(n)$ pour une matrice à p diagonales non nulles (ou q sous-diagonales non nulles) en fonction de n.
On remarque que le nombre d'opérations nécessaires au calcul de

$$\ell_{n+1,n+1} = (a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}\ell_{n+1,k})^{1/2} > 0,$$

$$\text{et } \ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k}\ell_{n+1,k}\right) \frac{1}{\ell_{n+1,n+1}},$$

est toujours inférieur à $2q+1$, car la somme $\sum_{k=1}^{n}$ fait intervenir au plus q termes non nuls.
De plus, pour chaque colonne $n+1$, il y a au plus $q+1$ coefficients $\ell_{i,n+1}$ non nuls, donc au plus $q+1$ coefficients à calculer. Donc le nombre d'opérations pour chaque colonne peut être majoré par $(2q+1)(q+1)$.
On peut donc majorer le nombre d'opérations z_q pour les q premières colonnes et les q dernières par $2q(2q+1)(q+1)$, qui est indépendant de n (on rappelle qu'on cherche une estimation en fonction de n, et donc le nombre z_q est $O(1)$ par rapport à n.)
Calculons maintenant le nombre d'opérations x_n nécessaires une colonne $n = q+1$ à $n-q-1$. Dans (1.50) et (1.51), les termes non nuls de la somme sont pour $k = i-q, \ldots, n$, et donc on a $(n-i+q+1)$ multiplications et additions, une division ou extraction de racine. On a donc

Exercice 04

- Si $a = 0$, alors $A = Id$, donc A est s.d.p. et la méthode de Jacobi converge.
- Si $a \neq 0$, posons $a\mu = (1 - \lambda)$, et calculons le polynôme caractéristique de la matrice A en fonction de la variable μ.

$$P(\mu) = det \begin{vmatrix} a\mu & a & a \\ a & a\mu & a \\ a & a & a\mu \end{vmatrix} = a^3 det \begin{vmatrix} \mu & 1 & 1 \\ 1 & \mu & 1 \\ 1 & 1 & \mu \end{vmatrix} = a^3(\mu^3 - 3\mu + 2).$$

On a donc $P(\mu) = a^3(\mu - 1)^2(\mu + 2)$. Les valeurs propres de la matrice A sont donc obtenues pour $\mu = 1$ et $\mu = 2$, c'est–à–dire : $\lambda_1 = 1 - a$ et $\lambda_2 = 1 + 2a$.
La matrice A est définie positive si $\lambda_1 > 0$ et $\lambda_2 > 0$, c'est–à–dire si $-\frac{1}{2} < a < 1$.
La méthode de Jacobi s'écrit :

$$X^{(n+1)} = D^{-1}(D - A)X^{(n)},$$

avec $D = Id$ dans le cas présent ; donc la méthode converge si et seulement si $\rho(D - A) < 1$.
Les valeurs propres de $D - A$ sont de la forme $\nu = 1 - \lambda$ où λ est valeur propre de A. Les valeurs propres de $D - A$ sont donc $\nu_1 == -a$ (valeur propre double) et $\nu_2 = 2a.$. On en conclut que la méthode de Jacobi converge si et seulement si $-1 < -a < 1$ et $-1 < 2a < 1$, *i.e.* $\frac{1}{2} < a < \frac{1}{2}$.
La méthode de Jacobi ne converge donc que sur l'intervalle $]-\frac{1}{2}, \frac{1}{2}[$ qui est strictement inclus dans l'intervalle $]-\frac{1}{2}, 1[$ des valeurs de a pour lesquelles la matrice A est s.d.p..

Exercice 05

1. On pose $L = \begin{pmatrix} 1 & 0 \\ \gamma & 1 \end{pmatrix}$ et $D = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$.
Par identification, on obtient $\alpha = 2$, $\beta = -\frac{1}{2}$ et $\gamma = \frac{1}{2}$.

Si maintenant on essaye d'écrire $A = LL^t$ avec $L = \begin{pmatrix} a & 0 \\ b & c \end{pmatrix}$, on obtient $c^2 = -\frac{1}{2}$ ce qui est impossible dans $\mathbb{R}$.

En fait, on peut remarquer qu'il est normal que A n'admette pas de décomposition LL^t, car elle n'est pas définie positive. En effet, soit $x = (x_1, x_2)^t \in \mathbb{R}^2$,, alors $Ax \cdot x = 2x_1(x_1 + x_2)$, et en prenant $x = (1, -2)^t$, on a $Ax \cdot x < 0$.

2. 2. Reprenons en l'adaptant la démonstration du théorème 1.3. On raisonne donc par récurrence sur la dimension.

1. Dans le cas $N = 1$, on a $A = (a_{1,1})$. On peut donc définir $L = (\ell_{1,1})$ où $\ell_{1,1} = 1$, $D = (a_{1,1})$, $d_{1,1} \neq 0$, et on a bien $A = LDL^t$.
2. On suppose que, pour $1 \leq p \leq N$, la décomposition $A = LDL^t$ s'obtient pour $A \in \mathcal{M}_p(\mathbb{R})$ symétrique définie positive ou négative, avec $d_{i,i} \neq 0$ pour $1 \leq i \leq N$ et on va démontrer que la propriété est encore vraie pour $A \in \mathcal{M}_{N+1}(\mathbb{R})$ symétrique définie positive ou négative. Soit donc $A \in \mathcal{M}_{N+1}(\mathbb{R})$ symétrique définie positive ou négative ; on peut écrire A sous la forme :

$$A = \left[\begin{array}{c|c} B & a \\ \hline a^t & \alpha \end{array}\right]$$

où $B \in \mathcal{M}_N(\mathbb{R})$ est symétrique définie positive ou négative (calculer $Ax \cdot x$ avec $x = (y, 0)^t$, avec $y \in \mathbb{R}^N$ pour le vérifier), $a \in \mathbb{R}^N$ et $\alpha \in \mathbb{R}$.
Par hypothèse de récurrence, il existe une matrice $M \in \mathcal{M}_N(\mathbb{R})$ $M = (m_{i,j})_{i,j=1}^N$ et une matrice diagonale $\tilde{D} = \mathrm{diag}(d_{1,1}, d_{2,2}, \ldots, d_{N,N})$ dont les coefficients sont tous non nuls, telles que :

(a) $m_{i,j} = 0$ si $j > i$

(b) $m_{i,i} = 1$

(c) $B = M\tilde{D}M^t$.

On va chercher L et D sous la forme :

$$L = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right], \; D = \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right],$$

avec $b \in \mathbb{R}^N$, $\lambda \in \mathbb{R}$ tels que $LDL^t = A$. Pour déterminer b et λ, calculons LDL^t avec L et D de la forme (1.8.75) et identifions avec A :

$$LDL^t = \left[\begin{array}{c|c} M & 0 \\ \hline b^t & 1 \end{array}\right] \left[\begin{array}{c|c} \tilde{D} & 0 \\ \hline 0 & \lambda \end{array}\right] \left[\begin{array}{c|c} M^t & b \\ \hline 0 & 1 \end{array}\right] = \left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b + \lambda \end{array}\right]$$

On cherche $b \in \mathbb{R}^N$ et $\lambda \in \mathbb{R}$ tels que $LDL^t = A$, et on veut donc que les égalités suivantes soient vérifiées :

$$M\tilde{D}b = a \text{ et } b^t\tilde{D}b + \lambda = \alpha.$$

La matrice M est inversible (en effet, le déterminant de M s'écrit $det(M) = \prod_{i=1}^N 1 = 1$). Par hypothèse de récurrence, la matrice $\tilde{D}$ est aussi inversible. La première égalité ci-dessus donne : $b = \tilde{D}^{-1}M^{-1}a$. On calcule alors $\lambda = \alpha - b^tM^{-1}a$. Remarquons qu'on a forcément $\lambda \neq 0$, car si $\lambda = 0$,

$$A = LDL^t = \left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right]$$

qui n'est pas inversible. En effet, si on cherche $(x, y) \in \mathbb{R}^N \times \mathbb{R}$ solution de

$$\left[\begin{array}{c|c} M\tilde{D}M^t & M\tilde{D}b \\ \hline b^t\tilde{D}M^t & b^t\tilde{D}b \end{array}\right] \left[\begin{array}{c} x \\ \hline y \end{array}\right] = \left[\begin{array}{c} 0 \\ \hline 0 \end{array}\right],$$

on se rend compte facilement que tous les couples de la forme $(-M^{-t}by, y)^t$, $y \in \mathbb{R}$, sont solutions. Le noyau de la matrice n'est donc pas réduit à $\{0\}$ et la matrice n'est donc pas inversible. On a ainsi montré que $d_{N+1,N+1} \neq 0$ ce qui termine la récurrence.

3. On reprend l'algorithme de décomposition LL^t :
Soit $A \in \mathcal{M}_N(\mathbb{R})$ symétrique définie positive ou négative ; on vient de montrer qu'il existe une matrice $L \in \mathcal{M}_N(\mathbb{R})$ triangulaire inférieure telle que $\ell_{i,j} = 0$ si $j > i$, $\ell_{i,i} = 1$, et une matrice $D \in \mathcal{M}_N(\mathbb{R})$ diagonale inversible, telles que et $A = LDL^t$. On a donc :

$$a_{i,j} = \sum_{k=1}^{N} \ell_{i,k} d_{k,k} \ell_{j,k}, \quad \forall\, (i,j) \in \{1, \ldots, N\}^2.$$

1. Calculons la 1ère colonne de L ; pour $j = 1$, on a :

$$\begin{aligned} &a_{1,1} = d_{1,1} \text{ donc } d_{1,1} = a_{1,1}, \\ &a_{2,1} = \ell_{2,1} d_{1,1} \text{ donc } \ell_{2,1} = \frac{a_{2,1}}{d_{1,1}}, \\ &a_{i,1} = \ell_{i,1} \ell_{1,1} \text{ donc } \ell_{i,1} = \frac{a_{i,1}}{\ell_{1,1}} \ \forall i \in \{2, \ldots, N\}. \end{aligned}$$

2. On suppose avoir calculé les n premières colonnes de L. On calcule la colonne $(n+1)$ en prenant $j = n+1$ dans

   Pour $i = n+1$, $a_{n+1,n+1} = \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k} + d_{n+1,n+1}$ donc

$$d_{n+1,n+1} = a_{n+1,n+1} - \sum_{k=1}^{n} \ell_{n+1,k}^2 d_{k,k}.$$

   On procède de la même manière pour $i = n+2, \ldots, N$; on a :

$$a_{i,n+1} = \sum_{k=1}^{n+1} \ell_{i,k} d_{k,k} \ell_{n+1,k} = \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} + \ell_{i,n+1} d_{n+1,n+1} \ell_{n+1,n+1}$$

   et donc, comme on a montré dans la question 2 que les coefficients $d_{k,k}$ sont tous non nuls, on peut écrire :

$$\ell_{i,n+1} = \left(a_{i,n+1} - \sum_{k=1}^{n} \ell_{i,k} d_{k,k} \ell_{n+1,k} \right) \frac{1}{d_{n+1,n+1}}.$$

Exercice 06

1. $J = D^{-1}(E+F)$ peut ne pas être symétrique, même si A est symétrique :

   En effet, prenons $A = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$.

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

   donc J n'est pas symétrique.

2. On applique l'exercice précédent pour l'application linéaire T de matrice D, qui est, par hypothèse, définie positive (et évidemment symétrique puisque diagonale) et $S = E + F$, symétrique car A est symétrique.
   Il existe donc $(f_1 \dots f_N)$ base de E et $(\mu_1 \dots \mu_N) \in \mathbb{R}^N$ tels que

   $$Jf_i = D^{-1}(E+F)f_i = \mu_i f_i, \quad \forall i = 1, \dots, N, \text{ et } (Df_i, f_j) = \delta_{ij}.$$

3. Par définition de J, tous les éléments diagonaux de J sont nuls et donc sa trace également. Or $TrJ = \sum_{i=1}^{N} \mu_i$.

   Si $\mu_i > 0 \ \forall i = 1, \dots, N$, alors $TrJ > 0$, donc $\exists i_0; \ \mu_i \leq 0$ et comme $\mu_1 \leq \mu_{i_0}$, on a $\mu_1 \leq 0$. Un raisonnement similaire montre que $\mu_N \geq 0$.

4. La méthode de Jacobi converge si et seulement si $\rho(J) < 1$ (théorème 1.27 page 28). Or, par la question précédente, $\rho(A) = max(-\mu_1, \mu_N)$. Supposons que $\mu_1 \leq -1$, alors $\mu_1 = -\alpha$, avec $\alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_1$ ou encore $(E+F)f_1 = -\alpha D f_1$, ce qui s'écrit aussi $(D+E+F)f_1 = D(1-\alpha)f_1$ c'est–à–dire $(2D-A)f_1 = \beta D f_1$ avec $\beta \leq 0$. On en déduit que $((2D-A)f_1, f_1) = \beta \leq 0$, ce qui contredit le fait que $2D - A$ est définie positive. En conséquence, on a bien $\mu_1 \geq -1$.
   Supposons maintenant que $\mu_N = \alpha \geq 1$. On a alors $D^{-1}(E+F)f_1 = -\alpha f_N$, soit encore $(E+F)f_N = -\alpha D f_N$. On en déduit que $Af_N = (D-E-F)f_N = D(1-\alpha)f_N = D\beta f_N$ avec $\beta \leq 0$. On a alors$(Af_N, f_N) \leq 0$, ce qui contredit le fait que A est définie positive.

5. Par définition, on a :$D\tilde{x}^{(n+1)}s = (E+F)x^{(n)} + b$ et $x^{(n+1)} = \omega\tilde{x}^{(n+1)} + (1-\omega)x^{(n)}$. On a donc $x^{(n+1)} = \omega[D^{-1}(E+F)x^{(n)} + D^{-1}b] + (1-\omega)x^{(n)}$ c'est–à–dire $x^{(n+1)} = [Id - \omega(Id - D^{-1}(E+F))]x^{(n)} + \omega D^{-1}b$,, soit encore $\frac{1}{\omega}Dx^{(n+1)} = [\frac{1}{\omega}D - (D - (E+F))]x^{(n)} + b$. On en déduit que $M_\omega x^{(n+1)} = N_\omega x^{(n)} + b$ avec $M_\omega = \frac{1}{\omega}D$ et $N_\omega = \frac{1}{\omega}D - A$.

6. La matrice d'itération est donc maintenant $J_\omega = M_\omega^{-1}N_\omega$ qui est symétrique pour le produit scalaire $(\cdot,\cdot)_{M_\omega}$ donc en reprenant le raisonnement de la question 2, il existe une base $(\tilde{f}_1, \dots, \tilde{f}_N) \in (\mathbb{R}^N)^N$ et $(\tilde{\mu}_1, \dots \tilde{\mu}_N) \subset \mathbb{R}^N$ tels que

   $$J_\omega \tilde{f}_i = {M_\omega}^{-1}N_\omega \tilde{f}_i = \omega D^{-1}\left(\frac{1}{\omega}D - A\right)\tilde{f}_i = \tilde{\mu}_i \tilde{f}_i, \quad \forall i = 1, \dots, N,$$

   $$\text{et } \frac{1}{-}D\tilde{f}_i \cdot \tilde{f} \ = \delta \quad , \quad \forall i, = 1, \dots, N, \ \forall j, = 1, \dots, N.$$

   Supposons $\tilde{\mu}_1 \leq -1$, alors $\tilde{\mu}_1 = -\alpha$, avec $\alpha \geq 1$ et $\omega D^{-1}(\frac{1}{\omega}D - A)\tilde{f}_1 = -\alpha\tilde{f}_1$, ou encore $\frac{1}{\omega}D - A\tilde{f}_1 = -\alpha\frac{1}{\omega}D\tilde{f}_1$. On a donc $\frac{2}{\omega}D - A\tilde{f}_1 = (1-\alpha)\frac{1}{\omega}D\tilde{f}_1$, ce qui entraîne $(\frac{2}{\omega}D - A)\tilde{f}_1 \cdot \tilde{f}_1 \leq 0$. Ceci contredit l'hypothèse $\frac{2}{-}D - A$ définie positive.

De même, si $\tilde{\mu}_N \geq 1$, alors $\tilde{\mu}_N = \alpha$ avec $\alpha \geq 1$. On a alors

$$(\frac{1}{\omega}D - A)\tilde{f}_N = \alpha\frac{1}{\omega}D\tilde{f}_N,$$

et donc $A\tilde{f}_N = (1-\alpha)\frac{1}{\omega}D\tilde{f}_N$ ce qui entraîne en particulier que $A\tilde{f}_N \cdot \tilde{f}_N \leq 0$; or ceci contredit l'hypothèse A définie positive.

7. On cherche une condition nécessaire et suffisante pour que

$$\left(\frac{2}{\omega}D - A\right) x \cdot x > 0, \quad \forall x \neq 0,$$

ce qui est équivalent à

$$\left(\frac{2}{\omega}D - A\right) f_i \cdot f_i > 0, \quad \forall i = 1, \ldots, N,$$

où les $(f_i)_{i=1,N}$ sont les vecteurs propres de $D^{-1}(E+F)$. En effet, la famille $(f_i)_{i=1,\ldots,N}$ est une base de $\mathbb{R}^N$, et

$$\begin{aligned}\left(\frac{2}{\omega}D - A\right) f_i &= \left(\frac{2}{\omega}D - D + (E+F)\right) f_i \\ &= \left(\frac{2}{\omega} - 1\right) Df_i + \mu_i Df_i \\ &= \left(\frac{2}{\omega} - 1 + \mu_i\right) Df_i.\end{aligned}$$

On a donc en particulier $\left(\frac{2}{\omega}D - A\right) f_i \cdot f_j = 0$ is $i \neq j$, ce qui prouve que (1.8.86) est équivalent à (1.8.87). De (1.8.87), on déduit, grâce au fait que $(Df_i, f_i) = 1$,

$$\left(\left(\frac{2}{\omega}D - A\right) f_i, f_i\right) = \left(\frac{2}{\omega} - 1 + \mu_i\right).$$

On veut donc que $\frac{2}{\omega} - 1 + \mu_1 > 0$ car $\mu_1 = \inf \mu_i$, c'est–à–dire : $-\frac{2}{\omega} < \mu_1 - 1$, ce qui est équivalent à : $\omega < \dfrac{2}{1-\mu_1}$.

8. La matrice d'itération J_ω s'écrit :

$$J_\omega = \quad \frac{1}{\omega}D \quad^{-1} \quad \frac{1}{\omega}D - A \quad = \omega I_\omega, \text{ avec } I_\omega = D^{-1}(\frac{1}{\omega}D - A).$$

Soit λ une valeur propre de I_ω associée à un vecteur propre u ; alors :

$$D^{-1}\left(\frac{1}{\omega}D - A\right) u = \lambda u, \ \textit{i.e.} \ \left(\frac{1}{\omega}D - A\right) u = \lambda Du.$$

On en déduit que

$$(D-A)u + \left(\frac{1}{\omega} - 1\right) Du = \lambda Du, \text{ soit encore}$$

$$D^{-1}(E+F)u = \left(1 - \frac{1}{\omega} + \lambda\right) u.$$

Or f_i est vecteur prore de $D^{-1}(E+F)$ associée à la valeur propre μ_i (question 2). On a donc :

$$D^{-1}(E+F)f_i = \mu_i f_i = \left(1 - \frac{1}{\omega} + \lambda\right) f_i,$$

ce qui est vrai si $\mu_i = 1 - \frac{1}{\omega} + \lambda$, c'est–à–dire $\lambda = \mu_i - 1 - \frac{1}{\omega}$. Donc $\mu_i^{(\omega)} = \omega\left(\mu_i - 1 - \frac{1}{\omega}\right)$ est valeur propre de J_ω associée au vecteur propre f_i.
On cherche maintenant à minimiser le rayon spectral

$$\rho(J_\omega) = \sup_i |\omega(\mu_i - 1 - \frac{1}{\omega})|$$

On a

$$\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq \omega(\mu_i - 1 - \frac{1}{\omega}) \leq \omega(\mu_N - 1 - \frac{1}{\omega}),$$

et

$$-\omega(\mu_N - 1 - \frac{1}{\omega}) \leq -\omega(\mu_1 - 1 - \frac{1}{\omega}) \leq -\omega(\mu_i - 1 - \frac{1}{\omega}),$$

donc

$$\rho(J_\omega) = \max\left(|\omega(\mu_N - 1 - \frac{1}{\omega})|, |-\omega(\mu_1 - 1 - \frac{1}{\omega}|)\right)$$

dont le minimum est atteint (voir Figure 1.8) pour

$$\omega(1-\mu_1) - 1 = 1 - \omega(1-\mu_N) \text{ c'est–à–dire } \omega = \frac{2}{2 - \mu_1 - \mu_N}.$$

Exercice 07

1. Pour montrer l'égalité, prendre x tel que $x_j = sign(a_{i_0,j})$ où i_0 est tel que $\sum_{j=1,...,N} |a_{i_0,j}| \geq \sum_{j=1,...,N} |a_{i,j}|$, $\forall i = 1, ..., N$, et sign(s) désigne le signe de s.
2. Pour montrer l'égalité, prendre x tel que $x_{j_0} = 1$ et $x_j = 0$ si $j \neq j0$, où $j0$ est tel que $\sum_{j=1,...,N} |a_{i,j_0}| = \sum_{i=1,...,N} |a_{i,j}|$.
3. Utiliser le fait que A^tA est une matrice symétrique positive pour montrer l'inégalité, et pour l'égalité, prendre pour x le vecteur propre associé à la plus grande valeur propre de A.

Exercice 08

1. Montrer que si $\rho(A) < 1$, alors 0 n'est pas valeur propre de $Id + A$ et $Id - A$.
2. Utiliser le résultat de Rayon spectral.

Exercice 09

Le système (1) s'écrit encore : $Ax = b$ avec

$$A = \begin{pmatrix} \boxed{2} & 1 & 0 & 4 \\ -4 & -2 & 3 & -7 \\ 4 & 1 & -2 & 8 \\ 0 & -3 & -12 & -1 \end{pmatrix} ; \quad b = \begin{pmatrix} 2 \\ -9 \\ 2 \\ 2 \end{pmatrix} \quad et \quad x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}$$

$\underline{1^{ere} \text{ étape :}}$

$\star$ Le pivot $a_{11}^{(1)} = 2 \neq 0$

$$\begin{pmatrix} \boxed{2} & 1 & 0 & 4 & 2 \\ -4 & -2 & 3 & -7 & -9 \\ 4 & 1 & -2 & 8 & 2 \\ 0 & -3 & -12 & -1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 1 & 0 & 4 & 2 \\ 0 & \boxed{0} & 3 & 1 & -5 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & -3 & -12 & -1 & 2 \end{pmatrix}$$

$$\left[A^{(1)} \vdots b^{(1)}\right] \longrightarrow \left[A^{(2)} \vdots b^{(2)}\right]$$

$\underline{2^{eme} \text{ étape :}}$

Dans $\left[A^{(2)} \vdots b^{(2)}\right]$ on constate que le pivot $a_{22}^{(2)} = 0$. D'où on fait une permutation des *lignes* 2 et 3 (par exemple). Ce qui revient à considerer $\left[\tilde{A}^{(2)} \vdots \tilde{b}^{(2)}\right]$ avec :

$$\begin{cases} \tilde{A}^{(2)} = P^{(2)} A^{(2)} \\ \tilde{b}^{(2)} = P^{(2)} b^{(2)} \end{cases} \qquad où \quad P^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

On a alors : Le pivot $a_{22}^{(2)} \neq 0$

$$\left(a_{22}^{(2)} = \tilde{a}_{22}^{(2)} = -1\right)$$

et

$$\begin{pmatrix} 2 & 1 & 0 & 4 & 2 \\ 0 & \boxed{-1} & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & -3 & -12 & -1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 1 & 0 & 4 & 2 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & 0 & -6 & -1 & 8 \end{pmatrix}$$

$$\left[\tilde{A}^{(2)} \vdots \tilde{b}^{(2)}\right] \longrightarrow \left[A^{(3)} \vdots b^{(3)}\right]$$

$\underline{3^{eme} \text{ étape :}}$

$\star$ Le pivot $a_{33}^{(3)} = 3 \neq 0$

$$\begin{pmatrix} 2 & 1 & 0 & 4 & 2 \\ 0 & \boxed{-1} & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & -3 & -12 & -1 & 2 \end{pmatrix} \longrightarrow \begin{pmatrix} 2 & 1 & 0 & 4 & 2 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & 0 & -6 & -1 & 8 \end{pmatrix}$$

$$\left[A^{(3)} \vdots\, b^{(3)}\right] \longrightarrow \left[A^{(4)} \vdots\, b^{(4)}\right]$$

Résolution de $A'x = b'$:
Posons $\left[A' \vdots\, b'\right] = \left[A^{(4)} \vdots\, b^{(4)}\right]$. On a alors

$$A'x = b' \iff \begin{pmatrix} 2 & 1 & 0 & 4 \\ 0 & -1 & -2 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 2 \\ -2 \\ -5 \\ -2 \end{pmatrix}$$

$$\iff \begin{cases} 2x_1 + x_2 - 0x_3 + 4x_4 = 2 \\ -x_2 - 2x_3 + 0x_4 = -2 \\ 3x_3 + x_4 = -5 \\ x_4 = -2 \end{cases} \iff \begin{cases} x_1 = 3 \\ x_2 = 4 \\ x_3 = -1 \\ x_4 = -2 \end{cases}$$

Chapitre 2

Calcul des valeurs et vecteurs propres

Les valeurs propres d'une matrice sont un outil précieux pour l'ingénieur. Comme application principale, on notera le calcul des oscillations propres d'un système qu'il soit mécanique ou électrique. Les valeurs propres peuvent également donner l'information contenue dans un grand ensemble de données, telles les directions principales d'un nuage de points, ou l'information contenue dans un grand graphe. Et la liste ne s'arrête évidemment pas là. Rappelons tout d'abord la définition mathématique d'une valeur propre et de son vecteur propre associé.

Définition 2.0.1 *Soit $A \in \mathbb{R}^{n\times n}$ une matrice réelle. La valeur $\lambda \in \mathbb{C}$ est une valeur propre de A s'il existe $v \in \mathbb{C}^n$ $v \neq 0$ (appelé vecteur propre associé λ) tel que $Av = \lambda v$.*

La question du calcul des valeurs propres d'une matrice est fondamentale. Il est cependant peu pratique de devoir calculer les racines du polynôme caractéristique d'une matrice $det(A-\lambda I)$ afin d'en connaître les valeurs propres.

Dans cette section, nous allons montrer comment on peut obtenir très rapidement quelques valeurs propres et leurs vecteurs propres associés en appliquant une méthode itérative, connue sous le nom de *méthode de la puissance*. La question du calcul de toutes les valeurs propres d'une matrice, bien qu'importante, déborde du cadre de ce cours.

2.1 Méthode de la puissance

Soit une matrice réelle $A \in \mathbb{R}^{n\times n}$. Dans cette section, nous supposons que les valeurs propres de A sont telles que

$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq ... \geq |\lambda_n|$

et que chaque valeur propre λ_i a un vecteur propre associé $v^{(i)}$.La méthode de la puissance est une méthode itérative qui sert a trouver une approximation de λ_1. Notons l'importance de la condition de stricte inégalité $|\lambda_1| > |\lambda_2|$.

Nous supposons d'abord que les vecteurs propres de A forment une base linéaire de $\mathbb{C}^n$. Nous partons d'un vecteur complexe arbitraire $w^{(0)} \in \mathbb{C}^n$.

Celui-ci peut s'écrire comme une combinaison linéaire des différents vecteurs propres de A. On a donc

$$w^{(0)} = \alpha_1 v^{(1)} + \alpha_2 v^{(2)} + ... + \alpha_n v^{(n)} \tag{01}$$

avec $\alpha_i \in \mathbb{C}$ pour tout i : Supposons encore que $\alpha_1 \neq 0$: Nous procédons à présent aux différentes itérations de la m ethode de la puissance. On calcule successivement

$w^{(1)} = Aw^{(0)}$

$w^{(2)} = Aw^{(1)} = A^2w^{(0)}$

...

$w^{(k)} = Aw^{(k-1)} = A^kw^{(0)}$.

Si on reprend (1), on peut egalement écrire

$w^{(k)} = A^kw^{(0)}$

$= A^k(\alpha_1v^{(1)} + ... + \alpha_nv^{(n)})$

$= \alpha_1\lambda_1^kv^{(1)} + ... + \alpha_n\lambda_n^kv^{(n)}$

en utilisant la propriété des vecteurs propres. Finalement, la dernière expression se rééecrit

$$w^{(k)} = \lambda_1^k \left(\alpha_1 v^{(1)} + \alpha_2 (\frac{\lambda_2}{\lambda_1})^k v^{(2)} + ... + \alpha_n (\frac{\lambda_n}{\lambda_1})^k v^{(n)} \right)$$

Comme $|\lambda_1| > |\lambda_j|$ pour tout $j \neq 1$, tous les termes $(\frac{\lambda j}{\lambda_1})^k$ tendent vers 0 quand k tend vers l'infini. A l'infini, la quantité $w^{(k)}$ tend donc vers la

direction du vecteur propre dominant de A. En général évidemment, soit si $|\lambda_1| > 1$, la quantité tend vers l'infini, soit si $|\lambda_1| < 1$, la quantité tend vers 0, et il sera difficile de localiser la vraie direction. Dans la pratique, on procède donc à la normalisation des vecteurs après chaque étape :

$$z^{(k)} = \frac{w^{(k)}}{\|w^{(k)}\|},\ w^{(k+1)} = Az^{(k)}.$$

Le processus converge alors vers le vecteur propre dominant. On peut obtenir la valeur propre dominante en calculant à chaque étape

$$\sigma_k = z^{(k)^T} w^{(k+1)}$$

qui converge vers λ_1. En effet, on a

$$\sigma_k = \frac{z^{(k)^T} A z^{(k)}}{z^{(k)^T} z^{(k)}}$$

qui converge vers λ_1 si z_k converge vers $v^{(1)}$. Remarquons que le processus que nous venons de décrire converge vers le vecteur propre avec une vitesse dépendant du ratio $|\lambda_2| \,/\, |\lambda_1|$. Plus ce quotient est petit, plus la convergence est rapide. Le processus ne convergera donc pas vite pour des matrices pour lesquelles les deux valeurs propres dominantes sont proches. Une façon d'accélérer le processus est de travailler avec une matrice $B = A - mI$. En effet toutes les valeurs propres de B sont exactement $\lambda_i - m$: Si on connaît une approximation des valeurs propres, il peut être possible d'améliorer le ratio $|\lambda_2 - m| \,/\, |\lambda_1 - m|$. En règle générale, on n'a évidemment pas accès aux valeurs propres et il est donc difficile de trouver le m optimal. Remarquons qu'on peut aussi se servir de cette astuce pour calculer une autre valeur propre. En effet, si on applique la méthode de la puissance à $B := A - mI$, celle-ci va converger vers la valeur propre la plus éloignée de m.

Exemple

Soit la matrice A :

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

On se donne le vecteur : $V_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

à partir duquel on construit la suite de vecteurs AV_1, $A^2V_1, ...$, reportée dans le tableau ci-dessous.

λ		1	5	13/5	41/13	121/41	365/121	.....
	1	1	1	1	1	1	1	
v	0	2	$\frac{4}{5}$	$\frac{14}{13}$	$\frac{40}{41}$	$\frac{122}{121}$	$\frac{364}{365}$	.....
	0	0	0	0	0	0	0	

Les composantes des vecteurs de ce tableau ont été divisées par la première composante.

Il est clair que la suite des vecteurs tend vers le vecteur : $u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

et que la suite des valeurs de λ converge vers la valeur $\lambda_1 = 3$.

La matrice A considérée étant symétrique, le vecteur propre de sa transposée, correspondant à la valeur propre λ_1 sera :

$$v_1 = u_1 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

Construisons à présent la matrice A_1 qui est telle que :

$$A_1 = A - \lambda_1 \frac{u_1 \, {}^t v_1}{{}^t v_1 u_1} = \begin{pmatrix} -1/2 & 1/2 & 0 \\ 1/2 & -1/2 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

On se donne à nouveau un vecteur V_1 : $V_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

et on procède de la même manière qu'avec la matrice A

λ		$\frac{-1}{2}$	5	13/5	41/13
	1	1	1	1	
v	0	-1	-1	-1	.....
	0	0	0	0	

On voit que cette suite de vecteurs converge vers le vecteur propre : $u_2 = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

et que la valeur propre qui lui correspond est $\lambda_2 = -1$.

On recommence le processus en définissant une matrice A_2 à partir de A_1, u_1 et $v_2 = u_2$:

$A_2 = A_1 - \lambda_1 \frac{u_2\ {}^t v_2}{{}^t v_2 u_2} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -1 \end{pmatrix}$

Il est clair que la valeur propre de cette matrice est $\lambda_3 = -1$ et que le vecteur propre est :

$u_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

2.2 Calcul de la valeur propre de plus petit module

Il est également possible de calculer la valeur propre de plus petit module (à condition qu'elle soit non nulle) avec la méthode de la puissance. En effet,

si A est inversible, on peut voir que si λ est une valeur propre de A, alors on a

$$Ax = \lambda x \Leftrightarrow x = A^{-1}(\lambda x) \Leftrightarrow A^{-1}x = \frac{1}{\lambda}x$$

Dès lors, si est une valeur propre de A, $\frac{1}{\lambda}$ est une valeur propre de A^{-1}. On en déduit aussi que si λ est la valeur propre de plus petit module de A, $\frac{1}{\lambda}$ sera la valeur propre de plus grand module deA^{-1}. On peut donc appliquer la méthode de la puissance de manière totalement similaire avec A^{-1} et ecrire

$$z^{(k)} = \frac{w^{(k)}}{\|w^{(k)}\|},\ w^{(k+1)} = A^{-1}z^{(k)}.$$

Remarquons que dans la dernière expression, il n'est pas nécessaire d'éffectuer la coûteuse opération de l'inversion de la matrice A. On peut tout à fait se contenter d'une factorisation LU de la matrice et résoudre le système $Aw^{(k+1)} = z^{(k)}$ a chaque itération. Finalement, on peut remarquer que l'on peut se servir de la méthode inverse de la puissance, associée à un changement de matrice $B = A - mI$ pour trouver la valeur propre qui est la plus proche d'un scalaire m.

2.3 Calcul d'autres valeurs propres

Supposons que l'on ait trouvé la valeur propre dominante λ_1 de A. On souhaite à présent calculer la deuxième valeur propre de plus grand module, à savoir λ_2.

La méthode que nous décrirons en premier lieu ne convient que pour une matrice A symétrique. Si λ_1 et v^1 sont respectivement la valeur propre et le vecteur propre déj-à calculés, on forme la matrice

$$A = A - \lambda_1 v^1 {v^1}^T \tag{02}$$

Comme la matrice A est symétrique, A_1 l'est aussi. On calcule que $A_1 v^1 = 0$ et que $A_1 v^j = \lambda_j v^j$ pour tout vecteur propre v^j associé à une valeur propre λ_j , $j = 2; 3, ..., n$. Par conséquent,

A_1 a tous les vecteurs propres de A et toutes ses valeurs propres excepté λ_1 qui est remplacée par zéro.

Lorsque λ_2 et v^2 ont été calculés a partir de A_1, le processus peut être répété en formant $A_2 = A_1 - \lambda_2 v^2 v^{2^T}$, et ainsi de suite pour la détermination des valeurs propres et vecteurs propres restant.

Une autre méthode de déflation consiste à trouver une matrice nonsingulière P telle que $Pv^1 = e_1$ où e_1 est le vecteur canonique $e1 = (1, 0...0)^T$.

On obtient alors de $Av^1 = \lambda_1 v^1$ que

$$PAP^{-1}Pv^1 = \lambda_1 Pv^1$$

$$(PAP^{-1})e_1 = \lambda_1 e_1.$$

La dernière égalité signifie que la matrice PAP^{-1}, qui a les mêmes valeurs propres que A, doit être de la forme

$$PAP^{-1} = \left(\begin{array}{c|c} \lambda_1 & b^T \\ \hline 0 & A_1 \\ 0 & \end{array}\right)$$

et la matrice d'ordre $(n-1)$ occupant le coin inférieur droit de PAP^{-1} possède donc bien les propriétés recherchées. Comme pour l'autre méthode de déflation, on peut répéter le processus en calculant A_2 a partir de A_1 une fois λ_2 et v^2 calculés.

2.4 Algorithme QR

La méthode que nous allons étudier maintenant s'est révélée dans la pratique comme l'une des plus efficaces pour la recherche de toutes les valeurs propres d'une matrice symétrique ou non-symétrique.

L'algorithme **QR** consiste a construire une suite de matrices $A = A_1, A_2, A_3, ...$ au moyen des relations

$$A_k = Q_k R_k, A_{k+1} = R_k Q_k, ... \tag{03}$$

Où les matrices Q_k sont orthogonales et les matrices R_k sont triangulaires supérieures. Rappelons ce théorème d'algèbre concernant la décomposition

QR découlant du principe d'orthogonalisation de Gram-Schmidt.

Théorème 2.4.1 *Toute matrice $A \in \mathbb{R}^{n\times n}$ carrée non singulière peut être ecrite sous la forme $A = QR$ où R désigne une matrice non singulière triangulaire supérieure et où Q désigne une matrice unitaire, c'est- à-dire $QQ^T = Q^TQ = I$.*

On peut montrer que la matrice A_k tend vers une matrice triangulaire supérieure dont les éléments diagonaux sont les valeurs propres de A.

Théorème 2.4.2 (Convergence de l'algorithme QR) *Si les valeurs propres λ_i, $i = 1; 2; ...; n$ de A satisfont les conditions*

$$|\lambda_1| > |\lambda_2| > |\lambda_3| > ... > |\lambda_n| \tag{04}$$

alors la matrice A_k définie en (4) tend vers une matrice triangulaire supérieure dont les éléments diagonaux sont les valeurs propres de A, rangées dans l'ordre des modules décroissants.

Preuve. Puisque les valeurs propres de A sont toutes différentes (et réelles), il existe une matrice non-singulière (et réelle) X, telle que :

$$A = XDX^{-1} \tag{05}$$

où

$$D := diag(\lambda_1, \lambda_2, ..., \lambda_n)$$

Définissons les matrices Q, R, L et U par les relations

$$X = QR \quad X^{-1} = LU. \tag{06}$$

Les matrices R et U sont triangulaires supérieures, la matrice L est triangulaire inférieure avec tous ses éléments diagonaux égaux à 1 et la matrice Q est orthogonale. La matrice R est non-singulière puisque X l'est également. La décomposition QR existe toujours, tandis que la décomposition LU existe seulement si tous les mineurs principaux de X^{-1} sont non nuls.

Analysons maintenant en détail une étape de l'algorithme QR. On a

$$A_{k+1} = R_k Q_k = Q_k^T A_k Q_k \tag{07}$$

A partir de cette dernière relation on déduit

$$A_{k+1} = P_k^T A P_k \tag{08}$$

où

$$P_k := Q_1 Q_2 ... Q_k \tag{09}$$

Si l'on pose alors

$$U_k := R_k R_{k-1} ... R_1 \tag{10}$$

on calcule

$$\begin{aligned} P_k U_k &= Q_1 Q_2 ... Q_{k-1} \left(Q_k R_k\right) R_{k-1} ... R_2 R_1 \\ &= P_{k-1} A_k U_{k-1} \\ &= A P_{k-1} U_{k-1} \end{aligned} \tag{11}$$

où la dernière égalité est obtenue grâce à (08). On obtient alors par récurrence de (11)

$$P_k U_k = A^k \tag{12}$$

Cette dernière relation montre que P_k et U_k sont les facteurs de la d écomposition QR de la matrice A^k.

Si on reprend a présent (06), on a successivement d'après (08), (05) et (06)

$$A_{k+1} = P_k^T A P_k$$

$$= P_k^T X D X^{-1} P_k$$

$$= P_k^T Q R D R^{-1} Q^T P_k \tag{13}$$

La matrice R etant triangulaire supérieure, la matrice R^{-1} l'est aussi et a pour éléments diagonaux les inverses des éléments diagonaux de R. Le produit RDR^{-1} est donc une matrice triangulaire supérieure dont la diagonale est égale à D. Il suffit donc pour établir le théorème de montrer que $P_k \to Q$.

Pour établir ce dernier fait, considérons la matrice A^k qui, étant donné (05) et (06) peut s'écrire sous les formes

$$A^k = X D^k X^{-1} = Q R D^k L U = Q R \left(D^k L D^{-k} \right) D^k U \tag{14}$$

On constate que la matrice $D^k L D^{-k}$ est une matrice triangulaire inférieure dont les éléments diagonaux sont égaux à 1, tandis que l'élément (i, j) est égal à $l_{ij}(\lambda_i/\lambda_j)^k$ lorsque $i > j$, et nous pouvons donc écrire

$$D^k L D^{-k} = I + E_k \quad \text{où} \quad \lim_{k \to \infty} E_k = 0$$

L'équation (14) donne alors

$$A^k = Q R \left(I + E_k \right) D^k U$$

$$= Q \left(I + R E_k R^{-1} \right) R D^k U$$

$$= Q \left(I + F_k \right) R D^k U$$

où

$$\lim_{k \to \infty} F_k = 0$$

On peut alors effectuer la décomposition QR de la matrice $(I + F_k)$, soit

$(I + F_k) = \tilde{Q}_k \tilde{R}_k$

où $\tilde{Q}_k$ et $\tilde{R}_k$ tendent vers I puisque F_k tend vers zéro. On a donc finalement

$$A_k = \left(Q \tilde{Q}_k \right) \left(\tilde{R}_k R D^k U \right) \tag{15}$$

Le premier des facteurs de (15) est orthogonal et le second est triangulaire supérieur : on a donc bien obtenu une décomposition QR de A^k. Mais cette décomposition est unique puisque A_k est non-singulière. Comparant alors (15) et (12) on a $P_k = Q\tilde{Q}_k$ et donc $P_k \to Q$ puisque $\tilde{Q}_k \to I$. ■

2.5 Méthode de Jacobi

Soit une matrice carrée symétrique A d'ordre n dont on cherche à déterminer les valeurs propres. La méthode consiste en des transformations successives du type $T^{-1}AT$ qui amènent la matrice A sous la forme diagonale. Comme les transformations de ce type ne modifient pas les valeurs propres, ces dernières se trouvent, en fin de calcul, sur la diagonale de la matrice transformée. De plus, il est possible de calculer aussi les valeurs propres V_i , $i = 1, ..., n$ de A en multipliant les valeurs propres ξ_i de la matrice finale, qui ne sont autres que les colonnes de la matrice identité, par le produit $T_1T_2...T_k$ des matrices de transformations successives, soit :

$$V_i = T_1T_2...T_k\xi_i \quad i = 1, ..., n$$

Dans la méthode de Jacobi, la matrice T est du type :

$$T = \begin{array}{c} \begin{array}{ccccccc} & & \overset{(p)^{\text{ieme}}}{\text{colonne}} & & & & \overset{(q)^{\text{ieme}}}{\text{colonne}} \\ & & \downarrow & & & & \downarrow \end{array} \\ \begin{pmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 & \cdots & 0 & 0 \\ & & & \cdots & & & & & \\ 0 & \cdots & \cos\varphi & 0 & \cdots & 0 & \sin\varphi & \cdots & 0 \\ & & & \cdots & & & & & \\ \cdots & & \sin\varphi & 0 & \cdots & 0 & -\cos\varphi & \cdots & 0 \\ & & & \cdots & & & & & \\ 0 & \cdots & 0 & 0 & \cdots & & & & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ \longleftarrow (p)^{\text{ieme}} \text{ ligne} \\ \\ \longleftarrow (q)^{\text{ieme}} \text{ ligne} \\ \\ \end{array}$$

qui est égale à la matrice unité du même ordre que la matrice A, à l'exeption des éléments t_{pp}, t_{qq}, t_{pq} et t_{qp}. p et q sont tels que l'élément a_{pq} soit le plus grand élément extra-diagonal et où φ est tel que :

$$\tan(2\varphi) = \frac{2a_{pq}}{a_{pp} - a_{qq}}$$

Exemple

Soit la matrice A :

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

Le plus grand élément extra-diagonal de la matrice A étant l'élément a_{12}, l'angle φ sera tel que :

$$\tan(2\varphi) = \frac{2a_{12}}{a_{11} - a_{22}}$$

Comme $a_{11} = a_{22}$, $\varphi = \frac{\pi}{4}$, la matrice transformation T_{12} s'écrit alors :

$$T_{12} = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ \sqrt{2}/2 & -\sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

d'où

$$A_1 = T_{12} A T_{12} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & -1 \end{pmatrix}$$

qui est diagonale. Les valeurs propres de A_1, et par conséquent de A sont donc : $x_1 = 3$ et $x_2 = x_3 = -1$.

Quant aux vecteurs propres, puisque le calcul s'est opéré en une seule étape, ils sont donnés directement par les colonnes de la matrice T. On a donc :

$$V_1 = T_{12} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \approxeq \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \; V_2 = T_{12} \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \approxeq \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}, \; V_3 = T_{12} \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \approxeq \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$$

Travaux dirigés 2

Exercice 01

1. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice symétrique. Soit $\lambda_N \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_N| = \rho(A)$ et soit $x^{(0)} \in \mathbb{R}^N$. On suppose que $-\lambda_N$ n'est pas une valeur propre de A et que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_N Id)$. On définit la suite $(x^{(n)})_{n\in\mathbb{N}}$ par $x^{(n+1)} = Ax^{(n)}$ pour $n \in \mathbb{N}$. Montrer que

   (a) $\frac{x^{(n)}}{(\lambda_N)^n} \to x$, quand $n \to \infty$, avec $x \neq 0$ et $Ax = \lambda_N x$.

   (b) $\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} \to \rho(A)$ quand $n \to \infty$.

   Cette méthode de calcul s'appelle "méthode de la puissance".

2. Soit $A \in \mathcal{M}_N(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^N$. Pour calculer x t.q. $Ax = b$, on considère la méthode itérative appelée "méthode I" en cours, et on suppose B symétrique. Montrer que, sauf cas particuliers à préciser,

   (a) $\frac{\|x^{(n+1)}-x\|}{\|x^{(n)}-x\|} \to \rho(B)$ quand $n \to \infty$ (ceci donne une estimation de la vitesse de convergence).

   (b) $\frac{\|x^{(n+1)}-x^{(n)}\|}{\|x^{(n)}-x^{(n-1)}\|} \to \rho(B)$ quand $n \to \infty$ (ceci permet d'estimer $\rho(B)$ au cours des itérations).

Exercice 02

Soit $A = \begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & 0 \end{bmatrix}$

1. Calculer les valeurs propres de la matrice A.
2. Effectuer la dÈcompositionQR de la matrice A.
3. Calculer $A_1 = RQ$ et $\tilde{A}_1 = RQ - bId$ *où* b est le terme a_{22}^1 de la matrice A_1
4. Effectuer la dÈcomposition QR de A_1 *et* $\tilde{A}_1$, et calculer les matrices $A_2 = R_1Q_1$ *et* $\tilde{A}_2 = \tilde{R}_1\tilde{Q}_1$.

Exercice 03

1. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice symétrique. Soit $\lambda_n \in \mathbb{R}$ valeur propre de A t.q. $|\lambda_n| = \rho(A)$ et soit $x^{(0)} \in \mathbb{R}^n$. On suppose que $-\lambda_n$ n'est pas une valeur propre de A et que $x^{(0)}$ n'est pas orthogonal à $\mathrm{Ker}(A - \lambda_n Id)$, ce qui revient à dire que . lorsqu'on écrit le vecteur propre $x^{(0)}$ dans la base des vecteurs propres, la composante sur le vecteur propre associé à λ_n est non nulle. On définit la suite $(x^{(k)})_{n\in\mathbb{N}}$ par $x^{(k+1)} = Ax^{(k)}$ pour $n \in \mathbb{N}$. Montrer que

   (a) $\frac{x^{(k)}}{(\lambda_n)^n} \to x$, quand $k \to \infty$, avec $x \neq 0$ et $Ax = \lambda_n x$.

   (b) $\frac{\|x^{(k+1)}\|}{\|x^{(k)}\|} \to \rho(A)$ quand $n \to \infty$.

   Cette méthode de calcul de la plus grande valeur propre s'appelle "méthode de la puissance".

2. Soit $A \in \mathcal{M}_n(\mathbb{R})$ une matrice inversible et $b \in \mathbb{R}^n$. Pour calculer x t.q. $Ax = b$, on considère un méthode itérative : on se donne un choix initial $x^{(0)}$, et on construit la suite $x^{(k)}$ telle que $x^{(k+1)} = Bx^{(k)} + c$ avec $c = (Id - B)A^{-1}b$, et on suppose B symétrique. On rappelle que si $\rho(B) < 1$, la suite tend vers x. Montrer que, sauf cas particuliers à préciser,

   (a) $\frac{\|x^{(k+1)}-x\|}{\|x^{(k)}-x\|} \to \rho(B)$ quand $k \to \infty$ (ceci donne une estimation de la vitesse de convergence de la méthode itérative).

   (b) $\frac{\|x^{(k+1)}-x^{(k)}\|}{\|x^{(k}-x^{(k-1)}\|} \to \rho(B)$ quand $k \to \infty$ (ceci permet d'estimer $\rho(B)$ au cours des itérations).

Exercice 04

Soient u et v deux vecteurs de $\mathbb{R}^n$. On rappelle que la projection orthogonale $\mathrm{proj}_u(v)$ du vecteur v sur la droite vectorielle engendrée par u peut s'écrire de la manière suivante :

$$\mathrm{proj}_u(v) = \frac{v \cdot u}{u \cdot u} u,$$

où $u \cdot v$ désigne le produit scalaire des vecteurs u et v. On note $\|\cdot\|$ la norme euclidienne sur $\mathbb{R}^n$.

1. Soient $(a_1, \ldots, a_n)$ une base de $\mathbb{R}^n$. On rappelle qu'à partir de cette base, on peut obtenir une base orthogonale $(v_1, \ldots, v_n)$ et une base orthonormale $(q_1, \ldots, q_n)$ par le procédé de Gram-Schmidt qui s'écrit :

$$\begin{aligned}
v_1 &= a_1, & q_1 &= \frac{a_1}{\|a_1\|} \\
v_2 &= a_2 - \mathrm{proj}_{v_1}(a_2), & q_2 &= \frac{v_2}{\|v_2\|} \\
v_3 &= a_3 - \mathrm{proj}_{v_1}(a_3) - \mathrm{proj}_{v_2}(a_3), & q_3 &= \frac{v_3}{\|v_3\|} \\
v_4 &= a_4 - \mathrm{proj}_{v_1}(a_4) - \mathrm{proj}_{v_2}(a_4) - \mathrm{proj}_{v_3}(a_4), & q_4 &= \frac{v_4}{\|v_4\|} \\
&\vdots & &\vdots \\
v_k &= a_k - \sum_{j=1}^{k-1} \mathrm{proj}_{v_j}(a_k), & q_k &= \frac{v_k}{\|v_k\|}
\end{aligned}$$

On a donc

$$v_k = a_k - \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j, \qquad q_k = \frac{v_k}{\|v_k\|}.$$

1. Montrer par récurrence que la famille $(v_1, \ldots, v_n)$ est une base orthogonale de $\mathbb{R}^n$.

2. Soient A la matrice carrée d'ordre n dont les colonnes sont les vecteurs a_j et Q la matrice carrée d'ordre N dont les colonnes sont les vecteurs q_j définis par le procédé de Gram-Schmidt (1.132), ce qu'on note :

$$A = \begin{bmatrix} a_1 & a_2 & \ldots & a_n \end{bmatrix}, \qquad Q = \begin{bmatrix} q_1 & q_2 & \ldots & q_n \end{bmatrix}.$$

Montrer que

$$a_k = \|v_k\| q_k + \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{\|v_j\|} q_j.$$

En déduire que $A = QR$, où R est une matrice triangulaire supérieure dont les coefficients diagonaux sont positifs.

3. Montrer que pour toute matrice $A \in \mathcal{M}_n(\mathbb{R})$ inversible, on peut construire une matrice orthogonale Q (c.à. d. telle que $QQ^t = Id$) et une matrice triangulaire supérieure R à coefficients diagonaux positifs telles que $A = QR$.

4. Donner la décomposition QR de $A = \begin{bmatrix} 1 & 4 \\ 1 & 0 \end{bmatrix}$.

5. On considère maintenant l'algorithme suivant (où l'on stocke la matrice Q orthogonale cherchée dans la matrice A de départ (qui est donc écrasée)

Algorithme 1.63 (Gram-Schmidt modifié)**.**

Pour $k = 1, \ldots, n$,

Calcul de la norme de a_k

$r_{kk} := (\sum_{i=1}^{n} a_{ik}^2)^{\frac{1}{2}}$

Normalisation

Pour $\ell = 1, \ldots, n$

$a_{\ell k} := a_{\ell k} / r_{kk}$

Fin pour ℓ

Pour $j = k+1, \ldots, n$

Produit scalaire correspondant à $q_k \cdot a_j$

$r_{kj} := \sum_{i=1}^{n} a_{ik} a_{ij}$

On soustrait la projection de a_k *sur* q_j *sur tous les vecteurs de* A *après* k.

Pour $i = k+1, \ldots, n$,

$a_{ij} := a_{ij} - a_{ik} r_{kj}$

Fin pour i

Fin pour j

Montrer que la matrice A résultant de cet algorithme est identique à la matrice Q donnée par la méthode de Gram-Schmidt, et que la matrice R est celle de Gram-Schmidt. (Cet algorithme est celui qui est effectivement implanté, car il est plus stable que le calcul par le procédé de Gram-Schmidt original.)

Suggestions et Corrigés

Exercice 01

1. Comme A est une matrice symétrique, A est diagonalisable dans $\mathbb{R}$. Soit $(f_1, \ldots, f_N) \in (\mathbb{R}^N)^N$ une base orthonormée de vecteurs propres de A associée aux valeurs propres $(\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N$. On décompose $x^{(0)}$ sur $(f_i)_{i=1,\ldots,N}$: $x^{(0)} = \sum_{i=1}^N \alpha_i f_i$. On a donc $Ax^{(0)} = \sum_{i=1}^N \lambda_i \alpha_i f_i$ et $A^n x^{(0)} = \sum_{i=1}^N \lambda_i^n \alpha_i f_i$.
On en déduit :
$$\frac{x^{(n)}}{\lambda_N^n} = \sum_{i=1}^N \left(\frac{\lambda_i}{\lambda_N}\right)^n \alpha_i f_i.$$
Comme $-\lambda_N$ n'est pas valeur propre,
$$\lim_{n\to+\infty} (\frac{\lambda_i}{\lambda_N})^n = 0 \text{ si } \lambda_i \neq \lambda_N.$$
Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres différentes de λ_N, et $\lambda_{p+1}, \ldots, \lambda_N = \lambda_N$. On a donc
$\lim_{n\to+\infty} \frac{x^{(n)}}{\lambda_N^n} = \sum_{i=p+1}^N \alpha_i f_i = x$, avec $Ax = \lambda_N x$.
De plus, $x \neq 0$: en effet, $x^{(0)} \notin (Ker(A - \lambda_N Id))^\perp = Vect\{f_1, \ldots, f_p\}$, et donc il existe $i \in \{p+1, \ldots, N\}$ tel que $\alpha_i \neq 0$.
Pour montrer (b), remarquons que :
$$\|x^{(n+1)}\| = \sum_{i=1}^N \lambda_i^{n+1} \alpha_i \text{ et } \|x^{(n)}\| = \sum_{i=1}^N \lambda_i^n \alpha_i$$
car $(f_1, \ldots, f_N)$ est une base orthonormée. On a donc
$$\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} = \lambda_N^n \frac{\|\frac{x^{(n+1)}}{\lambda_N^{n+1}}\|}{\|\frac{x^{(n)}}{\lambda_N}\|} \to \lambda_N \frac{\|x\|}{\|x\|} = \lambda_N \text{ lorsque } n \to +\infty.$$

2. a) La méthode I s'écrit à partir de $x^{(0)}$ connu : $x^{(n+1)} = Bx^{(n)} + c$ pour $n \geq 1$, avec $c = (I - B)A^{-1}b$.
On a donc
$$\begin{aligned} x^{(n+1)} - x &= Bx^{(n)} + (Id - B)x - x \\ &= B(x^{(n)} - x). \end{aligned}$$
Si $y^{(n)} = x^{(n)} - x$, on a donc $y^{(n+1)} = By^{(n)}$, et d'après la question 1a) si $y^{(0)} \not\perp \text{Ker}(B - \mu_N Id)$ où μ_N est la plus grande valeur propre de B, (avec $|\mu_N| = \rho(B) et - \mu_N$ non valeur propre), alors
$$\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty,$$
c'est–à–dire
$$\frac{\|x^{(n+1)} - x\|}{\|x^{(n)} - x\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty.$$

Exercice 03

1. Comme A est une matrice symétrique, A est diagonalisable dans $\mathbb{R}$. Soit $(f_1, \ldots, f_n) \in (\mathbb{R}^n)^n$ une base orthonormée de vecteurs propres de A associée aux valeurs propres $(\lambda_1, \ldots, \lambda_n) \in \mathbb{R}^n$. On décompose $x^{(0)}$ sur $(f_i)_{i=1,\ldots,n}$: $x^{(0)} = \sum_{i=1}^n \alpha_i f_i$. On a donc $Ax^{(0)} = \sum_{i=1}^n \lambda_i \alpha_i f_i$ et $A^n x^{(0)} = \sum_{i=1}^n \lambda_i^n \alpha_i f_i$.
   On en déduit :
$$\frac{x^{(n)}}{\lambda_n^n} = \sum_{i=1}^n \left(\frac{\lambda_i}{\lambda_n}\right)^n \alpha_i f_i.$$
   Comme $-\lambda_n$ n'est pas valeur propre,
$$\lim_{n\to+\infty} (\frac{\lambda_i}{\lambda_n})^n = 0 \text{ si } \lambda_i \neq \lambda_n.$$
   Soient $\lambda_1, \ldots, \lambda_p$ les valeurs propres différentes de λ_n, et $\lambda_{p+1}, \ldots, \lambda_n = \lambda_n$. On a donc
   $\lim_{n\to+\infty} \frac{x^{(n)}}{\lambda_n^n} = \sum_{i=p+1}^n \alpha_i f_i = x$, avec $Ax = \lambda_n x$.
   De plus, $x \neq 0$: en effet, $x^{(0)} \notin (Ker(A - \lambda_n Id))^\perp = Vect\{f_1, \ldots, f_p\}$, et donc il existe $i \in \{p+1, \ldots, n\}$ tel que $\alpha_i \neq 0$.
   Pour montrer (b), remarquons que :
$$\|x^{(n+1)}\| = \sum_{i=1}^n \lambda_i^{n+1} \alpha_i \text{ et } \|x^{(n)}\| = \sum_{i=1}^n \lambda_i^n \alpha_i$$
   car $(f_1, \ldots, f_n)$ est une base orthonormée. On a donc
$$\frac{\|x^{(n+1)}\|}{\|x^{(n)}\|} = \lambda_n^n \frac{\|\frac{x^{(n+1)}}{\lambda_n^{n+1}}\|}{\|\frac{x^{(n)}}{\lambda_n}\|} \to \lambda_n \frac{\|x\|}{\|x\|} = \lambda_n \text{ lorsque } n \to +\infty.$$

2. a) La méthode I s'écrit à partir de $x^{(0)}$ connu : $x^{(n+1)} = Bx^{(n)} + c$ pour $n \geq 1$, avec $c = (I - B)A^{-1}b$. On a donc
$$\begin{aligned} x^{(n+1)} - x &= Bx^{(n)} + (Id - B)x - x \\ &= B(x^{(n)} - x). \end{aligned}$$
   Si $y^{(n)} = x^{(n)} - x$, on a donc $y^{(n+1)} = By^{(n)}$, et d'après la question 1a) si $y^{(0)} \not\perp \mathrm{Ker}(B - \mu_n Id)$ où μ_n est la plus grande valeur propre de B, (avec $|\mu_n| = \rho(B) et - \mu_n$ non valeur propre), alors
$$\frac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty,$$
   c'est–à–dire
$$\frac{\|x^{(n+1)} - x\|}{\|x^{(n)} - x\|} \longrightarrow \rho(B) \text{ lorsque } n \to +\infty.$$

b) On applique maintenant 1a) à $y^{(n)} = x^{(n+1)} - x^{(n)}$ avec

$$y^{(0)} = x^{(1)} - x^{(0)} \text{ où } x^{(1)} = Ax^{(0)}.$$

On demande que $x^{(1)} - x^{(0)} \notin \mathrm{Ker}(B - \mu_n Id)^{\perp}$ comme en a), et on a bien $y^{(n+1)} = By^{(n)}$, donc $\dfrac{\|y^{(n+1)}\|}{\|y^{(n)}\|} \longrightarrow \rho(B)$ lorsque $n \to +\infty$.

Exercice 04

1. Par définition de la projection orthogonale, on a $v_1 \cdot v_2 = a_1 \cdot (a_2 - \mathrm{proj}_{a_1}(a_2)) = 0$.
Supposons la récurrence vraie au rang $N-1$ et montrons que v_n est orthogonal à tous les v_i pour $i = 1, \ldots, N-1$.
Par définition, $v_n = a_n - \sum_{j=1}^{n-1} \frac{a_k \cdot v_j}{v_j \cdot v_j} v_j$, et donc

$$v_n \cdot v_i = a_n \cdot v_i - \sum_{j=1}^{n-1} \frac{a_n \cdot v_j}{v_j \cdot v_j} v_j \cdot v_i = a_n \cdot v_i - \frac{a_n \cdot v_i}{v_i \cdot v_i}$$

par hypothèse de récurrence. On en déduit que $v_n \cdot v_i = 0$ et donc que la famille $(v_1, \ldots v_n)$ est une base orthogonale.

2. De la relation (1.132), on déduit que :

$$a_k = v_k + \sum_{j=1}^{k-1} \frac{w_k \cdot v_j}{v_j \cdot v_j} v_j, \qquad q_k = \frac{v_k}{\|v_k\|},$$

et comme $v_j = \|v_j\| a_j$, on a bien :

$$a_k = \|v_k\| q_k + \sum_{j=1}^{k-1} \frac{a_k \cdot v_j}{\|v_j\|} q_j.$$

La k-ième colonne de A est donc une combinaison linéaire de la k-ème colonne de Q affectée du poids $\|v_k\|$ et des $k-1$ premières affectées des poids $\frac{a_k \cdot v_j}{\|l v_j\|}$. Ceci s'écrit sous forme matricielle $A = QR$ où R est une matrice carrée dont les coefficients sont $R_{k,k} = \|v_k\|$, $R_{j,k} = \frac{a_k \cdot v_j}{\|v_j\|}$ si $j < k$, et $R_{j,k} = 0$ si $j > k$. La matrice R est donc bien triangulaire supérieure et à coefficients diagonaux positifs.

3. Si A est inversible, par le procédé de Gram-Schmidt (1.132) on construit la matrice $Q = \begin{bmatrix} q_1 & q_2 & \ldots & q_n \end{bmatrix}$, et par la question 1.b, on sait construire une matrice R triangulaire supérieure à coefficients diagonaux positifs $A = QR$.

4. On a $a_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ et donc $q_1 = \frac{1}{2} \begin{bmatrix} \sqrt{2} \\ \sqrt{2} \end{bmatrix}$

Puis $a_2 = \begin{bmatrix} 4 \\ 0 \end{bmatrix}$ et donc $v_2 = a_2 - \frac{a_2 \cdot v_1}{v_1 \cdot v_1} v_1 = \begin{bmatrix} 4 \\ 0 \end{bmatrix} - \frac{4}{2} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -2 \end{bmatrix}$. Donc $\boldsymbol{q}_2 = \frac{1}{2} \begin{bmatrix} \sqrt{2} \\ -\sqrt{2} \end{bmatrix}$, et $Q = \frac{1}{2} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} \end{bmatrix}$.

Enfin, $R = \begin{bmatrix} \|v_1\| & \frac{a_2 \cdot v_1}{\|v_1\|} \\ 0 & \|v_1\| \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 2\sqrt{2} \\ 0 & 2\sqrt{2} \end{bmatrix}$, et $Q = \frac{1}{2} \begin{bmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & -\sqrt{2} \end{bmatrix}$.

Chapitre 3

Résolution d'équations et systèmes non linéaires

3.1 Racines de l'équation $f(x) = 0$

Définition 3.1.1 *Soit f une fonction de $\mathbb{R}$ dans $\mathbb{R}$ dont le domaine de définition est une partie D_f de $\mathbb{R}$. On dit que $\alpha \in D_f$ est une racine de l'équation*

$$f(x) = 0 \tag{1}$$

, si

$$f(\alpha) = 0 \tag{2}$$

Résoudre l'équation (1) c'est trouver tous les nombres réels α tels que (2) soit vérifiée.
En d'autres termes, on cherche à déterminer l'ensemble
$Z(f) = \{x \in D_f / f(x) = 0\}$, appelé les de f. $Z(f)$ est donc l'ensemble des racinees de $f(x) = 0$.
Il n'est pas toujours possible de résoudre complétement ce probléme pour toutes formes de fonctions f. $Z(f)$ peut en effet, avoir un grand nombre de structures possibles.

Exemples 2.1
1- Soit $f(x) = ax^2 + bx + c$, $a, b, c \in \mathbb{R}$, avec $D_f = \mathbb{R}$. Alors ker f contient au plus deux éléments et peut aussi être vide.
2- Soit $f(x) = \sin(x)$ et $D_f = \mathbb{R}_+$, alors les racines de l'équation $f(x) = 0$ sont en nombres infini dénombrable et $Z(f) = \{x \in \mathbb{R}_+ / x = k\pi,\ k = 0, 1, 2, 3, ...\}$
3- Pour f définie par

$$f(x) = \begin{cases} \sin(\frac{1}{x}) & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

Les racines de l'équation $f(x) = 0$ sont dans ce cas la demi droite négative $\mathbb{R}_-$ et les éléments de la suite $S = \left\{\frac{1}{k\pi};\ k = 1, 2, 3, ...\right\}$.
Ainsi $Z(f) = \mathbb{R}_- \cup \left\{x \in \mathbb{R} \ / \ x = \frac{1}{k\pi};\ k = 1, 2, 3, ...\right\}$
On peut remarquer qu'il existe (au moins) une racine dans S (différente de 0) aussi prés que l'on veut de 0. On dit que 0 est un point d'accumulation de la suite S.

Définition 3.1.2 *On dit qu'une racine α d'une équation $f(x) = 0$ est séparable si on peut trouver un intervalle $[a, b]$ tel que α soit la seule racine de cette équation dans $[a, b]$; ou encore si $Z(f) \cap [a, b] = \{\alpha\}$*
La racine α est alors dite séparée (on dit aussi racine isolée).

Remarque 3.1.1 *Dans les deux cas 1 et 2 de l'exemple 2.1 ci-dessus, toutes les racines sont séparables. Par contre dans le cas 3, les seules racines séparables sont les éléments de S. Les éléments de S qui sont les plus prés de 0 sont les plus difficile à séparer. La longueur de l'intervalle $[a, b]$ de la définition 2.2 devient en effet, de plus en plus petite au fur et à mesure que l'on s'approche de 0 dans S.*

3.2 Séparation des racines

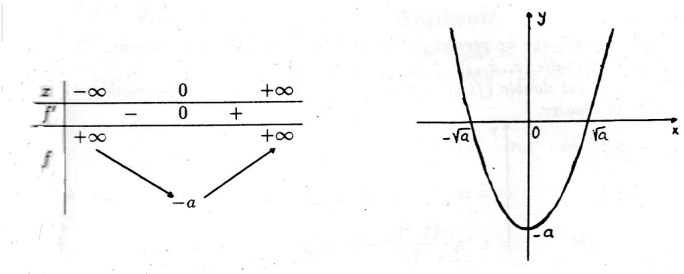
Il n'y a pas de méthode générale pour séparer les racines d'une équation $f(x) = 0$.
Pratiquement, en dehors de l'étude théorique directe de f si f est donnée analytiquement, on utilise deux types de méthodes : une méthode graphique et une méthode de balayage.

3.2.1 Méthode graphique

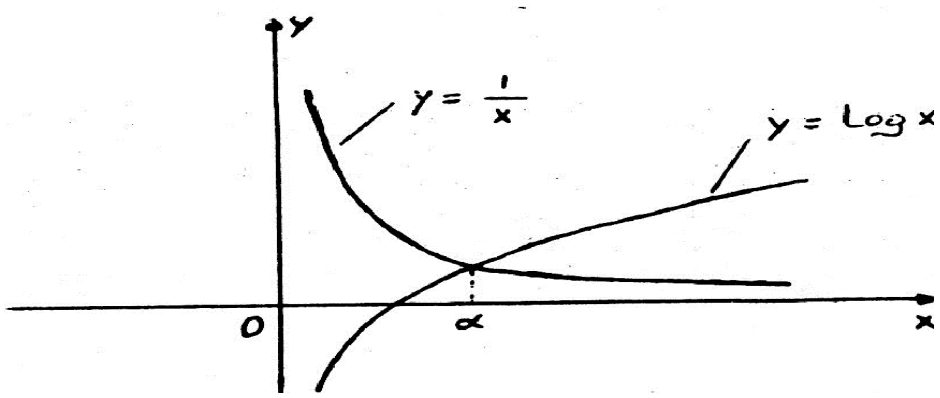
Soit on trace (expérimentalement ou par étude des variations de f) le graphe de la fonction f et on cherche son intersection avec l'axe Ox. Soit on décompose f en deux fonctions f_1 et f_2 simples à étudier, telles que : $f = f_1 - f_2$, et on cherche les points d'intersection des graphes de f_1 et f_2, dont les abscisses sont exactement les racines de l'équation $f(x) = 0$

Remarque 3.2.1 *On choisit souvent f_1 et f_2 de façon à ce que leur courbes soient des courbes connues.*

Exemples 2.2
1- *Soit à résoudre graphiquement l'équation : $x^2 - a = 0$, où $a > 0$, fixé. $(D = \mathbb{R})$*
Les variations et la courbe représentative de la fonction $f(x) = x^2 - a$ sont données par le tableau et le graphe suivants :



Bien que dans cet exemple les solutions soient évidement connues ($x = \pm\sqrt{a}$). on voit que l'intersection du graphe de la fonction $f(x) = x^2 - a$ avec l'axe Ox permet de localiser les racines de l'équation $f(x) = 0$.



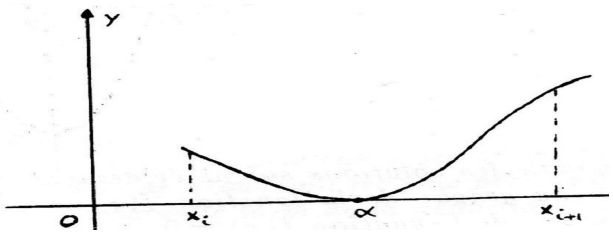
2- Soit l'équation $(*) \quad x \log x = 1, \quad x > 0 \quad (D = \mathbb{R}_+^*)$
Cette équation s'écrit encore sous la forme : $\log x = \frac{1}{x}$, En posant $f_1(x) = \log x$. $f_2(x) = \frac{1}{x}$ et $f(x) = f_1(x) - f_2(x) = \log x - \frac{1}{x}$, l'équation $(*)$ devient équivalente à $f(x) = f_1(x) - f_2(x) = 0$.
Les variations des fonctions f_1 et f_2 sont données par les courbes ci-dessous :

L'abscisse du point d'intersection des deux courbes permet de localiser la solution de $(*)$ et fournit même une (première) approximation de celle-ci (et ceci en utilisant, par exemple, du papier millimètré ou encore en graduant les deux axes).

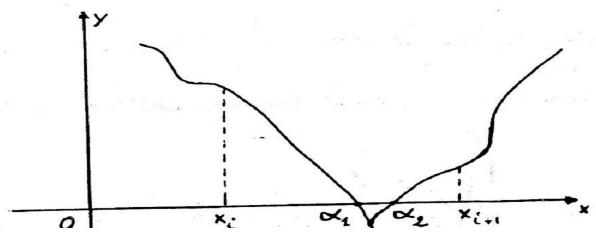
3.2.2 Méthode de balayage

On considère une suite croissante finie $\{x_i\}$, $(i = 0, 1, ..., n)$ de valeurs de x réparties sur l'intervale $[a, b]$ contenu dans le domaine de définition D de f. Si f est continue et si $f(x_i)f(x_{i+1}) < 0$ alors il existe entre x_i et x_{i+1} au moins une racine de $f(x) = 0$ (c'est le théorème classique des valeurs intermédiaires).
La méthode consiste donc à déterminer parmi les quantités $f(x_i)f(x_{i+1})$, $(i = 0, 1, ..., n)$ celles qui sont négatives.

Remarque 2.3
1- La méthode de balayage ne permet de conclure qu'à l'existence d'(au moins) une racine dans un intervalle $[x_i, x_{i+1}]$.
• Si une racine α est double ($f(\alpha) = f'(\alpha) = 0$ et $f"(\alpha) \neq 0$), cette méthode ne permet pas de la séparer.



• De même, si deux racines sont trés voisines, on risque de ne pas les séparer dans le processus ci-dessus comme le montre la figure suivante :



2- L'intervalle de départ $[a, b]$ doit être suffisamment grand afin de contenir les racines éventuelles de l'équation $f(x) = 0$; mais sauf dans des cas particulier, on ne peut pas estimer correctement sa longueur.

3.3 Approximation des racines : Méthodes itérative

Définition 3.3.1 *On appelle méthode itérative un procédé de calcul de la forme :*

$$x_{k+1} = F(x_k), \qquad k = 0, 1, 2, ... \qquad (a)$$

dans lequel on part d'une valeur (approchée) x_0 pour calculer x_1, puis à l'aide de x_1 on calcule x_2, etc...
La formule (a) est dite formule de récurrence.
Le procédé est dit convergent si la suite (x_k) est convergente.

Parmi les méthode numérique en général et les méthode itératives en particulier, les plus puissantes permettant la résolution approchée des équations de la forme $f(x) = 0$ figure la méthode suivante :

3.3.1 Méthode de Newton-Raphson

Notons par x^* une racine (exacte) recharchée et par x_0 une valeur approchée de x^*.
On suppose que f est de classe C^1 au voisinage de x^*. Le développement de Taylor d'ordre deux de f nous donne :

$$f(x^*) = f(x_0) + f'(x_0)(x^* - x_0) + \frac{f"(\xi)}{2}(x^* - x_0)^2 \qquad \text{où} \quad \xi \in (x^*, x_0)$$

et comme $f(x^*) = 0$, on supposant $f(x_0) \neq 0$, on aurra :

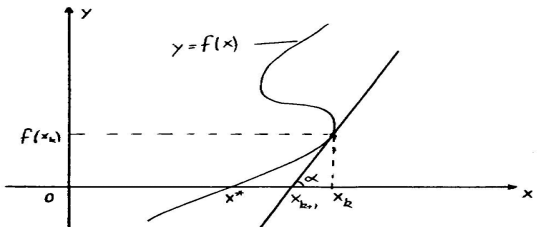
$$x^* = x_0 - \frac{f(x_0)}{f'(x_0)} - \frac{f"(\xi)}{2f'(x_0)}(x^* - x_0)^2 \qquad (b)$$

et en négligeant le reste $R_2 = -\frac{f"(\xi)}{2f'(x_0)}(x^* - x_0)^2$, la quantité $x_0 - \frac{f(x_0)}{f'(x_0)}$ dans (b) qu'on notera x_1, constitue alors une valeur approchée améliorée de x^*.
En itérant le procédé on trouve la formule de récurrence :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \qquad k = 0, 1, 2, ...$$

qu'on appelle : **Formule de récurrence de Newton-Raphson.**

Géométriquement :



$tg(\alpha) = \frac{f(x_k)}{\triangle x_k} = f'(x_k)$ où $\triangle x_k = x_k - x_{k+1}$

et donc $\triangle x_k = \frac{f(x_k)}{f'(x_k)} = x_k - x_{k+1} \Rightarrow x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$

Commentaire 2.1 Une méthode itérative ne présente de l'intérêt que si elle est convergente (vers les valeurs recherchées), et c'est dans ce sens que l'on étudiera la convergence de la méthode de Newton-Raphson de manière plus approfondie plus bas.

Exemple 2.3 (Calcul itérative de la valeur inverse)
Soit $a > 0$ donné. On désire calculer son inverse $\frac{1}{a}$.
On peut ramener le problème à la résolution de l'équation $f(x) = \frac{1}{x} - a = 0$. Selon la formule de Newton-Raphson, nous avons :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = 2x_k - a\ x_k^2$$

Remarquons que dans cette formule de récurrence on ne fait pas de division.
Pour $a = 7$ et on partant de $x_0 = 0.2$, le calcul donne :
$x_1 = 0.12$; $x_2 = 0.1392$; $x_3 = 0.14276352$; $x_4 = 0.142857081$; ... et cette suite converge vers $\frac{1}{7} = 0.142857142...$

Exemple 2.4 (Comparaison de différentes formules de récurrence)
Soit à résoudre l'équation $x = e^{\frac{1}{x}} = \exp(\frac{1}{x})$ (1)
L'équation (1) peut être transformée de différentes manières afin de se mettre sous la forme $f(x) = 0$.
i) $x = \exp(\frac{1}{x}) \Leftrightarrow x - \exp(\frac{1}{x}) = f(x) = 0$
La formule de Newton-Raphson donne :

$$x_{k+1} = x_k - x_k^2 \frac{x_k - \exp(\frac{1}{x_k})}{x_k^2 - \exp(\frac{1}{x_k})} \qquad (i)$$

ii) Posons $y = \frac{1}{x}$.
L'équation (1) devient alors $\frac{1}{y} - \exp(y) \Leftrightarrow 1 - y * \exp(y) = f(y) = 0$ et la formule de Newton-Raphson donne encore :

$$y_{k+1} = y_k + \frac{\exp(-y_k) - y_k}{1 + y_k} \qquad (ii)$$

iii) $x = \exp(\frac{1}{x}) \Leftrightarrow \log(x) = \frac{1}{x} \Leftrightarrow 1 - x\log(x) = f(x) = 0$ et la formule de Newton-Raphson donne enfin :

$$x_{k+1} = x_k + \frac{1 - x_k \log(x_k)}{1 + \log(x_k)} \qquad (iii)$$

Nous disposons ainsi de trois formules récurrentes différentes pour le même problème, et on constate que les formules (ii) et (iii) sont mieux adaptées au calcul que la formule (i)

En étudiant les variations de f, par exemple $f(x) = x - \exp(\frac{1}{x})$, on constate qu'il n'existe qu'une seule racine x^* de (1), et qu'elle se trouve au voisinage de $x_0 = 1.8$.
Et delà :
Pour le (ii) : En partant de $y_0 = 0.6 \simeq \frac{1}{1.8}$, nous obtenons :
$y_1 = 0.568007$; $y_2 = 0.567144$; $y_3 = 0.567143$; etc... et $x^* \simeq \frac{1}{y_3} \simeq 1.763223$
Pour le (iii) : En partant de $x_0 = 1.8$:
$x_1 = 1.763461$; $x_2 = 1.763223$; $x_3 = 1.763223$; etc... et $x^* \simeq 1.763223 = x_3$

Critère d'arrêt dans la méthode de Newton-Raphson

Soit x^* une racine isolée de l'équation $f(x) = 0$; x_0 une approximation de x^* et (x_k) la suite des approximations obtenue a l'aide de la formule de récurence de Newton-Raphson. On suppose que f est de classe C^∞ au voisinage de x^*.
Le développement de Taylor a l'ordre 1 donne, pour $k \in \mathbb{N}$:

$$f(x_k) = f(x^*) + f'(\xi)(x_k - x^*), \qquad \xi \in (x_k, x^*)$$

donc $\quad x_k - x^* = \frac{f(x_k)}{f'(\xi)}, \quad$ où $\frac{f(x_k)}{f'(x_k)} \in (x_k, x^*)$

D'autre part : $\quad x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} \Longrightarrow x_{k+1} - x_k = -\frac{f(x_k)}{f'(x_k)}$

et on fait l'approximation $f'(\xi) \simeq f'(x_k)$ pour obtenir enfin : $|x_k - x^*| \simeq |x_{k+1} - x_k|$.
Ainsi, si on veut calculer une approximation de x^* avec n décimales exactes, il suffit d'aller dans les itérations jusqu'a ce que k vérifie :

$$|x_{k+1} - x_k| \leq 0.5\ 10^{-n}.$$

3.3.2 Méthode de Newton-Raphson pour deux inconnues

Soit à résoudre le systéme d'équations non linéaires :

$$\begin{cases} F(x, y) = 0 \\ G(x, y) = 0 \end{cases}$$

où F et G sont des fonctions données des variables indépendantes x et y.
Soit $(x_{0,}y_0)$ une valeur approchée déune solution exacte (x^*, y^*).

Posons :

$$\begin{cases} x^* = x_0 + \varepsilon_0 \\ y^* = y_0 + \lambda_0 \end{cases} \quad \text{où } \varepsilon_0 \text{ et } \lambda_0 \text{ sont les erreurs absolues de } x_0 \text{ et } y_0.$$

En faisant un développement de Taylor à l'ordre 1 des deux fonctions F et G au point (x_0, y_0), on obtient :

$$0 = F(x^*, y^*) = F(x_0 + \varepsilon_0, y_0 + \lambda_0) = F(x_0, y_0) + \frac{\partial F}{\partial x}(x_0, y_0).\varepsilon_0 + \frac{\partial F}{\partial x}(x_0, y_0).\lambda_0 + R_0$$

où R_0 = reste = termes d'ordres supérieurs à 1 en ε_0 et λ_0.
et

$$0 = G(x^*, y^*) = G(x_0 + \varepsilon_0, y_0 + \lambda_0) = G(x_0, y_0) + \frac{\partial G}{\partial x}(x_0, y_0).\varepsilon_0 + \frac{\partial G}{\partial x}(x_0, y_0)\lambda_0 + R_1$$

où R_1 = reste = termes d'ordres superieurs à 1 en ε_0 et λ_0.
Ou bien sous forme matricielle :

$$\begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} F \\ G \end{pmatrix}_{(x_0,y_0)} + \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}_{(x_0,y_0)} \cdot \begin{pmatrix} \varepsilon_0 \\ \lambda_0 \end{pmatrix} + \begin{pmatrix} R_0 \\ R_1 \end{pmatrix}$$

On suppose l'existence de $\begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}^{-1}$, et donc :

$$\begin{pmatrix} \varepsilon_0 \\ \lambda_0 \end{pmatrix} = \begin{pmatrix} x^* \\ y^* \end{pmatrix} - \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = - \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}^{-1}_{(x_0,y_0)} \cdot \begin{pmatrix} F \\ G \end{pmatrix}_{(x_0,y_0)} + \begin{array}{c}\text{termes dordre supér-} \\ \text{ieurs à 1en } \varepsilon_0 \text{ et } \lambda_0.\end{array}$$

delà :

$$\begin{pmatrix} x^* \\ y^* \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}^{-1}_{(x_0,y_0)} \cdot \begin{pmatrix} F \\ G \end{pmatrix}_{(x_0,y_0)} + \begin{array}{c}\text{termes dordre supérieurs à} \\ 1 \text{ en } \varepsilon_0 \text{ et } \lambda_0.\end{array}$$

Et en négligeante le reste qui est formé des termes d'ordres supérieurs à 1, la quantité :

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}^{-1}_{(x_0,y_0)} \cdot \begin{pmatrix} F \\ G \end{pmatrix}_{(x_0,y_0)}$$

devient une approximation (x_1, y_1) de la valeur exacte (x^*, y^*).
Ainsi :

$$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}^{-1}_{(x_0,y_0)} \cdot \begin{pmatrix} F \\ G \end{pmatrix}_{(x_0,y_0)}$$

et en itérant le procédé, on trouve **la formule de récurrence de Newton-Raphson pour deux inconnues :**

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_0 \\ y_0 \end{pmatrix} - \begin{pmatrix} \frac{\partial F}{\partial x} & \frac{\partial F}{\partial y} \\ \frac{\partial G}{\partial x} & \frac{\partial G}{\partial y} \end{pmatrix}^{-1}_{(x_k,y_k)} \cdot \begin{pmatrix} F \\ G \end{pmatrix}_{(x_k,y_k)}, \quad k = 0, 1, 2, ...$$

Exemples 2.5 Soit à résoudre l'équation : $x = 2\sin x \qquad (a)$
$1^{\text{éme}}$ méthode : On pose $f\ (x) = x - 2\sin x$, et en utilisant la méthode récurrente de Newton-Raphson pour une seule variable, afin de résoudre l'équations $f(x) = 2\sin x = 0$, on aura :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)} = x_k - \frac{x_k - 2\sin x_k}{1 - 2\cos x_k}, \quad k = 0, 1, 2, ...$$

$2^{\text{éme}}$ méthode : On décompose (a) en un système de deux équations

$$\begin{cases} y = x \\ y = 2\sin x \end{cases} \Longleftrightarrow \begin{cases} F(x,y) = y - x = 0 \\ G(x,y) = y - 2\sin x = 0 \end{cases}$$

et la formule de récurrence de Newton-Raphson pour deux inconnues devient :

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} = \begin{pmatrix} x_k \\ y_k \end{pmatrix} - \begin{pmatrix} -1 & 1 \\ -2\sin x_k & 1 \end{pmatrix}^{-1} \cdot \begin{pmatrix} y_k - x_k \\ y_k - 2\sin x_k \end{pmatrix}, \quad k = 0, 1, 2, ...$$

3.3.3 La méthode de Newton-Raphson et les polynôme

On suppose que f est un polynôme P_n de dégré n à coefficients réels, n'ayant que des racines distinctes :

$$f(x) = P_n(x) = a_0x^n + a_1x^{n-1} + ... + a_{n-1}x + a_n, \qquad a_0 \neq 0$$

Question : Estimer les racines réelles de l'équation non linéaire $P_n(x) = 0$.

Définition 3.3.2 *On appelle suite de sturm la suite définie par :*

$$\begin{cases} S_0(x) = P_n(x) \\ S_1(x) = P_n'(x) \\ S_2(x) = -Reste(S_0(x)/S_1(x)) \\ \quad \vdots \\ S_i(x) = -Reste(S_{i-2}(x)/S_{i-1}(x)) \\ \quad \vdots \\ S_n(x) = -Reste(S_{n-2}(x)/S_{n-1}(x)) \end{cases}$$

Théorème 3.3.1 *(Théorème de Sturm : Nombres de racines réelles)*
Le nombre de solutions réelles (qui sont supposées simples) de l'équation $P_n(x) = 0$ est égale à $N(a) - N(b)$ ou $N(\xi)$ est le nombre de changement de signe de la suite $\{S_i(\xi)\}$. Les réelles a et b étant les extrémités de l'intervalle contenant les racines.

Théorème 3.3.2 (Localisation) *Les racines réelles de l'équation $P_n(x) = 0$ sont contenues dans l'intervalle $]-T, T[$ avec $T = 1 + \dfrac{1}{|a_0|}\left[\max_{i=\overline{1,n}} |a_i|\right]$.*

Propriété 2.1 Si la suite S_i est d'ordre p ie : $S_{p+1}(x) = 0, \quad (p < n)$,
alors : les racines multiples de $P_n(x)$ sont les racines simples de $S_p(x)$.

Remarque 3.3.1 *La divergence, dans le cas des polynômes, de la méthode de Newton-Raphson est due à deus raisons :*
1- Soit au mauvais choix de x_0.
2- Soit qu'on n'a pas de racines réelles.

Exemples 2.6 Résoudre par la méthode de Newton-Raphson, l'équation :

$$x^3 - 3x^2 + x - 3 = 0$$

• Localisation des solutions :
$T = 1 + \dfrac{1}{|a_0|}\left[\underset{i=\overline{1,n}}{max}|a_i|\right] = 4$. Et donc les solution de l'équation considérée, si elles existent, se trouvent contenues dans l'intervalle $]-4, 4[$.
• Nombre de racines :
Conciderons pour celà la suite de Sturm associée au polynôme $P_3(x) = x^3 - 3x^2 + x - 3$.

$$\begin{cases} S_0(x) = P_3(x) = x^3 - 3x^2 + x - 3 \\ S_1(x) = P_3'(x) = 3x^2 - 6x + 1 \\ S_2(x) = -\text{Reste}(S_0(x)/S_1(x)) = -\text{Reste}(x^3 - 3x^2 + x - 3/3x^2 - 6x + 1) \\ \quad = x + 2 \quad \text{(à un facteur multiplicatife canstant prés)} \\ S_3(x) = -\text{Reste}(S_1(x)/S_2(x)) = -\text{Reste}(3x^2 - 6x + 1/x + 2) = -25 \end{cases}$$

Nous avons alors :

$S_0(x)$	$S_1(x)$	$S_2(x)$	$S_3(x)$	$N(-)$
$-$	$+$	$-$	$-$	2
$+$	$+$	$+$	$-$	1

Le nombre de racines réelles de l'équation concidérée est donc $2 - 1 = 1$.
• Résoudre par la méthode de Newton-Raphson :
Pour cela précisons encore plus l'intervalle contenant la racine cherchée x^*.
Nous avons $P_3(0) = -3$ et $P_3(4) \geq 0$, alors $x^* \in [0\ , 4]$. Prenons par exemple $x_0 = 0$. Nous avons alors :
$x_1 = x_0 - \dfrac{P_3(x_0)}{P_3'(x_0)} = 3$, et $x_2 = x_1 - \dfrac{P_3(x_1)}{P_3'(x_1)} = x_1$. De même $x_k = 3 = x_1$.
Ainsi, nous obtenons une suite stationnaire qui donne dans ce cas, non pas une racine approchée mais carrément la racine exacte égale à 3.

3.3.4 Méthode de point fixe

Définition 3.3.3 *Soit $f : \mathbb{R} \to \mathbb{R}$ une application continue. On dit que $x^* \in \mathbb{R}$ est un point fixe de f si $f(x^*) = x^*$.*

Commentaire 2.2 La résolution d'un problème à l'aide d'une formule de récurrence $x_{k+1} = f(x_k)$, $k \in \mathbb{N}$ peut etre considérée comme détemination d'un point fixe de la fonction f.

En effet : Soit (x_k) la suite défnie par $x_{k+1} = f(x_k)$, On suppose qu'elle converge vers une valeur qu'on notera x^*. Par passage à limite dans l'expression $x_{k+1} = f(x_k)$, on obtient :

$$\lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} f(x_k) \iff \lim_{k\to\infty} x_{k+1} = f(\lim_{k\to\infty} x_k) \text{ (par continuité de } f)$$

$$\Longleftrightarrow \quad x^* = f(x^*) \quad (\text{car } \lim_{k\to\infty} x_k = x^*)$$

et donc x^*est un point fixe de f.

Théorème 3.3.3 (Théorème du point fixe) *Supposons que f est définie sur l'intervalle* $[a\ ,b]$ *et satisfait aux conditions suivantes :*
i) f $([a\ ,b]) \subset [a,b]$ *ie :* $\forall\ x \in [a,b]$, $a \leq f(x) \leq b$
ii) f *est contractante ie :*$\exists L \in \mathbb{R}$, $0 \leq L < 1$ *tel que*

$$\forall x, y \in [a,b] : |f(x) - f(y)| \leq L\,|x - y|$$

Alors : f *admet un point fixe unique* $x^* \in [a,b]$. *De plus, pour tout point* $x_0 \in [a,b]$, *la suite* (x_k) *définie par* $x_{k+1} = f(x_k)$ *converge vers* x^*.

Commentaire 2.3 La condition $ii)$ entraine(x_k) :
1. La continuité de f dans $[a,b]$
2. En divisant l'inégalité par $|x - y|$ et en passant à la limite quand y tend vers x, on obtient :$\left|f^{'}(x)\right| \leq L < 1$, et ceci en supposant que f est dérivable au point x.

Démonstration du théorème du point fixe

Existence du point fixe : Soient x_0 un point initial dans $[a,b]$ et (x_k) la suite associée à f .
Pour montrer que la suite (x_k) converge, on va montrer qu'elle est de Cauchy.
Le point fixe sera alors la limite de (x_k).
Soit $k \in \mathbb{N}$, alors
$|x_{k+1} - x_k| = |f(x_k) - f(x_{k-1})| \leq L\ |x_k - x_{k-1}|$.
et par récurrence, on démontre que :

$$|x_{k+1} - x_k| \leq L\,|x_k - x_{k-1}| \leq \cdots \leq L^k\,|x_1 - x_0|$$

et pour $n > k$, en écrivant :

$$x_n - x_k = (x_n - x_{n-1}) + (x_{n-1} - x_{n-2)} + \cdots + (x_{k+1} - x_k),$$

nous aurons :

$$\begin{aligned} |x_n - x_k| &\leq |x_n - x_{n-1}| + |x_{n-1} - x_{n-2}| + \cdots + |x_{k+1} - x_k| \\ &\leq (L^{n-1} + L^{n-2} + \cdots + L^k)\,|x_1 - x_0| \\ &\leq L^k \frac{1 - L^{n-k}}{1 - L}\,|x_1 - x_0| \end{aligned}$$

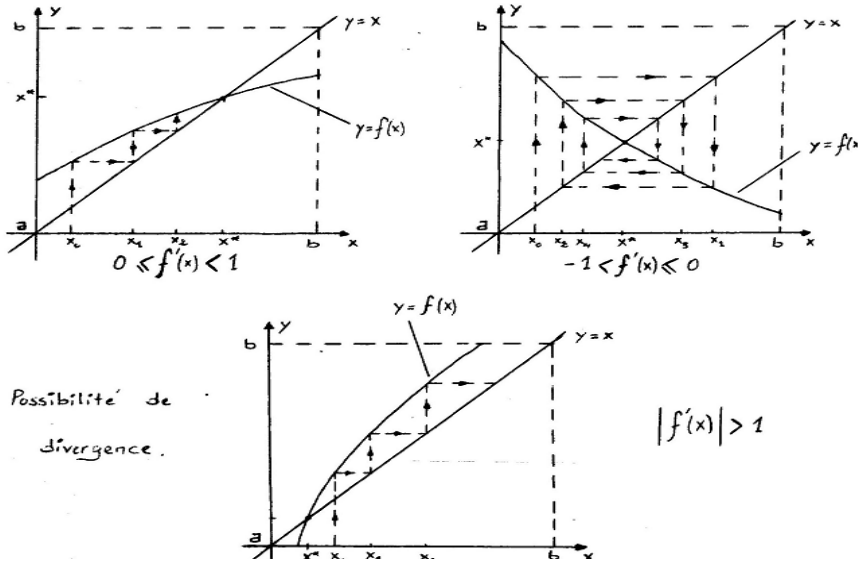
$$\leq \frac{L^k}{1-L}\,|x_1 - x_0|$$

et delà on déduit que (x_k) est une suite de Cauchy, donc elle converge vers $x^* \in [a,b]$.
Montrons que x^* est un point fixe.
En effet : l'égalité $x_{k+1} = f(x_k)$ et la continuité de f entrainent que
$x^* = f(x^*)$.
Unicité de point fixe : Supposons qu'il existe un autre point fixe y^* avec
$y^* \neq x^*$.
Alors :

$$0 < |y^* - x^*| = |f(y^*) - f(x^*)| \leq L\,|y^* - x^*| < |y^* - x^*| \quad \text{car } L \leq 1$$

ce qui est absurde x^* est donc unique.

Interpprétation géométrique du théorème du point fixe



Critére d'arrêt $n°1$

Soit $f : [a, b] \to \mathbb{R}$ satisfaisant aux hypothèses du théorème du point fixe, et soit (x_k) la suite des approximations définie par $x_{k+1} = f(x_k), \quad k = 0, 1, 2, ...$
Pour $k, n \in \mathbb{N}$, $n > k$:

$$|x_k - x_n| = |f(x_{k-1}) - f(x_{n-1})| \leq L\,|x_{k-1} - x_{n-1}|$$

et par récurrence :

$$|x_k - x_n| \leq L\,|x_{k-1} - x_{n-1}| \leq L^2\,|x_{k-2} - x_{n-2}| \leq \cdots \leq L^k\,|x_0 - x_{n-k}| \leq L^k\,|b - a|$$

En passant à la limite sur n, on obtient : $|x_k - x^*| \leq L^k(b - a)$.
Ainsi, si on désire calculer une approximation x_k de x^* avec n décimales exactes, il suffit d'arrêter les itérations à k vérifiant :

$$k \geq \frac{\log(\frac{0{,}5\ 10^{-n}}{b-a})}{\log L}$$

Critére d'arrêt $n°2$

Soit $k \in \mathbb{N}$:
$|x_{k+1} - x^*| = |f(x_k) - f(x^*)| \leq L\,|x_k - x^*| \leq L\,|x_k - x_{k+1}| + L\,|x_{k+1} - x^*| \qquad \leq L\,|x_k - x_{k+1}| +$
$L\,|x_{k+1} - x^*|$
$\Rightarrow |x_{k+1} - x^*|\,(1 - L) \leq L\,|x_k - x_{k+1}|$
$\Rightarrow |x_{k+1} - x^*| \leq \frac{L}{1-L}\,|x_k - x_{k+1}|$
et donc pour avoir n décimales exactes, on arrête les itérations lorsque :

$$|x_k - x_{k+1}| \leq \frac{1 - L}{L} 0.5 \ \ 10^{-n}$$

Remarque 3.3.2 *Généralement L est assez petit pour que $2L < 1$, et delà $L < 1 - L$, et donc $1 < (1 - L)/L$. Conséquence : au lieu du critère d'arrête ci dessus on impose la condition*

$$|x_k - x_{k+1}| \leq 0.5 \ \ 10^{-n}.$$

Exemple 2.7 (Résolution itérative de l'équation quadratique)
Soit à résoudre, à l'aide de la méthode du point fixe, l'équation

$$(a) \qquad \varphi(x) = x^2 - 100x + 1 = 0$$

On remarquera que (a) admet deux racines réelles x_1^* et x_2^*, l'une au voisinage de 10^{-2} et l'autre au voisinage de 10^2.
Pour x_1^* : En tenant compte du fait que $x_1^* \simeq 10^{-2}$, l'équation (a) peut se mettre sous la forme :

$x = f(x) = \frac{1}{100}(x^2 + 1)$ et la formule de récurrence est donnée par :
$x_{k+1} = f(x_k) = \frac{1}{100}(x_k^2+1)$. Comme $f'(x) = \frac{2}{100}x = 0.02x$ est petit, prenons l'intervalle $[10^{-2}, 1]$ comme voisinage de x_1^*.
Nous avons, en effet :

$$\left.\begin{array}{c}\varphi(10^{-2}) = 10^{-4} > 0 \\ \varphi(1) = -98 < 0\end{array}\right) \Rightarrow x_1^* \in [10^{-2}, 1].$$

et on vérifie aisement que :
i)- $f([10^{-2}, 1]) = [f(10^{-2}), f(1)] \subset [10^{-2}, 1]$
ii)- Pour $L = 0.02$, f est contractante
Ainsi : $\forall x_0 \in [10^{-2}, 1]$, la suite (x_k) converge vers $x_1^* \in [10^{-2}, 1]$.
Prenons par exemple, $x_0 = 1$. On calcule :
$x_1 = 0.02\,; x_2 = 0.010004\,; x_3 = 0.0100010008\,; x_4 = 0.0100010020017$ donc : $x_1^* \simeq 0.010001002002$ (11 c.s.e)

Pour x_2^* : En tenant compte du fait que $x_2^* \simeq 10^2$, l'équation (a) peut se méttre sous la forme : $x^2 = 100x - 1$, puis en divisant par x en trouve la formule de récurrence :
$x_{k+1} = 100 - \frac{1}{x_k} = f(x_k)$, où $f(x) = 100 - \frac{1}{x}$. Comme $f'(x) = \frac{1}{x^2}$ est petit pour x assez grand, prenons l'intervalle $[10, 10^2]$ comme voisinage de x_2^*.
Nous avons, en effet :

$$\left.\begin{array}{c}\varphi(10) = -990 < 0 \\ \varphi(10^2) = 1 > 0\end{array}\right) \Rightarrow x_2^* \in [10, 10^2].$$

et on vérifie que :
i)- $f([10, 10^2]) = [f(10), f(10^2)] \subset [10, 10^2]$
ii)- Pour $L = 0.01$, f est contractante
Ainsi : $\forall x_0 \in [10, 10^2]$, la suite (x_k) converge vers $x_2^* \in [10, 10^2]$.
Prenons par exemple, $x_0 = 10$. On calcule : (avec 5 décimales exactes)
$x_1 = 99.9\,; x_2 = 99.989990\,; x_3 = 99.989990\,; x_4 = 99.989999$
donc : $x_2^* \simeq 99.99000 \pm 10^{-5}$

3.3.5 Accélération de la convergence

Lemme 3.3.1 *On suppose que f' existe. Plus $f'(x^*)$ est petite, plus la convergence est rapide.*

Preuve.
On pose E
$_k = x_k - x^*$, et un développement de Taylor de f nous donne :

$$x_{k+1} = f(x_k) = f(x^* + E_k) = f(x^*) + E_k f'(x^*) + \frac{E_k^2}{2!} f''(x^*) + \cdots$$

En négligeant les termes d'ordre supérieur à 1 en E_k, il vient :
$x_{k+1} - f(x^*) = x_{k+1} - x^* \simeq f'(x^*).E_k \Rightarrow = E_{k+1} = x_{k+1} - x^* \simeq f'(x^*).E_k$

Comme (E_k) est une suite convergente vers zéro (0), on voit que plus $f'(x^*)$ est petit, plus la convergence est rapide.

Accélération

Soit $\lambda \neq -1$. posons $G(x) = \dfrac{\lambda x + f(x)}{1+\lambda}$.
Comme $f(x) = x \Leftrightarrow G(x) = x$: alors, x^* est un point fixe de f si et suelement si il l'est pour G.
On a $G'(x) = \dfrac{\lambda + f'(x)}{1+\lambda}$; donc $G'(x^*) = \dfrac{\lambda + f'(x^*)}{1+\lambda}$. Il suffit alors de prendre λ voisin de $-f'(x^*)$ pourvu que $f'(x^*)$ soit petit. Et donc d'aprés le lemme 2.1, la convergence sera plus rapide que celle donnée par $f(x) = x$.

Algorithme d'accélération ■

1. Calculer x_0, x_1,..., x_n par la méthode lente :

$$\begin{cases} x_0 \in [a,b] \\ x_{k+1} = f(x_k),\ k = 0,1,...,n-1 \end{cases}$$

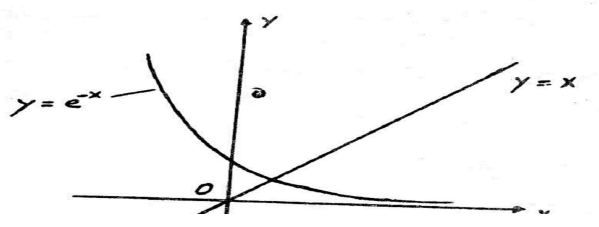
   où n est choisi selon chaque cas
2. On pose $\lambda = -f'(x_n)$
3. On calcule les approximations suivantes par

$$x_{k+1} = G(x_k), \quad k = n,n+1,... \quad \text{où} \quad G(x) = \frac{\lambda x + f(x)}{1+\lambda}$$

Exemple 2.8 Soit à résoudre l'équation

$$(1) \qquad x = e^{-x}.$$

L'équation (1) admet un point fixe dans $\mathbb{R}$.
On vérifie même qu'il est dans
l'intervalle [0.5,0.6].



1- Posons $x_0 = 0.5$. On a alors, $x_1 = f(x_0) = e^{-0.5} = 0.6065306 \simeq 0.61$ et $x_2 = f(x_1) = e^{-0.61} =$

$0.5433508 \simeq 0.54$
On constate que les différences $|x_0 - x_1|$ et $|x_1 - x_2|$ (etc..) sont assez importantes, alors on arrête là la méthode est lente et :
2- $n = 2 : -f'(x_2) = e^{-0.54} = 0.5827482 \simeq 0.58 = \lambda$
3- Puis, on pose : $G(x) = \frac{0.58x + e^{-x}}{1.58}$ et si on veut calculer x^* avec $sept(07)$ décimales exactes, on applique, par exemple, le critére d'arrêt $n°1$, pour l'algorithme :

$$\begin{cases} \widetilde{x}_0 = 0.54 \in [0.5, 0.6] \\ \widetilde{x}_{k+1} = G(\widetilde{x}_k) = \dfrac{0.58\widetilde{x}_k + e^{-\widetilde{x}_k}}{1.58} \end{cases}$$

alors k védifie $k \geq \dfrac{\log(\dfrac{0.5\ 10^{-7}}{0.6 - 0.5})}{\log 0.02} \simeq 3.7$ où $0.02 \geq \max\limits_{[0.5,0.6]} |G'(x)|$ et donc, on prend $k = 4$.
Ainsi :

$$\begin{aligned} \widetilde{x}_1 &= G(\widetilde{x}_0) = 0.56705684 \\ \widetilde{x}_2 &= G(\widetilde{x}_1) = 0.56714259 \\ \widetilde{x}_3 &= G(\widetilde{x}_2) = 0.56714328 \\ \widetilde{x}_4 &= G(\widetilde{x}_3) = 0.56714329 \end{aligned}$$

d'où $x^* = 0.5671433 \pm 0.5\ 10^{-7}$
Question : Déterminer le nombre d'itérations correspondant au cas de la méthode lente (Réponse : $k = 30$).

3.3.6 Convergence de la méthode de newton-Raphson

Reprenons la formule de Newton-Raphson :

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad k = 0, 1, 2, ...$$

On a donc : $x_{k+1} = g(x_k)$, où $g(x) = x_k - \dfrac{f(x_k)}{f'(x_k)}$.
Et comme $g'(x) = 1 - \dfrac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \dfrac{f(x)f''(x)}{(f'(x))^2}$; alors pour
$x = x^*$, x^* étant une racine séparée de l'équation $f(x) = 0$, on a $g'(x) = 0$ (puisque $f(x^*) = 0$) et delà $|g'(x)| < 1$ au "voisinage" de x^*. Conséquence : les conditions $i)$ et $ii)$ du théorème du point fixe sont alors réalisées, et par suite, on a convergence de la suite (x_k) vers x^*.
Et c'est dans ce but qu'on choisit comme voisinage de x^* un intervalle $[a, b]$ telle que :

1. $x^* \in [a, b]$
2. f est de classe C^2 sur $[a, b]$
3. $f' \neq 0$ sur $[a, b]$

Et on verfiera ci-après qu'avec une condition supplémentaire sur l'intervalle $[a, b]$, la suite (x_k)

converge vers x^*, $\forall x_0 \in [a,b]$.

Commentaire 2.4

- f étant de classe C^2 sur $[a,b]$, f'' est continue sur $[a,b]$ et donc bornée sur $[a,b]$ par une constante $M > 0$ de manière à ce qu'on ait : $\forall x \in [a,b]$, $|f''(x)| \leq M$

- f' ne s'annulant pas sur $[a,b]$, il existe $m > 0$ tel que : $\forall\ x \in [a,b]$, $|f'(x)| \geq m$

Sous les hypothèses 1), 2) et 3), faisons un developpement de taylor de f à l'ordre 1 au voisinage de x_k. On obtien :

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(\xi_k)(x - x_k)^2, \text{ où } \xi_k \in (x, x_k)$$

comme $f(x) = 0$ pour $x = x^*$, il s'ensuit :

$$0 = f(x_k) + f'(x_k)(x^* - x_k) + \frac{1}{2}f''(\xi_k)(x^* - x_k)^2, \text{ où } \xi_k \in (x, x_k)$$

en divisant par $f'(x_k) \neq 0$ et on translatant, on aura :

$$-\frac{f(x_k)}{f'(x_k)} = x^* - x_k + \frac{f''(\xi_k)(x^* - x_k)^2}{2f'(x_k)}$$

or : $\quad x_{k+1} = x_k - \dfrac{f(x_k)}{f'(x_k)} \Rightarrow -\dfrac{f(x_k)}{f'(x_k)} = x_{k+1} - x_k$

D'où, après simplification : $x_{k+1} - x^* = \dfrac{f''(\xi_k)(x^* - x_k)^2}{2f'(x_k)}$

et comme $|f''(x)| \leq M$ et $|f'(x)| \geq m > 0$, on aboutit alors à :

$$|x_{k+1} - x^*| \leq \frac{M\,|x^* - x_k|^2}{2m} \quad (*)$$

On constate que l'erreur absolue de la $(k+1)^{eme}$ approximation est proportionnelle au carré de l'erreur absolue de la k^{eme} approximation. On dit que la méthode de Newton-Raphson est une méthode itérative du deuxième ordre.

Posons $C = \dfrac{M}{2m}$. L'inégalité (*) entraîne par induction :

$$|x_{k+1} - x^*| \leq C\,|x_k - x^*|^2 \leq C^3\,|x_{k-1} - x^*|^4 \leq ... \leq C^{2^{k+1}-1}\,|x_0 - x^*|^{2^{k+1}}$$

ou encore $\quad |x_k - x^*| \leq C^{2^k-1}\,|x_0 - x^*|^{2^k}$

comme : $|x_0 - x^*| \leq |b - a| = b - a$, alors $|x_k - x^*| \leq C^{2^k-1}(b-a)^{2^k}$

ou encore :

$$|x_k - x^*| \leq \frac{[C(b-a)]^{2^k}}{C},\ k = 0, 1, 2, ...$$

Il y a donc convergence vers x^* si $C(b-a) = \frac{M}{2m}(b-a) < 1$, et c'est pour cette raison que l'on choisit $[a, b]$ de longueur $b - a < \frac{2m}{M}$.

Conclusion :
Si en plus des conditions 1), 2) et 3), on choisit $[a, b]$ tel que $b - a < \frac{2m}{M}$, alors : $\forall\ x_0 \in [a, b]$, l'algorithme de Newton-Raphson converge (vers x^*).

Remarque 3.3.3 (choix de l'approximation initiale x_0) *Le point x_0 est généralement choisi de manière à ce que $f(x_0)f''(x_0) > 0$ (en particulier $x_0 = a$ ou $x_0 = b$ si cette condition est satisfaite).*

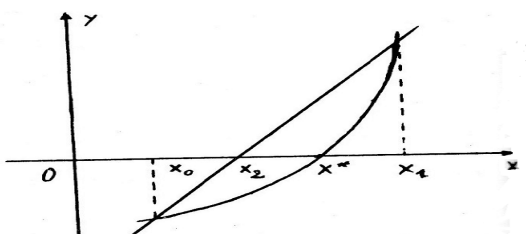
3.3.7 Méthode de la sécante

Dans certaines circonstances (si f provient d'un calcul expérimental par exemple), on ne peut calculer f'. L'idée est alors de remplacer f' par le taux d'accroissement de f sur un petit intervalle.
Supposons que x_0 et x_1 soient deux valeurs approchées de la racine x^* de l'équation

$$f(x) = 0 \quad \text{avec} \quad x_0 < x^* < x_1.$$

Le taux d'accroissement de f sur $[x_0, x_1]$ est :
$\tau_1 = \dfrac{f(x_1) - f(x_0)}{x_1 - x_0}$
et l'équation de la sécante (en x_0 et x_1) est :
$y = \tau_1(x - x_1) + f(x_1) \qquad (1)$
On obtient une nouvelle approximation x_2 de x^* en calculant l'abscisse du point d'intersection de la sécante avec l'axe O_x :

$$x_2 = x_1 - \frac{f(x_1)}{\tau_1} \quad \text{(obtenu en posant } y = 0 \text{ dans(1))}$$



et, on itère ce procédé pour obtenir :

$$\tau_k = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}} \quad \text{et} \quad x_{k+1} = x_k - \frac{f(x_k)}{\tau_k}$$

L'algorithme correspondant est : **Algorithme de la sécante :**

$$\begin{cases} x_0 \ , \ x_1 \ \text{ donnés} \\ x_{k+1} = x_k - \dfrac{f(x_k)}{f(x_k) - f(x_{k-1})}(x_k - x_{k-1}) \ , \ k \in \mathbb{N}^* \end{cases}$$

Critère d'arrêt
C'est le même que celui correspondant à la méthode de Newton-Raphson.

3.3.8 Méthode de dichotomie

L'idée : Construction d'une suite d'intervalles de plus en plus petits contenant une racine isolée de l'équation $f(x) = 0$.
L'outil utilisé : Théorème des valeurs intermédiaires.

Théorème 3.3.4 (Théorème des valeurs intermédiaires) *Soit $f : [a, b] \to \mathbb{R}$ continue, avec $f(a)f(b) < 0$. Alors : il existe au moins $x^* \in]a, b[$ tel que : $f(x^*) = 0$*

Remarque 3.3.4 *Si de plus f est injective, alors x^* est unique.*

Algorithme
Supposons que a et b soient tels que : $x^* \in [a, b]$, avec $f(a)f(b) < 0$. On pose $a = a_0$. $b = b_0$ et $[a, b] = [a_0, b_0] = I_0$.
On divise l'intervalle $I_0 = [a_0, b_0]$ en deux, et on construit l'intervalle
$I_1 = [a_1, b_1]$ comme suit :
Pour $x_0 = \dfrac{a_0 + b_0}{2}$ (milieu du segment $[a_0, b_0]$) on fait le test suivant :

Si $f(a_0)f(x_0) < 0$ alors $[a_0, x_0] = [a_1, b_1] = I_1$
sinon $[x_0, b_0] = [a_1, b_1] = I_1$

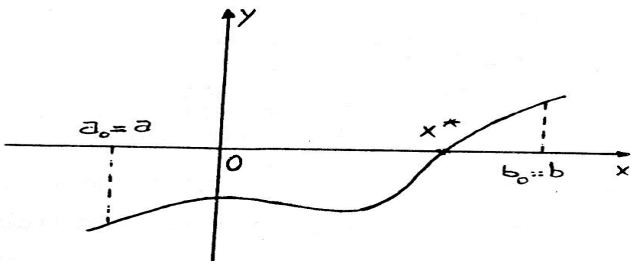
On itère le procédé pour obtenir une suite d'intervalles emboités :
$I_k = [a_k, b_k]$, $k = 1, 2, 3, ...$
comme suit :
On pose : $x_k = \dfrac{a_k + b_k}{2}$

Si $f(a_k)f(x_k) < 0$ alors $[a_k, b_k] = [a_{k+1}, b_{k+1}] = I_{k+1}$
sinon $[x_k, b_k] = [a_{k+1}, b_{k+1}] = I_{k+1}$.

Et on prend comme approximation de x^* la valeur x_k.

Critère d'arrêt

Soit $k \in \mathbb{N}$. Nous avons :



$$b_{k+1} - a_{k+1} = \frac{b_k - a_k}{2} = \frac{b_{k-1} - a_{k-1}}{2^2} = ... = \frac{b_0 - a_0}{2^{k+1}}$$

et si x^* est la racine de l'équation $f(x) = 0$, nous aurons :

$$|x^* - x_k| \leq b_{k+1} - a_{k+1} = \frac{b_k - a_k}{2} = \frac{b_0 - a_0}{2^{k+1}}$$

Et donc, si on désire calculer une approximation x_k de x^* avec n décimales exactes, il suffit de poser $\frac{b_0 - a_0}{2^{k+1}} \leq 0.5\ 10^{-n}$.
Ce qui revient à ce qu'on aille dans les itérations jusqu'à ce que k vérifie l'inégalité :

$$k \geq \frac{\log \frac{b_0 - a_0}{0.5\ 10^{-n}}}{\log 2} - 1.$$

Travaux dirigés 3

Exercice 1 :
Parmi les fonctions suivantes lesquelles sont contractantes et sur quel intervalle si celui-ci n'est pas indiqué :

(a) $g(x) = 5 - \frac{1}{4}\cos 3x$, $0 \leq x \leq \frac{2\pi}{3}$;

(b) $g(x) = 2 + \frac{1}{2}|x|$, $-1 \leq x \leq 1$;

(c) $g(x) = 3 - \frac{1}{2}\sin 3x$

(d) $g(x) = \sqrt{x+2}$.

Exercice 2 :
Voir si chacune des fonctions suivantes admet zero, un ou plusieurs points fixes, puis donner pour chacun un intervalle de separation :
$g(x) = \frac{1}{\sqrt{x}}$, $g(x) = e^{-x}$, $g(x) = x + (x-2)^3$, $g(x) = (x-2)^2 + x - \frac{e^x}{\pi}$

Exercice 3 :
Montrer que l'équation $x = \omega - \epsilon \sin(x)$ admet une unique racine dans l'intérvalle $[\omega - \pi, \omega + \pi]$; $\omega \in \mathbb{R}$, et $|\epsilon| \leq 1$.

Exercice 4 :
Déterminer à l'aide de la méthode du point fixe les deux racines réelles de $x^2 - 100x + 1 = 0$ avec une erreur $e \leq 10^{-3}$. Utiliser pour l'une de ces racines la méthode itérative $x = f(x) = \frac{x^2+1}{100}$, et pour l'autre $x = g(x) = 100 - \frac{1}{x}$.

Exercice 5 :
Soit la fonction $F(x) = 2x^3 - x - 2$, on se propose de trouver les racines réelles de F par la méthode des approximations successives.
Montrer que F possède une seule racine réelle $\bar{x} \in [1; 2]$.
Etudier la convergence des trois méthodes itératives suivantes : $x_0 \in [1; 2]$ donné et

(α) $x_{n+1} = 2x_n^3 - 2$;

(β) $x_{n+1} = \frac{2}{2x_n^2 - 1}$;

(γ) $x_{n+1} = \sqrt[3]{1 + \frac{x_n}{2}}$.

Si l'une de ces méthodes converge l'utiliser pour déterminer $\bar{x}$ à 10^{-3} près.

Exercice 6 :
Soit l'équation $x = \ln(1+x) + 0.2$ dans $\mathbb{R}^+$.
Montrer que la méthode itérative définie par $g(x) = \ln(1+x) + 0.2$ est convergente (vérifier les hypothèses du théorème du point fixe). Choisir x_0, condition initiale de l'itération, dans l'intervalle de convergence puis trouver $\bar{x}$ limite de la suite. Donner l'ordre de la méthode.

Exercice 7 :
On veut résoudre dans $\mathbb{R}^+$ l'équation $x = g(x)$ où, $g(x) = -\ln x$,
a) 1) Montrer qu'elle admet une seule racine $\bar{x}$, montrer que $\bar{x} \in I = [0; 1]$.
2) Montrer que la méthode itérative : $x_{n+1} = g(x_n)$ diverge.
3) on considère alors $g^{-1}(x) = g^{-1}og(x) = x$, (remarquer que g^{-1} existe), montrer que la méthode itérative : $x_{n+1} = g^{-1}(x)$ converge.
En posant $e_n = x_n - \bar{x}$ montrer que e_{n+1} est de signe opposé à e_n, qu'en conclut-on ?
Donner le bon test d'arrêt des itérations pour avoir $\bar{x}$ à $\epsilon = 10^{-4}$ près, puis donner cette racine approchée.
b) Retrouver $\bar{x}$ à l'aide de la méthode de Newton.
Remarquer que Newton est d'ordre 2.

Suggestions et Corrigés

Exercice 1

(a) $g(x) = 1 - \frac{1}{5}\sin(4x), x \in \mathbb{R}$.
Montrons que g est contractante sur $\mathbb{R}$, on a :

$$g'(x) = -\frac{4}{5}\cos(4x) \quad \text{et } |g'(x)| \leq \frac{4}{5},$$

donc, d'après la proposition 1, g est contractante de rapport de contraction inférieur ou égal à$\frac{4}{5}$.

(b) $g(x) = 2 + \frac{1}{2}|x|, x \in [-1, 1]$.
Soient $x, y \in [-1, 1]$, montrons que $|g(x) - g(y)| \leq \frac{1}{2}|x - y|$. On a

$$\begin{aligned} |g(x) - g(y)| &= \left|2 + \frac{1}{2}|x| - 2 - \frac{1}{2}|y|\right| \\ &= \frac{1}{2}\big||x| - |y|\big| \\ &\leq \frac{1}{2}|x - y|. \end{aligned}$$

En effet, d'une manière générale, on peut montrer que $\big||x| - |y|\big| \leq |x-y|$:
Supposons que $|x| \geq |y|$ alors

$$\begin{aligned} \big||x| - |y|\big| &= |x - y + y| - |y| \\ &\leq |x - y| + |y| - |y| \text{d'après l'inégalité triangulaire} \\ &\leq |x - y|. \end{aligned}$$

On fait de même si $|x| \leq |y|$, d'où le résultat $\big||x| - |y|\big| \leq |x - y|$.
Ainsi, $|g(x) - g(y)| \leq \frac{1}{2}|x - y|$ et le rapport de contraction est $k = \frac{1}{2}$.

(c) $g(x) = \frac{1}{x}, x \in [2, 3]$.
On a :

$$g'(x) = -\frac{1}{x^2} \quad \text{et} \quad 4 < x^2 < 9 \Leftrightarrow \frac{1}{9} \leq \left|-\frac{1}{x^2}\right| \leq \frac{1}{4}$$

donc, $\forall x \in [2, 3], \quad |g'(x)| \leq \frac{1}{4}$.
Ainsi, g est contractante de rapport $k \leq \frac{1}{4}$.

(d) $g(x) = \sqrt{x + 2}$.
g est définie sur $[-2, +\infty[$ mais n'y est pas lipschitzienne.

En effet, g est lipschitzienne sur I s'il existe une constante réelle $L > 0$, telle que $\forall (x, y) \in I^2$, $|g(x) - g(y)| \leq L|x - y|$, c'est à dire que le rapport $\left|\dfrac{g(x) - g(y)}{x - y}\right|$, pour $x \neq y$, est borné.
Posons $y = -2$. Ce rapport vaut

$$\left|\frac{g(x) - g(-2)}{x + 2}\right| = \frac{\sqrt{x+2}}{x+2} = \frac{1}{\sqrt{x+2}} \quad \xrightarrow[x \to -2]{} \quad +\infty$$

donc non bornable sur tout intervalle contenant -2 ; ainsi g ne peut être lipschitzienne sur $[-2, +\infty[$.
En fait, on montre que g est lipschitzienne sur tout intervalle $[a, +\infty[$ avec $a > -2$. Sur cet intervalle, $g'(x) = \dfrac{1}{2\sqrt{x+2}} \leq \dfrac{1}{2\sqrt{a+2}}$. Ainsi, grâce à la proposition 1, g est lipschitzienne de constante $L \leq \dfrac{1}{2\sqrt{a+2}}$.
En outre, g est contractante si $L < 1$, donc si $\dfrac{1}{2\sqrt{a+2}} < 1$, c'est à dire pour $a > \dfrac{-7}{4}$.
En conclusion, g est contractante sur $]\dfrac{-7}{4}, +\infty[$.

Exercice 2

(a) Points fixes de $g(x) = \dfrac{1}{\sqrt{x}}$. Rappelons qu'un point fixe de g est un point d'abscisse $\bar{x}$ vérifiant $g(\bar{x}) = \bar{x}$. Par abus de langage, et dans tous les exercices qui suivent, on dira que $\bar{x}$ est le point fixe de g (au lieu de l'abscisse du point fixe de g).
Ici g est définie sur R^{+*} et on a

$$g(x) = x \quad \Rightarrow \quad x\sqrt{x} = 1 \quad \Rightarrow x = 1.$$

$x = 1$ est clairement la seule solution sur $\mathbb{R}^{+*}$ de cette équation et est par conséquent le seul point fixe de g.
Démontrons le autrement :

$$g(x) = x \quad \Rightarrow \quad \frac{1}{\sqrt{x}} - x = 0 \quad \Rightarrow \quad x\sqrt{x} - 1 = 0 \text{ et } x > 0.$$

Posons $F(x) = x\sqrt{x} - 1$; F est continue sur $\mathbb{R}^+$ et dérivable sur $\mathbb{R}_*^+$ et $F'(x) = \frac{3}{2}\sqrt{x} > 0$, donc F est strictement croissante sur $\mathbb{R}_*^+$. D'autre part, $F(0.1) < 0$ et $F(2) \geq 0$, donc $F(0.1)F(2) < 0$. Ainsi, d'après le théorème

de la valeur intermédiaire, il existe un et un seul réel $c \in [0.1, 2]$ tel que $F(c) = 0$; celui-ci est donc le seul point fixe de g sur $[0.1, 2]$. Le lecteur pourra aisément démontrer que cela reste vrai sur tout $\mathbb{R}^{+*}$.

(b) Points fixes de $g(x) = e^{-x}$.

Posons $F(x) = e^{-x} - x$. F est continue et dérivable sur $\mathbb{R}$, et $F'(x) = -e^{-x} - 1 < 0$, donc F est strictement décroissante. D'autre part, $F(0) = 1$ et $F(1) = \frac{1}{e} - 1 < 0$. D'après le théorème de la valeur intermédiaire, il existe un et un seul réel $c \in [0, 1]$ tel que $F(c) = 0$. Ce réel est donc l'unique point fixe de g sur $[0, 1]$. De même, on peut aisément démontrer que cela reste vrai sur tout $\mathbb{R}$.

(c) Points fixes de $g(x) = x + (x-2)^3$.

$$g(x) = x \quad \Rightarrow \quad (x-2)^3 = 0 \quad \Rightarrow \quad x = 2.$$

Donc 2 est l'unique point fixe de g sur $\mathbb{R}$; ce point fixe est dit triple à cause de la puissance 3 du terme $(x-2)$.

(d) Points fixes de $g(x) = (x-2)^2 + x - \dfrac{e^x}{\pi}$.

$$g(x) = x \quad \Rightarrow \quad (x-2)^2 = \frac{e^x}{\pi}.$$

Appliquons le théorème de la valeur intermédiaire à

$$F(x) = (x-2)^2 - \frac{e^x}{\pi}$$

. F est continue et dérivable sur R.

$$F'(x) = 2(x-2) - \frac{e^x}{\pi}.$$

Montrons que $\forall x \in \mathbb{R}, F'(x) < 0$.

Pour cela, on étudie le signe de F'', on a :

$$F''(x) = 2 - \frac{e^x}{\pi} > 0 \quad \Rightarrow \quad e^x < 2\pi \quad \Rightarrow \quad x < \ln(2\pi),$$

En conséquence, F' est strictement croissante sur $]-\infty, \ln(2\pi)[$, strictement croissante sur $]\ln(2\pi, \infty)[$ et $F'(\ln(2\pi)) < 0$. Ainsi, $\forall x \in \mathbb{R}, F'(x) < 0$, donc F est strictement décroissante.

D'après le théorème de la valeur intermédiaire, il existe un et un seul réel $c \in [A, \ln(2\pi)[$ tel que $F(c) = 0$. Ce dernier est l'unique point fixe de g.

Exercice 3

$x = \omega - \varepsilon \sin(x)$, $\omega \in \mathbb{R}$, $|\varepsilon| \leq 1$.

Montrons qu'il existe un unique réel $\bar{x} \in [\omega - \pi, \omega + \pi]$ solution de l'équation

$$F(x) = x - \omega + \varepsilon \sin x = 0.$$

On a

$$F'(x) = 1 - \varepsilon \cos x \geq 0, \text{ car} |\varepsilon| \leq 1,$$

ainsi F est monotone. Or,

$$\begin{aligned} F(\omega - \pi) &= -\pi + \varepsilon \sin(\omega - \pi) < 0 \\ F(\omega + \pi) &= \pi + \varepsilon \sin(\omega - \pi) > 0 \end{aligned}$$

donc, $F(\omega - \pi)F(\omega + \pi) < 0$, et d'après le théorème de la valeur intermédiaire, il existe un unique réel $\bar{x} \in [\omega - \pi, \omega + \pi]$ tel que $F(\bar{x} = 0)$.

Exercice 4

Soit l'équation $F(x) = x^2 - 100x + 1 = 0$.

a) Posons $g_1(x) = \dfrac{x^2 + 1}{100}$

Étudions donc la méthode itérative $x_{n+1} = \dfrac{x_n^2 + 1}{100}$.

Remarquons tout d'abord que si cette méthode converge, elle converge bien vers une des racines de $F(x) = 0$, si $\bar{x}$ est la limite de la suite (x_n), alors $\bar{x} = \dfrac{\bar{x}^2 + 1}{100}$ donc $\bar{x}^2 - 100\bar{x} + 1 = 0$, c'est à dire que $F(\bar{x}) = 0$.

Localisons la racine $\bar{x}$ de cette équation. On a $F(0) = 1$ et $F(1) = -98$ donc $F(0)F(1) < 0$, et comme F est dérivable sur $\mathbb{R}$ et $F'(x) = 2x - 100 < 0$ sur $[0, 1]$ alors, grâce au théorème de la valeur intermédiaire, il existe un unique réel $\bar{x} \in [0, 1]$ solution de l'équation $F(x) = 0$.

On a $g_1(0) = \frac{1}{100} > 0$ et $g_1(1) = \frac{1}{50} < 1$. Comme g_1 est monotone sur $\mathbb{R}^+$ (puisque $g_1'(x) = x/50 > 0$ sur R^+), on a donc $g_1([0, 1]) \subset [0, 1]$.

Démontrons que $g_1(x) = \dfrac{x^2 + 1}{100}$ est contractante sur $[0, 1]$:

$$\begin{aligned} |g_1(x) - g_1(y)| &= \left| \frac{x^2 + 1}{100} - \frac{y^2 + 1}{100} \right| \\ &= \frac{1}{100} \left| x^2 - y^2 \right| \\ &= \frac{|x + y|}{100} |x - y| \leq \frac{1}{50} |x - y|. \end{aligned}$$

On aurait pu aussi la proposition 1, puisque $g_1'(x) = \dfrac{x}{50}$ et $max_{[0,1]}|g_1'(x)| = 1/50$ donc g_1 contractante de rapport $1/50$.

Ainsi, $\forall x_0 \in [0,1]$, $x_{n+1} = g_1(x_1) = \frac{x_n^2+1}{100}$ converge vers $\bar{x}$, unique solution de $x^2 - 100x = 0 + 1$ dans $[0,1]$.

Calculons cette racine partant de $x_0 = 0$, on a

$$\begin{aligned} x_0 &= 0 \\ x_1 &= g_1(x_0) = g_1(0) = \frac{1}{100} = 0.010001 \\ x_2 &= g_1(x_1) = g_1(0.010001) = 0.01000100020001 \\ x_3 &= g_1(x_2) = g_1(0.01000100020001) = ... \\ x_4 &= g_1(x_3) = ... \end{aligned}$$

Si on cherche $\bar{x}$ à ε près, on arrêtera les calculs à l'itération p telle que $|x_{p+1} - x_p| \leq \varepsilon$. Ainsi la solution $\bar{x}$ ici vaut 0.010001 à 10^{-6} près.

b) L'autre solution est obtenue grâce à la méthode itérative $x_{n+1} = g_2(x_n) = 100 - \frac{1}{x_n}$. Cette question est laissée en exercice.

Exercice 5

Soit l'équation $F(x) = 2x^3 - x - 2 = 0$. Il est clair que F est continue et dérivable sur $\mathbb{R}$.

On a $F(1) = -1$, $F(2) = 12$, donc $F(1)F(2) < 0$. D'autre part, $F'(x) = 6x^2 \geq 0$ sur $[1,2]$. Donc, d'après le théorème de la valeur intermédiaire, il existe une seule solution $\bar{x} \in [1,2]$ telle que $F(\bar{x}) = 0$.

(a) Etudions la convergence de la suite $x_{n+1} = g_1(x_n) = 2x_n^3 - 2$. Tout d'abord, cette suite, si elle converge, conduit bien à une racine de $F(x) = 0$ car si $\bar{x}$ est la limite de la suite (x_n), alors

$$\bar{x} = 2\bar{x}^3 - 2 \quad \text{donc} \quad F(\bar{x}) = 2\bar{x}^3 - \bar{x} - 2 = 0.$$

Par ailleurs, $g_1'(x) = 6x^2 \geq 6$ sur $[1,2]$. Par conséquent, grâce au théorème des accroissements finis, il existe ξ_n compris entre x_n et x_{n+1} tel que

$$|g_1(x_{n+1}) - g_1(x_n)| = g_1'(\xi_n)|x_{n+1} - x_n|.$$

Donc

$$\begin{aligned} |g_1(x_{n+1}) - g_1(x_n)| &\geq 6|x_{n+1} - x_n| \\ &\geq 6^2|x_n - x_{n-1}| \\ &\vdots \\ &\geq 6^n|x_1 - x_0|. \end{aligned}$$

Ainsi, cette suite diverge et la méthode est à rejeter.

(b) Étudions la convergence de $x_{n+1} = g_2(x_n) = \dfrac{2}{2x_n^2 - 1}$. Cette méthode, si elle converge conduit vers la racine $\bar{x}$ de $F(x)$ dans $[1, 2]$, car si $\bar{x}$ est la limite de la suite (x_n), alors

$$\bar{x} = \frac{2}{2\bar{x}^2 - 1} \quad \text{donc} \quad F(\bar{x}) = 2\bar{x}^3 - 2\bar{x} - 1 = 0.$$

$$\begin{aligned} g_2'(x) &= \frac{-8x}{(2x^2-1)^2} \\ g_2''(x) &= \frac{8(6x^2+1)}{(2x^2-1)^3} \end{aligned}$$

	1		2
g_2''		+	
g_2'	-8	$\nearrow$	$\frac{16}{49}$

En conséquence, on ne peut conclure sur la monotonie de g_2. Cependant on a

$$g_2'(\bar{x}) = \frac{-8\bar{x}}{(2\bar{x}^2-1)^2} = -2\bar{x}\left(\frac{2}{2\bar{x}-1}\right)^2,$$

or $\bar{x}$ le point fixe de F vérifie

$$\frac{2}{2\bar{x}-1} = \bar{x}.$$

Donc $g_2'(\bar{x}) = -2\bar{x}^3$, et comme g_2' est continue, il existe un voisinage V de $\bar{x}$ tel que $V \subset [1, 2]$, et $\forall x \in V, |g_2'(x)| > 2$. Donc cette méthode ne peut pas converger d'après la proposition 3. En effet, grâce au théorème des accroissements finis, on a $|x_{n+1} - x_n| \geq 2^n|x_1 - x_0|$.

(c) Étudions la convergence de $x_{n+1} = g_3(x_n) = \sqrt[3]{1 + \dfrac{x_n}{2}}$. Si elle converge, cette méthode conduit à la racine de $F(x) = 0$ dans $[1, 2]$ car si $\bar{x}$ est la limite de la suite (x_n), alors

$$\bar{x} = \sqrt[3]{1 + \frac{\bar{x}}{2}} \quad \text{donc} \quad \bar{x}^3 = 1 + \frac{\bar{x}}{2} \quad \text{et} \quad F(\bar{x}) = 2\bar{x}^3 - 2\bar{x} - 1 = 0.$$

On a

$$0 < g_3'(x) \;\; = \;\; \frac{1}{6\sqrt[3]{(1 + \frac{x}{2})^2}} < 1,$$

donc g_3 est strictement contractante d'après la proposition 1. D'autre part, $g_3(1) = \sqrt[3]{\frac{3}{2}} > 1$, $g_3(2) = \sqrt[3]{2} < 2$, or g_3 est monotone, donc $g_3([1, 2]) \subset [1, 2]$. Donc d'après le théorème du point fixe, la suite

$$\begin{cases} x_0 \in [1, 2] \\ x_{n+1} = g_3(x_n) \end{cases}$$

converge vers l'unique racine $\bar{x} \in [1, 2]$ de l'équation $x = g_3(x)$.

Calcul numérique de cette racine à 10^{-3} près, à partir de $x_0 = 1$:

n	0	1	2	3	4
x_n	1	1.144	1.**16**2	1.**165**	0.**165**

Donc $\bar{x} = 1.165$ est solution de l'équation à 10^{-3} près.

Exercice 6

Soit l'équation $x = \ln(1 + x) + 0.2$ dans R^+.

Considérons la méthode itérative définie par :

$$x_{n+1} = g(x_n) = \ln(1 + x_n) + 0.2$$

Montrons d'abord l'existence d'une solution pour cette équation.

Soit $F(x) = \ln(1 + x) + 0,2 - x = 0$, on a $F'(x) = \frac{-x}{1+x} < 0$ sur $\mathbb{R}^+$, donc l'équation $F(x) = 0$ admet au plus une racine. D'autre part on a $F(0) = 0.2$ et $F(1) = \ln 2 - 0.8 < 0$, donc $F(0)F(1) < 0$; ainsi, d'après le théorème de la valeur intermédiaire, il existe une unique racine $\bar{x} \in [0, 1]$ solution de l'équation $F(x) = 0$.

Appliquons la Méthode du point fixe pour $g(x) = \ln(1 + x) + 0,2$.

g est contractante sur $I = [a, b] \subset]0, 1]$ car

$$\forall x \in I, 0 < g'(x) = \frac{1}{1+x} < 1.$$

Donc, si $g([a, b]) \subset [a, b]$, d'après le théorème du point fixe, il existe une unique racine $\bar{x} \in [a, b]$ solution de l'équation $F(x) = 0$.

Par exemple, on vérifie que $g([0.7, 0.8]) \subset [0.7, 0.8]$. En effet,$g(0.7) = 0.73... > 0.7$ et $g(0.8) = 0.78... < 0.8$.

Calcul numérique de cette racine à 10^{-2} et 10^{-3} près :

n	0	1	2	3	4	5	6	7	8	9
x_n	0.7	0.730	0.748	0.758	0.**76**4	0.**76**7	0.769	0.770	0.**771**	0.**771**

Ainsi la racine cherchée est $\bar{x} = 0.76$ à 10^{-2} près, et $\bar{x} = 0.771$ à 10^{-3} près.

Exercice 7

Soit l'équation $x = g(x)$ où $g(x) = -\ln x$.

1. 1) Posons $F(x) = x - g(x) = x + \ln(x)$, $\forall x \in \mathbb{R}^{*+}$. Appliquons le théorème de la valeur intermédiaire à F sur $[a, 1]$ où $0 < a \leq 1$.
   F est continue sur $[a, 1]$, $\forall a \in]0, 1]$. $\forall x \in [a, 1]$, $F'(x) = 1 + \frac{1}{x}$ et $F'(x) > 0$, donc F est strictement monotone sur $[a, 1]$.
   D'autre part on a $F(1) = 1 > 0$, et comme $\lim\limits_{x \longrightarrow 0^+} \ln(x) = -\infty$, alors il existe $a \in]0, 1]$ tel que $\ln(a) < -a < 0$; par conséquent, $F(1)F(a) < 0$, et d'après le théorème de la valeur intermédiaire, il existe un unique $\bar{x} \in [a, 1]$ (d'ou $\bar{x} \in]0, 1]$) tel que $F(\bar{x}) = 0$, et donc tel que $\bar{x} = g(\bar{x}) = \ln \bar{x}$.
   2) Si $x_{n+1} = g(x_n)$ converge, elle conduit bien à la racine de l'équation car cette dernière vérifie $\bar{x} = g(\bar{x})$ donc

   $$\bar{x} = -\ln(\bar{x}) \implies \bar{x} + \ln(\bar{x}) = 0$$

   Mais, $\forall x \in [a, 1[, g'(x) = -\frac{1}{x}$, et $|g'(x)| > 1$ donc la méthode $x_{n+1} = -\ln(x_n)$ diverge pour tout $x_0 \in [a, 1[$.
   3) g^{-1} existe car g est continue et strictement croissante donc bijective. Montrons que $x_{n+1} = g^{-1}(x_n)$ est convergente.
   Attention à la notation utilisée : g^{-1} désigne la réciproque de g et non $\frac{1}{g(x)}$.
   g^{-1} est dérivable et on a $(g^{-1})' = \frac{1}{g' \circ g^{-1}}$.
   En effet, d'une manière générale $(u \circ v)' = (u' \circ v)v'$, par conséquent $(g \circ g^{-1})'(x) = (g' \circ g^{-1})(x)(g^{-1})'(x)$.

Or, $g \circ g^{-1} = Id$, donc $(g \circ g^{-1})' = 1$. On a bien alors $(g^{-1})'(x) = \dfrac{1}{g' \circ g^{-1}}$.
Or, on a montré que $\forall x \in [a, 1[$, $|g'(x)| > 1$. Mais $g^{-1} = e^{-x}$, donc $y = g^{-1}(x) \in]e^{-1}, e^{-a}]$, car e^{-x} est décroissante. Puisque $a > 0$, alors $e^{-a} < e^0 = 1$, donc $0 < y < 1$ et $|g'(y)| = \left|-\dfrac{1}{y}\right| > 1$
Par conséquent $|(g^{-1})'(x)| < 1$.
Ainsi la méthode $x_{n+1} = g^{-1}(x_n)$ converge au voisinage de $\bar{x}$, d'après la proposition 2.

4) Posons $e_n = x_n - \bar{x}$ et $h(x) = e^{-x}$. La méthode $x_{n+1} = e^{-x_n}$ converge (d'ailleurs, on a $h'(x) = -e^{-x}$ et $|h'(x)| = e^{-x} < 1$, $\forall x \in \mathbb{R}_+^*$). D'autre part, grâce au théorème des accroissements finis on sait qu'au voisinage de $\bar{x}$, il existe ξ_n compris entre x_n et $\bar{x}$ tel que :

$$e_{n+1} = x_{n+1} - \bar{x} = h(x_n) - h(\bar{x}) = h'(\xi_n)(x_n - \bar{x}).$$

Ainsi, $e_{n+1} = h'(\xi_n)e_n$. Or, $h'(x) < 0$ $\forall x \in \mathbb{R}^{*+}$. Donc e_{n+1} et e_n sont de signes opposés, par conséquent deux itérations successives donnent un encadrement de $\bar{x}$.

5) Un test d'arrêt des itérations est : $|e_{n+1} - e_n| = |x_{n+1} - x_n| \leq \varepsilon = 10^{-4}$.
Prenons $I = [a, 1]$ avec $a = 0.1$. On a bien $|h'(x)| < 1$ sur I et $h(I) \subset I$ car :

$$\begin{aligned} h(0.1) = e^{-0,1} &> 0.1 \\ h(1) = \frac{1}{e} &< 1 \end{aligned}$$

et $h(x) = e^{-x}$ est monotone sur $[0.1, 1]$
Calcul numérique de la racine à 10^{-3} près. Soit donc la méthode $x_{n+1} = e^{-x_n}$ et $x_0 = 1$

n	0	1	2	3	4	5	6	7	8
x_n	1	0.367	0.692	0.500	0.606	0.**5**45	0.**5**79	0.**5**60	0.**5**71

n	9	10	11	12	13	14	15
x_n	0.**56**4	0.**56**8	0.**56**6	0.**56**7	0.**56**6	0.**567**	0.**567**

Ainsi la racine cherchée est $\bar{x} = 0.567$ à 10^{-3} près.

2. Méthode de Newton.
Soit $F(x) = x - e^{-x} = 0$, F est clairement indéfiniment dérivable. La

méthode de Newton s'écrit,

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

D'autre part on a $F(0) = -1 < 0$ et $F(1) = 1 - \frac{1}{e} > 0$, donc $F(1)F(0) < 0$ et la racine est située dans $[0, 1]$, elle est unique puisque F est strictemnet monotone, car $F'(x) = 1 + e^{-x} > 0$ pour tout $x \in \mathbb{R}$. On a aussi $F''(x) = -e^{-x} < 0$ pour tout $x \in \mathbb{R}$. Ainsi d'après le théorème de convergence globale de cette méthode (voir théorème 3), pour tout $x_0 \in [0, 1]$ tel que $F(x_0)F''(x_0) > 0$ l'itération de Newton converge. Prenons alors, par exemple, $x_0 = 0$, alors $F(0)F''(0) = 1/e > 0$, donc la méthode

$$x_{n+1} = x_n - \frac{x_n - e^{-x_n}}{1 + e^{-x_n}} x_0$$

convegera vers l'unique racine $\bar{x}$ de l'équation.

Chapitre 4

Résolution numérique des équations différentielles ordinaires d'ordre 1

4.1 Introduction

Soit le probléme de cauchy

$$\begin{cases} y' = f(t,y) \\ y(t_0) = y_0 \end{cases} \qquad (1)$$

ou $f : [t_0, t_0+T] \times \mathbb{R} \to \mathbb{R}$, le théorème ci-après nous donne des conditions qui assurent l'existence et l'unicitè de la solution (théorique) de ce problème :

Théorème 4.1.1 *Si f est continue et s'il existe une constante L strictement positive telle que pour toute $t \in [t_0, t_0+T]$,et tout $y_1, y_2 \in \mathbb{R}$, on ait $\left|f(t,y_1) - f(t,y_2)\right| \leq L\left|y_1 - y_2\right|$ alors, le poblème de Cauchy admet une solution unique ,quelque soit $y_0 \in \mathbb{R}$. On dit alors que f est L-lipchitzienne ,où L est la constante de Lipschitz.*

REMARQUE 7.1

Nous nous placerons toujours dans les conditions du théorème7.1

L'objectif de ce chapitre est de décrire un certain nombre de méthodes permettant de résoudre numériquement le probléme de **Cauchy** :

Etant donné une subdivition $t_0 < t_1 < ... < t_N = t_0 + T$ de $[t_0, t_0+T]$, on cherche à déterminer des valeurs approchées $y_0, y_{1,...,} y_N$des valeurs $y(t_n)$ prises par la solutionexacte y. On notera les pas succéssives

$$\begin{aligned} h_n &= t_{n+1} - t_n, \quad 0 \leq n \leq N-1 \\ et \quad h_{\max} &= \max(h_n) \quad le \text{ maximum des pas} \end{aligned}$$

On appelle méthode à un pas une méthode permettant de calculer Y_{n+1}à partir de la seule valeur antérieure y_n.Une méthode à r pas est au contraire une méthode ou le calcul de y_{n+1}nécessite la mémorisation des valeurs $y_n, y_{n-1}, ...y_{n-r+1}$.

4.2 Méthodes numériques à un pas

Les méthodes à un pas sont les méthodes de résolution numérique qui peuvent s'écrire sous la forme :

$$y_{n+1} = y_n + h_n\varphi(t_n, y_n, h_n),\ 0 \leq n \leq N \tag{1.1}$$

où $\varphi : [t_0, t_0 + T] \times \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$ est une fonction que l'on supposera continue .

Dans la pratique, la fonction $\varphi(t_n, y_n, h_n)$ peut n'être définie que sur une partie de la forme $[t_0, t_0 + T] \times J \times [0, \delta]$ où J est un intervalle de $\mathbb{R}$ (de sorte en particulier que $[t_0, t_0 + T] \times J$ soit contenu dans le domaine de définition de l'équation différentielle).

Dans toutes les méthodes numériques développées par la suite, on subdivise l'intervalle $[t_0, t_0 + T]$ en N intervalles de longueur

$$h = \frac{(t_0 + T) - t_0}{N} = \frac{T}{N}$$

limités par les points

$$t_n = t_0 + nh,\ 0 \leq n \leq N \tag{1.2}$$

4.2.1 Méthode D'EULER

En t_0 on connait y_0, donc aussi

$$y'(t_0) = f(t_0, y_0).$$

Si $y(t)$ est la solution exacte de (1). $y(t)$ est approchée sur l'intervalle $[t_0, t_1]$ par sa tangente au point t_0 .

Et ainsi, on a

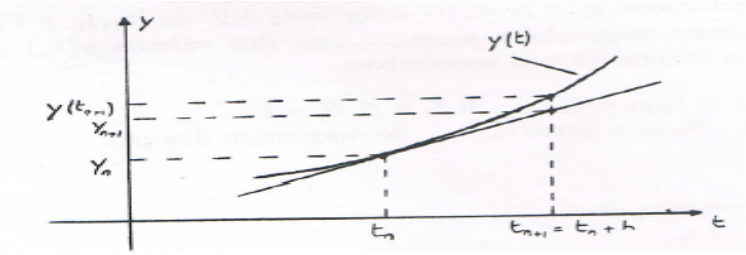
$$y_1 = y_0 + hf(t_0, y_0).$$

Sur l'intervalle $[t_1, t_2]$, $y(t)$ sera remplacée par la tangente au point (t_1, y_1). On trouve

$$y_2 = y_1 + hf(t_1, y_1).$$

Ceci conduit à l'**Algorithme d'Euler :** $\begin{cases} y_{n+1} = y_n + hf(t_n, y_n), 0 \leq n \leq N-1 \\ t_{n+1} = t_n + h, \end{cases}$

Précision de la méthode d'Euler



la méthode d'Euler est une méthode du premier ordre , c'est-à-dire que l'erreur au point t_n s'exprime par l'inégalité

$$|y_n - y(t_n)| \leq kh \tag{1.3}$$

où y_n est la valeur approchée définie par l'algorithme d'Euler, $y(t_n)$ est la valeur exacte de la soltion du problème de Cauchy au point

$$t = t_n = t_0 + nh$$

et k une constante indépendante de n et de h.

Application

Résolution d'une équation selon la méthode d'Euler

Soit le peobléme de Cauchy

$$\begin{cases} y' = t + y \\ y(0) = 1 \end{cases} \tag{1.4}$$

On veut approcher, à 10^{-3}, la solution de (1.4) en $t = 1$ à l'aide de la méthode d'Euler, en subdivisant l'intervalle [0,1] en dix parties égales.

Selon l'algorithme d'Euler :

$$y_{n+1} = y_n + hf(t_n + y_n), \qquad 0 \leq n \leq 9 \ \ et \ \ h = 0.1 \tag{1.5}$$

$$t_{n+1} = t_n + h \tag{1.6}$$

On calcule les valeurs du tableau :

n	0	1	2	3	4	5	6	7	8	9	10
t_n	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
y_n	1	1.1	1.22	1.362	1.5282	1.7210	1.9431	2.1974	2.4871	2.8158	3.1874

On trouve

$$y(1) \simeq 3.187$$

La solution exacte de l'équation (1) est donnée par $y(t) = 2\exp(t) - t - 1$, ce qui donne $y(1) = 3,437$.

L'approximation calculée est donc trés grossière.

Remarque 4.2.1 (1.1) *L'erreur dans la méthode d'Euler est relativement importante. Elle peut être améliorée en choisissant plus petit le pas h, ce qui augmente considérablement le volume des calculs à effectuer, ou en approchant la solution du probléme de Cauchy par des méthodes permettant de réduire cette erreur .*

Programme

En Matlab, on peut facilement programmer la méthode d'Euler avec la fonction suivante :

****

Programme d'Euler

```
**********************************************************************************************
function [t,y] = Euler(f,tmin,tmax,Nint,y0) % Méthode d'Euler
% Nint - nombre de sous intervalles N
% tmin - temps t 0
% tmax - temps t 0+ T
% f est une fonction avec comme arguments t et y(t) : f(t, y(t))
% y0 à l'entrée est composé des valeurs des conditions limites y 0= y (t 0) (E)
h = (tmax-tmin)/Nint ; % valeur du pas
t = linspace(tmin,tmax,Nint+1) ; % vecteur de t discrétisé t=[tmin,tmax]
y(1) = y0 ; % initialisation : y(1)=y(t 0) = y 0
for n = 2 : Nint+1
y(n) =y(n-1) + h*feval(f,t(n-1),y(n-1)) ; % Calcul d'Euler
end % for n
end % fonction Euler
**********************************************************************************************
**********************************************************************************************
```

4.2.2 Méthode de Taylor (d'ordre2)

Supposons que f soit de classe C^1 Alors $y(t)$ est de classe C^2, et le devloppement de Taylor d'ordre 2 implique :

$$y(t_{n+1}) = y(t_n + h) = y(t_n) + hy\prime(t_n) + \frac{h^2}{2!}y\prime\prime(t_n) + o(h^2) \tag{1.7}$$

et comme :
d'une part,

$$y\prime(t_n) = f\left(t_n, y(t_n)\right) \tag{1.8}$$

d'autre part

$$y\prime\prime(t_n) = \frac{d}{dt}(y\prime(t)) \mid_{t=t_n} = \frac{d}{dt}(f(t, y(t))) \mid_{t=t_n} = (\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}.\frac{dy}{dt}) \mid_{t=t_n}$$

$$= (\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y}.y\prime) \mid_{t=t_n} = \frac{\partial f}{\partial t}(t_n, y(t_n)) + \frac{\partial f}{\partial y}(t_n, y(t_n)).f(t_n, y(t_n)).$$

Il vient :

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \frac{h^2}{2}\left[\frac{\partial f}{\partial t}(t_n, y(t_n)) + \frac{\partial f}{\partial y}(t_n, y(t_n))f(t_n, y(t_n))\right] + o(h^2)$$

On est amené à considérer l'algorithme suivant , appelée : **Algorithme de Taylor(d'ordre2) :**

$$\begin{cases} y(t_{n+1}) = y(t_n) + hf(t_n, y_n) + \frac{h^2}{2}\left[\frac{\partial f}{\partial t}(t_n, y_n) + \frac{\partial f}{\partial y}(t_n, y_n).f(t_n, y_n)\right] \\ t_{n+1} = t_n + h \end{cases} \tag{T_2}$$

Remarque 4.2.2 (1.2) *En fait, la méthode de Taylor consiste à approcher la solution de l'équation (1) par des arcs de paraboles au lieu des segments de droits (des tangentes) utilisés dans la méthode d'Euler.*

Précision de la méthode de Taylor
On montre que la méthode de Taylor (d'ordre 2) , est une méthode du second ordre.
Autrement dit, l'uerreur au point t_n vérifie :

$$|y_n - y(t_n)| \leq kh^2 \tag{1.9}$$

où y_n est la valeur approchée définie par l'algorithme de Taylor (d'ordre 2), $y(t_n)$ est la valeur exacte de la solution du problème de Cauchy au point t_n, et k une constante indépendante de n et h.

Conséquence
L'algorithme de Taylor (d'ordre 2), est plus précis que celui d'Euler

Remarque 4.2.3 *Si l'on veut encore réduire la marge d'erreur, on tiendra compte d'un plus grand nombre de termes dans le développement de Taylor c'est-à-dire, on suppose f de class C^p, et donc y en sera en class C^{p+1} et sa dérivée k^{eme} est*

$$y^{(k)}(t) = f^{[k-1]}(t, y(t))$$

avec

$$f^{[1]} = f'_t + f'_y \cdot f.$$

*Le développement de Taylor d'ordre p permet d'aboutir à l'**algorithme de Taylor d'ordre p***

$$\begin{cases} y_{n+1} = y_n + hf(t_n, y_n) + \sum_{k=1}^{p} \frac{h^k}{k} f^{(k-1)}(t_n \cdot y_n) \\ t_{n+1} = t_n + h \end{cases}$$

qui est d'ordre p (au sens de la précision)

Applications

Résolution d'une équation avec la méthode de Taylor
Soit l'équation différentielle régissant le mouvement du pendule :

$$a\ddot{x} + y \sin x = 0 \tag{1.10}$$

où a est la longueur du pendule, g la gravitation terrestre et x l'écart du pendule avec la verticale au sol.
En posant $\dot{x} = \frac{dx}{dt} = y(x)$,on a : $\ddot{x} = y\prime\dot{x} = y\prime y$ et (1.10) devient :

$$ay\prime y + g \sin x = 0 \tag{1.11}$$

qui s'écrit encore :

$$y\prime = -\frac{g}{a} \cdot \frac{\sin x}{y} = f(x, y) \tag{1.12}$$

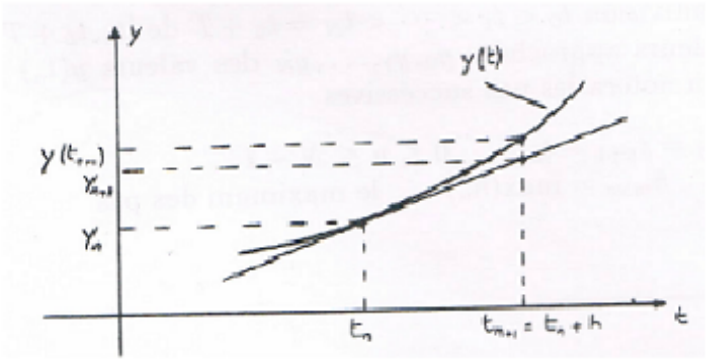
L'algoritme de Taylor (d'ordre2) correspondant s'écrit :

$$\begin{cases} y_{n+1} = \dot{x}_{n+1} = y_n - h\frac{g}{a}\frac{\sin x_n}{y_n} + \frac{1}{2}h^2\left[-\frac{g}{a}\frac{\cos x_n}{\mathbf{y}_n} - \frac{g^2}{a^2}\frac{\sin^2 x_n}{y_n^3}\right] \\ x_{n+1} = x_n + h \qquad 0 \leq n \leq N-1 \end{cases} \tag{1.13}$$

Programme

La fonction Matlab suivante calcule les approximations à solution de y = t - y faites avec la série de Taylor jusqu'à l'ordre 4.

```
**************************************************************************************
                    Programme de Taylor (d'ordre2)
**************************************************************************************
function Taylor(tmin,tmax,Nint,y0) % Approximations par série de Taylor
% Nint - nombre de sous intervalles N
% tmin - temps t 0 ; tmax - temps t 0+ T
% On trait ici l'équation y' = t - y(t)
h = (tmax-tmin)/Nint ; % valeur du pas
t = linspace(tmin,tmax,Nint+1) ; % vecteur de t discrétisé t=[tmin,tmax]
yt = inline('(y0+1)*exp(-t) + t - 1','t','y0') ; % solution exact (F)
texa = linspace(tmin,tmax,101) ; % discrétisation de t pour l'a ?chage de la solution exacte
yexact = yt(texa,y0) ;
(1)
y1 = inline('t - y','t','y') ; % y
(2)
y2 = inline('1 - t + y','t','y') ; % y
(3)
y3 = inline('-1 + t - y','t','y') ; % y
(4)
y4 = inline('1 - t + y','t','y') ; % y
yT1(1) = y0 ; % yT1 contient les solutions de y - Approximation d'Euler
yT2(1) = y0 ; % yT2 contient les solutions de y Approximation de Taylor ordre 2
yT3(1) = y0 ; % yT3 contient les solutions de y Approximation de Taylor ordre 3
yT4(1) = y0 ; % yT4 contient les solutions de y Approximation de Taylor ordre 4
for n = 2 :Nint+1 % Calcul approximation d'Euler
yT1(n) = yT1(n-1) + h*y1(t(n-1),yT1(n-1)) ;
end % for n
for n = 2 :Nint+1 % Calcul de l'approximation de Taylor ordre 2
yT2(n) = yT2(n-1) + h*y1(t(n-1),yT2(n-1)) + (h^2/2)*y2(t(n-1),yT2(n-1)) ;
end % for n
for n = 2 :Nint+1 % Calcul de l'approximation de Taylor ordre 3
```




```
yT3(n) = yT3(n-1) + h*y1(t(n-1),yT3(n-1)) + ...
+(h^2/2)*y2(t(n-1),yT3(n-1))+(h^3/6)*y3(t(n-1),yT3(n-1)) ;
end % for n
for n = 2 :Nint+1 % Calcul de l'approximation de Taylor ordre 4
yT4(n) = yT4(n-1) + h*y1(t(n-1),yT4(n-1)) + ...
+(h^2/2)*y2(t(n-1),yT4(n-1))+(h^3/6)*y3(t(n-1),yT4(n-1)) + ...
+ (h^4/24)*y4(t(n-1),yT4(n-1)) ;
end % for n
% plot : créer un graphique avec les résultats des 4 courbes
axes('FontSize',14) ;
plot(texa,yexact,t,yT1,t,yT2,t,yT3,t,yT4,'+','LineWidth',2)
xlabel(' t ','FontSize',16) ;
ylabel(' y(t) ','FontSize',16) ;
legend('y(t) exacte','Euler (ordre 1)','Taylor ordre 2','Taylor ordre 3','Taylor ordre 4') ;
tit = sprintf('Méthodes de Taylor y"=t-y : N=%d, h=%.2f',Nint,h) ;
title(tit) ;
end
```

****

****

4.2.3 Méthode du point milieu

L'idée est que la corde de la fonction $y(t)$ sur $[t, t+h]$ a une pente voisine de $y'(t+\frac{h}{2})$, alors que dans la méthode d'Euler on approxime brutalement cette pente par $y'(t)$.

On écrit donc :

$$y(t+h) \simeq y(t) + hy'(t+\frac{h}{2}) \tag{1.14}$$

On a par ailleurs

$$y'(t+\frac{y}{2}) = f(t+\frac{h}{2}, y'(t+\frac{h}{2})). \tag{1.15}$$

Comme la valeur de $y'(t+\frac{h}{2})$ n'est pas connue , on l'approxime par :

$$y'(t+\frac{h}{2}) \simeq y(t) + hf(t+\frac{h}{2}, y(t) + \frac{h}{2}f(t,y(t)). \tag{1.16}$$

D'ou, on définitife :

$$y(t+h) \simeq y(t) + hf(t+\frac{h}{2}, y(t) + \frac{h}{2}f(t,y(t)) \tag{1.17}$$

L'**Algorithme du point milieu** peut donc s'écrire :

$$\begin{cases} y_{n+\frac{1}{2}} = y_n + \frac{h}{2}f\,(t_n, h) \\ p_n = f(t_n + \frac{h}{2}, y_{n+\frac{1}{2}}) \\ y_{n+1} = y_n + hp_n \\ t_{n+1} = t_n + h \;\; 0 \leq n \leq N-1 \end{cases} \tag{1.18}$$

Précision de la méthode du point milieu

Comme la méthode de Taylor d'ordre 2, la méthode du point milieu est d'ordre 2.

4.2.4 Méthode de Runge-Kutta

Soit de nouveau le problèmè de Cauchy

$$\begin{cases} y' = f(t,y) \\ y(t_0) = y_0 \end{cases} \tag{1.19}$$

On reprend l'algorithme de Taylor en écrivant la seconde équation de la manière suivante :

$$y_{n+1} = y_n + \frac{h}{2}f(t_n,y_n) + \frac{h}{2}\left[f(t_n,y_n) + h\frac{\partial f}{\partial t}(t_n,y_n) + h\frac{\partial f}{\partial y}(t_n,y_n)\cdot f(t_n,y_n)\right]$$

Selon le développement de Taylor on a, à des termes en h prés,

$$\left[f(t_n,y_n) + h\frac{\partial f}{\partial t}(t_n,y_n) + h\frac{\partial f}{\partial y}(t_n,y_n)\cdot f(t_n,y_n)\right] = f(t_n+h, y_n + hf(t_n,y_n)) \tag{1.21}$$

Ainsi, on obtient L'**Algorithme de Runge-kutta d'ordre 2 :**

$$\begin{cases} t_{n+1} = t_n + h \qquad 0 \leq n \leq N-1 \\ p_{n,1} = f(t_n, y_n) \\ p_{n,2} = f(t_{n+1}, y_n + \frac{h}{2}p_{n,1}) \\ y_{n+1} = y_n + \frac{h}{2}(p_{n,1} + hp_{n,2}) \end{cases} \tag{1.22}$$

La méthode de Rung-kutta la plus utilisée est d'ordre 4(on néglige les dérivées du 4^{eme} ordre dans le dévlopement de Taylor) ; Il s'agit de l'algorithme suivant :

Algorithme de Rung- Kutta d'ordre 4 :

$$\begin{cases} t_{n+1} = t_n + h \qquad 0 \leq n \leq N-1 \\ p_{n,1} = f(t_n, y_n) \\ p_{n,2} = (t_n + \frac{h}{2}, y_n + \frac{h}{2} p_{n,1}) \\ p_{n,3} = (t_n + \frac{h}{2}, y_n + \frac{h}{2} p_{n,2}) \\ p_{n,4} = f(t_{n+1}, y_n + h p_{n,3}) \\ y_{n+1} = y_n + \frac{h}{6}(p_{n,1} + 2p_{n,2} + 2p_{n,3} + p_{n,4}) \end{cases} \tag{1.23}$$

Ainsi,la méthode de Runge-kutta d'ordre 4 consiste à évaluer 4 valeurs intermédiaires $p_{n,1}, p_{n,2}, p_{n,3}$,et $p_{n,4}$ et à faire la moyenne pondérée .

Précision de la méthode de Rung -kutta

La méthode de runge -kutta d'ordre 2 est une méthode du second ord, i.e.

$$|y_n - y(t_n)| \leq kh^2 \tag{1.24}$$

et la méthode de Runge-kutta 4 est une méthode du quatriéme ordre, i.e.

$$|y_n - y(t_n)| \leq kh^4 \tag{1.25}$$

où k est une constante qui ne dépend pas du pas h.

Applications

Soit le problème de Cauchy

$$(1) \begin{cases} y' = y - \frac{2t}{y} \\ y(0) = 1 \end{cases} \tag{1.26}$$

On désire approcher, en effectuant le calcul avec six (06) décimales, la slution de (1) en $t = 0.2$ à l'aide des méthodes de Runge-kutta d'ordre 2 et d'ordre 4.

La solution exacte étant

$$y = \sqrt{2x+1}, \tag{1.27}$$

on estimera alors les résultats obtenus .

On considére donc, l'inetrvalle $[t_0, t_1]$ avec

$$t_0 = 0, t_1 = t_0 + h = 0.2$$

et

$$h = 0.2,$$

ie :

$$[t_0, t_1] = [0, 0.2]$$

1.Méthode de Rung-Kutta d'ordre 2 :

En utilisant l'algorithme correspondant, on obtient :

$$\begin{cases} t_1 = t_0 + h = 0.2 \\ p_{0,1} = f(t_0, y_0) = f(0,1) = 1 \\ p_{0,2} = f(t_1, y_0 + hp_{0,1}) = f(0.2, 1.2) = 0.866667 \\ y_1 = y_0 + \frac{h}{2}(p_{0.1} + p_{0.2}) = 1.186667 \end{cases} \tag{1.28}$$

Ainsi

$$y(0.2) \simeq y_1 = 1.186667$$

La valeur exacte étant

$$y(0.2) = \sqrt{1.4} = 1.183216$$

L'erreur commise est :

$$|y(0.2) - y_1| = |1.186667 - 1.183216| = 0.003451 < 0.510^{-2}$$

et donc

$$y(0.2) \simeq 1.19 \pm 0.01$$

ie :

y_1 approche $y(0.2)$ avec trois (03) c.s.e

2.Méthode de Runge-Kutta d'ordre 4 :

De l' algorithme correspondant, il vient :

$$\begin{cases} t_1 = t_0 + h \quad = 0.2 \\ p_{0,1} = f(t_0, y_0) = f(o.1, 1.1) = 1 \\ p_{0,2} = (t_0 + \frac{h}{2}, y_0 + \frac{h}{2}p_{0,1}) \\ p_{0,3} = (t_0 + \frac{h}{2}, y_0 + \frac{h}{2}p_{0,2}) = f(0.1, 1.0918182) = 0.0908637 \\ p_{0,4} = f(t_1, y_0 + hp_{0,3}) = f(0.2, 1.181727) = 0.843239 \\ y_1 = y_0 + \frac{h}{6}(p_{0,1} + 2p_{0,2} + 2p_{0,3} + p_{0,4}) = 1.183229 \end{cases} \tag{1.29}$$

Ainsi,

$$y(0.2) \simeq y_1 = 1.183229$$

La valeur exacte, calculée plus haut, étant

$$y(0.2) = \sqrt{1.4} = 1.183216$$

L'erreur est :

$$|y(0.2) - y_1| = |1.183216 - 1.183229| = 0.000013 < 0.510^{-4},$$

Donc,

$$y(0.2) \simeq 1.1832 \pm 0.0001$$

ie :

y_1 approche $y(0.2)$ avec cinq (05) c.s.e

Remarque Les méthodes à un pas considérées jusque là se programme facilement et en cours du calcul elles s'aprêtent sans problémes au changement de pas.

Programme

Une façon de programmer la méthode Runge Kutta d'ordre 2 est la suivante :

```
**************************************************************************************
                    Programme Runge Kutta d'ordre 2
**************************************************************************************
function [t,y] = RK2(f,tmin,tmax,Nint,y0) % Méthode de Runge Kutta d'ordre 2
% Nint - nombre de sous intervalles
% tmin - temps t0
% tmax - temps t0 + T
% f est une fonction avec comme arguments t et y : f(t,y(t))
% y0 contient les valeurs des conditions limites
h = (tmax-tmin)/Nint ; % valeur du pas (G)
t = linspace(tmin,tmax,Nint+1) ; % vecteur de t discrétisé t=[tmin,tmax]
y(1) = y0 ; % y contient les solutions de y(tn)n = 1, ...,Nint + 1
for n = 2 :Nint+1
k1 = h*feval(f,t(n-1),y(n-1)) ;
k2 = h*feval(f,t(n-1)+h/2,y(n-1)+k1/2) ;
y(n) = y(n-1) + k2 ;
end % for n
end
**************************************************************************************
**************************************************************************************
```

Comme la méthode d'Euler, les méthodes de Runge Kutta peuvent être appliquées à une fonction arbitraire.

4.3 Méthode numériques à pas multiples

Les méthodes à pas multiples sont les méthodes de résolution numérique où pour évaluer y_{n+1}en utilisant plusieurs y_{n-k} , $k = 0, 1, ..., r$. Au départ on doit alors calculer un premier ensemble de valeurs $y_1, ...y_n$ avec une autre méthode et puis commencer à itérer à partir de y_{r+1}en utilisant les valeurs précédantes. On suppose ici que le pas $h_n = h$ est constant. Les calculs alors effectués suivant le schéma suivant :

$$(\alpha) \qquad y_{n+1} = \sum_{i=0}^{r} \alpha_i y_i + h\beta_{i-1} f(t_{n+1}, y_{n+1}) + h \sum_{i=0}^{r} \beta_i f(t_{n-i}, y_{n-i}) \tag{2.1}$$

où les α_i, β_isont des constantes réelles .

Si $\beta_{-1} = 0$, le schéma est dit explicite (ou bien forme ouverte) : y_{n+1} est oubtenu directement par l'application de la formule .

Si $\beta_{-1} \neq 0$, le schéma estimpicite car il faut résoudre une équation de la forme

$$y_{n+1} = g(y_{n+1}) \tag{2.2}$$

pour obtenir y_{n+1}. Dans ce cas la formule est appelée fermée ou encore formule de prédiction , trouvée antérieurement .

La supérorité des méthode à pas multiples sur les méthodes à pas constant réside dans le fait qu 'elle ne nécessite pas d'évaluation de f en des points intermédiaires (sauf au démarage). Par rapport à une méthode de Runge-kutta du méme ordre, le temps de calcul est donc réduit dans une proportion importante .

Remarque Comme il a été montionné plus haut, le point initial (t_0, y_0) étant donné, l'algorithme ne peut démarrer que si les valeurs $(y_1, f(t_{1,}y_1)), ..., (y_r, f(y_r, f(t_r, y_r))$ ont déja calculées.

Le calcul ne peut être fait que par une méthode à un pas pour $(y_1, f(y_1, t_1))$, à au plus deux pas pour $(y_2, f(y_{2,}t_2))$, ... au plus r pas pour $(y_r, f(t_r, y_r))$.

L'initiation de des r premières valeurs $(y_i, f(y_i, t_i))$, $1 \leq i \leq r$, sera généralement faite à l' aide d'une méthode de Rung- kutta d'ordre supérieur ou égal à celui de la méthode (α), ou à la rigueur un de moins.

4.3.1 Méthode d'Adams-Bashforth

1er Cas : r=1

Si y(t) est une solution exacte associée au problème de Cauchy (1), on écrit :

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt \tag{2.3}$$

Supposons qu' on ait déja calculé les points $y(t_{n-1})$ et $y(t)$ et les pentes $f_{n-1} = f(t_{n-1,}y(t_{n-1}))$ et $f_n = f(t_n, y(t_n))$. La méthode d'Adams-Bashforth consiste à approximer la fonction $f(t, y(t))$, en tant que fonction de t, par son polynôme d'interpolation aux points t_{n-1} et t_n.Soit $P(t)$ ce polynôme, $P(t)$ est la fonction affine définie par :

$$P(t) = f_n + \frac{f_n - f_{n-1}}{t_n - t_{n-1}}(t - t_n) \tag{2.4}$$

On écrit alors ,

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} f(t, y(t))dt \tag{2.5}$$

$$\simeq y(t_n) + \int_{t_n}^{t_{n+1}} p(t)dt \tag{2.6}$$

$$= y(t_n) + h(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}) \qquad \text{(après calcul)} \tag{2.7}$$

L'algorithme d 'Adams-bashforth à 2 pas va donc s'écrire :

$$\begin{cases} y_0, y_1 \text{ donnés} \\ y_{n+1} = y_n + h(\frac{3}{2}f_n - \frac{1}{2}f_{n-1}) \\ t_{n+1} = t_n + h \\ f_{n+1} = f(t_n, y_n) \end{cases} \tag{2.8}$$

Précision :

La méthode d'Adams-Bashforth 2 pas est ordre 2 , ie l' erreur au point t_n est donnée par :

$$|y(t_n) - y(t)| \leq kh^2 \tag{2.9}$$

où k est indépendant de n et de h.

Initialisation : pour calculer y_i il est préférable d' utilser une méthode de même ordre (par exemple la méthode de Runge -Kutta d'ordre 2).

Remarque 4.3.1 *On peut augmenter la précision de la méthode d'Adams-Bashforth en considérant le cas $r = 2$.*

2 ème cas : $r = 2$

Un raisonnement analogue au cas où $r = 1$, nous amène à l'**Algorthme d'Adams-Bashforth à 3 pas** qui s'écrit :

$$\begin{cases} y_0, y_{1,} y_2 \text{ donnés} \\ y_{n+1} = y_n + h(\frac{23}{12}f_n - \frac{4}{3}f_{n-1} + \frac{5}{12}f_{n-2}) \\ t_{n+1} = t_n + h \\ f_{n+1} = f(t_n, y_n) \end{cases} \tag{2.10}$$

Précision de la méthode d'Adams-Bashforth

La méthode d'Adams-Bashforth à 3 pas est d'ordre 3.

4.3.2 Méthode d'Adams-Moulton

Soit y(t) une solution exacte .Le développement de Taylor de y(t) s'écrirait :

$$y(t) = y(t_{n+1} - h) = y(t_{n+1}) - hy'(t_{n+1}) + o(h^2) \tag{2.11}$$

$$= y(t_{n+1}) - hf(t_{n+1,} y(t_{n+1})) + o(h^2) \tag{2.12}$$

En négligeant les termes en h^2, l'**Algorithme d'Adams -Moulton** d'ordre 2 s'en déduit alors :

$$\begin{cases} y_0 \text{donnée} \\ y_{n+1} = y_n + hf(t_{n+1} + y_{n+1}) \\ t_{n+1} = t_n + h \end{cases} \tag{2.13}$$

La première équaton de l'algorithme (2.13) s'écrit aussi :

$$y_{n+1} - hf(t_{n+1} + y_{n+1}) \quad = y_n \tag{2.14}$$

Remarque 4.3.2 *On augmente la précision de la méthode d'Adams-moulton en négligeant les termes en h^3 dans le developement de Taylor. On aboutit alors à :*

Algorithme de la méthode d'**Adams-moulton d'**ordre 3, dite aussi méthode des trapèzes (ou méthode de Crank-Nicolson) :

$$\left\{ \begin{array}{c} y_0 \text{ donnée} \\ y_{n+1} = y_n + \frac{h}{2}(f_{n+1} + f_n) \\ t_{n+1} = t_n + h \end{array} \right. \tag{2.15}$$

où

$$f_{n+1} = f(t_{n+1}, y_{n+1})$$

et

$$f_n = f(t_n, y_n)$$

La première équation de l'algorithme (2.15) peut s'écrire encore

$$y_{n+1} - \frac{h}{2} f(t_{n+1} + y_{n+1}) \quad = y_n + \frac{h}{2} f_n$$

On fait de même pour obtenir l'**Algorithme d'Adams -Moulton** d'ordre 4 :

$$\left\{ \begin{array}{c} y_0, y_1 \text{ donnée} \\ y_{n+1} = y_n + h(\frac{5}{12} f_{n+1} + \frac{8}{12} f_n - \frac{1}{12} f_{n-1}) \\ t_{n+1} = t_n + h \end{array} \right. \tag{2.16}$$

où

$$f_{n+1} = f(t_{n+1}, y_{n+1}), f_n = f(t_n, y_n) \tag{2.17}$$

et

$$f_{n-1} = f(t_{n-1}, y_{n-1}). \tag{2.18}$$

Initialisation :Comme pour la méthode d'Adams-Bashforth, on initialise y_1à l'aide d'une méthode de même ordre (par exemple la méthode de Rnng-Kutta corresponante).

4.3.3 Méthode de prédiction-correction

On ce donne une méthode dit de prédiction fournissant (explicitement) une premire valeur approchée py_{n+1} du point y_{n+1} à atteindre :

$$py_{n+1} = \text{prédiction de } \ y_{n+1}$$

$$pf_{n+1} = f(t_{n+1,}py_{n+1}) = \text{prédiction de } f_{n+1} \tag{2.20}$$

En substituant la valeur pf_{n+1} ainsi trouvée à f_{n+1} dans la formule d'Adams-Moulton, on obtient alors une nouvelle valeur corrigée y^*_{n+1}qui est retenue en vue des calculs ultérieurs .

La première approximaiton py_{n+1} de y_{n+1} est dite le prédicteur de $y_{n+1,}$ et la valeur corrigée y^*_{n+1}, le correcteur. D'ou le nom (de cette méthode) : prédicteur -correcteur .

Application

Méthode de prédiction-correction d'Adams-Moulton d'ordre 3 :

$$\begin{cases} \text{prédiction} : py_{n+1} = y_n + hf_n(\text{méthode d'Euler }) \\ t_{n+1} = t_n + h \\ \text{Evaluation} : \ pf_{n+1} = f(t_{n+1,}py_{n+1}) \\ \text{Correction} : y^*_{n+1} = y_n + h(\frac{1}{2}pf_{n+1} + \frac{1}{2}f_n) \\ \text{Evaluation} : \ f_{n+1} = f(t_{n+1,}y_{n+1}) \end{cases} \tag{2.21}$$

4.4 Autres Méthodes

4.4.1 Méthode d'Adams

Soit y(t) une solution exacte du problème de Cauchy (1.1). Supposons qu' on ait déjà calculé, par un procédé quelconque, les trois valeurs suivantes de $y(t)$:

$$y_1 = y(t_1) = y(t_0 + h) \tag{3.1}$$

$$y_2 = y(t_2) = y(t_0 + 2h) \tag{3.2}$$

$$y_3 = y(t_1) = y(t_0 + 3h) \tag{3.3}$$

Compte tenu de la condition initiale

$$y(t_0) = y_0 \tag{3.4}$$

et à l'aide des nombres :

$$t_0, t_1, t_2, t_3 \text{ et } y_0, y_1, y_2, y_3$$

On calcule les grandeurs $u_{0,}u_{1,}u_2, u_3$ où

$$u_0 = hf(t_0, y_0) \tag{3.5}$$

$$u_1 = hf(t_1, y_1) \tag{3.6}$$

$$u_2 = hf(t_2, y_2) \tag{3.7}$$

et

$$u_3 = hf(t_3, y_3) \tag{3.8}$$

En suite on forme le tableau des différences finies des grandeurs y_i et u_i :

x_i	y_i	$\triangle y_i$	u_i	$\triangle u_i$	$\triangle^2 u_i$	$\triangle^3 u_i$
t_0	y_0		u_0			
		$\triangle y_0$		$\triangle u_0$		
t_1	y_1		u_1		$\triangle^2 u_0$	
		$\triangle y_1$		$\triangle u_1$		$\triangle^3 u_0$
t_2	y_2		u_2		$\triangle^2 u_1$	
		$\triangle y_2$		$\triangle u_2$		$\triangle^3 u_1$
t_3	y_3		u_3		$\triangle^2 u_2$	
		$\triangle y_3$		$\triangle u_3$		$\triangle^3 u_2$
t_4	y_4		u_4		$\triangle^2 u_3$	
.	.	$\triangle y_4$	.	$\triangle u_4$	.	$\triangle^3 u_3$
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
t_{n-2}	y_{n-2}	.	u_{n-2}	.	.	.
		$\triangle y_{n-2}$		$\triangle u_{n-2}$	.	$.\triangle^3 u_{n-3}$
t_{n-1}	y_{n-1}		u_{n-1}		$\triangle^2 u_{n-2}$	
		$\triangle y_{n-1}$		$\triangle u_{n-1}$		
t_n	y_n		u_n			

Au bas du tableau, la conaissance des nombres de larangée oblique composée de

$u_n \quad \triangle u_{n-1} \qquad \triangle^2 u_{n-2} \qquad\qquad \triangle u_{n-3}$

permet de trouver y_n par la formule d'Adams :

$$\triangle y_n = u_n + \frac{1}{2} \triangle u_{n-1} + \frac{5}{12} \triangle^2 u_{n-2} + \frac{3}{8} \triangle^3 u_{n-3} \tag{3.9}$$

qui s'écrit encore (car$\triangle y_n = y_{n+1} - y_n$):
formule d'Adams,

$$y_{n+1} = y_n + u_n + \frac{1}{2} \triangle u_{n-1} + \frac{5}{12} \triangle^2 u_{n-2} + \frac{3}{8} \triangle^3 u_{n-3} \tag{A}$$

Pour $n = 3$: On obtient ,

$$y_4 = y_3 + u_3 + \frac{1}{2} \triangle u_2 + \frac{5}{12} \triangle^2 u_1 + \frac{3}{8} \triangle^3 u_{0.} \tag{3.10}$$

Si on connait y_4 on peut calculer

$$u_4 = hf(t_4, y_4) \tag{3.11}$$

ce qui permet d'écrire la rangée oblique suivante :

$$u_4 = hf(t_4, y_4),\ \triangle u_3 = u_4 - u_3, \triangle^2 u_2 = \triangle u_3 - \triangle u_2,\ \ \triangle^3 u_1 = \triangle^2 u_2 - \triangle^2 u_1 \tag{3.12}$$

La nouvelle diagonale donne la possibilité de calculer par la formule d'Adams (A), la valeurs de

$$y_5 = y_4 + u_4 + \frac{1}{2} \triangle u_3 + \frac{5}{12} \triangle^2 u_2 + \frac{3}{8} \triangle^3 u_{1,} \tag{3.13}$$

et ainsi de suite ....

Ainsi , la formule (A) résou le probléme posé. Lorsqu 'on a estimé $y(t)$ pour les valeurs $t_{1,}..., t_n$ elle fournit l'approximation y_{n+1} de $y(t_{n+1})$. On peut encore calculer

$$u_{n+1} = hf(t_{n+1,} y_{n+1}) \tag{3.14}$$

compléter le tableau, et recommencer les mèmes opérations en remplaçant n par $n + 1$.

Précision : La méthode d'Adams, représentée par la formule (A), est d'ordre 5,c'est-à-dire :

$$|y_n - y(t_n)| \leq kh^5 \tag{3.15}$$

où y_n est la valeur calculée par la formule (A), $y(t_n)$ la valeur exacte de la solution $y(t)$ du problème de Cauchy (1.1) au point t_n et k une constante qui ne dépend pas de h.

Applications

Afin de calculer la valeur approchée, au point $t = 2$, de la solution de l'équation différentielle $t^2y' - ty = 1$, vérifiant la condition initiale $y(1) = 0$,appliquons la méthode d'Adams, avec $h = 0.2$.

Pour cela, soit

$$y_0 = y(1) = 0 \tag{3.16}$$
$$y_1 = y(1.2) = 0.1834$$
$$y_2 = y(1.4) = 0.3429$$

et

$$y_3 = y(1.6) = 0.4898$$

(calculés préalablement dans notre cas par la méthode de **Runnge-kutta** d'ordre 4).

L'équation differentielle associée au problème s'écrit encore :

$$t^2y' - ty = 1 \Longleftrightarrow y' = \frac{1}{y}(y + \frac{1}{t}) = f(t, y) \tag{3.17}$$

Calculons alors $u_{0,}u_{1,}u_2$ et u_3. Nous avons :

$$u_0 = hf(t_{0,}y_0) = 0.2$$

$$u_1 = hf(t_{1,}y_1) = 0.1695$$

$$u_2 = hf(t_{2,}y_2) = 0.1510$$

$$u_3 = hf(t_{3,}y_3) = 0.1394$$

Et le tableau des différences fnies des grandeurs y_i et u_i s'écrit :

x_i	y_i	$\triangle y_i$	u_i	$\triangle u_i$	$\triangle^2 u_i$	$\triangle^3 u_i$
1.0	0		0.2			
		0.1834		-0.0305		
1.2	0.1834		0.1695		0.120	
		0.1595		-0.0105		-0.0051
1.4	0.3429		0.1510		0.0069	
		0.1469		-0.0116		
1.6	0.4898		0.1394			

Alors, d'après la formule d'Adams (A), on obtient :

$$y_4 = y_3 + u_3 + \frac{1}{2} \triangle u_2 + \frac{5}{12} \triangle^2 u_1 + \frac{3}{8} \triangle^3 u_0 \tag{3.18}$$

$$= 0.4898 + 0.1394 - 0.0058 + 0.0029 - 0.0019 = 0.6244$$

La solution exacte étant

$$y(t) = \frac{t}{2} - \frac{1}{2t}. \tag{3.19}$$

Estimons la valeur obtenue. L'erreur commise est égale à :

$$y_4 - y(1.8) = 0.6244 - 0.6222 = 0.0022 \leq 0.510^{-2}$$

Ainsi,

$$y_4 = 0.6244 \simeq 0.62 \pm 10^{-2}$$

Et comme on connnait y_4, on peut calculer

$$u_4 = h.f(t_4, y) = 0.1311 \tag{3.20}$$

Cela permet de compléter le tableau par la diagonale inférieure suivantes :

$$u_4, \triangle u_3, \triangle^2 u_{2,} \triangle^3 u_1 \tag{3.21}$$

Où

$$\triangle u_3 = u_4 - u_3 = 0.1311 - 0.1394 = -0.0083$$

$$\triangle^2 u_2 = \triangle u_3 - \triangle u_2 = -0.0083 + 0.0116 = 0.0033$$

$$\triangle^3 u_1 = \triangle^2 u_2 - \triangle^2 u_1 = 0.0033 - 0.0069 = -0.0036$$

La nouvelle diagonale donne la possibilité de calculer par la formule d'Adams (A), la valeur :

$$y_4 = y_4 + u_4 + \frac{1}{2} \triangle u_3 + \frac{5}{12} \triangle^2 u_2 + \frac{3}{8} \triangle^3 u_1 \tag{3.22}$$

$$= 0.6244 + 0.1311 - 0.0042 + 0.0014 - 0.0014 = 0.7513$$

l'erreur commise dans ce cas est égale à :

$$y_5 - y(2) = 0.7513 - 0.7500 = 0.0013 \leq 0.510^{-2}$$

Ainsi ,une valeur approchée, au point $t = 2$, de la solution exacte du problème de Cauchy ci-dessus est donnée par :

$$y_5 = 0.7513 \simeq 0.75 \pm 10^{-2}.$$

4.4.2 Méthode des approximations successives (Picard)

La méthode des approximaton successives de Picard est une méthode analytique qui permet de trouver une solution approchée du problème de Cauchy (1.1).Si $y(t)$
est la solution exacte de (1), elle s'écrirait :

$$y(t) = y_0 + \int_{t_0}^{t} f(s, y(s))ds \tag{3.23}$$

En substituant $y_0 = y(t_0)$ à $y(t)$ sous le signe intégrle (car la fonction $y(t)$ est inconnue), on obtient la premiére approximation :

$$y_1(t) = y_0 + \int_{t_0}^{t} f(s, y_0)ds \tag{3.24}$$

La deuxième itération est obtenu en portant dans la formule (3.23) la valeur trouvée $y_1(t)$ à la place de la fonction inconnue $y(t)$:

$$y_2(t) = y_0 + \int_{t_0}^{t} f(s, y_1(s))ds \tag{3.25}$$

et ainsi de suite ...
L'algorithme de la **méthode des approximations successives** s'ecrit alors :

$$\begin{cases} y_0(t) = y_0 \\ y_n(t) = y_0 + \int_{t_0}^{t} f(s, y_{n-1}(s))ds & n = 1, 2, ... \end{cases} \tag{3.26}$$

Les $y_n(t)$ sont appelées les approximations successives(de $y(t)$, solution de (3.26)).

Si

$$f'_y(t,y) > 0 \tag{3.27}$$

et

$$f(t,y_0) > 0 \tag{3.28}$$

Les approximations de Picard forment une suite croissante de fonctions inférieures à

$$y(t) : y_0 < y_1 < ... < y_n < y(t). \tag{3.29}$$

Par contre si $f(t,y_0) < 0$, alors la suite des approximations est décroissante :

$$y_0 > y_1 > ... > y_n > y(t). \tag{3.30}$$

Et ainsi lorsque $f'_y(t,y) > 0$, les approximations de Picard forment une suite d'approximations unilatérales.

Si$f'_y(t,y) < 0$,les approximations forment une suite bilatérales.

Exemple 7.5 En appliquant la méthode des approximations successives à la résolution du probléme de Cauchy :

$$\begin{cases} y' = y + t & t \leq 0 \\ y(0) = 0 \end{cases}$$

on trouve, en partant de $y_0 = y(0) = 0$:

$$y_{n+1} = \int_0^t (s + y_n(s))ds \qquad n = 1, 2, ...$$

donc

$$\begin{cases} y_1(t) = \frac{t^2}{2!} \\ y_2(t) = \frac{t^2}{2!} + \frac{t^3}{3!} \\ y_3(t) = \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} \\ \vdots \\ y_n(t) = \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots + \frac{t^{n+1}}{(n+1)!} \end{cases}$$

Ici, $f(t,y_0) = t + y_0 \geq 0$ et $f'_y(t,y) = 1 \rangle 0$**.** Par conséquent la suite des approximations de Picard(y_n), forme une suite de fonctions inférieurs (suite décroissante)**.**

La solution exacte du problème est obtenue par :

$$\begin{aligned} y(t) &= \lim_{n\to+\infty} (\frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots + \frac{t^{n+1}}{(n+1)!}) \\ &= \lim_{n\to+\infty} (1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \frac{t^4}{4!} + \cdots + \frac{t^{n+1}}{(n+1)!}) - t - 1 \\ &= e^t - t - 1 \end{aligned}$$

4.5 Stabilité des solutions

On se propose ici d'étudier le comportement des solutions d'une équation différentielle et des lignes intégrales d'un champ de vecteurs lorsque le temps t tend vers l'infini. On s'intéresse essentiellement au cas des équations linéaires ou voisines de telles équations. Dans ce cas, le comportement des solutions est gouverné par lesigne de la partie réelle des valeurs propres de la matrice associée à la partie linéaire de l'équation : une solution est dite stable si les solutions associées à des valeurs voisines de la donnée initiale restent proches de la solution considérée jusqu'à l'infini.

Cette notion de stabilité (dite aussi stabilité au sens de Lyapunov) ne devra pas être confondue avec la notion de stabilité d'une méthode numérique, qui concerne la stabilité de l'algorithme sur un intervalle de temps fixé. On étudie finalement les différentes configurations possibles des lignes intégrales au voisinages des points singuliers non dégénérés d'un champ de vecteurs plan.

Définition 4.5.1 *Soit $y(t,z)$ la solution maximale de (1) tel que $y(t0,z)=z$. On dira que la solution $y(t,z_0)$ est stable s'il existe une boule $\overline{B}(z_0,r)$ et une constante $C\geq 0$ telles que*

(i) Pour tout $z\in\overline{B}(z_0,r)$, $t\to y(t,z)$ est définie sur $[t_0,+\infty[$;

(ii) Pour tous $z\in\overline{B}(z_0,r)$ et $t\geq t_0$ on a $\|y(t,z)-y(t,z_0)\|\leq C\|z-z_0\|$.

La solution $y(t,z_0)$ est dite asymptotiquement stable si elle est stable et si la condition (ii') plus forte que (ii) est satisfaite :

(ii') Il existe une boule $\overline{B}(z_0,r)$ et une fonction $\gamma:[t_0,+\infty[\to\mathbb{R}_+$ continue avec $\gamma(t)=0$ telles que pour tous $z\in B(z_0,r)$ et $t\geq t_0$ on ait

$$\|y(t,z)-y(t,z_0)\|\leq\gamma(t)\|z-z_0\|.$$

4.5.1 Cas d'un système linéaire à coefficients constants

Nous étudierons d'abord le cas le plus simple, à savoir le cas d'un système linéaire sans second membre

$$Y'=AY,\ Y=\begin{pmatrix}y_1\\ \vdots\\ y_m\end{pmatrix},\ A=\begin{pmatrix}a_{11}&\cdots&a_{1m}\\ \vdots&\ddots&\vdots\\ a_{m1}&\cdots&a_{mm}\end{pmatrix}\qquad((E))$$

avec $y_j,\ a_{ij}\in\mathbb{C}$; le cas réel peut bien entendu être vu comme un cas particulier du cas complexe. La solution du problème de Cauchy de condition initiale $Y(t_0)=Z$ est donnée par $Y(t,Z)=e^{(t-t_0)A}\cdot Z$. On a donc

$$Y(t,Z)-Y(t,Z_0)=e^{(t-t_0)A}\cdot(Z-Z_0)$$

et la stabilité est liée au comportement de $e^{(t-t_0)A}$ quand t tend vers $+\infty$, dont la norme $\left\|e^{(t-t_0)A}\cdot\right\|$ doit rester bornée.

Théorème 4.5.1 *Soient $\lambda_1,...,\lambda_m$ les valeurs propres complexes de la matrice A.*

Alors les solutions du système linéaire $Y'=AY$ sont

• *asymptotiquement stables si et seulement si* $Re(\lambda_j) < 0$ *pour tout* $j = 1, ..., m$.

• *stables si et seulement si pour tout* j, *ou bien* $Re(\lambda_j) < 0$, *ou bien* $Re(\lambda_j) = 0$ *et le bloc correspondant est diagonalisable.*

4.5.2 Ptite perturbation d'un système linéaire

On considère dans $\mathbb{k}^m = \mathbb{R}^m$ ou $\mathbb{C}^m$ un système de la forme

$$Y' = AY + g(t, Y) \qquad (E)$$

où $g : [t_0, +\infty[\times \mathbb{k}^m \to \mathbb{k}^m$ est une fonction continue. On se propose de montrer que si la partie linéaire est asymptotiquement stable et si la perturbation g est suffisamment petite, en un sens à préciser, alors les solutions de (E) sont encore asymptotiquement stables.

Théorème 4.5.2 *On suppose que les valeurs propres complexes* λ_j *de* A *sont de partie réelle* $Re\lambda_j < 0$.

(a) S'il existe une fonction $k : [t_0, +\infty[\to \mathbb{R}_+$ *continue telle que* $\lim_{t\to\infty} k(t) = 0$ *et*

$$\forall t \in [t_0, +\infty[,\ \forall Y_1, Y_2 \in \mathbb{k}^m, \qquad \|g(t, Y_1) - g(t, Y_2)\| \leq k(t) \|Y_1 - Y_2\|,$$

alors toute solution de (E) *est asymptotiquement stable.*

(b) Si $g(t, 0) = 0$ *et s'il existe* $r_0 > 0$ *et une fonction continue* $k : [0, r_0] \to R+$ *telle que* $\lim_{r\to\infty} k(r) = 0$ *et*

$$\forall t \in [t_0, +\infty[,\ \forall Y_1, Y_2 \in \overline{B}(0, r), \qquad \|g(t, Y_1) - g(t, Y_2)\| \leq k(r) \|Y_1 - Y_2\|,$$

pourr $\leq r_0$, *alors il existe une boule* $B(0, r_1) \subset B(0, r_0)$ *telle que toute solution* $Y(t, Z_0)$ *de valeur initiale* $Z_0 \in B(0, r_1)$ *soit asymptotiquement stable.*

Travaux dirigés 4

Exercice 01 :
Soit l'équation différentielle à condition initiale $y'(t) = y(t) + t$ et $y(0) = 1$. Approcher la solution de cette équation en $t = 1$ à l'aide de la méthode d'Euler en subdivisant l'intervalle de travail en 10 parties égales. Comparer à la solution exacte.

Exercice 02 :
Approcher la solution de l'équation différentielle ci-dessous en $t_1 = 0.2$ en utilisant RK2, avec un pas $h = 0.2$

$$y'(t) = y(t) - \frac{2t}{y} \quad et \quad y(0) = 1$$

Comparer à la solution exacte.

Exercice 03 :
Soit le problème de Cauchy suivant

$$\begin{cases} \text{y'(t)} & = \text{t + y(t)}, \, t \in [0,1] \\ \text{y(0)} & = 1 \end{cases}$$

1. Trouver la solution exacte de ce problème.
2. Appliquer la méthode d'Euler à ce problème, avec $h = 0.1$, puis evaluer la solution en $t = 0.3$. Comparer à la solution exacte.

Exercice 04 :
En donnant les solutions de l'équation différentielle ci-dessous avec la condition initiale $y(0) = 1$ puis $y(0) = 1 + \epsilon$, ϵ réel non nul, vérifier qu'elle conduit à des schémas instables.

$$y'(t) = 36y(t) - 37e^{-t}.$$

Suggestions et Corrigés

Exercice 01

$$\begin{cases} y'(t) &= y(t)+t = f(t,y) \\ y(0) &= 1 \end{cases} \quad (1)$$

L'intervalle d'intégration est $[0,1]$.

Remarquons tout d'abord que f étant continue et lipshitzienne par rapport à y le problème de Cauchy (1) admet une solution unique (théorème 9 de Cauchy-Lipshitz).

Méthode d'Euler Elle s'écrit :

$$\begin{aligned} y_{n+1} &= y_n + hf(t_n, y_n) \\ &= y_n + h(t_n + y_n) \\ &= y_n(1+h) + ht_n \end{aligned}$$

On a aussi $y(0) = y_0 = 1$, $h = \frac{1-0}{10} = 0.1$ $t_0 = 0$ et $t_n = t_0 + nh = \frac{n}{10}$.
D'où le tableau,

n	0	1	2	3	4	5	6	7	8	9	10
t_n	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
y_n	1	1.1	1.22	1.362							3.1874

C'est à dire que l'approximation en $t = 1$ de $y(t)$, est $y_{10} = 3.1874$.

Solution exacte de cette équation Appliquons la méthode de la variation de la constante.

1ère étape : équation sans second membre.

$$\frac{dy(t)}{dt} = y'(t) = y$$

$y = 0$ est une solution evidente. Les autres solutions sont données par $\frac{dy}{y} = dt$. D'où $y = ke^t$ avec $k \in \mathbb{R}$ (2).

2nde étape : une solution particulière On applique la méthode de la variation de la constante $y = ke^t$ d'où $y' = k'e^t + ke^t$ que l'on reporte dans (1) : $k'e^t + ke^t = ke^t + t$ ainsi, $k' = te^{-t}$.
$k = \int te^{-t}\,dt$ en intégrant par parties on trouve

$$\begin{aligned} k &= -te^{-t} + \int e^{-t}\,dt \\ &= -te^{-t} - e^{-t} + c \\ &= e^{-t}(-1-t) + c \qquad (3) \end{aligned}$$

avec $c \in \mathbb{R}$.

3ème étape : solution générale. On remplace k donné par (3) dans (2) :

$$y = \big(e^{-t}(-1-t) + c\big)e^{t}$$

donc,

$$y = -1 - t + ce^{t}$$

Finalement, grâce à la condition initiale $y(0) = 1$, on détermine c, d'où

$$y(0) = 1 = -1 - 0 + ce^{0} \Rightarrow c = 2$$

Ainsi, la solution exacte de (1) est $y = -1 - t + 2e^{t}$.

Estimation de l'erreur. La solution exacte ci-dessus donne $y(1) = -1-1+2e = 3.4366$. Ainsi, l'erreur effectivement commise lors de l'application de la méthode d'Euler est $|E_e| = |3.4366 - 3.1874| = 0.25$. Cherchons l'erreur théorique qui est donnée par :

$$E_t \leq (e^{L(b-a)} - 1)\frac{M_2 \times h}{2L}$$

Où $M_2 = max_{[a,b]}|y''(t)|$ et L est la constante de Lipschitz de f par rapport à y, qui se calcule aisément :

$$|f(t,y) - f(t,z)| = |y - z| \Rightarrow L = 1.$$

De même, on a

$$\begin{aligned} y''(t) &= 1 + t + y \\ &= 1 + t + (-1 - t + 2e^{t}) \\ &= 2e^{t} \end{aligned}$$

Ainsi $M_2 = 2e$. Donc,

$$\begin{aligned} |E_t| &\leq (e^{1(1-0)} - 1)\frac{2e10^{-1}}{2 \times 1} \\ &\leq (e-1)\frac{e}{10} = 0.4673 \end{aligned}$$

Clairement, $|E_e| \leq |E_t|$, donc la méthode d'Euler donne une bonne approximation de la solution de ce problème de Cauchy en $t = 1$.

Exercice 02

$y'(t) = y(t) - \frac{2t}{y} = f(t,y)$ et $y(0) = 1$

Remarquons tout d'abord que f étant continue et lipshitzienne par rapport à y ce problème de Cauchy admet une solution unique (théorème 9 de Cauchy-Lipshitz).

L'intervalle d'intégration est $[0, 0.2]$ et le pas d'intégration est $h = 0.2$.

Méthode de RK2 Elle s'écrit :

$$y_{n+1} = y_n + \frac{h}{2}(M_1 + M_2),$$

soit

$$y_1 = y_0 + \frac{h}{2}(M_1 + M_2)$$

, avec $M_1 = f(0.1) = 1$ et $M_2 = f(0.2, 1 + 0.2 \cdot 1)$. Donc $f(0.2, 1.2) = 0.8666$.

Ainsi, l'approximation en $t = 0.2$ de $y(t)$, est

$$\begin{aligned} y_1 &= 1 + \frac{h}{2}(M_1 + M_2) = 1 + 0.1(1.8666) \\ y_1 &= 1.18666 \end{aligned}$$

Solution exacte de cette équation

$$(1) \quad y' = y - \frac{2t}{y}, \qquad y \neq 0$$

En multipliant (1) par y on a $y'y = y^2 - 2t$ d'où $\frac{1}{2}(y^2) = y^2 - 2t$. On pose donc $u = y^2$ d'où :

$$(2) \quad \frac{1}{2}u' = u - 2t$$

Ce qui est une équation différentielle linéaire du 1^{er} ordre. On l'intègre par la méthode de la variation de la constante comme à l'exercice précédent.

L'équation sans second membre est $u' = 2u$ de solution $u = 0$ ou $u = ke^{2t}$.

Une solution particulière par la variation de la constante. On a

$$(3) \quad u' = \lambda' e^{2t} + 2\lambda e^{2t}$$

avec $\lambda \in \mathbb{R}$. D'où, dans (2) $\frac{1}{2}\lambda' e^{2t} = -2t$ ce qui implique $\lambda' = -4te^{-2t}$.

Intégrons λ par parties :

$$\begin{aligned}\lambda &= -4\left[-\frac{1}{2}te^{-2t}+\frac{1}{2}\int e^{-2t}dt\right]\\ &= 2te^{-2t}+e^{-2t}+c\\ &= (2t+1)e^{-2t}+c\end{aligned}$$

avec $c \in \mathbb{R}$. La solution générale est donc $u = 2t+1+ce^{2t}$ comme $y(0) = 1$, $c = 0$. Finalement, $u = y^2 = 2t+1$. Ainsi, $y = \pm\sqrt{2t+1}$. Comme $y(0) = 1 > 0$ alors $y = \sqrt{2t+1}$.

Estimation de l'erreur. La solution exacte est $y(0.2) = \sqrt{2\cdot 0.2+1} = 1.18321$. Donc l'erreur commise est $|E_e| = |1.18321-1.18666| = 0.00345$. On peut comparer cette erreur effective à l'erreur théorique sur RK2, donnée par : ————-
(le théorème du cours)

Exercice 03

Soit le problème de Cauchy suivant

$$\begin{cases} y'(t) &= 2t - y(t) = f(t,y)\\ y(0) &= 1\end{cases} \quad (1)$$

Remarquons tout d'abord que f étant continue et lipshitzienne par rapport à y le problème de Cauchy (1) admet une solution unique (théorème 9 de Cauchy-Lipshitz).

Méthode d'Euler. Elle s'écrit :

$$\begin{aligned}y_{n+1} &= y_n + hf(t_n,y_n)\\ &= y_n + h(2t_n - y_n)\\ &= y_n(1-h)+2ht_n\end{aligned}$$

On a aussi $y(0) = y_0 = 1$, $h == 0.1$ $t_0 = 0$ et $t_n = nh$.
Donc $y_{n+1} = 0.9y_n + 0.02n$, d'où le tableau,

n	0	1	2	3
t_n	0	0.1	0.2	0.3
y_n	1	0.92	0.868	0.8412

C'est à dire que l'approximation en $t = 0.3$ de $y(t)$ avec le pas $h = 0.1$, est $y_{10} = 0.8412$.

Solution exacte de cette équation. En appliquant la méthode de la variation de la constante, comme au premier exercice, on trouve la solution générale, $y(t) = 2t - 2 + 3e^{-t}$.

Estimation de l'erreur. La solution exacte ci-dessus donne $y(0.3) = 0.822$. Ainsi, l'erreur effectivement commise lors de l'application de la méthode d'Euler est $|E_e| = |0.822 - 0.841| = 0.019$. Cherchons l'erreur théorique qui est donnée par :

$$E_t \leq (e^{L(b-a)} - 1)\frac{M_2 \times h}{2L}$$

Où $M_2 = max_{[a,b]}|y''(t)|$ et L est la constante de Lipschitz de f par rapport à y, qui est ici clairement égale à 1. On a

$$y''(t) = 3e^{-t}.$$

Ainsi $M_2 = 3$. Donc,

$$\begin{aligned} |E_t| &\leq (e^{1(0.3-0)} - 1)\frac{3 \times 10^{-1}}{2 \times 1} \\ &\leq 0.15(e^{0.3} - 1) \approx 0.05247. \end{aligned}$$

Clairement, $|E_e| \leq |E_t|$, donc la méthode d'Euler donne une bonne approximation de la solution de ce problème de Cauchy en $t = 0.3$.

Exercice 04

$$(1) \begin{cases} y'(t) &= 36y(t) - 37e^{-t} \\ y(0) &= 1, \text{ puis } 1 + \varepsilon \end{cases}$$

En appliquant la méthode de la variation de la constante, comme au premier exercice, on trouve la solution générale, $y(t) = (e^{-37} + c)e^{36t} = e^{-t} + ce^{36t}$ où c est la constante d'intégration.

1. Si $y(0) = 1$, alors $c = 0$, donc la solution du problème est $y(t) = e^{-t}$.
2. Si $y(0) = 1 + \varepsilon$, alors $c = \varepsilon$, donc la solution du problème est $y_\varepsilon(t) = e^{-t} + \varepsilon e^{36t}$.

Conclusion : En comparant $y(t)$ et $y_\varepsilon(t)$, on voit que la différence $|y(t) - y_\varepsilon(t)| = \varepsilon e^{36t}$. Même si ε est très petit, cet écart tend vers $+\infty$, les deux solutions divergent l'une de l'autre. Ce problème est donc très sensible aux Conditions Initiales.

Bibliographie

[1] K. ARBENTZ, A. WOHLHAUSER : *Analyse Numérique.* Suisse, 1980.

[2] J. BASTIEN : *Introduction àl'analyse numérique :Applications sous Matlab*, Dunod, 2003.

[3] E. CANON : *Analyse numérique, Cours et exercices corrigés - Licence 2 et 3 Mathématiques*, Dunod, 2012.

[4] P. DEUFLHARD, A. HOHMANN : *Numerical Analysis in Modern Scientific Computing, An Introduction*, Springer-Verlag, 2003, 2^{e} édition.

[5] M. FELLAH, N.H. ALLAL : *Exercices corrigés en Analyse Numérique Elémentaire.* O.P.U, Alger, Réimpression 2005.

[6] F. FILBET : *Analyse numérique, algorithme et étude mathématique*, Dunod, 2013.

[7] F. JEDRZEJEWSKI : *Introduction aux Méthodes Numériques,* Deuxième édition. Springer-Verlag France, Paris 2005.

[8] R. HERBIN : *Cours d'Analyse numérique, polycopié*, Université Aix Marseille, 2014.

[9] M. LAKRIB : *Cours d'Analyse Numérique.* O.P.U, Alger, 2005.

[10] A. QUARTERONI, R. SACCO ET F. SALERI : *Méthodes Numériques, Algorithmes, Analyse et Applications*, Springer, 2006.

[11] R. THEODOR : *Initiation à l'analyse numérique*, CNAM, Masson, 1989.

[12] D. ZILL : *Dierential Equations with Boundary Value Problems*, PWSKent Pub, 1989.