Université Abderrahmane Mira de Béjaia

Faculté des sciences exactes

Département de mathématiques

Analyse Numérique

Cours, $2^{ème}$ année licence mathématiques

Karima MEBARKI[1].

[1] version 1.0 mebarqi_karima@hotmail.fr pour toute remarque

Table des matières

Introduction

En analyse numérique, et pour un problème posé (P), on étudie toutes les méthodes de résolution de (P), au moyen du calcul arithmétique. L'étude peut englober aussi bien les conditions d'existence et d'unicité de la solution du problème (P), et aussi les performances et l'efficacité du procédé choisi (précision, convergence, stabilité, rapidité,...).

Ordinateur et analyse numérique

Les calculatrices et les ordinateurs nous permettent de faire beaucoup d'opérations et ce très rapidement. Mais pour que les machines soient capables de faire ces calculs, il faut les programmer. C'est l'objet essentiel de l'analyse numérique qui s'est développée avec l'apparition des ordinateurs. Les caractéristiques des ordinateurs (fidélité, rapidité, précision,...ect) ont permis d'améliorer plusieurs méthodes numérique connues, et ont facilité la création d'algorithmes relatifs à des problèmes difficilement maîtrisés par l'homme jusque-là.

Mais faire beaucoup d'opérations ne veut pas dire faire n'importe quoi : les méthodes ont un **coût** (nombre d'opérations arithmétiques élémentaires), lié d'une part au temps de calcul, et d'autre part à la capacité de mémoire nécessaire pour stocker les données et les résultats. D'où l'étude de la complexité (efficacité) d'un algorithme.

Complexité d'un algorithme

Soit (P) un problème à N variables. On dit d'un algorithme, relatif à une méthode de résolution de (P), qu'il est de complexité exponentielle si le coût $f(N)$ nécessaire à sa résolution croit comme : $N!$, N^N, α^N(où $\alpha > 1$),... *etc.* On écrit simplement :

$$f(N) = O(N!),\ f(N) = O(N^N),\ f(N) = O(\alpha^N), \ldots etc.$$

D'autre part, un algorithme est de complexité polynômiale si

$$f(N) = O(N),\ f(N) = O(N^2),\ f(N) = O(N^\beta)\ \beta > 0, \ldots etc.$$

L'intérêt, en analyse numérique, est -entre autres objectifs- de mettre au point des algorithmes de complexité polynômiale.

En général, afin de choisir le meilleur algorithme possible, il faut choisir l'algorithme :

1. le moins coûteux possible en place mémoire,
2. le moins coûteux possible en temps de calcul : c'est à dire celui qui a de complexité polynômiale.
3. le plus stable possible : c'est à dire le moins sensible aux erreurs d'arrondi,
4. le plus précis possible : c'est à dire celui qui permet d'estimer l'erreur.

Ce cours est dispensé depuis 2009 aux étudiants de $2^{ème}$ année licence mathématiques de l'université Abderrahmane Mira de Béjaia. Il a pour objectif de présenter aux étudiants une variété d'outils numériques (algorithmes) permettant la résolution effective d'un certain nombre de problèmes. Les chapitres de ce cours sont illustrés par des exemples d'applications, et une série d'exercices est proposée à la fin de chacun d'entre eux. La plupart de ces exercices étaient proposés lors des séances de travaux dirigés ou des épreuves de moyenne durée.

Ce cours se compose de neuf chapitres. Il est divisé en deux parties couvrant le programme des modules d'analyse numérique I et analyse numérique II destinés aux étudiants de $2^{ème}$ année licence mathématiques. Particulièrement, ce cours traite les sujets suivants :

- Notions sur les erreurs,
- Interpolation et approximation polynomiale,
- Intégration et dérivation numériques,
- Résolution des systèmes linéaires,
- Résolution des équations non linéaires,
- Calcul des valeurs et vecteurs propres,
- Résolution des équations différentielles ordinaires.

Enfin merci de me communiquer toute erreur éventuelle dans le fond ou dans la forme de ce premier essai.

Bibliographie

1. Cours de licence de Mathématiques d'Analyse Numérique de N. Akroune (Université Abderrahmane Mira, Béjaia).
2. Cours de licence de Mathématiques d'Analyse Numérique de S. Salmon (Université Louis Pasteur, Strasbourg).
3. Cours d'Analyse Numérique de P. Goatin (Université du Sud Toulon-Var, France).
4. Analyse numérique pour ingénieurs (Deuxième édition). Auteur : André Fortin, Editeur : Presses internationales Polytechnique.
5. Introduction aux méthodes numériques (Deuxième édition). Auteur : Franck Jedrzejewski, Editeur : Springer-Verlag France, Paris 2005.

Première partie

Analyse numérique I

Chapitre 1

Notions sur les erreurs

1.1 Introduction

En général, la résolution des problèmes scientifiques passe par une représentation mathématique des phénomènes mis en jeu. Ces phénomènes sont en général compliqués et multiples. Pour les représenter, on est amené à négliger certains paramètres et à simplifier d'autres. Même avec ces simplifications, les équations obtenues sont souvent insolubles par les méthodes analytiques connues. Par exemple, on ne sait pas trouver analytiquement, la solution des équations $x^5 + 3x^4 + 7x + 8 = 0$, $x = e^{-x}$, $\sin x + e^x = 0, ...etc.$

C'est ici que l'analyse numérique se distingue des autres champs plus classiques des mathématiques. En effet, pour un problème donné, il est possible d'utiliser différents algorithmes de résolution. Ces algorithmes dépendent de certains paramètres qui influent sur la précision du résultat. De plus, on utilise en cours de calcul des approximations plus ou moins précises. Par exemple, on peut remplacer une dérivée par une différence finie de façon à transformer une équation différentielle en une équation algébrique. Le résultat final et son degré de précision dépendent des choix que l'on fait.

Une partie importante de l'analyse numérique consiste donc à contenir les effets des erreurs ainsi introduites, qui proviennent de trois sources principales :

- les erreurs de modélisation ;
- les erreurs de représentation sur ordinateur ;
- les erreurs de troncature.

Ce chapitre traite principalement des erreurs numériques. La première source d'erreurs dans les calculs faits par un ordinateur provient d'abord des erreurs d'arrondi sur les données, puis des opérations effectuées sur les donnés. Il devrait donc permettre au lecteur de mieux gérer les erreurs au sein des processus numériques afin d'être en mesure de mieux interpréter les résultats.

1.2 Erreurs absolue et relative

$$\text{Nombres exacts} \begin{cases} \text{dans } \mathbb{N} : 1, 3, 9; \\ \text{dans } \mathbb{Q} : \frac{2}{3}, \frac{1}{7}, \frac{10}{3}; \\ \text{dans } \mathbb{R} : \sqrt{5}, \pi, e. \end{cases}$$

Soit x un nombre exact et x^* une valeur approchée de x, on écrit

$$x \simeq x^* \quad \text{ou} \quad x \approx x^*.$$

- Si $x^* > x$, x^* est dite valeur approchée par excès.
- Si $x^* < x$, x^* est dite valeur approchée par défaut.

Exemples : $\frac{2}{3} \simeq 0,6666$, $\pi \simeq 3,14$, $\sqrt{5} \simeq 2,23$, sont des approximations par défaut mais $e \simeq 2,72$ est une approximation par excès.

1.2.1 Erreur absolue

Définition 1.1. *On appelle erreur absolue du nombre approché x^* de x la quantité réelle positive, notée $\Delta(x)$, définie par*

$$\Delta(x) = |x - x^*| .$$

Commentaire *:* Plus l'erreur absolue est petite, plus x^* est précis.

Exemple 1.1. *Pour la valeur exacte $x = \frac{2}{3}$, la valeur approchée $x_1^* = 0.666667$ est 1000 fois plus précise que la valeur approchée $x_2^* = 0.667$. En effet, nous avons :*

$$\begin{array}{rcl} \Delta_1(x) & = & |x - x_1^*| = |\frac{2}{3} - 0.666667| = |\frac{2}{3} - \frac{666667}{10^6}| = \frac{1}{3}10^{-6}, \\ \Delta_2(x) & = & |x - x_2^*| = |\frac{2}{3} - 0.667| = |\frac{2}{3} - \frac{667}{10^3}| = \frac{1}{3}10^{-3}. \end{array}$$

1.2.2 Erreur relative

Définition 1.2. *On appelle erreur relative du nombre approché x^* de x la quantité réelle positive, notée $r(x)$, définie par*

$$r(x) = \frac{|x - x^*|}{|x|} = \frac{\Delta(x)}{|x|}$$

Commentaire *:* l'erreur relative est souvent exprimée en pourcentage (précision relative) par :

$$r\% = r(x) \times 100$$

Exemple 1.2. *Pour les valeurs exactes $x = \frac{2}{3}$ et $y = \frac{1}{15}$ on considère les valeurs approchées $x^* = 0.67$ et $y^* = 0.07$, respectivement. Les erreurs absolues correspondantes sont :*

$$\begin{array}{rcl} \Delta(x) & = & |x - x^*| = |\frac{2}{3} - 0.67| = |\frac{200-201}{3.10^2}| = \frac{1}{3}10^{-2}, \\ \Delta(y) & = & |y - y^*| = |\frac{1}{5} - 0.07| = |\frac{100-105}{15.10^2}| = \frac{1}{3}10^{-2}. \end{array}$$

Les erreurs relatives correspondantes sont :

$$\begin{array}{rcl} r(x) & = & \frac{|x-x^*|}{|x|} = \frac{\Delta(x)}{|x|} = 0.5 \times 10^{-2} = 0.5\%, \\ r(y) & = & \frac{|y-y^*|}{|y|} = \frac{\Delta(y)}{|y|} = 5 \times 10^{-2} = 5\%. \end{array}$$

Ainsi, bien que les erreurs absolues soient égales, x^* est une approximation 10 fois plus précise pour x que y^* l'est pour y.

1.2.3 Majorants des erreurs absolue et relative

Si la valeur exacte est connue on peut déterminer les erreurs absolue et relative. Mais dans la majorité des cas, elle ne l'est pas. Les erreurs absolue et relative deviennent alors inconnues, et pour les estimer on introduit la notion de majorant de l'erreur absolue et de l'erreur relative.

Définition 1.3. *On appelle majorant de l'erreur absolue d'une valeur approchée x^* de x tout nombre réel positif noté Δx vérifiant :*

$$\Delta(x) = |x - x^*| \leqslant \Delta x$$

ou de manière équivalente : $x^* - \Delta x \leqslant x \leqslant x^* + \Delta x$, et on écrit :

$$x = x^* \pm \Delta x \quad \textit{qui veut dire :} \quad x \in [x^* - \Delta x, x^* + \Delta x]$$

Remarque 1.1. *1. Plus Δx est petit, plus l'approximation x^* est précise. D'où, en pratique, on prend le plus petit Δx possible.*

2. Comme $x \simeq x^*$, en pratique on prend $r_x \simeq \dfrac{\Delta x}{|x^*|}$ qui est un majorant de l'erreur relative de x^* et on écrit $x = x^* \pm |x^*|\, r_x$.

3. A défaut de l'erreur absolue (l'erreur relative) effective, Δx (r_x) est appelé par abus de langage, erreur absolue (erreur relative) de x^*.

1.3 Représentation décimale des nombres approchés

On sait que tout nombre réel positif x peut être représenté sous la forme d'une représentation décimale de développement limité ou illimité :

$$x = a_m 10^m + a_{m-1}10^{m-1} + ... + a_{m-n+1}10^{m-n+1} + ...$$

avec $a_i \in \{0, 1, 2, ..., 9\}$ pour $i \neq m$ et $a_m \neq 0$ où m est le rang supérieur de x (la plus grande puissance de 10).

Exemple 1.3. $5406,3080 = 5.10^3 + 4.10^2 + 0.10^1 + 6.10^0 + 3.10^{-1} + 0.10^{-2} + 8.10^{-3} + 0.10^{-4}$
$\pi = 3.14159265358... = 3.10^0 + 1.10^{-1} + 4.10^{-2} + 1.10^{-3} + 5.10^{-4} + ... + 5.10^{-10} + 8.10^{-11} + ...$
$\frac{68}{3} = 2.10^1 + 2.10^0 + 6.10^{-1} + 6.10^{-2} + 6.10^{-3} + 6.10^{-4} + ...$

Remarque 1.2. *Dans la pratique, les nombres utilisés x ont des représentations décimales limitées (car, en général, ce sont des nombres approchés).*

i. *Tous les chiffres conservés a_i s'appellent chiffres significatifs du nombre approché x.*

ii. *Certains des a_i peuvent être nuls.*

1.3.1 Chiffre significatif (c.s)

Définition 1.4. *On appelle chiffre significatif d'un nombre approché, tout chiffre dans sa représentation décimale différent de zéro ; et un zéro s'il se trouve entre deux chiffres significatifs, ou s'il constitue un chiffre conservé.*

Exemple 1.4. *Une approximation à 6 décimales de* 0.00301045 *est :*

$$\underbrace{0.00}_{(1)}\underbrace{3}_{(2)}\underbrace{0}_{(2)}\underbrace{1}_{(3)}\underbrace{0}_{(3)}$$

$$\begin{cases} (1) : \textit{Ne sont pas significatifs car ils ne servent qu'à indiquer les rangs des autres chiffres.} \\ (2) : \textit{Etant placé entre les chiffres significatifs } 1 \textit{et } 3, \textit{ zéro est lui même un chiffre significatif.} \\ (3) : \textit{ Ce zéro traduit le fait que le nombre approché a conservé la décimale } 10^{-6} \\ \qquad\qquad \textit{est un chiffre significatif.} \end{cases}$$

Exemple 1.5. *Les valeurs approchées $x^* =$ 0.0301009 , 400357 ont 6 chiffres significatifs (6 c.s.).*

Remarque 1.3. *1. On ne peut pas connaître le nombre de chiffres significatifs de $x = 45800$ donné sous cette forme. Pour savoir, il faut soit sa représentation décimale, ou encore -d'une façon équivalente- connaître l'écriture sous la forme $(p, q) \times 10^s$. En effet ;*
$x = x_1 = 4,58 \times 10^4$ (x_1 a 3 c.s.) ou encore $x = x_2 = 4,5800 \times 10^4$ (x_2 a 5 c.s.).

2. *Sur un ordinateur, les nombres sont représentés en virgule flottante comme suit : soit x un réel non nul, en virgule flottante x s'écrit sous la forme :*

$$x = \pm 0.a_1...a_N.b^E,$$

avec $b \in \mathbb{N}$ est la base, $a = 0.a_1...a_N$ que l'on appelle la mantisse, $0 \leq a_i < b, a_1 \neq 0, E \in \mathbb{Z}$, l'exposant compris entre deux entiers m et $M(-m \leq E \leq M)$ et $N \in \mathbb{N}$ le nombre de chiffres significatifs.

1.3.2 Chiffre significatif exact (c.s.e)

Soit x un nombre exact, x^* une valeur approchée de x dont sa représentation décimale (prise de gauche à droite) est :

$$x^* = \underbrace{a_m}_{1^{er}c.s}10^m + \underbrace{a_{m-1}}_{2^{eme}c.s}10^{m-1} + ... + \underbrace{a_{m-n+1}}_{n^{eme}c.s}10^{m-n+1} + a_{m-n} \underbrace{10^{m-n}}_{rang\,du(n+1)^{eme}c.s} ... + a_k 10^k,\ a_m \neq 0,\ k \in \mathbb{Z}.$$

Définition 1.5. *(importante) On dit que les n premiers chiffres significatifs d'un nombre x^* sont exacts si l'erreur absolue de ce nombre ne dépasse pas la moitié du rang du $n^{ème}$ chiffre significatif. c'est à dire*

$$\Delta x \leq \frac{1}{2} \cdot 10^{m-n+1}.$$

Proposition 1.1. *Si un nombre approché possède n c.s. exacts alors :*

$$r_x \leq 5.10^{-n}.$$

Exercice 1.1. *Donner une borne supérieure de l'erreur absolue et estimer l'erreur relative, si tous les chiffres significatifs des nombres approchés suivants sont exacts.*

$$x_1 = 0,0019,\ x_2 = 99,200,\ x_3 = -34508,\ x_4 = 0,000805.$$

Propriétés

a. Si un chiffre significatif est exact, tous les chiffres à sa gauche sont exacts.

b. Si un chiffre n'est pas exact, tous ceux à sa droite ne les sont pas.

Exemple 1.6. *Soit $x = 35.97$ et $x^* = 36.00$, ici $m = 1$ car $x^* = 3.10^1 + 6.10^0 + 0.10^{-1} + 0.10^{-2}$.*

$$\Delta x = |x - x^*| = 0,03 = 0,3.10^{-1} < 0,5.10^{-1}.$$

Alors $\begin{cases} m - n + 1 = -1 \\ m = 1 \end{cases} \Longrightarrow n = 3$. *Donc, x^* est une approximation de x avec trois chiffres significatifs exacts.*

Remarque 1.4. *La notion de chiffre significatif exact est purement mathématique, elle ne veut pas dire que les n premiers c.s. de x^* coïncident avec les n premiers c.s. de x; l'exemple ci-dessus l'illustre bien.*

1.4 Troncature et arrondissement d'un nombre

★ Pour approximer $\pi = 3.141592653589....$, on peut considérer la valeur approchée 3.14 ou encore 3.14159, etc... et cela selon le besoin. Dans le premier cas on a tronqué (couper en éliminant une partie) le nombre π après 2 décimales. Dans le second cas on l'a tronqué après 5 décimales.

★ Une méthode habituelle pour tronquer un nombre pour ne garder qu'un nombre fini de chiffres significatifs est l'arrondi :

Règle d'arrondissement :

Pour arrondir un nombre jusqu'à n chiffres significatifs, il faut éliminer les chiffres à droite du $n^{ième}$ chiffre significatif conservé si on se trouve après la virgule, sinon on remplace par des zéros :

1. Si le $(n+1)^{ième}$ chiffre significatif est > 5, on augmente le $n^{ième}$ chiffre de 1.
2. Si le $(n+1)^{ième}$ chiffre significatif est < 5, les chiffres retenus restent inchangés.
3. Si le $(n+1)^{ième}$ chiffre significatif est 5, alors deux cas sont possibles :
   - **i)** Tous les chiffres rejetés, situés après le $(n+1)^{ième}$ c.s., sont des zéros : On applique la règle du chiffre pair, i.e. : le $n^{ième}$ chiffre reste inchangé s'il est pair. On lui ajoute 1 s'il est impair.
   - **ii)** Parmi les chiffres rejetés, situés après le $(n+1)^{ième}$ c.s., il existe au moins un qui soit non nul : On ajoute 1 au $n^{ième}$ chiffre.

Remarque 1.5.

1. *On n'arrondit que le résultat final, jamais les résultats intermédiaires.*
2. *Un résultat final n'a jamais plus de précision que la précision des données.*
3. *Un nombre correctement arrondi ne possède que des chiffres significatifs exacts.*

Définition 1.6. *On appelle la borne supérieure de l'erreur d'arrondi du nombre approché le nombre noté Δ_{arr} vérifiant*

$$\Delta_{arr} \leq 0,5 \times 10^{m-n+1}.$$

après l'arrondissement de la valeur approchée de x on aura

$$x = x^*_{arr} \pm [\Delta x + \Delta_{arr}].$$

En général, on prend $\Delta x \simeq \Delta_{arr}$. Par suite

$$x = x^*_{arr} \pm 2\Delta x,$$

où x^*_{arr} est le nombre approché arrondi.

$2\Delta x$ est l'erreur absolue d'arrondi.

Exercice 1.2. *Arrondir les nombres suivants à 4 c.s.e. et indiquer l'erreur absolue d'arrondi :* $x_1 = 33,789$, $x_2 = 0,00489001$, $x_3 = 199993,99$, $x_4 = 0,0346750060$, $x_5 = 89765,5000$, *et* $x_6 = 9,007500$.

1.5 Propagation des erreurs

Soient x et y deux quantités exactes , x^* et y^* des approximations de x et y, respectivement, Δx et Δy des erreurs absolues sur x et y respectivement, r_x et r_y des erreurs relatives sur x et y respectivement.

1.5.1 Erreurs d'une addition

Erreur absolue : $\Delta_{x+y} = \Delta x + \Delta y.$

Erreur relative : $r_{x+y} = \dfrac{\Delta x + \Delta y}{|x^* + y^*|} \leq \max(r_x, r_y)$

Démonstration. Supposons que $x^*, y^* \in \mathbb{R}_+^*$. Nous avons

$$x^* - \Delta x \leq x \leq x^* + \Delta x \quad \text{et} \quad y^* - \Delta y \leq y \leq y^* + \Delta y. \text{ Donc}$$

a)

$$(x^* + y^*) - (\Delta x + \Delta y) \leq x + y \leq (x^* + y^*) + (\Delta x + \Delta y)$$

c'est à dire $(\Delta x + \Delta y)$ est l'erreur absolue de $x + y$, d'où $\Delta(x + y) = \Delta x + \Delta y$.

b)

$$\begin{aligned} r_{x+y} &= \tfrac{\Delta x + \Delta y}{x^* + y^*} \\ &= \tfrac{\Delta x}{x^*}\tfrac{x^*}{x^*+y^*} + \tfrac{\Delta y}{y^*}\tfrac{y^*}{x^*+y^*} \\ &= r_x.\lambda_1 + r_y.\lambda_2, \ (\lambda_1 = \tfrac{x^*}{x^*+y^*} > 0, \ \lambda_2 = \tfrac{y^*}{x^*+y^*} > 0 \text{ et } \lambda_1 + \lambda_2 = 1) \\ &\leq \max(r_x, r_y).\lambda_1 + \max(r_x, r_y).\lambda_2 \\ &\leq \underbrace{(\lambda_1 + \lambda_2)}_{=1} \max(r_x, r_y) \leq \max(r_x, r_y). \end{aligned}$$

□

1.5.2 Erreurs d'une soustraction

Erreur absolue : $\Delta_{x-y} = \Delta x + \Delta y.$

Erreur relative : $r_{x-y} = \dfrac{\Delta x + \Delta y}{|x^* - y^*|} \leq \dfrac{x^* + y^*}{|x^* - y^*|} \max(r_x, r_y)$

Démonstration. Exercice. □

Remarque 1.6. *La soustraction est l'opération qui fait perdre le plus de précision.*

Exemple 1.7. *Soient* $x^* = 255$ *et* $y^* = 250$ *avec* $r_x = r_y = 0,1\% = 10^{-3}$.
Question : $r_{(x-y)} = ?$
Nous avons d'abord : $\Delta_x = x^*.r_x = 0,255, \ \Delta y = y^*.r_y = 0,250.$
et puis $x^* - y^* = 5$ *avec une erreur relative :*

$$r_{(x-y)} = \frac{\Delta(x-y)}{|x^* - y^*|} = \frac{\Delta x + \Delta y}{|x^* - y^*|}) = 10,1.10^{-2} = 10,1\%.$$

On remarque que x^* *et* y^* *sont 101 fois plus précis pour* x *et* y *(respectivement) que* $x^* - y^*$ *l'est pour* $x - y$.

1.5.3 Erreurs d'une multiplication

Erreur absolue : $\Delta_{xy} = |x^*|\,\Delta y + |y^*|\,\Delta x.$

Erreur relative : $r_{xy} = r_x + r_y$

Démonstration. Supposons que $x^*, y^* \in \mathbb{R}_+^*$. Nous avons

$$x^* - \Delta x \leq x \leq x^* + \Delta x \quad \text{et} \quad y^* - \Delta y \leq y \leq y^* + \Delta y.$$

a) En supposant $x^* - \Delta x > 0$ et $y^* - \Delta y > 0$, on aura

$$(x^* - \Delta x)(y^* - \Delta y) \leq xy \leq (x^* + \Delta x)(y^* + \Delta y).$$

Si on néglige l'erreur de second ordre $\Delta x \Delta y$, on obtient

$$x^*y^* - [x^*\Delta y + y^*\Delta x] \leq xy \leq x^*y^* - [x^*\Delta y + y^*\Delta x].$$

b)

$$r_{(xy)} = \frac{\Delta(xy)}{x^*y^*} = \frac{x^*\Delta y + y^*\Delta x}{x^*y^*} = \frac{\Delta x}{x^*} + \frac{\Delta y}{y^*}.$$

□

1.5.4 Erreurs d'une division

Erreur absolue : $\Delta_{\frac{x}{y}} = \dfrac{|x^*|\,\Delta y + |y^*|\,\Delta x}{|y^*|^2}.$

Erreur relative : $r_{\frac{x}{y}} = r_x + r_y.$

Démonstration. Exercice. □

1.5.5 Erreurs d'une puissance

Erreur relative : $r_{x^n} = n\, r_x.$
Erreur absolue : $\Delta_{x^n} = n\,|(x^*)^{n-1}|\Delta x.$

Démonstration. Exercice. □

1.6 Exercices

Exercice 1.3. *Soit H la hauteur d'un barrage. Sa valeur mesurée h est $74,260m$. Si l'erreur relative commise sur h est de $0,3$, trouver a, b tels que $a \leq H \leq b$.*

Exercice 1.4. *Avec combien de c.s.e. faut-il calculer $e^{\sqrt{2}}$ pour que l'erreur relative ne dépasse pas 1%?*

Exercice 1.5. *On désire approcher $I = \int_0^{\frac{1}{p}} f(t)\,dt$, où $f(t) = t^3$ et $p \in \mathbb{N}^*$, par la surface S du triangle de sommets $A(0,0)$, $B(\frac{1}{p},0)$, $C(\frac{1}{p}, f(\frac{1}{p}))$. Déterminer la valeur minimale de p pour que l'erreur absolue d'approximation ne dépasse pas $\varepsilon = 10^{-2}$ puis $\varepsilon = 10^{-4}$.*

Exercice 1.6. *On considère l'équation*

$$x^2 - 1634x + 2 = 0. \tag{1.1}$$

1. *Résoudre l'équation (1.1) en effectuant les calculs avec $N = 10$ chiffres significatifs. (Utiliser le discriminant)*
2. *Commenter le résultat obtenu et proposer une autre méthode de calcul pour contourner le problème posé.*

Exercice 1.7. *Soient les trois nombres réels :*

$$x = 8,22 \quad y = 0,00317 \quad z = 0,00432.$$

1. *Représenter les nombres x, y et z avec virgule flottante.*
2. *En effectuant les calculs avec $N = 3$ c.s, calculer la somme $x + y + z$ en faisant :*
   i) $(x + y) + z$;
   ii) $x + (y + z)$.
3. *Commenter les résultats obtenus. Conclure.*

Exercice 1.8.
On veut calculer la surface d'un disque $S = \pi R^2$ où $R = 2,3400$, $\pi = 3,1416$.
En admettant que tous les chiffres de R et π sont exacts.

1. *Estimer les erreurs absolue et relative de S.*
2. *Calculer S en arrondissant au nombre de chiffres significatifs exacts.*

Exercice 1.9. *Soit l'approximation $\Delta f(x) \simeq |df(x)|$*

1. *Montrer que $\Delta ln(x) = \frac{\Delta x}{x}$.*
2. *En déduire l'erreur relative d'une puissance $u = x^n$ est telle que $\delta_u = n\delta_x$.*
3. *Déterminer l'erreur absolue et l'erreur relative de la racine énième $v = \sqrt[n]{x}$.*
4. *Soient $x = 22,123002$ et $y = 1,252468$ où tous les c.s. sont exacts.*
   Calculer $ln(\frac{x}{y})$ en déduire la valeur de $\sqrt[3]{\frac{x}{y}}$ arrondir au dernier c.s.e.

Exercice 1.10. *Soit T la période des petites oscillations du pendule qui est donnée par $T = 2\pi\sqrt{\frac{l}{g}}$ où l est la longueur du pendule et g la gravité. Supposons que les mesures faites sur T et l ont donné les résultats suivants :*

$$T = T^* \pm \Delta T = 1,936 \pm 0,002\ s \qquad et\ l = l^* \pm \Delta l = 92,95 \pm 0,10\ cm$$

1. *Calculer la gravité g en arrondissant au dernier chiffre significatif exact.*
2. *Donner l'erreur relative sur g en pourcentage.*

Chapitre 2

Interpolation polynomiale

2.1 Position du problème d'interpolation

Soient (x_i, y_i), $i = 0, \ldots, n$, $(n+1)$ couples de valeurs réelles. Des telles valeurs peuvent être le résultat de mesures effectuées expérimentalement. Le but du problème d'interpolation est de déterminer une fonction F -appartenant à une certaine classe- qui passe par les points (x_i, y_i) donnés, c'est à dire

$$F(x_i) = y_i,\ i = 0, \ldots, n.$$

Les points (x_i, y_i) sont appelés *les points d'interpolation.*

Exemple 2.1. *En physique, on mesure expérimentalement la température d'un objet qui refroidit au cours du temps. On obtient une suite de valeurs à différents temps t_i. On cherche alors à tracer la courbe de refroidissement la plus proche possible des points mesurés, et ainsi à estimer des valeurs de la fonction en des points non mesurés.*

Les fonctions les plus faciles à évaluer numériquement sont les polynômes. Il est donc important de savoir approximer une fonction arbitraire par des polynômes.
L'interpolation polynomiale consiste à chercher la fonction F sous forme d'un polynôme. C'est à ce cas qu'on va s'intéresser dans ce chapitre.

$$\textit{Existe-t-il un polynôme } P \textit{ tel que } P(x_i) = y_i,\ i = 0, \ldots, n?$$

Proposition 2.1. *Si les xi sont tous distincts alors il existe un unique polynôme P_n de degré inférieur ou égal à n vérifiant*

$$P_n(x_i) = y_i,\quad i = 0, \ldots, n.$$

Démonstration. On peut écrire le polynôme P_n comme suit :

$$P_n(x) = a_0 x^n + a_1 x^{n-1} + \ldots + a_n.$$

Du fait que $P_n(x_i) = y_i$, $i = \overline{0, n}$, les coefficients a_i, $i = \overline{0, n}$ vérifient le système :

$$(S)\ \begin{cases} a_0 x_0^n + a_1 x_0^{n-1} + \ldots + a_n &= y_0 \\ a_0 x_1^n + a_1 x_1^{n-1} + \ldots + a_n &= y_1 \\ & \vdots \\ a_0 x_n^n + a_1 x_n^{n-1} + \ldots + a_n &= y_n, \end{cases}$$

(S) est un système linéaire de déterminant

$$\Delta = \begin{vmatrix} x_0^n & x_0^{n-1} & . & . & . & x_o & 1 \\ x_1^n & x_1^{n-1} & . & . & . & x_1 & 1 \\ . & . & & & & . & . \\ . & . & & & & . & . \\ . & . & & & & . & . \\ x_n^n & x_n^{n-1} & . & . & . & x_n & 1 \end{vmatrix} = \prod_{i=0, j>i}^{n} (x_i - x_j) \tag{2.1}$$

(c'est le déterminant de Vandermonde). On a $\Delta \neq 0,$ car les x_i, sont tous distincts. Donc, le système (S) admet une et une seule solution $(a_0, a_1, \ldots, a_n)$, d'où le polynôme d'interpolation existe et il est unique. □

Exercice 2.1. *Soient la fonction $f(x) = x^4 - 2x^3 + x$. Parmi les polynômes suivants quel est celui qui interpole f aux points d'abscisses $x_0 = -1, x_1 = 0$ et $x_2 = 2$:*
$P_1(x) = 2x^4 - x^2 - x, \quad P_2(x) = x^2 + x + 1,$
$P_3(x) = x^3 - 3x, \quad P_4(x) = x^2 - x.$

Remarque 2.1. *Dans le cas général, on montre que la résolution du système plein (S), permettant le calcul des coefficients du polynôme d'interpolation, nécessite un nombre d'opérations en $O(n^3)$. On utilisera plutôt d'autres méthodes, moins coûteuses en nombre d'opérations, dès que n devient grand. Par exemple, l'utilisation des polynômes de Lagrange, présentée ci-dessous, nécessite un nombre d'opérations en $O(n^2)$ (voir la section 2.6).*

2.2 Interpolation de Lagrange

• Résolvons d'abord le problème partiel suivant :
Construire un polynôme $L_i = L_i(x)$ de degré n tel que

$$L_i(x_j) = \delta_{i,j} = \begin{cases} 1, & \text{si } j = i \\ 0, & \text{si } j \neq i \end{cases} \qquad i = 0, \ldots, n \text{ fixé.}$$

Le polynôme L_i s'annule en $x_0, x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$, il s'écrit alors sous la forme :

$$L_i(x) = K_i(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n), \quad K_i = C^{te}.$$

Donc,
pour $x = x_j$: $L_i(x_j) = 0,\ (j = \overline{0, n},\ j \neq i),$

pour $x = x_i$: $L_i(x_i) = 1 \Rightarrow K_i(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n) = 1,$
d'où la valeur de K_i

$$K_i = \frac{1}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)}.$$

Donc,

$$L_i(x) = \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)},$$

ou encore

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

Pour chaque $i = \overline{0, n},\ L_i$ est appelé polynôme élémentaire de Lagrange au point x_i.
Remarquons que $deg(L_i) = n$.

Proposition 2.2. *Les polynômes $L_0, L_1, \ldots, L_n$ forment une base de l'espace vectoriel $\mathbb{P}_n$.*

Démonstration. La famille $\{L_0, L_1, \ldots, L_n\}$ est composée de (n + 1) éléments. Pour montrer qu'elle forme une base de $\mathbb{P}_n$, qui est de dimension (n+1), il faut et il suffit d'établir que les Li, $i = \overline{0,n}$ sont linéairement indépendants. Etant donné $n+1$ scalaires α_i, $i = \overline{0,n}$.
Si $\sum_{i=0}^{n} \alpha_i L_i(x) = 0$, alors en particulier $\sum_{i=0}^{n} \alpha_i L_i(x_j) = \sum_{i=0}^{n} \alpha_i \delta_{ij} = 0$, pour tout $j = \overline{0,n}$.
D'où, $\alpha_0 = \alpha_1 = \ldots = \alpha_n = 0$. □

• Passant à la résolution du problème général qui consiste à former P_n vérifiant les conditions indiquées plus haut. Ce polynôme est de la forme :

$$P_n(x) = \sum_{i=0}^{n} y_i \, L_i(x).$$

On a bien

1. $deg(P_n) \leq n$,
2. $P_n(x_j) = \sum_{i=0}^{n} y_i L_i(x_j) = \sum_{i=0}^{n} y_i \delta_{ij} = y_j, \ j = 0, 1, \ldots, n.$

Par unicité, le polynôme P_n est le polynôme cherché. Il s'appelle le polynôme d'interpolation de Lagrange[1] qui interpole les points $(x_i, y_i), i = 0, \ldots, n$.

Exemple 2.2. *Déterminer le polynôme d'interpolation P_3 de la fonction f dont on connait les valeurs suivantes :*

x_i	0	1	3	4
$f(x_i)$	1	3	0	5

Sous la forme de Lagrange, ce polynôme s'écrit

$$P_3(x) = \sum_{i=0}^{3} f(x_i) Li(x) = L_0(x) + 3L_1(x) + 5L_3(x)$$

où

$$\begin{aligned} L_0(x) &= \frac{(x-x_1)(x-x_2)(x-x_3)}{(x_0-x_1)(x_0-x_2)(x_0-x_3)} = -\frac{1}{12}(x-1)(x-3)(x-4), \\ L_1(x) &= \frac{(x-x_0)(x-x_2)(x-x_3)}{(x_1-x_0)(x_1-x_2)(x_1-x_3)} = \frac{1}{6}x(x-3)(x-4), \\ L_3(x) &= \frac{(x-x_0)(x-x_1)(x-x_2)}{(x_3-x_0)(x_3-x_1)(x_3-x_2)} = \frac{1}{12}x(x-1)(x-3). \end{aligned}$$

Donc, $P_3(x) = -\frac{1}{12}(x-1)(x-3)(x-4) + \frac{1}{2}x(x-3)(x-4) + \frac{5}{12}x(x-1)(x-3)$.

Remarque 2.2.

1. *La formule de Lagrange est intéressante du point de vue numérique, car elle ne dépend que des abscisses x_i, $i = 0, 1, \ldots, n$. Soit à calculer plusieurs polynômes d'interpolation où les abscisses correspondant à ces polynômes restent fixés (il n'y a que les points y_i, $i = 0, 1, \ldots, n$ qui changent). Les polynômes L_i, $i = 0, 1, \ldots, n$ sont calculés une fois pour toute et servent ensuite pour tous les polynômes.*
2. *La méthode d'interpolation de Lagrange présente deux inconvénients majeurs :*
   - **i)** *L'erreur d'approximation peut ne pas diminuer si on augmente le nombre de points d'interpolation (voir TP).*

[1] Joseph-Louis Lagrange, français né en Italie, 1736-1813

ii) *La méthode n'est pas récurrente : connaissant P_n le polynôme d'interpolation de Lagrange de degré au plus n en les points (x_i, y_i), $i = \overline{0,n}$, si on rajoute le point d'interpolation (x_{n+1}, y_{n+1}), il existe un unique polynôme P_{n+1} de degré au plus $n+1$. Il est impossible de déduire P_{n+1} à partir de P_n, car dans la méthode de Lagrange, l'introduction d'un nouveau point x_{n+1} nécessite un nouveau calcul de tous les polynômes L_i de Lagrange aux points x_i, $i = 0, 1, \ldots, n+1$.*

Il est intéressant de mettre P_n sous une forme **récurrente** qui permet de compléter les valeurs déja obtenues sans refaire tous les calculs. On arrive donc au polynôme d'interpolation de Newton.

2.3 Interpolation de Newton

L'interpolation de Newton[2] utilise les différences divisées ou finies de f aux points donnés.

2.3.1 Différences divisées

Définition 2.1. *Soient $x_0, x_1, \ldots, x_n$, $(n+1)$ points (abscisses) distincts de $[a,b]$ et f une fonction réelle définie sur $[a,b]$ connue uniquement en x_i donnés. On définit les différences divisées d'ordres successifs $0, 1, 2, \ldots, n$ par :*

$$(I)\begin{cases} \text{ordre } 0 & : & \delta^0[x_i] & = & f(x_i), \\ \text{ordre } 1 & : & \delta[x_i, x_{i+1}] & = & \frac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i} \\ \text{ordre } 2 & : & \delta^2[x_i, x_{i+1}, x_{i+2}] & = & \frac{\delta[x_{i+1},x_{i+2}]-\delta[x_i,x_{i+1}]}{x_{i+2}-x_i} \\ & . & & & \\ & . & & & \\ & . & & & \\ \text{ordre } k & : & \delta^k[x_i, x_{i+1}, \ldots, x_{i+k}] & = & \frac{\delta^{k-1}[x_{i+1},\ldots,x_{i+k}]-\delta^{k-1}[x_i,\ldots,x_{i+k-1}]}{x_{i+k}-x_i}. \end{cases}$$

La dernière relation est appelée la différence divisée d'ordre k $(k = 0, 1, 2, \ldots, n)$ de la fonction f en $x_i, x_{i+1}, \ldots, x_{i+k}$.

Remarque 2.3. *Comme $\delta[a,b] = \frac{f(a)-f(b)}{a-b} = \delta[b,a]$, la différence divisée de n'importe quel ordre est indépendante de la position des points (abscisses) sur lesquels elle est prise.*

Calcul des différences divisées

Pour calculer la différence divisée d'ordre n de la fonction f aux points $x_0, \ldots, x_n$ on forme le tableau suivant, en appliquant les formules (I) colonne après colonne.

x_i	$f(x_i)$	DD1	DD2	DD3	DDn
x_0	$f(x_0)$				
x_1	$f(x_1)$	$\delta[x_0, x_1]$			
x_2	$f(x_2)$	$\delta[x_1, x_2]$	$\delta^2[x_0, x_1, x_2]$		
x_3	$f(x_3)$	$\delta[x_2, x_3]$	$\delta^2[x_1, x_2, x_3]$	$\delta^3[x_0, x_1, x_2, x_3]$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
x_n	$f(x_n)$	$\delta[x_{n-1}, x_n]$	$\delta^2[x_{n-2}, x_{n-1}, x_n]$	$\delta^3[x_{n-3}, \ldots, x_n] \ldots$	$\ldots \delta^n[x_0, x_1, \ldots, x_n]$

[2] Sir Isaac Newton, anglais, 1643-1727

2.3.2 Polynôme d'interpolation de Newton

Théorème 2.1. *Soient $f : [a,b] \to \mathbb{R}$ et $x_0, x_1, \ldots, x_n$, $(n+1)$ points distincts de $[a,b]$, le polynôme d'interpolation de f aux points x_i $i = \overline{0,n}$ peut être mis sous la forme :*

$$\begin{aligned} P_n(x) &= f(x_0) + \delta[x_0,x_1](x-x_0) + \delta^2[x_0,x_1,x_2](x-x_0)(x-x_1) \\ &\quad + \ldots + \delta^n[x_0,x_1,\ldots,x_n](x-x_0)(x-x_1)\ldots(x-x_{n-1}), \end{aligned}$$

ou encore

$$P_n(x) = f(x_0) + \sum_{i=1}^{n} \delta^i[x_0,x_1,\ldots,x_i] \prod_{j=0}^{i-1}(x-x_j). \tag{2.2}$$

Remarque 2.4. *On peut définir P_n par la relation de récurrence :*

$$\begin{cases} P_0(x) = f(x_0), \\ P_n(x) = P_{n-1}(x) + \delta^n[x_0,x_1,\ldots,x_n] \prod_{i=0}^{n-1}(x-x_i), \ n \geq 1. \end{cases}$$

Démonstration. Soit $x \in [a,b]$. D'après la définition des différence divisées on peut écrire :

$$\delta[x_0,x] = \frac{f(x)-f(x_0)}{x-x_0}$$

alors,

$$f(x) = f(x_0) + \delta[x_0,x](x-x_0),$$

puis

$$\delta^2[x_0,x_1,x] = \delta^2[x_1,x_0,x] = \frac{\delta[x_0,x]-\delta[x_0,x_1]}{x-x_1}$$

d'où,

$$f(x) = f(x_0) + \delta[x_0,x_1](x-x_0) + \delta^2[x_0,x_1,x](x-x_0)(x-x_1).$$

En continuant ainsi on obtient la formule des différences divisées, on aura donc

$$\begin{aligned} f(x) &= f(x_0) + \delta[x_0,x_1](x-x_0) + \delta^2[x_0,x_1,x_2](x-x_0)(x-x_1) \\ &\quad + \ldots + \delta^n[x_0,x_1,\ldots,x_n](x-x_0)(x-x_1)\ldots(x-x_{n-1}) \\ &\quad + \delta^{n+1}[x_0,x_1,\ldots,x_n,x](x-x_0)(x-x_1)\ldots(x-x_n). \end{aligned}$$

Si on pose $P_n(x) = f(x_0) + \delta[x_0,x_1](x-x_0) + \sum_{i=1}^{n} \delta^i[x_0,x_1,\ldots,x_i] \prod_{j=0}^{i-1}(x-x_j)$,
et $E(x) = \delta^{n+1}[x_0,x_1,\ldots,x_n,x](x-x_0)(x-x_1)\ldots(x-x_n)$, alors

$$f(x) = P_n(x) + E(x).$$

Montrons que P_n est le polynôme d'interpolation de f en x_i, $i = \overline{0,n}$.

i) $deg(P_n) \leq n$ (évident)

ii) $f(x_i) = P_n(x_i) + E(x_i)$ avec $E(x_i) = 0$, $i = \overline{0,n}$, donc $P_n(x_i) = f(x_i)$ $i = \overline{0,n}$.

□

Remarque 2.5.

1. *L'expression (2.4) s'appelle polynôme d'interpolation de Newton de f aux points (abscisses) x_i, $i = \overline{0,n}$.*
2. *L'expression de P_n est aussi valable si les points x_i, $i = \overline{0,n}$ ne sont pas ordonnés.*

3. *Si on rajoute un point d'interpolation x_{n+1} supplémentaire, le polynôme d'interpolation de f aux points x_i, $i = \overline{0, n+1}$ est donné par :*

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

Exemple 2.3. 1. *Déterminer P_3 le polynôme d'interpolation de la fonction passant par les points* $(0,1),(1,2),(2,9),(3,28)$.

2. *En déduire P_4 le polynôme d'interpolation passant par les points* $(0,1),(1,2),(2,9),(3,28)$ *et* $(5,54)$.

Solution : *On a 4 points, donc $deg(P_3) \leq 3$. Posons $x_i = i$, $i = \overline{0,4}$.*

$$\begin{aligned} P_3(x) &= f(x_0) + \delta[x_0, x_1](x - x_0) + \delta^2[x_0, x_1, x_2](x - x_0)(x - x_1) \\ &+ \delta^3[x_0, x_1, x_2, x_3](x - x_0)(x - x_1)(x - x_2). \end{aligned}$$

Le calcul des différences divisées se fait comme suit :

x_i	$f(x_i)$	*DD1*	*DD2*	*DD3*
$x_0 = 0$	$f(x_0) = 1$			
$x_1 = 1$	$f(x_1) = 2$	$\delta[x_0, x_1] = 1$		
$x_2 = 2$	$f(x_2) = 9$	$\delta[x_1, x_2] = 7$	$\delta^2[x_0, x_1, x_2] = 3$	
$x_3 = 3$	$f(x_3) = 28$	$\delta[x_2, x_3] = 19$	$\delta^2[x_1, x_2, x_3] = 6$	$\delta^3[x_0, x_1, x_2, x_3] = 1$

On obtient, $P_3(x) = 1 + x + 3x(x-1) + x(x-1)(x-2)$.
et

$$\begin{aligned} P_4(x) &= P_3(x) + \delta^4[x_0, x_1, x_2, x_3, x_4](x - x_0)(x - x_1)(x - x_2)(x - x_3) \\ &= P_3(x) - \tfrac{3}{5}x(x-1)(x-2)(x-3). \end{aligned}$$

2.3.3 Cas particulier : points équidistants

On considère $(n+1)$ points d'interpolation $(x_i, y_i)\, i = \overline{0,n}$ où les x_i sont équidistants, soit h la distance entre deux points consécutifs x_{i_0}, x_{i_0+1} (h est appelé pas d'interpolation).

Différences finies (non divisées) progressives

à l'ordre 1 : $\Delta f(x) = f(x+h) - f(x)$, alors

$$\Delta f(x_i) = f(x_{i+1}) - f(x_i),\ i = \overline{0, n-1}.$$

à l'ordre $k > 1$: $\Delta^k f(x) = \Delta(\Delta^{k-1} f)(x)$, alors

$$\Delta^k f(x_i) = \Delta^{k-1} f(x_{i+1}) - \Delta^{k-1} f(x_i),\ i = \overline{0, n-k}.$$

Table des différences finies progressives

$f(x_i)$	$\Delta f(.)$	$\Delta^2 f(.)$	$\Delta^3 f(.)$	$\Delta^n f(.)$
$f(x_0)$				
$f(x_1)$	$\Delta f(x_0)$			
$f(x_2)$	$\Delta f(x_1)$	$\Delta^2 f(x_0)$		
$f(x_3)$	$\Delta f(x_2)$	$\Delta^2 f(x_1)$	$\Delta^3 f(x_0)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(x_n)$	$\Delta f(x_{n-1})$	$\Delta^2 f(x_{n-2})$	$\Delta^3 f(x_{n-3}) \ldots$	$\ldots\ \Delta^n f(x_0)$

Proposition 2.3. *Pour tout $i = 0, \dots, n$, les égalités suivantes sont valables :*

1. $\Delta f(x_i) = f(x_{i+1}) - f(x_i),\ i = 0, \dots, n-1.$
2. $\Delta^2 f(x_i) = f(x_{i+2}) - 2f(x_{i+1}) + f(x_i),\ i = 0, \dots, n-2.$
3. $\Delta^3 f(x_i) = f(x_{i+3}) - 3f(x_{i+2}) + 3f(x_{i+1}) - f(x_i),\ i = 0, \dots, n-3.$
4. *En général, les différences finies progressives satisfont la relation*

$$\Delta^k f(x_i) = \sum_{j=0}^{k} (-1)^j C_k^j f(x_{i+k-j}),\ k = 0, \dots, n,\ i = 0, \dots, n-k.$$

Démonstration. Exercice. □

Relation entre les différences finies progressives et les différences divisées

Lemme 2.1.

$$\frac{\Delta^j f(x_i)}{j! h^j} = \delta^j[x_i, x_{i+1}, \dots, x_{i+j}],\ j = \overline{1, n},\ i = \overline{0, n-1}.$$

Démonstration. On montre cette relation, pour $i = 0$, par récurrence :
Pour $j = 1$: $\delta[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{\Delta f(x_0)}{h}$.
Supposons que $\delta^j[x_0, x_1, \dots, x_n] = \frac{\Delta^j f(x_0)}{j!\, h^j}$ et montrons que $\delta^{j+1}[x_0, \dots, x_{j+1}] = \frac{\Delta^{j+1} f(x_0)}{(j+1)!\, h^{j+1}}$. On a

$$\begin{aligned}
\delta^{j+1}[x_0, x_1, x_2, \dots, x_{j+1}] &= \frac{\delta^j[x_1, x_2 \dots, x_{j+1}] - \delta^j[x_0, x_2, \dots, x_j]}{x_{j+1} - x_0} \\
&= \frac{1}{(j+1)h} \left[\frac{\Delta^j f(x_1)}{j!\, h^j} - \frac{\Delta^j f(x_0)}{j!\, h^j} \right] \\
&= \frac{\Delta^j f(x_1) - \Delta^j f(x_0)}{(j+1)!\, h^{j+1}} \\
&= \frac{\Delta^{j+1} f(x_0)}{(j+1)!\, h^{j+1}}.
\end{aligned}$$

□

Formule de Newton progressive

Prenons les points d'interpolation dans l'ordre $x_0, x_1, \dots, x_n$.
D'après la formule de Newton avec les différences divisées et le lemme 2.1, le polynôme d'interpolation de Newton progressif s'écrit :

$$\begin{aligned}
P_n(x) &= f(x_0) + \frac{\Delta f(x_0)}{h}(x - x_0) + \frac{\Delta^2 f(x_0)}{2! h^2}(x - x_0)(x - x_1) \\
&\quad + \dots + \frac{\Delta^n f(x_0)}{n! h^n}(x - x_0)(x - x_1) \dots (x - x_{n-1}),
\end{aligned}$$

ou encore sous la forme de récurrence

$$\begin{cases} P_0(x) = f(x_0), \\ P_n(x) = P_{n-1}(x) + \frac{\Delta^n f(x_0)}{n! h^n}(x - x_0)(x - x_1) \dots (x - x_{n-1}),\ n \geq 1. \end{cases}$$

Exemple 2.4. *Déterminer le polynôme d'interpolation de Newton progressif associé à la fonction f passant par les points $(0, 1), (1, 2), (2, 9), (3, 28)$.*

Solution : *Posons $x_i = i$, $i = 0, 1, 2, 3$. Remarquons que les points donnés sont équidistants avec le pas d'interpolation $h = 1$, alors sous la forme de Newton progressive, le polynôme d'interpolation de f est donné par :*

$P_3(x) = f(x_0) + \frac{\Delta f(x_0)}{h}(x - x_0) + \frac{\Delta^2 f(x_0)}{2!h^2}(x - x_0)(x - x_1) + \frac{\Delta^3 f(x_0)}{3!h^3}(x - x_0)(x - x_1)(x - x_2)$.
Le calcul des différences finies progressives se fait comme suit :

$f(x_i)$	$\Delta f(.)$	$\Delta^2 f(.)$	$\Delta^3 f(.)$	$\Delta^n f(.)$
$f(x_0) = 1$				
$f(x_1) = 2$	$\Delta f(x_0) = 1$			
$f(x_2) = 9$	$\Delta f(x_1) = 7$	$\Delta^2 f(x_0) = 6$		
$f(x_3) = 28$	$\Delta f(x_2) = 19$	$\Delta^2 f(x_1) = 12$	$\Delta^3 f(x_0) = 6$	

D'où, $P_3(x) = 1 + x + 3x(x - 1) + x(x - 1)(x - 2)$.

Proposition 2.4. *Posons* $x = x_0 + th$, *le polynôme de Newton progressif s'écrit :*

$$\tilde{P}_n(t) = f(x_0) + \sum_{k=1}^{n} \frac{t(t-1)\dots(t-k+1)}{k!}\Delta^k f(x_0).$$

Démonstration. Exercice. □

Différences finies (non divisées) régressives

à l'ordre 1 : $\nabla f(x) = f(x) - f(x - h)$, alors

$$\nabla f(x_i) = f(x_i) - f(x_{i-1}),\ i = \overline{1, n}.$$

à l'ordre $k > 1$: $\nabla^k f(x) = \nabla(\nabla^{k-1} f)(x)$, alors

$$\nabla^k f(x_i) = \nabla^{k-1} f(x_i) - \nabla^{k-1} f(x_{i-1}),\ i = \overline{k, n}.$$

Table des différences finies régressives

$f(x_i)$	$\nabla f(.)$	$\nabla^2 f(.)$	$\nabla^3 f(.)$	$\nabla^n f(.)$
$f(x_0)$				
$f(x_1)$	$\nabla f(x_1)$			
$f(x_2)$	$\nabla f(x_2)$	$\nabla^2 f(x_2)$		
$f(x_3)$	$\nabla f(x_3)$	$\nabla^2 f(x_3)$	$\nabla^3 f(x_3)$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(x_n)$	$\nabla f(x_n)$	$\nabla^2 f(x_n)$	$\nabla^3 f(x_n) \dots$	$\dots \nabla^n f(x_n)$

Proposition 2.5. *Pour tout* $i = 0, \dots, n$, *les égalités suivantes sont valables :*

1. $\nabla f(x_i) = f(x_i) - f(x_{i-1}),\ i = 1, \dots, n.$
2. $\nabla^2 f(x_i) = f(x_i) - 2f(x_{i-1}) + f(x_{i-2}),\ i = 2, \dots, n.$
3. $\nabla^3 f(x_i) = f(x_i) - 3f(x_{i-1}) + 3f(x_{i-2}) - f(x_{i-3}),\ i = 3, \dots, n.$
4. *En général, les différences finies progressives satisfont la relation*

$$\nabla^k f(x_i) = \sum_{j=0}^{k} (-1)^{j+1} C_k^j f(x_{i-k+j}),\ k = 0, \dots, n,\ i = k, \dots, n.$$

Démonstration. Exercice. □

Le lemme suivant donne la relation entre les différences finies (non divisées) régressives et les différences divisées :

Lemme 2.2.

$$\frac{\nabla^j f(x_{i+j})}{j!\,h^j} = \delta^j[x_i, x_{i+1}, \ldots, x_{i+j}], \quad j = \overline{1,n},\ i = \overline{0,n-1}.$$

Formule de Newton régressive

Prenons les points d'interpolation dans l'ordre $x_n, x_{n-1}, \ldots, x_0$.
D'après la formule de Newton et le lemme 2.2, en remplaçant les différences divisées par les différences non divisées régressives et les points $\{x_0, x_1, \ldots, x_n\}$ par $\{x_n, x_{n-1}, \ldots, x_0\}$, le polynôme d'interpolation de Newton régressif s'écrit :

$$\begin{aligned} P_n(x) &= f(x_n) + \frac{\nabla f(x_n)}{h}(x - x_n) + \frac{\nabla^2 f(x_n)}{2!\,h^2}(x - x_n)(x - x_{n-1}) \\ &\quad + \ldots + \frac{\nabla^n f(x_n)}{n!\,h^n}(x - x_n)(x - x_{n-1})\ldots(x - x_1). \end{aligned}$$

ou encore sous la forme récurrente

$$\begin{cases} P_0(x) = f(x_n), \\ P_k(x) = P_{k-1}(x) + \frac{\nabla^k f(x_n)}{k!h^k}(x - x_n)\ldots(x - x_{n-k+1}),\ 1 \le k \le n. \end{cases}$$

Exemple 2.5. *On considère la même fonction donnée dans l'exemple 2.4. Déterminer le polynôme d'interpolation de la fonction f en $x_3 = 3, x_2 = 2, x_1 = 1, x_0 = 0$ sous la forme de Newton régressive.*
Solution : *Ce polynôme s'écrit :*

$$P_3(x) = f(x_3) + \frac{\nabla f(x_3)}{h}(x - x_3) + \frac{\nabla^2 f(x_3)}{2!h^2}(x - x_3)(x - x_2) + \frac{\nabla^3 f(x_3)}{3!h^3}(x - x_3)(x - x_2)(x - x_1).$$

Le calcul des différences finies progressives se fait comme suit :

$f(x_i)$	$\nabla f(.)$	$\nabla^2 f(.)$	$\nabla^3 f(.)$	$\nabla^n f(.)$
$f(x_0) = 1$				
$f(x_1) = 2$	$\nabla f(x_1) = 1$			
$f(x_2) = 9$	$\nabla f(x_2) = 7$	$\nabla^2 f(x_2) = 6$		
$f(x_3) = 28$	$\nabla f(x_3) = 19$	$\nabla^2 f(x_3) = 12$	$\nabla^3 f(x_3) = 6$	

D'où, $P_3(x) = 28 + 19(x - 3) + 6(x - 3)(x - 2) + (x - 3)(x - 2)(x - 1)$.

Proposition 2.6. *Posons $x = x_n + th$, le polynôme de Newton régressif s'écrit :*

$$\tilde{P}_n(t) = f(x_n) + \sum_{k=1}^{n} \frac{t(t+1)\ldots(t+k-1)}{k!}\,\nabla^k f(x_n).$$

Démonstration. Exercice. □

2.4 Erreur d'interpolation

Dans la pratique, l'interpolation polynomiale sert à remplacer une fonction f qui est soit inconnue, soit trop compliquée, par une fonction plus simple, en l'occurrence un polynôme. On dit que l'on approxime f par le polynôme d'interpolation P_n.

Comme c'est le cas dans de nombreuses méthodes d'analyse numérique, il est fondamental d'étudier l'erreur d'approximation. Naturellement, sauf cas particulier, l'expression de l'erreur ne permet pas son calcul exact ; elle peut cependant être très utile pour en calculer une majoration.

L'erreur d'interpolation $E(x) = f(x) - P_n(x),\ x \in [a,b]$ $(a \leq x_0 \leq \ldots \leq x_n \leq b)$, en utilisant les différences divisées, quelle que soit la formule de P_n, puisqu'il est unique est donnée par :

$$E(x) = \delta^{n+1}[x_0, x_1, \ldots, x_n, x](x - x_0)(x - x_1)\ldots(x - x_n). \tag{2.3}$$

Cette formule générale peut être modifiée lorsque la fonction f est $(n+1)$ fois continument dérivable sur l'intervalle $[a,b]$, qui est le plus petit intervalle contenant les x_i, $i = \overline{0,n}$, $f \in C^{n+1}([a,b])$.

Théorème 2.2. *Si $f \in C^n([a,b]), (n+1)$ fois dérivables sur $]a,b[$. Alors*

$$\forall\, x \in [a,b], \exists\, \xi = \xi(x) \in [a,b]/\ \ E(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\prod_{i=0}^{n}(x - x_i).$$

Démonstration. Soit $x \neq x_i$, sinon $E(x) = f(x) - P_n(x) = 0$.
D'après (2.3), il suffit de montrer que

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

On considère sur $[a,b]$ le fonction g définie par :

$$g(t) = f(t) - P_n(t) - \prod_{i=0}^{n}(t - x_i)\delta^{n+1}[x_0, x_1, \ldots, x_n, x].$$

La fonction g s'annule pour $t \in \{x_0, x_1, \ldots, x\}$, donc $g \in C^{n+1}([a,b])$ et s'annule au moins en $(n+2)$ points dans $[a,b]$. Alors, d'après le théorème de Rolle[3], $g' \in C^n([a,b])$ et s'annule au moins en $(n+1)$ points dans $[a,b]$, $g^{(k)}$ a au moins $(n-k+2)$ racines dans $[a,b]$ avec $0 \leq k \leq n+1$, et pour $k = n+1$, $g^{(n+1)}$ a au moins une racine dans $[a,b]$.
Notons par ξ cette racine de $g^{(n+1)}$, on a

$$\begin{cases} g^{(n+1)}(\xi) = 0, \\ g^{(n+1)}(t) = f^{(n+1)}(t) - 0 - (n+1)!\delta^{n+1}[x_0, x_1, \ldots, x_n, x], \end{cases}$$

car
- $\frac{d^{n+1}P_n(t)}{dt^{n+1}} = 0$, $(deg(P_n) \leq n)$,
- $\prod_{i=0}^{n}(t - x_i) = t^{n+1} + Q_n(t) \Rightarrow \frac{d^{n+1}(\prod_{i=0}^{n}(t-x_i))}{dt^{n+1}} = \frac{d^{n+1}(\prod_{i=0}^{n} t^{n+1}+Q_n(t))}{dt^{n+1}} = (n+1)!$,

donc $f^{(n+1)}(\xi) - (n+1)!\delta^{n+1}[x_0, x_1, \ldots, x_n, x] = 0$, d'où

$$\delta^{n+1}[x_0, x_1, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

[3] Michel Rolle, français, 1652-1719

Conséquences : Si $f^{(n+1)}$ est continue sur $[a,b]$, alors

1. $|E(x)| = \frac{|f^{(n+1)}(\xi)|}{(n+1)!}|\prod_{i=0}^{n}(x-x_i)| \leq \frac{M_{n+1}}{(n+1)!}|\prod_{i=0}^{n}(x-x_i)|$ où $M_{n+1} = \max_{t\in[a,b]}|f^{(n+1)}(t)|$.
   En effet ; une fonction continue sur un compact est bornée est atteint ses bornes.
2. Si f est un polynôme de degré $\leq n$, alors $f^{(n+1)}(x) = 0 \Rightarrow E(x) = 0,\ \forall x \in [a,b]$ donc, $P_n(x) = f(x),\ \forall x \in [a,b]$.

□

Remarque 2.6. *(Remarques importantes)*

1. *L'erreur est composée de deux termes, le terme* $\frac{|\prod_{i=0}^{n}(x-x_i)|}{(n+1)!}$ *qui dépend du choix des points* x_i *et le terme* $\max_{t\in[a,b]}|f^{(n+1)}(t)|$ *lié à la régularité de* f.
2. *Estimation grossière :* $|f(x) - P_n(x)| \leq \frac{|a-b|^{n+1}}{(n+1)!}\ \|f^{(n+1)}\|_\infty, \forall x \in [a,b]$.
   En conclusion, plus f *est régulière, moins l'erreur est grande.*
3. *Dans la pratique,* f *est rarement connue, et quand elle l'est, son appartenance à* $C^{n+1}([a,b])$ *n'est pas toujours réalisée, et à fortiori si le nombre* $(n+1)$ *de points* $(x_i, y_i)_{i=\overline{0,n}}$ *est grand !.*
4. *Pour* x *donné,* ξ_x *est souvent difficile à trouver.*
5. *Dans les conditions "idéales" (*$f \in C^{n+1}([a,b])$ *et* ξ_x *connu) minimiser* $|E(x)|$ *revient à faire de même pour* $\left|\prod_{i=0}^{n}(x-x_i)\right|$, *alors on a intérêt à considérer des abscisses* x_i, $i = \overline{0,n}$ *proches de* x *(donc réparties -de préférence- de part et d'autre de* x*).*
6. *Pratiquement, on peut montrer qu'il existe un choix optimal des points* x_i *qui permet de minimiser l'erreur d'approximation. Ce choix optimal est de prendre pour points d'interpolation les zéros d'un certain polynôme de Tchebychev (voir section 2.7).*

2.5 Interpolation de Gauss

Soient $(2n+1)$ points équidistants

$$x_{-n},\ x_{-(n-1)}, \ldots, x_{-1},\ x_0,\ x_1, \ldots, x_{n-1},\ x_n \ \text{ tels que } h = x_{i+1} - x_i,\ \forall i$$

Différences finies (non divisées) centrales

On définit les différences finies centrales par :
à l'ordre 1 : $\delta f(x) = f(x + \frac{h}{2}) - f(x - \frac{h}{2})$, alors

$$\delta f(x_{\frac{i+j}{2}}) = f(x_i) - f(x_j),\ i,j = \overline{0,n}.$$

à l'ordre $k > 1$: $\delta^k f(x) = \delta(\delta^{k-1}f)(x)$, alors

$$\delta^k f(x_{\frac{i+j}{2}}) = \delta^{k-1}f(x_i) - \delta^{k-1}f(x_j),\ i,j = \overline{0,n}.$$

Table des différences finies centrales

$f(x_i)$	$\delta f(.)$	$\delta^2 f(.)$	$\delta^3 f(.)$	$\delta^4 f(.)$	$\dots$	$\delta^{2n} f(.)$
$f(x_n)$						
$f(x_{-n+1})$	$\delta f(x_{\frac{-2n+1}{2}})$					
$f(x_{-n+2})$	$\delta f(x_{\frac{-2n+3}{2}})$	$\delta^2 f(x_{-n+1})$				
$\vdots$	$\vdots$	$\vdots$	$\ddots$			
$f(x_{-2})$				$\ddots$		
$f(x_{-1})$	$\delta f(x_{\frac{-3}{2}})$					
$f(x_0)$	$\delta f(x_{\frac{-1}{2}})$	$\delta^2 f(x_{-1})$				
$f(x_1)$	$\delta f(x_{\frac{1}{2}})$	$\delta^2 f(x_0)$	$\delta^3 f(x_{-\frac{1}{2}})$			
$f(x_2)$	$\delta f(x_{\frac{3}{2}})$	$\delta^2 f(x_1)$	$\delta^3 f(x_{\frac{1}{2}})$	$\delta^4 f(x_0))$	$\dots$	$\dots\ \delta^{2n} f(x_0))$
$\vdots$	$\vdots$	$\vdots$				
$f(x_{n-2})$						
$f(x_{n-1})$	$\delta f(x_{\frac{2n-3}{2}})$	$\delta^2 f(x_{n-1})$				
$f(x_n)$	$\delta f(x_{\frac{2n-1}{2}})$					

Le lemme suivant donne la relation entre les trois différences finies (non divisées) :

Lemme 2.3. $\Delta^j f(x_i) = \nabla^j f(x_{i+j}) = \delta^j f(x_{i+\frac{j}{2}}),\ \ j = \overline{1,n},\ \ i = \overline{-n, n-1}.$

Démonstration. exercice. □

Formule de Gauss progressive

Le 1^{er} pas étant progressif. c'est à dire, en prenant les points d'interpolation dans l'ordre :

$$x_0,\ x_1,\ x_{-1},\ x_2,\ x_{-2}, \dots, x_n,\ x_{-n}.$$

Le polynôme d'interpolation s'écrit :

$$\begin{aligned} P_n(x) &= f(x_0) + \frac{\delta f(x_{\frac{1}{2}})}{h}(x - x_0) + \frac{\delta^2 f(x_0)}{2!\,h^2}(x - x_0)(x - x_1) \\ &+ \frac{\delta^3 f(x_{\frac{1}{2}})}{3!\,h^2}(x - x_0)(x - x_1)(x - x_{-1}) \\ &+ \dots + \frac{\delta^{2n} f(x_0)}{(2n)!\,h^{2n}}(x - x_0)(x - x_1)(x - x_{-1}) \dots (x - x_n). \end{aligned}$$

Formule de Gauss régressive

Le 1^{er} pas étant progressif. c'est à dire, en prenant les points d'interpolation dans l'ordre :

$$x_0,\ x_{-1},\ x_1,\ x_{-2},\ x_2, \dots, x_{-n},\ x_n.$$

Le polynôme d'interpolation s'écrit :

$$\begin{aligned} P_n(x) &= f(x_0) + \frac{\delta f(x_{-\frac{1}{2}})}{h}(x - x_0) + \frac{\delta^2 f(x_0)}{2!\,h^2}(x - x_0)(x - x_{-1}) \\ &+ \frac{\delta^3 f(x_{-\frac{1}{2}})}{3!\,h^2}(x - x_0)(x - x_{-1})(x - x_1) \\ &+ \dots + \frac{\delta^{2n} f(x_0)}{(2n)!\,h^{2n}}(x - x_0)(x - x_{-1})(x - x_1) \dots (x - x_{-n}). \end{aligned}$$

Remarque 2.7. *Soit une table des points $(x_i)_{i\in\mathbb{N}^*}$. Afin d'avoir une bonne approximation de f en $\bar{x} \in [\min x_i, \max x_i]$,*

1. *la formule de Newton progressive est adaptée au début de la table,*
2. *la formule de Newton régressive est adaptée à la fin de la table,*
3. *les formules de Gauss sont adaptées au milieu de la table.*

Exemple 2.6. *Exercice 2.9*

2.6 Evaluation des polynômes

2.6.1 Cas d'un polynôme quelconque

Il est important de pouvoir évaluer des polynômes rapidement et de la façon la plus stable possible. Pour évaluer un polynôme de la forme :

$$P(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0,$$

chaque terme $a_j x^j$ nécessite j multiplications. Donc, en tout on a $\frac{n}{2}(n+1)$ multiplications et n additions.

On peut diminuer le nombre d'opérations en stockant le terme x^{j-1} de sorte que pour calculer $a_j x^j$, il suffit de multiplier x^{j-1} par x puis par a_j.

D'où 2 multiplications pour chaque monôme à partir de a_2x^2, on aura donc, $2n-1$ multiplications et n additions.

On peut encore diminuer ce coût, en utilisant le schéma d'Horner, qui regroupe les termes de P de la façon suivante :

$$\begin{aligned} P(x) &= (a_n x^{n-1} + a_{n-1}x^{n-2} + \ldots + a_2 x + a_1)x + a_0 \\ &= ((a_n x^{n-2} + a_{n-1}x^{n-3} + \ldots + a_3 x + a_2)x + a_1)x + a_0 \\ &= (\ldots((a_n x + a_{n-1})x + a_{n-2})x + \ldots + a_1)x + a_0. \end{aligned}$$

Ce qui donne l'algorithme de récurrence d'Horner :

$$\begin{aligned} T_0(x) &= a_n \\ T_1(x) &= T_0(x)x + a_{n-1} \\ &\vdots \\ T_k(x) &= T_{k-1}(x)x + a_{n-k} \quad \text{pour} \quad k = 1\ldots n, \\ &\vdots \\ T_n(x) &= P(x). \end{aligned}$$

2.6.2 Cas d'un polynôme d'interpolation de Newton

Soit P le polynôme d'interpolation de Newton de f aux points x_i, $i = \overline{0,n}$.

$$\begin{aligned} P(x) &= f(x_0) + \delta[x_0,x_1](x-x_0) + \delta^2[x_0,x_1,x_2](x-x_0)(x-x_1) \\ &\quad + \ldots + \delta^n[x_0,\ldots,x_n](x-x_0)(x-x_1)\ldots(x-x_{n-1}) \\ &= (\ldots(\delta^n[x_0,\ldots,x_n](x-x_{n-1}) + \delta^{n-1}[x_0,\ldots,x_{n-1}])\,(x-x_{n-2}) \\ &\quad + \ldots + \delta[x_0,x_1])(x-x_0) + f(x_0). \end{aligned}$$

Ce qui donne l'algorithme d'Horner :

$$\begin{aligned} T_0(x) &= \delta^n[x_0,\ldots,x_n] \\ T_1(x) &= (x-x_{n-1})T_1(x)+\delta^{n-1}[x_0,\ldots,x_{n-1}] \\ &\vdots \\ T_k(x) &= (x-x_{n-k})T_{k-1}(x)+\delta^{n-k}[x_0,\ldots,x_{n-k}] \quad \text{pour} \quad k=1,\ldots,n, \\ &\vdots \\ T_n(x) &= P(x). \end{aligned}$$

2.6.3 Cas d'un polynôme d'interpolation de Lagrange

Neville[4] et Aitken[5] ont proposé un algorithme récurrent de calcul du polynôme d'interpolation de Lagrange sur n points à partir d'une expression portant sur $(n-1)$ points.
Soit P le polynôme d'interpolation de Lagrange de f aux points $x_i,\ i=\overline{0,n}$.

$$P(x)=\sum_{i=0}^{n} f(x_i)\ L_i(x), \quad \text{où} \quad L_i(x)=\prod_{j=0,j\neq i}^{n}\frac{(x-x_j)}{(x_i-x_j)}.$$

Le calcul d'un des polynômes élémentaires de Lagrange de degré n nécessite :

$$2(n-1) \text{ multiplications, } 2n \text{ additions et une division.}$$

Donc pour évaluer le polynôme P, les opérations nécessaires sont :
$(n+1)$ polynômes élémentaires de Lagrange, $(n+1)$ multiplications et n additions. D'où,

$$\text{Coût}=(2n-1)(n+1)\ \text{(mulitplications)}+\ (2n(n+1)+n)\ \text{(additions)}\ +(n+1)\text{(divisions)}.$$

On peut diminuer de moitié ce coût en utilisant l'algorithme de Neville-Aitken suivant :

$$\begin{aligned} P_{0,j} &= f(x_j) \quad \text{pour} \quad j=0,\ldots,n, \\ P_{i+1,j}(x) &= \frac{(x_{i+1}-x)P_{i,j}(x)-(x_j-x)P_{i,j+1}(x)}{x_{i+j+1}-x_j} \quad \text{pour} \quad i=0,\ldots,n-1 \quad \text{et } j=0,\ldots,n-i-1, \\ P_{n,0}(x) &= P(x). \end{aligned}$$

2.7 Complément du cours

2.7.1 Interpolation d'Hermite

Exercice :

Soient $f:[a,b]\longrightarrow\mathbb{R}$ une fonction dérivable sur $[a,b]$ et $x_0,x_1,\ldots,x_n,\ (n+1)$ points distincts de $[a,b]$. On pose $y_i=f(x_i),\ z_i=f'(x_i),\ i=0,1,\ldots,n$. Afin d'améliorer l'interpolation de Lagrange, on propose de déterminer l'unique polynôme P_{2n+1} de degré inférieur ou égal à $2n+1$ qui vérifie

$$P_{2n+1}(x_i)=y_i \quad \text{et} \quad P'_{2n+1}(x_i)=z_i,\ \forall\, i\in\{0,1,\ldots,n\}, \tag{2.4}$$

il s'agit de l'interpolation dite de Hermite.

[4] Eric Harold Neville, 1989-1961
[5] Alexander Craig Aitken, 1895-1967

1. On considère les polynômes H_i et K_i, $i = 0, 1, \ldots, n$ définis par :

$$H_i(x) = L_i^2(x)\left[1 - 2L_i'(x_i)(x - x_i)\right], \quad K_i(x) = L_i^2(x)(x - x_i),$$

où $\quad L_i(x) = \prod_{j=0,\, j\neq i}^{n} \frac{(x-x_j)}{(x_i-x_j)}$.

a) Montrer que

i) $deg(H_i) = deg(K_i) = 2n + 1$, $\forall\, i \in \{0, 1, \ldots, n\}$.

ii) $H_i(x_j) = \begin{cases} 1, & \text{si } \ j = i; \\ 0, & \text{si } j \neq i, \end{cases} \qquad K_i(x_j) = 0, \ \ \forall\, i, j \in \{0, 1, \ldots, n\}.$

ii) $K_i'(x_j) = \begin{cases} 1, & \text{si } \ j = i; \\ 0, & \text{si } j \neq i, \end{cases} \qquad H_i'(x_j) = 0, \ \ \forall\, i, j \in \{0, 1, \ldots, n\}.$

b) En déduire que

$$P_{2n+1}(x) = \sum_{i=0}^{n} \left[y_i H_i(x) + z_i K_i(x)\right] \tag{2.5}$$

est un polynôme de degré inférieur ou égal à $2n + 1$ vérifiant (2.4).

c) Montrer que P_{2n+1} donné dans (2.5) est le seul polynôme de $\mathbb{P}_{2n+1}$ qui satisfait (2.4).

2. On pose $f \in C^{2n+1}([a, b])$, dérivable $(2n + 2)$ fois sur $]a, b[$. Montrer que $\forall\, x \in [a, b]$, $\exists\, \xi \in]a, b[$ tel que

$$E(x) = f(x) - P_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n + 2)!} \prod_{i=0}^{n} (x - x_i)^2.$$

Indication : pour tout $x \in [a, b]$ avec $x \neq x_i$, $i = 0, \ldots, n$, on définit la fonction $t \mapsto \psi(t)$ par $\psi(t) = f(t) - P_{2n+1}(t) - \frac{f(x)-P_{2n+1}(x)}{\prod_{i=0}^{n}(x-x_i)^2} \prod_{i=0}^{n} (t - x_i)^2$.

3. Application numérique : Soit $f(x) = \frac{1}{1+x^2}$, $\ x \in [0, 1]$.

i) Déterminer le polynôme d'interpolation de Lagrange de f aux points $x_0 = 0$ et $x_1 = 1$.

ii) Déterminer le polynôme d'interpolation d'Hermite de f aux points $x_0 = 0$ et $x_1 = 1$.

iii) En déduire deux valeurs approchées de $f(\frac{1}{2})$. Comparer les résultats obtenus et conclure.

Solution :

1. On considère les polynômes H_i et K_i, $i = 0, 1, \ldots, n$ définis par :

$$H_i(x) = [L_i(x)]^2\left[1 - 2L_i'(x_i)(x - x_i)\right], \quad K_i(x) = [L_i(x)]^2(x - x_i),$$

où

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

a) Pour tout $i \in \{0, 1, \ldots, n\}$, on a $deg(L_i) = n$ et $L_i(x_j) = \begin{cases} 1, & \text{si } \ j = i; \\ 0, & \text{si } j \neq i. \end{cases}$ Alors

$$deg(H_i) = deg(K_i) = 2deg(L_i) + 1 = 2n + 1.$$

$$\begin{cases} H_i(x_i) = [L_i(x_i)]^2\,[1 - 2L_i'(x_i)(x_i - x_i)] = [L_i(x_i)]^2 = 1, \\ H_i(x_j) = [L_i(x_j)]^2\,[1 - 2L_i'(x_i)(x_j - x_i)] = 0, \quad \text{si } j \neq i. \end{cases}$$

$$\begin{cases} H_i'(x_i) = 2L_i(x_i)L_i'(x_i)\,[1 - 2L_i'(x_i)(x_i - x_i)] - 2L_i'(x_i)[L_i(x_i)]^2 = 0, \\ H_i'(x_j) = L_i(x_j)L_i'(x_j)\,[1 - 2L_i'(x_i)(x_j - x_i)] - 2L_i'(x_i)[L_i(x_j)]^2 = 0, \quad \text{si } j \neq i. \end{cases}$$

$$\begin{cases} K_i(x_i) = [L_i(x_i)]^2(x_i - x_i) = 0, \\ K_i(x_j) = [L_i(x_j)]^2(x_j - x_i) = 0, \quad \text{si } j \neq i. \end{cases}$$

$$\begin{cases} K_i'(x_i) = 2L_i'(x_i)L_i(x_i)(x_i - x_i) + [L_i(x_i)]^2 = [L_i(x_i)]^2 = 1, \\ K_i'(x_j) = 2L_i'(x_j)L_i(x_j)(x_j - x_i) + [L_i(x_j)]^2 = 0, \quad \text{si } j \neq i. \end{cases}$$

b) Posons $P_{2n+1}(x) = \sum_{i=0}^{n} [y_i H_i(x) + z_i K_i(x)]$. D'après la question précédente, on aura

i) $deg(P_{2n+1}) \leq 2n+1$, c'est évident car $(deg(H_i) = deg(K_i) = 2n+1)$,

ii) $P_{2n+1}(x_j) = \sum_{i=0}^{n} [y_i H_i(x_j) + z_i K_i(x_j)] = y_j H_j(x_j) = y_j,\ \forall\, j \in \{0, 1, \ldots, n\}$,

iii) $P_{2n+1}'(x_j) = \sum_{i=0}^{n} [y_i H_i'(x_j) + z_i K_i'(x_j)] = z_j K_j'(x_j) = z_j,\ \forall\, j \in \{0, 1, \ldots, n\}$.

c) On suppose qu'il existe un autre polynôme de degré au plus $2n+1$, noté Q_{2n+1}, vérifiant $Q_{2n+1}(x_i) = y_i$ et $Q_{2n+1}'(x_i) = z_i,\ \forall\, i \in \{0, 1, \ldots, n\}$.
Posons $R_{2n+1} = P_{2n+1} - Q_{2n+1}$, on obtient

$$R_{2n+1}(x_i) = P_{2n+1}(x_i) - Q_{2n+1}(x_i) = 0,\ i \in \{0, 1, \ldots, n\}$$

$$R_{2n+1}'(x_i) = P_{2n+1}'(x_i) - Q_{2n+1}'(x_i) = 0,\ i \in \{0, 1, \ldots, n\},$$

donc $\{x_0, \ldots, x_n\}$ sont des racines doubles de R_{2n+1}. D'où R_{2n+1} est un polynôme de degré au plus $2n+1$ possède au moins $2n+2$ racines , ce qui signifie que R est le polynôme nul. ce qui implique que Q_{2n+1} coïncide avec P_{2n+1}.

2. On pose $f \in C^{2n+1}([a, b])$, dérivable $(2n+2)$ fois sur $]a, b[$. Montrer que $\forall\, x \in [a, b],\ \exists\, \xi \in]a, b[$ tel que

$$E(x) = f(x) - P_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \prod_{i=0}^{n} (x - x_i)^2.$$

Remarquons que $E(x_i) = 0,\ \forall i \in \{0, 1, \ldots, n\}$.
On considère sur $[a, b]$ avec $x \neq x_i,\ i = 0, \ldots, n$, la fonction ψ définie par :

$$\psi(t) = f(t) - P_{2n+1}(t) - \frac{f(x) - P_{2n+1}(x)}{\prod_{i=0}^{n} (x - x_i)^2} \prod_{i=0}^{n} (t - x_i)^2.$$

La fonction ψ s'annule pour $t \in \{x_0, x_1, \ldots, x_n, x\}$, donc $\psi \in C^{2n+2}([a, b])$ et s'annule au moins $(n+2)$ fois dans $[a, b]$. Alors, d'après le théorème de Rolle a au moins $(n+1)$ racines dans $]a, b[$ différentes de $x_0, x_1, \ldots, x_n$. De plus on $\psi'(x_i) = 0,\ \forall i \in \{0, 1, \ldots, n\}$, donc ψ' a au moins $(2n+2)$ racines dans $]a, b[$. En appliquant le théorème de Rolle $(2n+2)$ fois à la fonction ψ sur $[a, b]$, on trouve $\psi^{(2n+2)}$ a au moins une racine $\xi = \xi(x) \in]a, b[$.
Par suite on aura

$$\begin{cases} \psi^{(2n+2)}(\xi) = 0, \\ \psi^{(2n+2)}(t) = f^{(2n+2)}(t) - 0 - \frac{f(x) - P_{2n+1}(x)}{\prod_{i=0}^{n} (x - x_i)^2} (2n+2)!, \end{cases}$$

car
• $\frac{d^{2n+2}P_{2n+1}(t)}{dt^{2n+2}} = 0, \ \ (deg(P_{2n+1}) \leq 2n+1),$
• $\prod_{i=0}^{n}(t-x_i)^2 = t^{2n+2} + S_{2n+1}(t), \ \text{avec } S_{2n+1} \in \mathbb{P}_{2n+1}.$
D'où, $\forall\, x \in [a,b], \ \exists\, \xi \in]a,b[$:

$$E(x) = f(x) - P_{2n+1}(x) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\prod_{i=0}^{n}(x-x_i)^2.$$

3. **Application numérique** : Soit $[a,b] = [0,1], x_0 = 0, \ x_1 = 1 (n=1)$ et $f(x) = \frac{1}{1+x^2}, \ f'(x) = \frac{-2x}{(1+x^2)^2}$, alors $y_0 = f(0) = 1, \ y_1 = f(1) = \frac{1}{2}, \ z_0 = f'(0) = 0$ et $z_1 = f'(1) = -\frac{1}{2}$. De plus $L_0(x) = 1-x, \ L_1(x) = x$ et $L_0'(x) = -1, \ L_1'(x) = 1$.

i) Sous la forme de Lagrange P_1 s'écrit :

$$P_1(x) = L_0(x) + \frac{1}{2}L_1(x) = \frac{x-1}{-1} + \frac{1}{2}x = -\frac{1}{2}x + 1.$$

ii) Sous la forme de Hermite, P_3 s'écrit : $P_3(x) = H_0(x) + \frac{1}{2}H_1(x) - \frac{1}{2}K_1(x)$, où

$$H_0(x) = [L_0(x)]^2(1 - 2L_0'(x_0)(x-x_0)) = 2x^3 - 3x^2 + 1,$$

$$H_1(x) = [L_1(x)]^2(1 - 2L_1'(x_1)(x-x_1)) = -2x^3 + 3x^2,$$

$$K_1(x) = [L_1(x)]^2(x-x_1) = x^3 - x^2.$$

D'où,

$$P_3(x) = \frac{1}{2}x^3 - x^2 + 1.$$

iii) On a $\frac{1}{2} \in [0,1]$, d'où les deux approximations de $f(\frac{1}{2})$

$$f(\frac{1}{2}) \simeq P_1(\frac{1}{2}) = \frac{3}{4} \text{ avec } |f(\frac{1}{2}) - P_1(\frac{1}{2})| = |\frac{4}{5} - \frac{3}{4}| = 0,05$$

et

$$f(\frac{1}{2}) \simeq P_3(\frac{1}{2}) = \frac{13}{16} \text{ avec } |f(\frac{1}{2}) - P_3(\frac{1}{2})| = |\frac{4}{5} - \frac{13}{16}| = 0,0125 < 0.05.$$

Donc, la approximation la plus précise de $f(\frac{1}{2})$ est celle obtenue par l'interpolation de Hermite.

2.7.2 Meilleur choix de points d'interpolation et polynômes de Tchebychev

Contrairement aux interpolations précédentes dans lesquelles l'utilisateur peut choisir sa subdivision, l'interpolation de Tchebychev impose une subdivision $\{x_0, x_1, \ldots, x_n\}$ de l'intervalle d'interpolation $[a,b]$ en des points appelés **points de Tchebychev** .

Exercice :

Afin de déterminer, pour n donné, la subdivision $\{x_0, x_1, \ldots, x_n\}$ de $[a,b]$ pour laquelle la quantité

$$L = \max_{x\in[a,b]} |(x-x_0)(x-x_1)\ldots(x-x_n)| \quad \text{soit minimale,}$$

On propose d'étudier les polynômes de Tchebychev définis pour tout $n \in \mathbb{N}$ par :

$$\begin{array}{rccl} T_n : & [-1,1] & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & T_n(x) = \cos(n \arccos x) \end{array}$$

1. **a)** Montrer que pour tout $x \in [-1,1]$, T_n satisfont la relation de récurrence :

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \ n \geq 1.$$

Indication : poser $x = \cos\theta$ et utiliser la relation

$$\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos\theta\cos n\theta.$$

b) Déterminer les trois premiers termes de la suite $(T_n)_n$.

c) En déduire que T_n est un polynôme de degré n dont le coefficient de x^n est 2^{n-1}.

2. Montrer que :

i) $|T_n(x)| \leq 1$ pour tout $x \in [-1,1]$.

ii) $T_n(\cos(\frac{k\pi}{n})) = (-1)^k$ pour $k = 0, 1, \ldots, n$.

iii) $T_n(\cos(\frac{(2k+1)\pi}{2n})) = 0$ pour $k = 1, \ldots, n-1$.

3. Montrer que pour tout polynôme Q de degré n dont le coefficient de x^n est 2^{n-1}, on a

$$\max_{x\in[-1,1]} |Q(x)| \geq \max_{x\in[-1,1]} |T_n(x)| = 1.$$

4. **a)** Soit $\{\bar{x}_0, \ldots, \bar{x}_n\}$ une subdivision de l'intervalle $[-1,1]$. Montrer que $\max_{x\in[-1,1]} |(x-\bar{x}_0)(x-\bar{x}_1)\ldots(x-\bar{x}_n)|$ est minimal si et seulement si

$$(x-\bar{x}_0)(x-\bar{x}_1)\ldots(x-\bar{x}_n) = 2^{-n}T_{n+1}(x).$$

b) En déduire que, pour n donné, la subdivision $\{\bar{x}_0, \ldots, \bar{x}_n\}$ pour laquelle

$\max_{x\in[-1,1]} |(x-\bar{x}_0)(x-\bar{x}_1)\ldots(x-\bar{x}_n)|$ soit minimal est

$$\bar{x}_k = \cos(\frac{(2k+1)\pi}{2(n+1)}), \ k = 0, \ldots, n.$$

5. Quels subdivision $\{x_0, \ldots, x_n\}$ faut-il choisir si l'intervalle d'interpolation est l'intervalle $[a,b]$? Commenter l'erreur d'interpolation.
Indication : utiliser le changement de variable $x_k = \frac{a+b}{2} + \frac{a-b}{2}\bar{x}_k$.

Solution :

Pour tout $n \in \mathbb{N}$, on définit les polynômes de Tchebychev par :

$$\begin{array}{rccl} T_n : & [-1,1] & \longrightarrow & \mathbb{R} \\ & x & \longmapsto & T_n(x) = \cos(n \arccos x) \end{array}$$

1. **a)** Posons $x = \cos\theta$, alors $\arccos x = \theta$ ce qui implique que $T_n(x) = \cos(n\theta)$, $\forall\, n \in \mathbb{N}$. D'autre part, les formules d'addition des fonctions trigonométriques donnent :

$$\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos\theta\cos(n\theta).$$

On en déduit que pour tout $x \in [-1,1]$ et $n \in \mathbb{N}^*$

$$\begin{aligned} T_{n+1}(x) &= \cos((n+1)\theta) \\ &= 2\cos\theta\cos(n\theta) - \cos((n-1)\theta) \\ &= 2xT_n(x) - T_{n-1}(x). \end{aligned}$$

b)

$$\begin{aligned} T_0(x) &= \cos(0) = 1, \\ T_1(x) &= \cos(\arccos x) = x, \\ T_2(x) &= 2xT_1(x) - T_0(x) = 2x^2 - 1. \end{aligned}$$

c) Le degré de T_n et le coefficient de x^n s'obtiennent par récurrence.

2. Montrer que :

i) $|T_n(x)| = |\cos(n\arccos x)| \leq 1$ pour tout $x \in [-1,1]$.

ii) $T_n(\cos(\frac{k\pi}{n})) = \cos(n\arccos(\cos(\frac{k\pi}{n}))) = \cos(k\pi) = (-1)^k,\ k = 0,1,\ldots,n.$
C'est à dire T_n atteint son extremum sur l'intervalle $[-1,1]$ aux $(n+1)$ points

$$y_k = \cos(\frac{k\pi}{n}),\ k = 0,1,\ldots,n.$$

pour lesquels il prend alternativement les valeurs 1 et -1.

iii) $T_n(\cos(\frac{(2k+1)\pi}{2n})) = \cos(n\arccos(\cos(\frac{(2k+1)\pi}{2n}))) = \cos(\frac{(2k+1)\pi}{2}) = 0,\ k = 0,\ldots,n-1.$
Constatons que pour $k = 0,\ldots,n-1$, $\frac{(2k+1)\pi}{2n} \in [0,\pi]$. C'est à dire T_n s'annule exactement n fois sur l'intervalle $[-1,1]$ en les points

$$z_k = \cos(\frac{(2k+1)\pi}{2n}),\ k = 0,\ldots,n-1.$$

3. Raisonnons par l'absurde, supposons qu'il existe un polynôme Q de degré n dont le coefficient de x^n est 2^{n-1} différent de T_n tel que

$$\max_{x\in[-1,1]} |Q(x)| < \max_{x\in[-1,1]} |T_n(x)| = 1.$$

Considérons le polynôme $R = T_n - Q$. C'est un polynôme de degré (n-1). De plus

$$\begin{aligned} R(\cos(\tfrac{k\pi}{n})) &= Q(\cos(\tfrac{k\pi}{n})) - T_n(\cos(\tfrac{k\pi}{n})) \\ &= Q(\cos(\tfrac{k\pi}{n})) - (-1)^k \\ &= \begin{cases} Q(\cos(\frac{k\pi}{n})) - 1, & \text{si } k \text{ est pair ;} \\ Q(\cos(\frac{k\pi}{n})) + 1, & \text{si } k \text{ est impair,} \end{cases} \quad k = 0,\ldots,n. \end{aligned}$$

Comme $|Q(x)| < 1, \forall\, x \in [-1,1]$, alors $R(\cos(\frac{k\pi}{n}))$ prend alternativement le signe (+)ou(-). On en déduit que la fonction R s'annule au moins une fois dans chacun des intervalles $[\cos(\frac{k\pi}{n}), \cos(\frac{(k+1)\pi}{n})], k = 0,\ldots,n-1$. Ainsi R possède n racines dans $[-1,1]$. Or R est un polynôme de degré $(n-1)$. Ceci n'est possible que si $R \equiv 0$, donc $Q \equiv T_n$. Contradiction.

4. **a)** Prenons $Q(x) = 2^n(x-\bar{x}_0)(x-\bar{x}_1)\dots(x-\bar{x}_n)$, on a Q est un polynôme de degré $n+1$ dont le coefficient de x^{n+1} est 2^n, ce qui entraîne, d'après la question précédente, que

$$\max_{x\in[-1,1]} |2^n(x-\bar{x}_0)(x-\bar{x}_1)\dots(\bar{x}-x_n)| \geq \max_{x\in[-1,1]} |T_{n+1}(x)|.$$

Donc, pour que $\max_{x\in[-1,1]} |(x-\bar{x}_0)(x-\bar{x}_1)\dots(x-\bar{x}_n)|$ soit minimal if faut et il suffit de prendre $2^n(x-\bar{x}_0)(x-\bar{x}_1)\dots(x-\bar{x}_n) = T_{n+1}(x)$. C'est à dire

$$(x-\bar{x}_0)(x-\bar{x}_1)\dots(x-\bar{x}_n) = 2^{-n}T_{n+1}(x). \tag{2.6}$$

b) De (2.6) on en déduit que, pour n donné, $\bar{x}_0,\dots,\bar{x}_n$ sont les (n+1) racines de polynôme de Tchebychev T_{n+1} et d'après la question (2.d) ces racines sont données par :

$$\bar{x}_k = \cos(\frac{(2k+1)\pi}{2(n+1)}),\ k=0,\dots,n.$$

5. Pour se ramener à un intervalle $[a,b]$ quelconque, on utilise le changement de variable suivant

$$x_k = \frac{a+b}{2} + \frac{a-b}{2}\,\bar{x}_k.$$

D'où pour que la quantité $L = \max_{x\in[a,b]} |(x-x_0)(x-x_1)\dots(x-x_n)|$ soit minimale, il faut et il suffit de prendre

$$x_k = \frac{a+b}{2} + \frac{b-a}{2}\,\cos(\frac{(2k+1)\pi}{2(n+1)}),\ k=0,\dots,n.$$

On obtient la majoration suivante le l'erreur d'interpolation de la fonction $f \in C^{(n+1)}([a,b])$ par le polynôme P_n

$$|f(x)-P_n(x)| \leq \frac{(b-a)^{n+1}}{(n+1)!2^{2n+1}} \max_{t\in[a,b]} |f^{(n+1)}(t)|,\ x\in[a,b].$$

C'est la meilleure majoration globale que l'on puisse obtenir. En effet ; soit $M_{n+1} = \max_{t\in[a,b]} |f^{(n+1)}(t)|$, pour tout $x\in[a,b]$ on a

$$\begin{aligned}
|f(x)-P(x)| &= |\tfrac{f^{(n+1)}(\xi)|}{(n+1)!}| \prod_{i=0}^{n}(x-x_k)|,\ \xi\in]a,b[\\
&\leq \tfrac{M_{n+1}}{(n+1)!}|(x-\tfrac{a+b}{2}-\tfrac{b-a}{2}\bar{x}_0)\dots(x-\tfrac{a+b}{2}-\tfrac{b-a}{2}\bar{x}_n)|,\\
&= \tfrac{M_{n+1}}{(n+1)!}(\tfrac{b-a}{2})^{n+1}|(\tfrac{2}{b-a}(x-\tfrac{a+b}{2})-\bar{x}_0)\dots(\tfrac{2}{b-a}(x-\tfrac{a+b}{2})-\bar{x}_n)|,\\
&= \tfrac{M_{n+1}}{(n+1)!}(\tfrac{b-a}{2})^{n+1}|(y-\bar{x}_0)\dots(y-\bar{x}_n)|, \quad \text{où } y=\tfrac{2}{b-a}(x-\tfrac{a+b}{2})\in[-1,1]\\
&= \tfrac{M_{n+1}}{(n+1)!}\tfrac{(b-a)}{2})^{n+1}|2^{-n}T_{n+1}(y)|\\
&\leq \tfrac{M_{n+1}}{(n+1)!}(\tfrac{b-a}{2})^{n+1}2^{-n}.
\end{aligned}$$

2.8 Conclusion

En fait, les méthodes d'interpolation polynomiale présentées dans ce cours (Lagrange, Newton, Hermite[6]) sont assez peu utilisées dans la pratique. D'abord, il est difficile d'interpoler avec des polynômes de degré très élevé : on voit alors apparaître un phénomène d'effets de bord dit **phénomène de Runge**[7]. Runge a montré en 1901 que quand le nombre de points d'interpolation croît indéfiniment, le polynôme d'interpolation ne converge pas toujours vers la fonction interpolée en tous points. La divergence s'observe aux bords de l'intervalle (la convergence n'est pas uniforme). On peut avoir une convergence uniforme, en choisissant judicieusement les points d'interpolation (racines d'un certain polynôme de Tchebychev[8]).

Néanmoins, ces méthodes d'interpolation ont un intérêt pour construire des formules de quadrature pour l'intégration numérique et des schémas pour résoudre des équations différentielles.

L'interpolation en utilisant des polynômes de degré élevé introduit aussi une grande sensibilité aux erreurs. On préfère alors faire de **l'interpolation par morceaux :** c'est à dire découper l'intevalle sur lequel on veut interpoler en petits intervalles et interpoler sur chacun des ces petits intervalles avec des polynômes de degré moindre. c'est la **méthode des splines**, la méthode d'interpolation la plus utilisée en pratique.

[6] Charles Hermite, français, 1822-1901
[7] Carle David Tolmé Runge, allemand, 1856-1927
[8] Pafnuty Lvovich Chebyshev, russe, 1821-1894

2.9 Exercices

Exercice 2.2. *Soient $x_0, x_1, \ldots, x_n$ (n+1) points distincts.*

1. *Montrer que, si $f \in \mathbb{P}_n$ ($\mathbb{P}_n$ est l'ensemble des polynômes de degré inférieur ou égal à n) alors le polynôme d'interpolation de f aux points x_i $(i = 0, 1, \ldots, n)$ est f lui même.*
2. *En déduire que, pour tout x réel, on a :*

   a. $\sum_{i=0}^{n} L_i(x) = 1$ **b.** $\sum_{i=0}^{n} x_i^k L_i(x) = x^k \quad (0 \leq k \leq n)$

   c. $\sum_{i=0}^{n} (x_i - x)^k L_i(x) = \delta_{0k} \quad (0 \leq k \leq n)$ *où* $\delta_{ij} = \begin{cases} 1, & si\ i = j; \\ 0, & si\ i \neq j. \end{cases}$

Exercice 2.3.

1. *Soient $x_0, x_1, \ldots, x_n$ (n+1) points réels distincts et F, G deux fonctions définies en ces points. On considère P et Q les polynômes d'interpolations de F et G (respectivement) aux points $x_0, x_1, \ldots, x_n$.*

   a. *Montrer que $P + Q$ est le polynôme d'interpolation de $F + G$ aux points $x_0, x_1, \ldots, x_n$.*

   b. *Donner une condition nécessaire et suffisante pour que PQ soit le polynôme d'interpolation de FG aux points $x_0, x_1, \ldots, x_n$.*
2. *Soit la fonction f donnée par la table suivante :*

x_i	0	$\frac{1}{2}$	1	$\frac{3}{2}$
$f(x_i)$	1	3	8	16

   a. *Déterminer le polynôme d'interpolation de Newton de f aux points x_i $(i = 0, 1, 2, 3)$.*

   b. *En déduire le polynôme d'interpolation de la fonction g définie par $g(x) = xf(x) + x^2$ aux points x_i $(i = 0, 1, 2, 3)$.*

Exercice 2.4. 1. *Soit la fonction réelle f définie par :*

$$f(x) = (x^3 - x^2 - 2x)e^{-x^4+6x^3-11x^2+6x}$$

   a. *Déterminer le polynôme d'interpolation de Newton de f aux points $0, 1, 2, 3$ en utilisant la formule de Newton progressive.*

   b. *Pouvait-on prévoir ce résultat ?*

2. *Soit g la fonction réelle donnée par la table suivante :*

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

   a. *Déterminer le polynôme d'interpolation de g aux points $0, 1, 2$:*
   i. sous la forme de Lagrange ii. sous la forme de Newton

   b. *En déduire le polynôme d'interpolation de g aux points $0, 1, 2$ et 3.*

3. **a.** *Déduire des questions précédentes, le polynôme d'interpolation de la fonction $F = f + g$ aux points $0, 1, 2$ et 3.*

   b. *En utilisant l'algorithme de Horner, calculer une valeur approchée de $F(\frac{1}{2})$.*

   c. *Estimer le résultat obtenu, sachant que $\max_{x \in [0,3]} |F^{(4)}(x)| \leq 1$.*

Exercice 2.5. *On considère une fonction impaire f connue aux points*

$$x_0 = -2,\ x_1 = -1,\ x_2 = 1 \quad et \quad x_3 = 2.$$

1. *Calculer les polynômes de Lagrange* $L_i (i = 0, 1, 2, 3)$ *en ces points.*
2. *Calculer* $L_3(x) - L_0(x)$ *et* $L_2(x) - L_1(x)$. *En déduire une expression du polynôme d'interpolation* P *de la fonction* f *aux points* x_0, x_1, x_2, x_3 *sous la forme :*

$$P(x) = R(x)f(1) + S(x)f(2)$$

où R *et* S *sont deux polynômes à déterminer.*

3. *Montrer que* P *est également le polynôme d'interpolation de* f *aux points* -2, -1, 0, 1, 2.
4. *Pour* $f(x) = \sin \frac{\pi}{2} x$, *en déduire une approximation de* $f(\frac{1}{4})$ *ainsi que l'erreur absolue commise.*

Exercice 2.6.

1. *Soient* $f : [-a, a] \to \mathbb{R}$ *une fonction,* $a \in \mathbb{R}_+^*$ *et* $x_0, x_1, \ldots, x_n$ *des points de* $[-a, a]$ *tels que :* $0 < x_0 < x_1 < \ldots < x_n < a$.
Montrer que le polynôme d'interpolation de f *aux points* $-x_n, \ldots, -x_1, -x_0, x_0, x_1, \ldots, x_n$ *est pair si* f *est paire.*
2. *Soit* $f(x) = \frac{1}{|x|+1}$, $x \in [-4, 4]$. *Déterminer le polynôme d'interpolation de Lagrange de* f *aux points* -3, -1, $-\frac{1}{2}$, $\frac{1}{2}$, 1, 3.

Exercice 2.7. *On considère la fonction* f *définie par* $f(x) = \sin(\pi x)$, $x \in [-1, 1]$.

1. *Déterminer le polynôme d'interpolation de* f, *noté* P_2, *aux points* $-1, -\frac{1}{2}, \frac{1}{2}$.
2. *Déterminer le polynôme d'interpolation de* f, *noté* Q_2, *aux points* $-\frac{1}{2}, \frac{1}{2}, 1$.
3. *Soit le polynôme* $P(x) = \frac{1}{2}[(x+1)Q_2(x) - (x-1)P_2(x)]$.
   a) *Montrer que* P *est le polynôme d'interpolation de* f *aux points* $-1, -\frac{1}{2}, \frac{1}{2}, 1$.
   b) *Peut-on affirmer que* P *est le polynôme d'interpolation de* f *aux points* $-1, -\frac{1}{2}, 0, \frac{1}{2}, 1$?
4. *Calculer* $P_2(\frac{1}{3}), Q_2(\frac{1}{3})$ *et* $P(\frac{1}{3})$ *puis comparer les résultats obtenus.*

Exercice 2.8. *On considère la fonction* $f(x) = \frac{16^x}{2}$ *et on se donne les points*

$$x_0 = 0,\ x_1 = \frac{1}{2},\ x_2 = 1,\ x_3 = \frac{3}{2}.$$

Les valeurs de f *en* x_i *sont résumées dans le tableau suivant :*

x_i	*0*	1/2	*1*	3/2
$f(x_i)$	1/2	*2*	*8*	*32*

1. *Déterminer le polynôme d'interpolation de* f *sous forme de Newton aux points* x_i, $i = 0, 1, 2, 3$, *en utilisant les différences finies.*
2. *En déduire une valeur approchée de* $\sqrt[3]{2}$. *(Indication :* $\sqrt[3]{2} = \frac{16^{\frac{1}{3}}}{2}$*)*

Exercice 2.9. *On donne la table des valeurs de la fonction* f *définie par* $f(x) = cos(\pi x)$:

x_i	*-4*	$-\frac{7}{2}$	*-3*	$-\frac{5}{2}$	*-2*	$-\frac{3}{2}$	*-1*	$-\frac{1}{2}$	*0*	$\frac{1}{2}$	*1*	$\frac{3}{2}$	*2*	$\frac{5}{2}$
$f(x_i)$	*1*	*0*	*-1*	*0*	*1*	*0*	*-1*	*0*	*1*	*0*	*-1*	*0*	*1*	*0*

En interpolant par un polynôme de degré inférieur ou égal à 4 et en utilisant les formules appropriées, calculer des valeurs approchées de :

$$i)\ cos\frac{\pi}{10}, \qquad ii)\ cos\frac{9\pi}{10}, \qquad iii)\ cos\frac{12\pi}{5}, \qquad vi)\ cos\frac{19\pi}{5}.$$

Estimer l'erreur absolue commise dans chacun des cas.

Exercice 2.10. *Soit la fonction f donnée par :*

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

1. **a)** *Déterminer le polynôme d'interpolation de Lagrange P de f aux points $-1, 0, 1$.*

   b) *Montrer que* $\forall\, x \in [-1,1]\ \ |E(x)| \leq \frac{\sqrt{3}}{27} M$;
   où $E(x)$ est l'erreur d'interpolation et $M = \sup_{x\in[-1,1]} |f^{(3)}(x)|$.

   c) *Calculer $f(\frac{1}{3})$ en arrondissant au dernier c.s.e, sachant que $M \leq 10^{-1}$.*

2. **a)** *Déterminer le polynôme d'interpolation Q de degré 4, qui permet de calculer une bonne valeur approchée de $f(\frac{1}{3})$.*

   b) *Calculer cette valeur approchée par l'algorithme d' Horner.*

Exercice 2.11. *On considère la fonction f définie par :* $f(x) = \cos(\pi x^2),\ \ x \in [-1,1]$.
Dans toute la suite, les expressions seront calculées en fonction de $\sqrt{2}$.

1. *Soit P le polynôme d'interpolation de f aux points $-1, -\frac{1}{2}, \frac{1}{2}, 1$.*

   a) *Donner le degré exact de P.*

   b) *Faire la table des différences divisées en ces points.*

   c) *Majorer dans $[-1,1]$ l'erreur d'interpolation $E(x) = f(x) - P(x)$ en fonction de* $M = \sup_{x\in[-1,1]} |f^{(k)}(x)|$, *$k$ à préciser (on ne calculera pas M).*

2. *Soit Q le polynôme d'interpolation de f aux points $-1, -\frac{1}{2}, 0, \frac{1}{2}, 1$.*
   Montrer que Q s'écrit sous la forme

$$Q(x) = P(x) + H(x)$$

   où H est un polynôme à calculer.

Exercice 2.12. *Soit la fonction définie par* $f(x) = \sqrt{x+1},\ x \in [0,1]$.

1. *Déterminer le polynôme d'interpolation de Newton P vérifiant :*

$$P(0) = f(0),\ P(\frac{1}{2}) = f(\frac{1}{2}),\ P(1) = f(1).$$

   Calculer $P(0.1)$ et $P(0.9)$. Comparer aux valeurs exactes.

2. *Déterminer le polynôme d'interpolation d'Hermite Q vérifiant :*

$$Q(0) = f(0),\ Q(\frac{1}{2}) = f(\frac{1}{2}),\ Q(1) = f(1).$$

$$Q'(0) = f'(0),\ Q'(\frac{1}{2}) = f'(\frac{1}{2}),\ Q'(1) = f'(1).$$

   Calculer $P(0.1)$ et $P(0.9)$. Comparer aux valeurs exactes.

Chapitre 3

Approximation au sens des moindres carrés

3.1 Définitions

Définition 3.1. *Soit E un espace vectoriel réel.*

1. *L'application*
$$\begin{array}{rccc} <.,.>: & E \times E & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & <x,y> \end{array}$$
est appelée produit scalaire sur E si elle vérifie :

   i) $<x,x> \geq 0,\ \forall x \in E,$

   ii) $<x,x> = 0 \Longleftrightarrow x = 0_E,$

   iii) $<x,y> = <y,x>,\ \forall x,y \in E,$

   vi) $<x,\alpha y + \beta z> = \alpha <y,x> + \beta <x,z>,\ \forall x,y,z \in E,\ \alpha,\beta \in \mathbb{R}.$

2. *L'application*
$$\begin{array}{rccc} \|.\| : & E & \longrightarrow & \mathbb{R}_+ \\ & x & \longmapsto & \|x\| = \sqrt{<x,x>} \end{array}$$
est appelée norme sur E associée au produit scalaire $<.,.>$.

3. *Un espace vectoriel réel muni par un produit scalaire est dit* ***préhilbertien sur $\boldsymbol{R}$****.*

4. *Un espace préhilbertien complet pour la norme associée à ce produit scalaire est dit* ***espace de Hilbert****.*

5. *Soit $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ une base de E.*

   a) *La famille $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ est appelée base orthogonale de E si*
$$<\varphi_i, \varphi_j> = 0 \ \ \forall i,j = \overline{0,n} \ \ \text{et } i \neq j.$$

   b) *La famille $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ est appelée base orthonormée de E si*
$$<\varphi_i, \varphi_j> = \begin{cases} 0, & si\ i \neq j; \\ 1, & si\ i = j. \end{cases}$$

Exemple 3.1. *1. Soit $E = \mathbb{R}^n,\ x = (x_1, \ldots, x_n),\ y = (y_1, \ldots, y_n) \in E.$*
L'application
$$\begin{array}{rccc} <.,.>: & \mathbb{R}^n \times \mathbb{R}^n & \longrightarrow & \mathbb{R} \\ & (x,y) & \longmapsto & <x,y> = \sum_{i=1}^{n} x_i y_i \end{array}$$

définit un produit scalaire sur $\mathbb{R}^n$ *et la norme associée à ce produit scalaire est définie par :*

$$x = \sqrt{< x, x >} = \left(\sum_{i=1}^{n} x_i^2 \right)^{\frac{1}{2}}.$$

2. *Soit* $E = C([a,b]),\ f,\ g \in E$.
L'application

$$\begin{array}{rccl} < ., . >: & C([a,b]) \times C([a,b]) & \longrightarrow & \mathbb{R} \\ & (x, y) & \longmapsto & < f, g >= \int_a^b f(x)\, g(x)\, dx \end{array}$$

définit un produit scalaire sur $C([a,b])$ *et la norme associée à ce produit scalaire est définie par :*

$$x = \sqrt{< x, x >} = \left(\int_a^b f(x)^2\, dx \right)^{\frac{1}{2}}.$$

3. *La famille* $\{1, x, \frac{1}{2}(3x^2 - 1)\}$ *est une base orthonormée de* $\mathbb{P}_2$ *pour le produit scalaire défini par*

$$< f, g >= \int_{-1}^{1} f(x)\, g(x)\, dx.$$

3.2 Position du problème

Soit E un espace vectoriel réel tel que $\mathbb{P}_n \subset E,\ f \in E$, où $\mathbb{P}_n$ est l'espace des polynômes de degré inférieur ou égal à n. L'approximation polynomiale au sens des moindres carrés consiste à déterminer le polynôme $P^* \in \mathbb{P}_n$ qui vérifie

$$\|f - P^*\| = \min_{P \in \mathbb{P}_n} \|f - P\|.$$

P^* est appelé **meilleur approximant au sens des moindres carrés de** f **dans** $\mathbb{P}_n$.

3.2.1 Existence et unicité de la meilleure approximation au s.m.c.

Théorème 3.1. *(**Théorème de la projection**) Soient* E *un espace de Hilbert et* $F \subseteq E$ *un convexe fermé non vide. Alors,*

$$\forall\, f \in E,\ \exists\, \psi^* \in F : \quad \|f - \psi^*\| = \min_{\psi \in F} \|f - \psi\|.$$

De plus, ψ^* *est caractérisé par :*

$$< f - \psi^*, \psi >= 0 \ \ \forall\, \psi \in F.$$

Dans notre cas : $F = \mathbb{P}_n$ et $dim(\mathbb{P}_n) = n + 1 < \infty$, alors d'une part, F est convexe, et d'autre part, F est complet et donc il est fermé. Donc, le meilleur approximant au sens des moindres carrés de f dans $\mathbb{P}_n$ existe et il est unique.

3.2.2 Détermination de la meilleure approximation au s.m.c.

Soit $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ une base de $\mathbb{P}_n$, $f \in E$.

$$P \in \mathbb{P}_n \Longrightarrow P = \sum_{j=0}^{n} a_j \varphi_j, \;\; a_j \in \mathbb{R}.$$

$$P^* \in \mathbb{P}_n \Longrightarrow P^* = \sum_{k=0}^{n} a_k^* \varphi_k, \;\; a_k^* \in \mathbb{R}.$$

D'après le théorème de la projection P^* vérifiant

$$\begin{array}{rl}
 & < f - P^*, P >= 0 \quad \forall P \in \mathbb{P}_n \\
\Longleftrightarrow & < f - P^*, \sum_{i=0}^{n} a_j \varphi_j >= 0 \;\; \forall a_j \in \mathbb{R}, j = 0, \ldots, n \\
\Longleftrightarrow & \sum_{j=0}^{n} a_j \; < f - P^*, \varphi_j >= 0 \;\; \forall a_j \in \mathbb{R}, j = 0, \ldots, n \\
\Longleftrightarrow & < f - P^*, \varphi_j >= 0 \; \forall j = 0, \ldots, n \\
\Longleftrightarrow & < P^*, \varphi_j >=< f, \varphi_j > \quad \forall j = 0, \ldots, n \\
\Longleftrightarrow & < \sum_{j=0}^{n} a_k^* \varphi_k, \varphi_j >=< f, \varphi_j > \quad \forall j = 0, \ldots, n \\
\Longleftrightarrow & \left\{ \begin{array}{l} \sum_{j=0}^{n} a_k^* \; < \varphi_k, \varphi_j >=< f, \varphi_j > \\ j = 0, \ldots, n. \end{array} \right.
\end{array}$$

C'est un système linéaire de $(n+1)$ équations à $(n+1)$ inconnues. On développe ce système, on aura le système matriciel :

$$\begin{pmatrix}
< \varphi_0, \varphi_0 > & < \varphi_1, \varphi_0 > & \cdots & \cdots & < \varphi_n, \varphi_0 > \\
< \varphi_0, \varphi_1 > & < \varphi_1, \varphi_1 > & \cdots & \cdots & < \varphi_n, \varphi_1 > \\
\vdots & \vdots & & & \vdots \\
\vdots & \vdots & & & \vdots \\
< \varphi_0, \varphi_n > & < \varphi_1, \varphi_n > & \cdots & \cdots & < \varphi_n, \varphi_n >
\end{pmatrix}
\begin{pmatrix} a_0^* \\ a_1^* \\ \vdots \\ \vdots \\ a_n^* \end{pmatrix}
=
\begin{pmatrix} < f, \varphi_0 > \\ < f, \varphi_1 > \\ \vdots \\ \vdots \\ < f, \varphi_n > \end{pmatrix}$$

Remarquons que ce système admet une et une seule solution car P^* existe et il est unique.

Remarque 3.1. *1. Si la famille $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ forme une base orthogonale de $\mathbb{P}_n$, on obtient le système diagonal*

$$\begin{pmatrix}
< \varphi_0, \varphi_0 > & 0 & \cdots & \cdots & 0 \\
0 & < \varphi_1, \varphi_1 > & \cdots & \cdots & 0 \\
\vdots & \vdots & & & \vdots \\
\vdots & \vdots & & & \vdots \\
0 & 0 & \cdots & \cdots & < \varphi_n, \varphi_n >
\end{pmatrix}
\begin{pmatrix} a_0^* \\ a_1^* \\ \vdots \\ \vdots \\ a_n^* \end{pmatrix}
=
\begin{pmatrix} < f, \varphi_0 > \\ < f, \varphi_1 > \\ \vdots \\ \vdots \\ < f, \varphi_n > \end{pmatrix}$$

D'où,

$$a_k^* = \frac{< f, \varphi_k >}{< \varphi_k, \varphi_k >} = \frac{< f, \varphi_k >}{\|\varphi_k\|^2}, \;\; k = 0, \ldots, n.$$

2. Si la famille $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ forme une base orthonormale de $\mathbb{P}_n$, on trouve immédiatement

$$a_k^* =< f, \varphi_k > \quad k = 0, \ldots, n.$$

3.2.3 Erreur d'approximation

$$\|f - P^*\| = \sqrt{\|f\|^2 - \sum_{i=0}^{N} a_k^* < f, \varphi_k >}.$$

En effef ;

$$\begin{aligned} \|f - P^*\|^2 &= < f - P^*, f - P^* > \\ &= < f - P^*, f > - < f - P^*, P^* > \\ &= < f, f > - < f, P^* > \\ &= \|f\|^2 - \sum_{i=0}^{N} a_k^* < f, \varphi_k > \cdot \end{aligned}$$

Cas particulier : si la base $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$ est orthonormale, on obtient

$$\|f - P^*\| = \sqrt{\|f\|^2 - \sum_{i=0}^{N} a_k^{*2}}.$$

3.2.4 Algorithme de Gram-Schmidt

Soit F un espace de dimension finie. En partant d'une base quelconque de F, notée $\{\varphi_0, \varphi_1, \ldots, \varphi_n\}$, on peut déterminer à la fois une base orthonormée de F, notée $\{u_0, u_1, \ldots, u_n\}$, et une autre base orthonormée, notée $\{h_0, h_1, \ldots, h_n\}$, à l'aide de l'algorithme de Gram-Schmidt suivant :

$$\begin{cases} h_0 = \frac{u_0}{\|u_0\|}, & \text{où} \quad u_0 = \varphi_0; \\ h_1 = \frac{u_1}{\|u_1\|}, & \text{où} \quad u_1 = \varphi_1 - < \varphi_1, h_0 > h_0; \\ h_2 = \frac{u_2}{\|u_2\|}, & \text{où} \quad u_2 = \varphi_2 - < \varphi_2, h_0 > h_0 - < \varphi_2, h_1 > h_1; \\ \vdots & \\ h_k = \frac{u_k}{\|u_k\|}, & \text{où} \quad u_k = \varphi_k - \sum_{m=0}^{k-1} < \varphi_k, h_m > h_m, \ 1 \leq k \leq n; \\ \vdots & \\ h_n = \frac{u_n}{\|u_n\|}, & \text{où} \quad u_n = \varphi_n - \sum_{m=0}^{n-1} < \varphi_n, h_m > h_m. \end{cases}$$

3.3 Application au cas discret

Soient $E = C([a,b])$, $x_0, x_1, \ldots, x_N$ $(N+1)$ points distincts de l'intervalle $[a,b]$. Soit $f \in E$ connue seulement aux points x_i, $i = 0, \ldots, N$. On définit sur E le produit scalaire

$$< g, h >= \sum_{i=0}^{N} \omega_i \, g(x_i) h(x_i), \ \forall \, g, h \in C([a,b])) \qquad \text{où } \omega_i \text{ sont des poids.}$$

ω_i sont des nombres positifs non nuls à la fois. En pratique, on prend des poids égaux.
La norme associée à ce produit scalaire est définie par :

$$\|g\| = \left(\sum_{i=0}^{N} \omega_i \, (g(x_i))^2 \right)^{\frac{1}{2}}, \ \forall \, g \in C([a,b])).$$

Posons $\varphi_k(x) = x^k, \;\; k = 0, \ldots, n$. Dans la base canonique de $\mathbb{P}_n$, la meilleure approximation au sens des moindres carrés P^* de f s'écrit

$$P^*(x) = a_0^* + a_1^* x + \ldots + a_n^* x^n,$$

où $(a_0^*, a_1^*, \ldots, a_n^*)$ est l'unique solution du système matriciel :

$$\begin{pmatrix} \sum_{i=0}^{N} \omega_i & \sum_{i=0}^{N} \omega_i\, x_i & \cdots & \cdots & \sum_{i=0}^{N} \omega_i\, x_i^n \\ \sum_{i=0}^{N} \omega_i\, x_i & \sum_{i=0}^{N} \omega_i\, x_i^2 & \cdots & \cdots & \sum_{i=0}^{N} \omega_i\, x_i^{n+1} \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \sum_{i=0}^{N} \omega_i\, x_i^n & \sum_{i=0}^{N} \omega_i\, x_i^{n+1} & \cdots & \cdots & \sum_{i=0}^{N} \omega_i\, x_i^{2n} \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ \vdots \\ \vdots \\ a_n^* \end{pmatrix} = \begin{pmatrix} \sum_{i=0}^{N} \omega_i f(x_i) \\ \sum_{i=0}^{N} \omega_i\, x_i\, f(x_i) \\ \vdots \\ \vdots \\ \sum_{i=0}^{N} \omega_i\, x_i^n\, f(x_i) \end{pmatrix}.$$

3.4 Application au cas continu

Soient $E = C([a,b])$ muni du produit scalaire

$$< g, h >= \int_a^b \omega(x)\, g(x) h(x)\, dx \;\; \forall\, g, h \in E \qquad \text{où } \omega \text{ est une fonction poids.}$$

$\omega : [a,b] \longrightarrow \mathbb{R}_+$ s'annule un nombre fini de fois sur $[a,b]$. En pratique, on prend $\omega \equiv 1$.
La norme associée à ce produit scalaire est définie par :

$$\|g\| = \left(\sum_{i=0}^{N} \omega(x)\, (g(x))^2 \right)^{\frac{1}{2}}, \;\; \forall\, g \in C([a,b])).$$

Posons $\varphi_k(x) = x^k, \;\; k = 0, \ldots, n$. Dans la base canonique de $\mathbb{P}_n$, la meilleure approximation au sens des moindres carrés P^*, d'une fonction $f \in E$, s'écrit

$$P^*(x) = a_0^* + a_1^* x + \ldots + a_n^* x^n,$$

où $(a_0^*, a_1^*, \ldots, a_n^*)$ est l'unique solution du système matriciel :

$$\begin{pmatrix} \int_a^b \omega(x)\, dx & \int_a^b \omega(x)\, x\, dx & \cdots & \cdots & \int_a^b \omega(x)\, x^n\, dx \\ \int_a^b \omega(x)\, x\, dx & \int_a^b \omega(x)\, x^2\, dx & \cdots & \cdots & \int_a^b \omega(x)\, x^{n+1}\, dx \\ \vdots & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ \int_a^b \omega(x)\, x^n\, dx & \int_a^b \omega(x)\, x^{n+1}\, x\, dx & \cdots & \cdots & \int_a^b \omega(x)\, x^{2n}\, dx \end{pmatrix} \begin{pmatrix} a_0^* \\ a_1^* \\ \vdots \\ \vdots \\ a_n^* \end{pmatrix} = \begin{pmatrix} \int_a^b \omega(x)\, f(x)\, dx \\ \int_a^b \omega(x)\, x\, f(x)\, dx \\ \vdots \\ \vdots \\ \int_a^b \omega(x)\, x^n\, f(x)\, dx \end{pmatrix}.$$

3.5 Exercices

Exercice 3.1.

1. *Soient les points* $x_0 = -1, x_1 = -\frac{1}{2}, x_2 = 0, x_3 = \frac{1}{2}, x_4 = 1$.
   Trouver le polynôme $P^* \in \mathbb{P}_2$ *qui réalise le minimum suivant :*

$$\min_{P\in\mathbb{P}_2} \sum_{i=0}^{4} \frac{1}{1+x_i^2} \left(|x_i| - P(x_i)\right)^2$$

   où $\mathbb{P}_2$ *est l'ensemble des polynômes de degé*≤ 2.

Exercice 3.2.

Soient $f(x) = |x|$ *et* $< g, h >= \sum_{i=0}^{4} g(x_i)h(x_i),\ \forall\, g, h \in C([-1,1],\mathbb{R})$ *où*

$$x_0 = -1,\ x_1 = -\frac{1}{2},\ x_2 = 0,\ x_3 = \frac{1}{2} \text{ et } x_4 = 1.$$

1. *Déterminer le polynôme* P *de* $\mathbb{P}_2$ *qui réalise la meilleure approximation de* f *au sens des moindres carrés.*
2. *Vérifier que* $Q(x) = \frac{7}{3}x^2 - \frac{4}{3}x^4$ *est le polynôme d'interpolation de* f *aux points* x_i, $i = \overline{0,4}$.
3. *Dans un même repère, tracer les graphes de* f, P *et* Q *sur* $[-1,1]$. *Commenter.*

Exercice 3.3. *On considère l'ensemble* $E = C([-1,1])$ *muni du produit scalaire*

$$< f, g >= \int_{-1}^{1} f(x)g(x)\,dx.$$

Soit $\mathbb{P}_1$ *l'ensemble des polynômes de degré* ≤ 1 *et* $f \in E$.

1. *Trouver en fonction de* f, *le polynôme* $P_1(x) = a + bx$ *qui réalise la meilleure approximation au sens des moindres carrés de* f *dans* $\mathbb{P}_1$.
2. *Montrer que* P_1 *peut se mettre sous la forme :*

$$P_1(x) = \frac{1}{2}\int_{-1}^{1} (1+3xt)f(t)\,dt.$$

3. *Trouver* P_1 *pour* $f(x) = -x^3 - 2x$. *Tracer les graphes de* f *et* P_1.
4. *Evaluer l'erreur et faire sa représentation graphique.*

Exercice 3.4. *On cherche le polynôme* $P_2(x) = a_0 + a_1x + a_2x^2$ *qui minimise l'expression*

$$\int_{-1}^{1} \left(f(x) - P_2(x)\right)^2 \frac{dx}{\sqrt{1-x^2}}$$

1. *On considère les polynômes orthogonaux, donnés par la relation*

$$T_0(x) = 1,\ T_1(x) = x,\ T_{n+1} = 2xT_n(x) - T_{n-1},\ n \geq 1;$$

   Calculer T_2, T_3, T_4.
2. *On prend* $f(x) = 2x^3 + x^4$.
   Déterminer P_2 *dans la base* $\{T_1, T_2, T_3\}$ *et déduire les valeurs de* a_0, a_1 *et* a_2.
3. *Evaluer l'erreur commise.*

Chapitre 4

Intégration numérique

4.1 Position du problème

On veut évaluer l'intégrale d'une fonction f sur un intervalle $[a, b]$. Si l'on connait sa primitive F, alors

$$\int_a^b f(x)\,dx = F(b) - F(a) \text{ où } F' = f.$$

Mais dans de nombreux cas, F ne peut pas être connue.

Exemple 4.1. *: $\int_a^b \frac{\sin x}{x}\,dx$, $\int_a^b e^{-x^2}\,dx$, $\int_a^b \sqrt{1-k\sin x}\,dx$, $\int_a^b \frac{1}{\sqrt{1-\sin(x)}}\,dx$.*

La plupart des fonctions qui interviennent dans les problèmes physiques sont des fonctions soit trop compliquées pour être intégrées, soit seulement données par une table de valeurs. Nous cherchons donc une valeur approchée à l'aide de sommes finies, qu'on appellera une ***formule de quadrature :***

$$I(f) = \int_a^b f(x)\,dx \simeq I_n(f) = \sum_{i=0}^{n} \alpha_i f(x_i),$$

où les x_i, $i = 0, \ldots, n$ sont appelés ***points*** *d'intégration et les α_i, $i = 0, \ldots, n$* ***poids*** *de la formule de quadrature.*

Définition 4.1. *Une formule de quadrature est dite* ***exacte*** *sur un ensemble V si*

$$\forall\, f \in V,\ R(f) = I(f) - I_n(f) = 0.$$

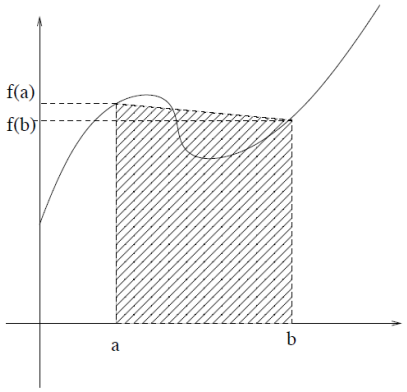
Définition 4.2. *Une formule de quadrature est dite de* ***degré de précision*** *n si elle est exacte pour x^k, $k = 0, \ldots, n$ et non exacte pour x^{n+1}.*

4.2 Formules de Newton-Côtes

4.2.1 Méthode des trapèzes

Pour évaluer numériquement $I(f) = \int_a^b f(x)dx$, *on divise l'intervalle borné* $[a,b]$ *en* n *parties* $[x_0,x_1],[x_1,x_2],...,[x_{n-1},x_n]$ *de même longueur* $h = \frac{b-a}{n}$ *et on considère les points d'intégration*

$$x_0 = a, x_1 = x_0 + h, ..., x_i = x_0 + ih \ (i = \overline{0,n}), x_n = b.$$



Si $n = 1,\ x_0 = a, x_1 = b,$ *on obtient*

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

C'est la formule simple des trapèzes sur l'intervalle $[a,b]$.
L'aire $I(f)$ *comprise entre* $[a,b]$ *et le graphe de* f *peut être approchée par la somme des aires des* n *trapèzes induits par les points* $x_0, x_1, ..., x_n$.

Sur chaque intervalle $[x_i, x_{i+1}],\ 0 \leq i \leq n-1$:

$$\int_{x_i}^{x_{i+1}} f(x)dx \simeq \frac{(x_{i+1} - x_i)}{2}(f(x_i) + f(x_{i+1})).$$

Donc

$$\begin{aligned} I(f) = \int_a^b f(x)dx &= \int_{a=x_0}^{x_1} f(x)dx + \int_{x_1}^{x_2} f(x)dx + \ldots + \int_{x_{n-1}}^{b=x_n} f(x)dx \\ &= \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} f(x)dx \\ &\simeq \frac{h}{2} \sum_{i=0}^{n-1} (f(x_i) + f(x_{i+1})) \\ &= \frac{h}{2}(f(x_0) + 2f(x_1) + 2f(x_2) + ... + 2f(x_{n-1}) + f(x_n)). \end{aligned}$$

D'où

$$I_n(f) = \int_a^b f(x)dx \simeq \frac{h}{2}\left[f(a) + 2\sum_{i=1}^{n-1} f(x_i) + f(b)\right].$$

C'est la formule des trapèze composite sur l'intervalle $[a,b]$.

Exemple 4.2. *Soit* $f(x) = x^2$, $a = 0, b = 1$, *on prend* $n = 3$ *subdivisions.*
Donc $h = \frac{b-a}{n} = \frac{1}{3}, x_0 = a = 0, x_1 = \frac{1}{3}, x_2 = \frac{2}{3}, x_3 = b = 1$ *avec*
$y_0 = f(x_0) = 0, y_1 = f(x_1) = \frac{1}{9}, y_2 = f(x_2) = \frac{4}{9}$ *et* $y_3 = f(x_3) = 1$.

$$I(f) = \int_0^1 x^2\,dx \simeq I_3(f) = \frac{h}{2}(y_0 + 2y_1 + 2y_2 + y_3) = \frac{19}{54} \simeq 0,351$$

Erreur d'approximation : $I(f) = \int_0^1 x^2\,dx = \frac{1}{3} \simeq 0,333$, *d'où,*

$$\text{l'erreur relative} \simeq \frac{|0,351-0,333|}{0,333} = 5,4\%.$$

Par contre, si $n = 6$, *l'erreur relative* $\simeq 1,5\%$.

On remarque numériquement que l'erreur a été divisée-approximativement-par 4 lorsque n *a doublé (de* $n = 3$ *à* $n = 6$*), l'erreur dépend donc de* h.

Recherche de l'erreur d'approximation si $f \in C^2([a, b])$

Rappel 1 : *Si* $g : [0, a] \to \mathbb{R}$ *dérivable sur* $]0, a[$, *alors on peut écrire :*

$$g(a) = g(0) + \int_0^a g'(t)\,dt.$$

Rappel 2 : *(le théorème de la moyenne) Soit* $f, g : [a, b] \to \mathbb{R}$ *continues, si* g *a un signe constant sur* $[a, b]$, *alors* $\exists\,\xi \in [a, b]$ *tel que :*

$$\int_a^b f(t)g(t)\,dt = f(\xi)\int_a^b g(t)\,dt.$$

Rappel 3 : *(théorème des valeurs intermédiaires) Soit* $f : [a, b] \to \mathbb{R}$ *continue de* $\max M$ *et* $\min m$. *Alors,* $\forall\, y \in [m, M]$, $\exists\, x \in [a, b]/f(x) = y$.

Posons

$$R = I(f) - I_n(f) = \int_a^b f(x)\,dx - \frac{h}{2}\left[f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right]$$

avec $x_i = x_0 + ih$ $(h = \frac{b-a}{n})$ $i = 0, 1, \ldots, n$.

Considérons d'abord R_1 *l'erreur commise sur le* 1^{er} *l'intervalle* $[x_0, x_1]$

$$R_1 = \int_{x_0}^{x_1} f(x)\,dx - \frac{h}{2}(f(x_0) + f(x_1)) = \int_{x_0}^{x_0+h} f(x)\,dx - \frac{h}{2}(f(x_0) + f(x_0 + h)),$$

f *étant donnée, on remarque que* R_1 *dépend seulement de* h, *donc* $R_1 = R_1(h)$.
Dérivons R_1 *une fois on obtient*

$$R_1'(h) = \frac{1}{2}(f(x_0 + h) - f(x_0)) - \frac{h}{2}f'(x_0 + h).$$

Notons, en passant, que $R_1(0) = R_1'(0) = 0$.
Dérivons R_1', *ceci est permis car* $f \in C^2([a, b])$, *on aura*

$$R_1''(h) = -\frac{h}{2}f''(x_0 + h).$$

Du rappel 1, on a

$$R_1'(h) = R_1'(0) + \int_0^h R_1''(t)\,dt = -\int_0^h \frac{t}{2} f''(x_0+t)\,dt,$$

et du rappel 2, pour $g(t) = t \geq 0\ \forall\, t \in [0,h], \exists\, \xi^* \in [0,h]$ *et donc* $\exists\, \xi_1 \in [x_0, x_0+h]$:

$$R_1'(h) = -\frac{f''(\xi_1)}{2}\int_0^h t\,dt,$$

alors $R_1'(h) = -\frac{h^2}{4} f''(\xi_1)$ *avec* $\xi_1 \in [x_0, x_1]$, *mais (du rappel 1) :*

$$R_1(h) = R_1(0) + \int_0^h R_1'(t)\,dt = -\frac{f''(\xi_1)}{4}\int_0^h t^2\,dt = -\frac{h^3}{12} f''(\xi_1),$$

$R_1(h)$ *est l'erreur d'approximation sur l'intervalle* $[x_0, x_1]$, *auquel appartient* ξ_1.
D'où l'erreur totale :

$$R(h) = \sum_{i=1}^{n} R_i(h) = -\frac{h^3}{12}(f''(\xi_1) + f''(\xi_2) + ... + f''(\xi_n)) \quad où\ \ \xi_i \in [x_i, x_{i+1}], i = \overline{0, n-1}.$$

Réécrivons $R(h)$, *en mulipliant et divisant par* n :

$$R(h) = \sum_{i=1}^{n} = -\frac{nh^3}{12}\frac{\sum_{i=1}^{n} f''(\xi_i)}{n} = -\frac{nh^3}{12}.\overline{m}$$

où $\overline{m}$ *est la moyenne arithmétique des* $f''(\xi_i)$ $i = 1, ..., n$. *Comme* $\xi_i \in [a,b]\ \forall\, i = \overline{0, n-1}$ *et* f'' *est continue sur* $[a,b]$; *alors d'après le théorème des valeurs intermédiaires,* $\exists\, \xi \in [a,b]$:
$\overline{m} = f''(\xi)$; *d'où :*

$$R(h) = -\frac{nh^3}{12}.f''(\xi) \Rightarrow |R(h)| \leq \frac{nh^3}{12} M_2$$

où $M_2 = \max_{t\in[a,b]} |f''(\xi)|$. *Comme* $h = \frac{(b-a)}{n}$, *on obtient finalement*

$$|R(h)| \leq \frac{(b-a)^3}{12n^2}\, M_2.$$

Ceci implique que la formule des trapèzes est exacte si f *est un polynôme de degré inférieur ou égal à 1. De plus, cette formule est de degré de précision égal à 1. En effet ;*
pour $f(x) = x^2$, $n = 1$, $x_0 = a$, $x_1 = b$, *on a*

$$I(f) = \int_a^b x^2\,dx = \left[\frac{x^3}{3}\right]_a^b = \frac{b-a}{3}\left(a^2 + b^2 + ab\right) \neq I_1(f) = \frac{b-a}{2}\left(a^2 + b^2\right).$$

Remarque 4.1. *Cette borne d'erreur confirme notre observation sur l'exemple donné précédemment ; en effet si n est doublé, l'erreur est alors divisée par 4 (approximativement !!). Mais le cumul d'erreurs d'arrondi (dûes au calcul numérique) fait que l'on n'atteint pas la valeur exacte de* $I(f)$ *lorsque n est de plus en plus grand.*

Exercice 4.1. *(exercice d'illustration) Reprendre l'exercice précédent avec un nombre des points* (x_i, y_i) $i = 0, \ldots, n$, *de plus en plus grand et observer la convergence de* $I_n(f)$ *vers* $I(f)$.

4.2.2 Méthode de Simpson

Dans le cadre de cette méthode, on interpole chaque trois points $(x_i, y_i), (x_{i+1}, y_{i+1}), (x_{i+2}, y_{i+2})$ *par un polynôme de degré inférieur ou égal à 2. Comme trois points induisent deux subdivisions, le nombre* n *de subdivisions doit être pris pair* $(n = 2m)$.
On peut prendre le polynôme d'interpolation de Lagrange ou celui de Newton. Considérons le polynôme d'interpolation de Newton qui passe par les points (x_0, y_0), (x_1, y_1) *et* (x_2, y_2), *notons le* $P_2(x)$. *Il s'écrit sous la forme :*

$$P_2(x) = y_0 + \frac{y_1 - y_0}{h}(x - x_0) + \frac{y_2 - 2y_1 + y_0}{2h^2}(x - x_0)(x - x_1).$$

Donc comme $f(x) \simeq P_2(x)\ \forall\, x \in [x_0, x_2]$, *on a*

$$\int_{x_0}^{x_2} f(x)\,dx \simeq \int_{x_0}^{x_2} P_2(x)\,dx = \frac{h}{3}(y_0 + 4y_1 + y_2).$$

C'est la première formule de Simpson[1] *simple sur l'intervalle* $[x_0, x_2]$. *D'où*

$$\begin{aligned}
\int_a^b f(x)\,dx &= \int_{a=x_0}^{x_2} f(x)\,dx + \int_{x_2}^{x_4} f(x)\,dx + \ldots + \int_{x_{2m-2}}^{b=x_{2m}} f(x)\,dx \\
&\simeq \tfrac{h}{3}(y_0 + 4y_1 + y_2) + \tfrac{h}{3}(y_2 + 4y_3 + y_4) + \ldots + \tfrac{h}{3}(y_{2m-2} + 4y_{2m-1} + y_{2m}) \\
&= \tfrac{h}{3}\left[y_0 + y_{2m=n} + 2(y_2 + y_4 + \ldots + y_{2m-2}) + 4(y_1 + y_3 + y_{2m-1})\right] \\
&= \tfrac{h}{3}[y_0 + y_n + 2\sum_{i\ pair} y_i + 4\sum_{i\ impair} y_i].
\end{aligned}$$

Donc,

$$I(f) = \int_a^b f(x)\,dx \simeq I_n(f) = \frac{h}{3}\left(y_0 + y_{n=2m} + 2\sum_{i\ pair} y_i + 4\sum_{i\ impair} y_i\right) \quad où\ h = \frac{b-a}{n}.$$

C'est la première de Simpson composite sur l'intervalle $[a, b]$.

Application à l'exemple précédent, avec $n = 4$:
$f(x) = x^2$, $a = 0, b = 1, h = 1/4$, *d'après la formule de Simpson*

$$\int_0^1 f(x)\,dx \simeq I_4(f) = \frac{h}{3}\left(y_0 + y_4 + 2y_2 + 4(y_1 + y_3)\right) = \frac{1}{12}\left(0 + 1 + \frac{5}{2} + \frac{1}{2}\right) = \frac{1}{3} = I(f)\ ?!$$

Recherche de l'erreur d'approximation si $f \in C^4([a, b])$

En suivant une démarche quasi-analogue au cas de la méthode des trapèzes, on tire que :

$$\exists\, \xi \in [a, b], \quad R(h) = I(f) - I_n(f) = -\frac{n\,h^5}{180} f^{(4)}(\xi). \tag{4.1}$$

Par consequent,

$$|R(h)| = |I(f) - I_n(f)| \leq \frac{(b-a)^5}{180 n^4} M_4, \quad où \quad M_4 = \max_{t \in [a,b]} |f^{(4)}(t)|.$$

[1] Thomas Simpson, anglais, 1710-1761

Ceci implique que la formule de Simpson est exacte si f est un polynôme de degré inférieur ou égal à 3. De plus, cette formule est de degré de précision égal à 3. En effet ;
pour $f(x) = x^4$, $n = 2$, $x_0 = a$, $x_1 = \frac{a+b}{2}$ et $x_2 = b$:

$$I(f) = \int_a^b x^4\,dx = \left[\frac{x^5}{5}\right]_a^b = \frac{b-a}{5}\left(b^4 + ab^3 + a^2b^2 + a^3b + a^4\right),$$

alors que

$$I_2(f) = \frac{b-a}{6}\left(a^4 + 4\left(\frac{a+b}{2}\right)^4 + b^4\right) = \frac{b-a}{24}\left(5b^4 + 4ab^3 + 6a^2b^2 + 4a^3b + 5a^4\right).$$

On voit clairement que $I(f) \neq I_2(f)$.

Remarque 4.2. *Si n a doublé, l'erreur est divisée par 16 (approximativement), donc la présente méthode est plus performante que celle des trapèzes.*

Exercice 4.2. *Notons par $R_1 = R_1(h)$ l'erreur commise sur la subdivision $[x_0, x_2]$ ou encore sur $[x_1 - h, x_1 + h]$:*

$$R_1(h) = \int_{x_1-h}^{x_1+h} f(x)\,dx - \frac{h}{3}\left(f(x_1 - h) + 4f(x_1) + +f(x_1 + h)\right).$$

En utilisant, les rappels 1, 2 et 3, démontrer l'inégalité (4.1).

Proposition 4.1. *Le degré de précision des formules de Newton-Côtes[2] à (n+1) points est*

$$\begin{cases} n, & \text{si } n \text{ est impair ;} \\ n+1, & \text{si } n \text{ est pair.} \end{cases}$$

L'erreur dans les formules de Newton-Côtes à $(n+1)$ points est en

$$\begin{cases} h^{n+2}, & \text{si } n \text{ est impair ;} \\ h^{n+3}, & \text{si } n \text{ est pair,} \end{cases} \qquad \text{où } h = \frac{b-a}{n}.$$

Remarque 4.3.

1. *On peut démonter que si le degré de précision est p alors l'erreur est en h^{p+2} et réciproquement. Comme le degré de précision est facile à évaluer, il est possible d'avoir une idée de l'erreur.*
2. *On peut construire des formules de Newton-Côtes qui ne comportent pas les bornes d'intégration comme points de la formule de quadrature.*
   Par exemple, pour intégrer la fonction $\frac{\sin x}{x}$ entre 0 et 1 on peut utiliser :

   La formule du point milieu : $\int_a^b f(x)\,dx \simeq (b-a) f\left(\frac{a+b}{2}\right)$.

   La formule variante des trapèzes : $\int_a^b f(x)\,dx \simeq \frac{(b-a)}{2}\left[f\left(a + \frac{b-a}{3}\right) + f\left(b - \frac{b-a}{3}\right)\right]$.

   La formule variante de Simpson : $\int_a^b f(x)\,dx \simeq \frac{(b-a)}{8}\left[3f\left(\frac{a+5b}{6}\right) + 2f\left(\frac{b-a}{2}\right) + 3f\left(\frac{5a+b}{6}\right)\right]$.

[2] Roger Côtes, anglais, 1682-1716

4.3 Formules de Gauss

On se pose la question suivante : comment choisir au mieux les points d'intégration x_i pour que la formule de quadrature soit de degré de précision le plus élevé possible ? Le problème revient donc à trouver à la fois les poids $\alpha_i^{(n)}, i = 0, \ldots, n$ et les points $x_i, i = 0, \ldots, n$ de la formule de quadrature :

$$\int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} \alpha_i^{(n)} f(x_i).$$

On a donc $2n+2$ inconnues à déterminer !
On cherche alors une formule de quadrature exacte sur $\mathbb{P}_{2n+1}$, ce qui donne les $2n+2$ équations :

$$\int_a^b x^k\,dx = \sum_{i=0}^{n} \alpha_i^{(n)} x_i^k, \quad k = 0, \ldots, 2n+1.$$

Remarque 4.4. *Le degré de précision de la formule proposée peut-il être $2n+2$?*
*La réponse est **non**. En effet ; soit le polynôme de degré $2n+2$,*

$$Q(x) = (x - x_0)^2 (x - x_1)^2 \ldots (x - x_n)^2.$$

Ce polynôme est strictement positif (si $x \neq x_i$, $i = 0, \ldots, n$) donc $\int_a^b Q(x)\,dx > 0$,
alors que $\sum_{i=0}^{n} \alpha_i^{(n)} Q(x_i) = 0$.

Théorème 4.1. *Les formules de type Gauss[3] sont stables et convergentes pour toute fonction continue.*

Remarque 4.5.
1. *Stable signifie les poids intervenant dans la formule de quadrature sont strictement positifs.*
2. *Convergente signifie $\lim_{n\to+\infty} \sum_{i=0}^{n} \alpha_i^{(n)} f(x_i) = \int_a^b f(x)\,dx$.*
3. *Les formules de type Gauss sont souvent utiles pour les intégrales impropres convergentes.*

Exemple 4.3. *Soit la formule de quadrature à 2 points :*

$$\int_{-1}^{1} f(x)\,dx = \alpha_0 f(x_0) + \alpha_1 f(x_1) + R(f).$$

Déterminons les poids α_0, α_1 et les points d'intégration x_0 et x_1 pour que la formule proposée soit exacte sur $\mathbb{P}_3$. Les quatre équations sont les suivantes :

$$\begin{aligned}
f \equiv 1,\ R(f) = 0 &\iff \textstyle\int_{-1}^{1} 1\,dx = 2 = \alpha_0 + \alpha_1,\\
f(x) = x,\ R(f) = 0 &\iff \textstyle\int_{-1}^{1} x\,dx = 0 = \alpha_0 x_0 + \alpha_1 x_1,\\
f(x) = x^2,\ R(f) = 0 &\iff \textstyle\int_{-1}^{1} x\,dx2 = \frac{2}{3} = \alpha_0 x_0^2 + \alpha_1 x_1^2,\\
f(x) = x^3,\ R(f) = 0 &\iff \textstyle\int_{-1}^{1} x\,dx = 0 = \alpha_0 x_0^3 + \alpha_1 x_1^3.
\end{aligned}$$

D'où,

$$\begin{cases} \alpha_0 x_0 = -\alpha_1 x_1, \\ \alpha_0 x_1 (x_1^2 - x_0^2) = 0. \end{cases}$$

De la dernière équation on tire :

[3] Carl Friedrich Gauss, allemand, 1777-1855

- *soit* $\alpha_1 = 0,\ \alpha_0 = 2$ *et* $x_0 = 0$ *ce qui est impossible puisque* $\alpha_0 x_0^2 + \alpha_1 x_1^2 = \frac{3}{2}$,
- *soit* $x_1 = 0,\ \alpha_0 = 0$ *ou* $x_0 = 0$ *ce qui est impossible,*
- *soit* $x_1 = -x_0$ *et dans ce cas, on obtient* $\alpha_0 = \alpha_1 = 1$ *et* $x_0^2 = \frac{1}{3}$.

D'où la formule de quadrature :

$$\int_{-1}^{1} f(x)\,dx = f(-\frac{1}{\sqrt{3}}) + f(\frac{1}{\sqrt{3}}) + R(f).$$

Remarque 4.6. *Il existe des formules de type Gauss où les extrémités de l'intervalle sont des points d'intégration :*

Formule de Gauss-Radau de degré de précision 2 :

$$\int_{-1}^{1} f(x)\,dx = \frac{1}{2}f(-1) + \frac{3}{2}f(\frac{1}{3}) + R(f).$$

Formule de Gauss-Labatto de degré de précision 3 :

$$\int_{-1}^{1} f(x)\,dx = \frac{1}{3}f(-1) + \frac{4}{3}f(0) + \frac{1}{3}f(1) + R(f).$$

Formule de Gauss-Labatto de degré de précision 5 :

$$\int_{-1}^{1} f(x)\,dx = \frac{1}{6}f(-1) + \frac{5}{6}f(-\frac{1}{\sqrt{5}}) + \frac{5}{6}f(-\frac{1}{\sqrt{5}}) + \frac{1}{6}f(1) + R(f).$$

4.4 Complément du cours

Exercice :

On se donne une fonction continue $f : [a,b] \times [c,d] \to \mathbb{R}$. *Approcher par la méthode des trapèzes*

$$I(f) = \int_a^b \int_c^d f(x,y)\,dy\,dx.$$

Solution :

Supposons que $I(f)$ *peut être écrit sous la forme*

$$I(f) = \int_a^b \left(\int_c^d f(x,y)\,dy \right) dx.$$

Posons $g(x) = \int_c^d f(x,y)\,dy$ *et divisons, par analogie au cas d'une seule variable, l'intervalle* $[a,b]$ *en* n *parties égales et l'intervalle* $[c,d]$ *en* m *parties égales. Ceci induit deux pas de discrétisation* $h = \frac{b-a}{n}$ *et* $k = \frac{d-c}{m}$, *alors*

$$I(f) = \int_a^b g(x)\,dx \simeq \frac{h}{2}\left(g(a) + 2\sum_{i=1}^{n-1} g(x_i) + g(b) \right). \tag{4.2}$$

D'autre part, on applique la formule des trapèzes pour approcher $g(x) = \int_c^d f(x,y)\,dy$ *avec* x *constante et* y *variable, on obtient*

$$g(x) \simeq \frac{k}{2}\left(f(x,c) + 2\sum_{j=1}^{m-1} f(x,y_j) + f(x,d)\right). \tag{4.3}$$

En substituant (4.3) dans (4.2), on aura

$$\begin{aligned} I(f) &\simeq \tfrac{h}{2}[\tfrac{k}{2}(f(a,c)+2\sum_{j=1}^{m-1} f(a,y_j)+f(a,d)) \\ &\quad +2\sum_{i=1}^{n-1}\tfrac{k}{2}(f(x_i,c)+2\sum_{j=1}^{m-1} f(x_i,y_j)+f(x_i,d)) \\ &\quad +\tfrac{k}{2}(f(b,c)+2\sum_{j=1}^{m-1} f(b,y_j)+f(b,d))] \\ &= \tfrac{hk}{4}[f(a,c)+f(a,d)+f(b,c)+f(b,d)+2(\sum_{j=1}^{m-1} f(a,y_j)+\sum_{j=1}^{m-1} f(b,y_j)) \\ &\quad +2(\sum_{i=1}^{n-1} f(x_i,c)+\sum_{i=1}^{n-1} f(x_i,d))+4\sum_{i=1}^{n-1}\sum_{j=1}^{m-1}\alpha_{ij}\,f(x_i,y_j)] \\ &= \sum_{i=0}^{n}\sum_{j=0}^{m}\alpha_{ij}\,f(x_i,y_j) \end{aligned}$$

où

$$x_i = a+ih,\ x_n = b \quad i=\overline{0,n}, \quad y_j = c+jk,\ y_m = d \quad j=\overline{0,m}$$

$$\alpha_{00} = \alpha_{n0} = \alpha_{m0} = \alpha_{nm} = \tfrac{hk}{4}$$

$$\alpha_{0j} = \alpha_{nj} = \alpha_{i0} = \alpha_{im} = \tfrac{hk}{2}, \quad i=\overline{1,n-1}\ \ j=\overline{1,m-1}$$

$$\alpha_{ij} = hk, \quad i=\overline{1,n-1}\ \ j=\overline{1,m-1}$$

Remarque 4.7. *Ce procédé peut être adapté au calcul numérique d'intégrales triples ou multiples.*

4.5 Conclusion

1. *Les formules des trapèzes et de Simpson sont les plus utilisées et souvent sous leurs formes composites.*
2. *Les formules de Gauss sont néanmoins plus précises et peuvent de la même façon être composées.*
3. *L'intégration numérique est aussi très utile pour calculer numériquement des intégrales doubles :*

$$\int_a^b \int_c^d f(x,y)\,dx\,dy.$$

4.6 Exercices

Exercice 4.3. *Calculer $Arctg(3)$ par les méthodes d'intégration des trapèzes et de Simpson pour $n=6$.*
Indication : *$Arctg(x) = \int_0^x \frac{dt}{1+t^2}$.*

Exercice 4.4. *On lance une fusée verticalement du sol et l'on mesure pendant les premières 80 secondes l'accéleration γ :*

t (en s)	*0*	*10*	*20*	*30*	*40*	*50*	*60*	*70*	*80*
γ(en m/s^2)	*30*	*31.63*	*33.44*	*35.47*	*37.75*	*40.33*	*43.29*	*46.70*	*50.67*

Calculer la vitesse V de la fusée à l'instant $t = 80s$, par la méthode des trapèzes puis par Simpson.

Exercice 4.5.

1. *Déterminer le nombre de subdivisions nécessaires de l'intervalle d'intégration $[-\pi, \pi]$ pour évaluer à 0.5×10^{-3} près, par la méthode de Simpson, l'intégrale*
$$\int_{-\pi}^{\pi} cosx\, dx.$$
2. *Déterminer le nombre de subdivisions nécessaires des intervalles d'intégration pour évaluer à $0,5 \times 10^{-6}$ près, par la méthode des trapèzes les intégrales :*
$$i)\ \int_0^1 \frac{dx}{1+e^x}, \qquad ii)\ \int_0^1 \frac{dx}{1+\sin x}.$$

Exercice 4.6.

1. *Construire un polynôme P qui passe par les points $(0,0)$, $(1,0)$ et $(-1,2)$.*
2. *Prendre n subdivisions sur $[0,1]$ et approcher $\int_0^1 P(x)\,dx$ par la méthode :*
   a) *Des trapèzes ($n = 3$); évaluer ensuite l'erreur relative commise.*
   b) *De Simpson ($n = 30$).*

Exercice 4.7. *Soit f une fonction continue donnée sur l'intervalle $[-1,1]$. Notons par P le polynôme de degré deux qui interpole f en les points $-1, 0$ et 1.*

1. *Exprimer $\int_{-1}^1 P(t)\,dt$ en fonction de $f(-1), f(0)$ et $f(1)$.*
2. *Vérifier que l'expression obtenue coïncide avec une formule d'intégration numérique dont on donnera le nom et la valeur du pas de discrétisation.*

Exercice 4.8. *Soit f une fonction réelle continue et définie sur l'intervalle $[a,b]$. On pose $h = \frac{b-a}{n}$ pour $n = 3$, $x_0 = a$, $x_3 = b$, $x_i = x_0 + ih$ ($i = \overline{0,3}$). On considère la formule de quadrature :*
$$I(f) = \int_a^b f(t)\,dt \simeq \alpha f(x_1) + \beta f(x_2).$$

1. *Déterminer les deux coefficients α et β sachant que la méthode est exacte sur $\mathbb{P}_1$.*
2. *Généraliser l'écriture précédente au cas où $n = 3m$.*
3. *Evaluer l'erreur relative commise en utilisant cette méthode pour approcher $I(\frac{\ln(x)}{x})$ sur $[1,2]$ en prenant $n = 3$ subdivisions.*

Exercice 4.9. *Soit la formule d'intégration numérique suivante :*

$$\int_0^1 f(x)\,dx = \left[af(0) + bf(\frac{1}{3}) + cf(\frac{2}{3}) + df(1)\right] + R(f).$$

1. *Déterminer a, b, c et d de sorte que $R(f) = 0$ quelle que soit f appartenant à l'espace vectoriel des polynômes de degré inférieur ou égal à 3.*
2. *Trouver p le degré de précision de la formule proposée.*
3. *En admettant que l'erreur d'approximation soit de la forme :*

$$R(f) = K f^{(p+1)}(\xi), \quad \xi \in [0, 1],$$

déterminer la constante K.

Exercice 4.10.

1. *Déterminer le polynôme d'interpolation de Lagrange P d'une fonction f construite sur les points :*

$$-1,\ -\frac{1}{3},\ \frac{1}{3},\ 1.$$

2. *Par intégration du polynôme obtenu, en déduire la formule d'intégration approchée suivante :*

$$\int_{-1}^1 f(x)\,dx \approx \frac{1}{4}f(-1) + \frac{3}{4}f(-\frac{1}{3}) + \frac{3}{4}f(\frac{1}{3}) + \frac{1}{4}f(1).$$

3. *En déduire la formule d'intégration approchée sur l'intervalle $[a, b]$.*

   Indication : utiliser le changement de variable $y = \frac{a+b}{2} + \frac{b-a}{2}x$.
4. *Application : Calculer, à l'aide de la formule de quadrature proposée, les intégrales*

$$\int_{-1}^1 e^x\,dx, \quad \int_{-5}^3 e^x\,dx.$$

   Calculer l' erreur commise dans chaque cas.

Exercice 4.11. ***(Méthode du point mileu)***
Soit $f : [a, b] \to \mathbb{R}$ une fonction continue. Afin d'approcher son intégrale sur $[a, b]$ en $(n = 2m)$ parties égales de longueur h. On pose $x_0 = a$, $x_n = b\, x_i = x_0 + ih$, $y_i = f(x_i)$ $(i = \overline{0, n})$.
La méthode proposée consiste à remplacer, pour $(i = 0, 2, 4, ..., n-2)$, l'intégrale $\int_{x_i}^{x_{i+2}} f(x)\,dx$ par l'aire du rectangle de côtés $f(x_{i+1})$ et $(x_{i+2} - x_i)$.

1. *Trouver les $(n+1)$ coefficients réels $(a_i = a_i(h))$ tels que*

$$I = \int_a^b f(x)\,dx \simeq \sum_{i=0}^{n} a_i\, y_i = V(h)$$

2. *Si f est suffisamment continûment dérivable, déterminer l'expression de l'erreur d'intégration $R(h) = I - V(h)$*
3. *Dans quel(s)cas la méthode est-elle exacte ?*
4. *Trouver $\lim R(h)$ quand h tend vers 0.*

5. *Quel est le nombre minimal de subdivisions à effectuer sur l'intervalle* $[1,2]$ *pour que l'erreur absolue commise sur l'intégrale de* $f(x)=\frac{1}{x}$ *ne dépasse pas* 2×10^{-4}.

Exercice 4.12. ***(Méthode des réctangles)***
On considère lintégrale $\int_a^b f(x)dx$. *On se donne une subdivision* $x_0, x_1, \ldots, x_n$ *de l'intervalle* $[a,b]$ *tel que* $x_i = a+ih$ *où* $h=\frac{b-a}{n}$. *Dans la méthode des rectangles, on remplace la fonction* f *par une fonction constante par morceaux. Soit* g *la fonction définie par*

$$g(x)=f(x_i) \quad pour\ x\in[x_i,x_{i+1}].$$

1. *Posons* $S=\int_a^b g(x)dx$. *Montrer que* $S=h\sum_{i=0}^{n-1} f(a+ih)$.
2. *On suppose que la fonction* f *est continue sur lintervalle* $[a,b]$ *et continûment dérivable sur lintervalle ouvert* $]a,b[$. *On pose*

$$\phi(h)=\int_\alpha^{\alpha+h} f(x)dx - hf(\alpha) \quad où\ a\le\alpha<\alpha+h\le b.$$

   Montrer quil existe un nombre $c\in]0,h[$ *tel que* $\phi(h)=\frac{h^2}{2}f'(c+\alpha)$.
3. *On suppose que* $|f'(x)|\le k,\ \forall\, x\in[a,b]$. *Montrer que*

$$|\int_a^b f(x)dx - S|\le k\frac{(b-a)^2}{2n}.$$

Exercice 4.13. *Soit* $f\in C^1([0,1],\mathbb{R})$. *On considère la formule de quadrature*

$$\int_0^1 f(x)\,dx \approx \alpha_0 f(0)+\alpha_1 f'(0)+\alpha_2 f(c)$$

où $c\in]0,1]$ *et* $\alpha_i\in\mathbb{R},\ i=0,1,2$.

1. *Calculer* $\alpha_0,\alpha_1,\alpha_2$ *et* c *sachant que la formule de quadrature soit exacte pour tout polynôme de degré* ≤ 3.
2. *Construire à partir de cette formule de quadrature sur* $[0,1]$ *une formule de quadrature sur* $[a,b]$.
3. *On subdivise* $[a,b]$ *en* n *parties égales de longeur* h. *Généraliser la formule de quadrature obtenue sur* $[a,b]$
4. *Application numérique.* $f(x)=e^x$. *Calculer* $\int_0^1 e^x\,dx$ *par la formule de quadrature. Estimer son erreur.*

Exercice 4.14. *Soient* a *et* h *deux réels donnés,* $h>0$. *On pose*

$$E(f)=\int_a^{a+h} f(x)dx - \alpha h f(a) - \beta h f(a+\tau h).$$

1. *Déterminer les valeurs de* α,β *et* τ *de sorte que l'on ait*

$$\forall\, p\in\mathbb{P}_2,\ \ E(p)=0.$$

2. *Supposons que* $\alpha=\frac{1}{4},\ \beta=\frac{3}{4},\ \tau=\frac{2}{3}$.
3. *Trouver l'ordre de précision de la méthode proposée.*

   a) *Montrer que pour toute fonction* $f\in C^3(\mathbb{R})$ *il existe* $\theta\in]0,1[$ *tel que*

$$E(f)=K\,h^4 f^{(3)}(a+\theta h) \quad où\ \ K\ est\ une\ constante\ indépendante\ de\ f$$

   Indication : *on pourra considérer l'application* $R:\ [0,h]\to\mathbb{R}$ *définie par :* $R(h)=E(f)$ *et on remarque que* $R(0)=R'(0)=0$.

   b) *Calculer* K *puis l'erreur relative commise lorsque* $f(x)=(x-a)^3$.

Chapitre 5

Dérivation numérique

5.1 Position du problème

Comme pour l'intégrale, on voudrait être en mesure d'évaluer numériquement la dérivée d'une fonction lourde à manipuler ou qui n'est connue que en un certain nombre de points. Ce problème de dérivation numérique est très commun en ingénierie et en analyse numérique (c'est la base des méthodes de différences finies).

On considère une fonction $f : [a, b] \longrightarrow \mathbb{R}$ *de classe suffisamment élevée,* $x \in]a, b[$ *fixé.*
On veut approcher (au mieux !) les dérivées de la fonction f *au point* x.
Or

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h},$$

avec $h > 0$ *(petit), on obtient*

$$f'(x) \simeq \frac{f(x+h) - f(x)}{h}.$$

Nous constatons que cette approximation de $f'(x)$ *fait intervenir la valeur de* f *en* x *et* $x + h$, *ce qui nous conduit à approcher les nombres* $f'(x), f''(x), \ldots, f^{(n)}(x)$ *en utilisant un ensemble discret de points.*

Une des méthodes les plus anciennes utilisées pour obtenir des formules de dérivation numérique consiste à construire des quotients différentiels à l'aide des développements de Taylor[1].

5.2 Approximation de la dérivée première

Fixons $h > 0$ *(petit) et introduisons les notations :*

$$\begin{aligned} \Delta_h f(x) &= f(x+h) - f(x), \\ \nabla_h f(x) &= f(x) - f(x-h), \\ \delta_{2h} f(x) &= f(x+h) - f(x-h). \end{aligned}$$

5.2.1 Formules à deux points

Effectuons un premier développement de Taylor d'ordre 1 de f *autour de* x :

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(\xi),\ \xi \in [x, x+h].$$

[1] Brook Taylor, anglais, 1685-1731

On obtient

$$\begin{cases} f'(x) \simeq f'_{hd}(x) = \frac{f(x+h)-f(x)}{h} = \frac{\Delta f(x)}{h}, \text{ c'est la formule de différences finies progressives } (\textbf{DFP}) \\ E = -\frac{h}{2}f''(\xi),\ \xi \in [x, x+h] \text{ c'est l'erreur commise} \end{cases}$$

Effectuons un deuxième développement de Taylor d'ordre 1 de f autour de x :

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(\xi),\ \xi \in [x-h, x].$$

On obtient

$$\begin{cases} f'(x) \simeq f'_{hg}(x) = \frac{f(x)-f(x-h)}{h} = \frac{\nabla f(x)}{h}, \text{ c'est la formule de différences finies régressive } (\textbf{DFR}), \\ E = \frac{h}{2}f''(\xi),\ \xi \in [x-h, x] \text{ c'est l'erreur commise.} \end{cases}$$

Les formules obtenues sont donc deux approximations d'ordre 1 de la dérivée première et l'erreur à chaque fois tend vers 0 quand h tend vers 0.

Augmentons l'ordre du développement de Taylor de f autour de x :

$$\begin{aligned} f(x+h) &= f(x) + hf'(x) + \tfrac{h^2}{2}f''(x) + \tfrac{h^3}{6}f'''(\xi_1),\ \xi_1 \in [x, x+h], \\ f(x-h) &= f(x) - hf'(x) + \tfrac{h^2}{2}f''(x) - \tfrac{h^3}{6}f'''(\xi_2),\ \xi_2 \in [x-h, x]. \end{aligned}$$

En soustrayant membre à membre, on obtient

$$f(x+h) - f(x-h) = 2hf'(x) + \frac{h^3}{6}(f'''(\xi_1) + f'''(\xi_2)).$$

Ce qui donne

$$\begin{cases} f'(x) \simeq f'_{hc}(x) = \frac{f(x+h)-f(x-h)}{2h} = \frac{\delta_{2h}f(x)}{2h}, \text{ c'est la formule de différences finies centrées } (\textbf{DFC}), \\ E = -\frac{h^2}{12}(f'''(\xi_1) + f'''(\xi_2)) = -\frac{h^2}{6}(f'''(\xi),\ \xi \in [x-h, x+h], \text{ c'est l'erreur commise.} \end{cases}$$

La formule centrée (symétrique) est une formule d'approximation d'ordre 2 et donc plus précise que les deux premières formules, même si elle nécessite la connaissance de f au même nombre de points. Il faut cependant noter que les points utilisés sont disposés symétriquement par rapport à celui où l'on calcule la dérivée.

Remarque 5.1. *Supposons que l'intervalle $[a, b]$ est découpé en N intervalles, on pose $h = \frac{b-a}{N}$ et on introduit* ***les points de grille*** *x_i de sorte que $x_i = a + ih, i = 0, \ldots, N$. Supposons que f est connue uniquement en les points de grille x_i. Alors pour approcher $f'(x_i)$, $i = 0, \ldots, N$:*

1. *La formule de différences finies progressive donne :*

$$f'(x_i) \simeq f'_{hd}(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h},\ i = 0, \ldots, N-1. \tag{5.1}$$

On remarque que la relation (5.1) n'est pas définie pour $i = N$, on perd donc le dernier point de grille quand on utilise cette relation pour approcher $f'(x_i)$.

2. *La formule de différences finies régressive donne :*

$$f'(x_i) \simeq f'_{hg}(x_i) = \frac{f(x_i) - f(x_{i-1})}{h},\ i = 1, \ldots, N. \tag{5.2}$$

On remarque que la relation (5.2) n'est pas définie pour $i = 0$, on perd donc le premier point de grille quand on utilise cette relation pour approcher $f'(x_i)$.

3. *La formule de différences finies centrées donne :*

$$f'(x_i) \simeq f'_{hc}(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h},\ i = 1, \ldots, N-1. \tag{5.3}$$

On remarque que la relation (5.3) n'est pas définie pour $i = 0$ et $i = N$, on perd donc le premier et le dernier point de grille quand on utilise cette relation pour approcher $f'(x_i)$. Constatons aussi que

$$f'_{hc}(x_i) = \frac{f'_{hg}(x_i) + f'_{hd}(x_i)}{2},\ i = 1, \ldots, N-1.$$

Exemple 5.1. *Pour illustrer les trois formules précédentes, considérons la fonction $x \mapsto f(x) = 2^x,; x \in [1,5]$ passant par les points $(x_0, y_0) = (1,2)$, $(x_1, y_1) = (2,4)$, $(x_2, y_2) = (3,8)$, $(x_3, y_3) = (4,16)$ et $(x_4, y_4) = (5,32)$.*

Nous voulons approcher le nombre $f'(x_2)$:

1. *La formule de DFP : $f'(x_2) \simeq f'_{hd}(x_2) = \frac{f(x_3)-f(x_2)}{h} = y_3 - y_2 = 8$.*
2. *La formule de DFR :$f'(x_2) \simeq f'_{hg}(x_2) = \frac{f(x_2)-f(x_1)}{h} = y_2 - y_1 = 4$.*
3. *La formule de DFC :$f'(x_2) \simeq f'_{hc}(x_2) = \frac{f(x_3)-f(x_1)}{2h} = \frac{y_3-y_1}{2} = 6$.*

Evaluons les erreurs d'approximation sachant que $f'(x) = \ln 2\ 2^x, x \in [1,5]$:

$$\begin{aligned} E_1 &= |f'(x_2) - f'_{hd}(x_2)| = |8\ln 2 - 8| \simeq 2,454. \\ E_2 &= |f'(x_2) - f'_{hg}(x_2)| = |8\ln 2 - 4| \simeq 1,545. \\ E_3 &= |f'(x_2) - f'_{hc}(x_2)| = |8\ln 2 - 6| \simeq 0,454. \end{aligned}$$

D'où, la meilleure formule pour approcher $f'(3)$ est celle de DFC.

5.2.2 Formules à trois points

Il est possible de développer d'autres formules, pour cela il suffit d'effectuer un développement de Taylor de f autour de x avec un pas $2h$ par exemple :

$$f(x+2h) = f(x) + 2hf'(x) + 2h^2 f''(x) + \frac{4h^3}{3} f'''(\xi_1),\ \xi_1 \in [x, x+2h].$$

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2} f''(x) + \frac{h^3}{6} f'''(\xi_2),\ \xi_2 \in [x, x+h].$$

En combinant ces deux équations de façon à faire disparaître la dérivée seconde, on obtient

$$\begin{cases} f'_{hd}(x) &= \frac{4f(x+h)-3f(x)-f(x+2h)}{2h} \text{ est une approximation de la dérivée première de } f \text{ en } x; \\ E &= \frac{h^2}{3} f''(\xi),\ \xi \in [x, x+2h] \text{ c'est l'erreur commise.} \end{cases}$$

5.2.3 Approximation de la dérivée seconde

Pour obtenir une approximation de la dérivée seconde, nous procédons de la même façon, mais à partir de développements de Taylor d'ordre 4 :

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(\xi_1),\ \xi_1 \in [x, x+h].$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(\xi_2),\ \xi_2 \in [x-h, x].$$

Après addition, on obtient

$$\begin{cases} f''_{hd}(x) &= \frac{f(x+h)-2f(x)+f(x-h)}{h^2} = \frac{\delta_h^2 f(x)}{h^2} \text{ est une approximation de la dérivée seconde de } f \text{ en } x; \\ E &= -\frac{h^2}{12}f^{(4)}(\xi),\ \xi \in [x-h, x+h] \text{ c'est l'erreur commise.} \end{cases}$$

Cette formule est donc d'ordre 2 (le terme d'erreur tend vers 0 quand h tend vers 0) et est très importante dans la pratique.

5.2.4 Approximation des dérivées d'ordre supérieur

On peut évidemment généraliser cette approche et déterminer des approximations des dérivées d'ordre supérieur.

On a déja vu que $\frac{\Delta_h f}{h}$, $\frac{\nabla_h f}{h}$ et $\frac{\delta_{2h} f}{2h}$ approximent la dérivée première de f (en x) avec une précision proportionnelle à h, h et h^2 respectivement. Ces trois opérateurs sont linéaires par rapport à f et sont appelés ***opérateurs aux différences finies****. Pour $n \in \mathbb{N}^*$, on peut généraliser ces concepts d'opérateurs pour approximer la dérivée d'ordre n de f en un point $x \in]a, b[$. Pour cela, on va définir récursivement :*

$$\Delta_h^n f(x) = \Delta_h(\Delta_h^{n-1} f)(x),\ \ \nabla_h^n f(x) = \nabla_h(\nabla_h^{n-1} f)(x),\ \ \delta_h^n f(x) = \delta_h(\delta_h^{n-1} f)(x).$$

De manière analogue au cas où $n = 1$, on peut montrer que

$$\frac{\Delta_h^n f}{h^n}, \quad \frac{\nabla_h^n f}{h^n} \quad \text{et} \quad \frac{\delta_h^n f}{h^n}$$

sont des approximations de $f^{(n)}$ (en x) avec une erreur proportionnelle à h, h et h^2, respectivement, dès que f est de classe C^{n+1}, C^{n+1} et C^{n+2}, respectivement.

Remarque 5.2. *La dérivation numérique est une opération très instable, c'est à dire très sensible aux erreurs d'arrondi (soustraction entre termes voisins). Prenons par exemple*

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + O(h^2),$$

où $f(x-h) = f^*(x-h) \pm e_1$ et $f(x+h) = f^*(x+h) \pm e_2$, alors

$$f'(x) = \frac{f^*(x+h) - f^*(x-h)}{2h} \pm \frac{e_2 + e_1}{2h} + O(h^2).$$

Si le pas h est trop réduit il y'aura beaucoup d'erreurs d'arrondi.

5.3 Exercices

Exercice :

Soit $f : [a,b] \longrightarrow \mathbb{R}$ une fonction de classe au moins C^5 sur l'intervalle $[a,b]$. On se fixe un nombre $(h>0)$ "petit" ainsi qu'un point quelconque $x \in]a,b[$. Soit le rapport

$$A = \frac{f(x+3h) - 3f(x+h) + 3f(x-h) - f(x-3h)}{8h^3}.$$

1. *Montrer que le rapport A approche une dérivée de f que l'on déterminera. Donner l'ordre de précision de cette approximation.*
2. *Vérifier que A coïncide avec une formule de dérivation numérique que l'on donnera.*

Solution :

1. *Effectuons les développements de Taylor de f au voisinage de x jusqu'à l'ordre 4 avec un reste en $O(h^5)$:*

$$\begin{aligned}
f(x+h) &= f(x) + hf'(x) + \tfrac{h^2}{2}f''(x) + \tfrac{h^3}{6}f'''(x) + \tfrac{h^4}{24}f^{(4)}(x) + O(h^5), && (1)\\
f(x-h) &= f(x) - hf'(x) + \tfrac{h^2}{2}f''(x) - \tfrac{h^3}{6}f'''(x) + \tfrac{h^4}{24}f^{(4)}(x) + O(h^5), && (2)\\
f(x+3h) &= f(x) + 3hf'(x) + \tfrac{9h^2}{2}f''(x) + \tfrac{9h^3}{2}f'''(x) + \tfrac{9h^4}{8}f^{(4)}(x) + O(h^5), && (3)\\
f(x-3h) &= f(x) - 3hf'(x) + \tfrac{9h^2}{2}f''(x) - \tfrac{9h^3}{2}f'''(x) + \tfrac{9h^4}{8}f^{(4)}(x) + O(h^5). && (4)
\end{aligned}$$

   Le numérateur du rapport A s'écrit alors :

$$(3) - 3\times(1) + 3\times(2) - (4) = 8h^3 f'''(x) + O(h^5).$$

   Autrement dit, $A = f'''(x) + O(h^2)$.
   D'où, A approxime la dérivée troisième de f en x, avec une erreur en $O(h^2)$.
2. *On a $\delta^3 f(x) = f(x+3h) - 3f(x+h) + 3f(x-h) - f(x-3h)$. Alors*

$$A = \frac{\delta^3_{2h} f(x)}{(2h)^3}.$$

   D'où, le rapport R n'est autre que la formule des différences finies centrées (symétriques) d'ordre 3 qui approxime $f'''(x)$ par $\frac{\delta^3_{2h} f(x)}{(2h)^3}$ avec une erreur en $O(h^2)$.

Exercice :

Soit $f : [a,b] \longrightarrow \mathbb{R}$ une fonction de classe au moins C^6 sur l'intervalle $[a,b]$. On se fixe un nombre $(h>0)$ "petit" ainsi qu'un point quelconque $x \in]a,b[$.

1. *Montrer que le rapport*

$$B = \frac{-2f(x+2h) + 32f(x+h) - 60f(x) + 32f(x-h) - 2f(x-2h)}{24h^2}$$

   approche une dérivée de f que l'on déterminera. Donner l'ordre de précision de cette approximation.
2. *Commenter les résultats obtenus.*

Solution :

1. *Effectuons les développements de Taylor de f au voisinage de x jusqu'à l'ordre 5 avec un reste en $O(h^6)$:*

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(x) + \frac{h^5}{120}f^{(5)}(x) + O(h^6), \quad (1)$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(x) - \frac{h^5}{120}f^{(5)}(x) + O(h^6), \quad (2)$$

$$f(x+2h) = f(x) + 2hf'(x) + 2h^2f''(x) + \frac{4h^3}{3}f'''(x) + \frac{2h^4}{3}f^{(4)}(x) + \frac{4h^5}{15}f^{(5)}(x) + O(h^6), \quad (3)$$

$$f(x-2h) = f(x) - 2hf'(x) + 2h^2f''(x) - \frac{4h^3}{3}f'''(x) + \frac{2h^4}{3}f^{(4)}(x) - \frac{4h^5}{15}f^{(5)}(x) + O(h^6). \quad (4)$$

Le numérateur du rapport B s'écrit alors :

$$32 \times [(1)+(2)] - 2 \times [(3)+(4)] - 60f(x) = 24h^2f''(x) + O(h^6).$$

Autrement dit, $B = f''(x) + O(h^4)$.
D'où, B approxime la dérivée seconde de f en x, avec une erreur en $O(h^4)$.

2. *Rappelons que $\frac{\Delta_h^n f}{h^n}$, $\frac{\nabla_h^n f}{h^n}$ et $\frac{\delta_h^n f}{(h)^n}$ approximent f'' en x avec une erreur proportionnelle à h, h et h^2 (respectivement). Donc le rapport B est beaucoup plus précis.*

Exercice :

Soit $f : [a,b] \longrightarrow \mathbb{R}$ une fonction de classe au moins C^6 sur l'intervalle $[a,b]$. On se fixe un nombre $(h > 0)$ "petit" ainsi qu'un point quelconque $x \in]a,b[$.

1. *Montrer que le rapport*

$$C = \frac{f(x+2h) - 4f(x+h) + 6f(x) - 4f(x-h) + f(x-2h)}{h^4}$$

approche une dérivée de f que l'on déterminera. Donner l'ordre de précision de cette approximation.

Solution :

1. *Effectuons les développements de Taylor de f au voisinage de x jusqu'à l'ordre 5 avec un reste en $O(h^6)$:*

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(x) + \frac{h^5}{120}f^{(5)}(x) + O(h^6), \quad (1)$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(x) + \frac{h^4}{24}f^{(4)}(x) - \frac{h^5}{120}f^{(5)}(x) + O(h^6), \quad (2)$$

$$f(x+2h) = f(x) + 2hf'(x) + 2h^2f''(x) + \frac{4h^3}{3}f'''(x) + \frac{2h^4}{3}f^{(4)}(x) + \frac{4h^5}{15}f^{(5)}(x) + O(h^6), \quad (3)$$

$$f(x-2h) = f(x) - 2hf'(x) + 2h^2f''(x) - \frac{4h^3}{3}f'''(x) + \frac{2h^4}{3}f^{(4)}(x) - \frac{4h^5}{15}f^{(5)}(x) + O(h^6). \quad (4)$$

En combinant ces quatre équations, on obtient

$$[(3)+(4)] - 4 \times [(1)+(2)] + 6f(x) = h^4f^{(4)}(x) + O(h^6).$$

Donc, $C = f^{(4)}(x) + O(h^2)$.
D'où, C est une approximation de la dérivée seconde de f en x, avec une précision en $O(h^2)$.

Deuxième partie

Analyse Numérique II

Chapitre 6

Résolution des systèmes linéaires

6.1 Position du problème

On appelle système linéaire d'ordre n, $(n \in \mathbb{N}^*)$, une expression de la forme

$$AX = b, \tag{6.1}$$

où $A = (a_{ij})$, $1 \leq i, j \leq n$, désigne une matrice carrée d'ordre n de nombres réels ou complexes, $b = (b_i)$, $1 \leq i \leq n$, un vecteur colonne réel ou complexe et $X = (x_i)$, $1 \leq i \leq n$, est le vecteur des inconnues du système. La relation (6.1) équivaut aux équations

$$\sum_{j=1}^{n} a_{ij} x_j = b_i, \quad i = 1, \ldots, n.$$

La matrice A est dite régulière (inversible) si $det(A) \neq 0$; on a existence et unicité de la solution X si et seulement si la matrice A est inversible.

On cherche à résoudre le système linéaire (6.1).
Théoriquement, si A est inversible, la solution du système $AX = b$ est donnée par la formule de Cramer[1] *:*

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad i = 1, \ldots, n,$$

où A_i est une matrice obtenue à partir de A en remplaçant la $i^{ième}$ colonne de A par le vecteur b. Cependant l'application de cette formule est inacceptable pour la résolution pratique des systèmes, car son coût (ou nombre d'opérations) est en $O((n+1)!)$. Par exemple, sur un ordinateur effectuant 10^9 opérations par seconde il faudrait au moins 10^{47} années pour résoudre un système linéaire de seulement 50 équations.

Il faut donc développer des algorithmes alternatifs avec un coût raisonnable. Ce problème est un des plus importants de l'analyse numérique.
Dans les sections suivantes plusieurs méthodes sont analysées. Ces méthodes se divisent en deux catégories :

a) Méthodes directes : *Ce sont des méthodes qui permettent d'obtenir la solution X de (6.1), si l'ordinateur faisait des calculs exacts, en un nombre fini (en relation avec n) d'opérations élémentaires.*

b) Méthodes itératives : *Ce sont des méthodes qui consistent à construire une suite de vecteurs $X^{(n)}$ convergeant vers la solution X.*

[1] Gabriel Cramer, 1704-1752

Dans toute la suite de ce chapitre, on suppose que A est une matrice inversible de $M_n(\mathbb{R})$, où $M_n(\mathbb{R})$ est l'ensemble des matrices carrée d'ordre n à coefficients réels.

Remarque 6.1. *Le fait que les coefficients de A soient réels n'est pas restrictif. En effet ; supposons que l'on veut résoudre (6.1) dans $\mathbb{C}$, c'est-à-dire*

$$(A+iB)(X+iY) = b+ic, \quad \text{où } A,B \in M_n(\mathbb{R}) \text{ and } X,Y,a,b \in \mathbb{R}^n.$$

On peut le réécrire sous la forme :

$$\begin{pmatrix} A & -B \\ B & A \end{pmatrix} \begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} b \\ c \end{pmatrix}$$

ou encore

$$A'X' = b' \tag{6.2}$$

avec (6.2) est un système réel de $(2n)$ équations à $(2n)$ inconnues.

6.2 Méthodes directes

6.2.1 Systèmes particuliers

Système diagonal

Soit $A = (a_{ij})_{1\le i,j\le n}$ où $a_{ij} = 0$, si $i \neq j$ et $a_{ii} \neq 0 \; \forall i = \overline{1,n}$.
Posons $X = (x_1, x_2, ..., x_n)^t$ et $b = (b_1, b_2, ..., b_n)^t$. Alors

$$AX = b \Leftrightarrow \begin{cases} a_{11}x_1 = b_1 \\ a_{22}x_2 = b_2 \\ \quad \cdot \\ \quad \cdot \\ \quad \cdot \\ a_{nn}x_n = b_n \end{cases} \Longrightarrow x_i = \frac{b_i}{a_{ii}}, i = \overline{1,n},$$

donc la solution est $X = (x_1, x_2, ..., x_n)^t$ où $x_i = \frac{b_i}{a_{ii}}$ avec $i = \overline{1,n}$.

Coût : *n divisions.*

Système triangulaire supérieur

Soit $A = (a_{ij})_{1\le i,j\le n}$ où $a_{ij} = 0$ si $i > j$ et $a_{ii} \neq 0 \; \forall i = \overline{1,n}$. Alors

$$AX = b \Leftrightarrow \begin{cases} a_{11}x_1 + a_{12}x_2 + ... + a_{1n}x_1 = b_1 \\ \qquad\qquad a_{22}x_2 + ... + a_{2n}x_n = b_2 \\ \qquad\qquad\qquad \ddots \qquad\qquad \vdots \\ \qquad\qquad\qquad\qquad \ddots \qquad \vdots \\ \qquad\qquad\qquad\qquad\qquad a_{nn}x_n = b_n \end{cases}$$

donc le système $AX = b$ se résout par la méthode ascendante, c'est-à-dire on trouve d'abord x_n puis $x_{n-1}, ...,$ puis x_1; d'où les relations :

$$\begin{cases} x_n = \frac{b_n}{a_{nn}}, \\ \\ x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right), \; i = n-1, ..., 1. \end{cases}$$

Coût : $\frac{n(n-1)}{2}$ *additions (ou soustractions)+* $\frac{n(n-1)}{2}$ *multiplications+*n *divisions=*n^2 *opérations ; en effet :*

- *Le nombre de divisions étant évident.*
- *Pour calculer* $x_i (i = 1, ..., n)$*, on fait* $(n - i)$ *additions et* $(n - i)$ *multiplications, d'où*

$$coût\ (+) = \ coût\ (\times) = \sum_{i=1}^{n}(n-i) = n^2 - \frac{n(n+1)}{2} = \frac{n(n-1)}{2}\cdot$$

Système triangulaire inférieur

Soit $A = (a_{ij})_{1\leq i,j\leq n}$ *où* $a_{ij} = 0$ *si* $i < j$ *et* $a_{ii} \neq 0\ \forall\, i = \overline{1,n}$*. Alors*

$$AX = b \Leftrightarrow \begin{cases} a_{11}x_1 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \\ \vdots \qquad \ddots \\ \vdots \qquad\qquad \ddots \\ a_{n1}x_1 + a_{n2}x_2 + ... + a_{nn}x_n = b_n \end{cases}$$

donc le système $AX = b$ *se résout par la méthode descendante, c'est-à-dire on trouve d'abord* x_1 *puis* $x_2, ...,$ *puis* x_n*; d'où les relations :*

$$\begin{cases} x_1 = \frac{b_1}{a_{11}}, \\ \\ x_i = \frac{1}{a_{ii}}\ (b_i - \sum\limits_{j=1}^{i-1} a_{ij}x_j),\ i = 2, \ldots, n. \end{cases}$$

Coût : $\frac{n(n-1)}{2}$ *additions (ou soustractions),* $\frac{n(n-1)}{2}$ *multiplications et* n *divisions.*

Remarque 6.2. *Les coûts des trois méthodes précédentes sont valables (point de vue calcul sur machine) pour* n *pas assez grand (*≤ 100*), on s'efforcera donc -dans toute la suite concernant la catégorie des méthodes directes- de transformer la matrice* A *du système* $AX = b$ *afin de nous ramener au cas triangulaire (le plus souvent) et même au cas diagonal.*

6.2.2 Méthode d'élimination de Gauss

Comme les systèmes triangulaires sont faciles et économiques à résoudre, l'objectif est de transformer tout système linéaire en système triangulaire équivalent. Bien que cette méthode soit parmi les plus anciennes qui ont été proposées pour résoudre les systèmes linéaires elle reste encore actuellement la plus utilisée ; en particulier son algorithme est d'exploitation aisée sur micro-ordinateur.

Principe de la méthode :

Déterminer une matrice M *inversible telle que la matrice* MA *soit triangulaire supérieure. Alors*

$$AX = b \Leftrightarrow (MA)X = Mb,$$

ensuite résoudre le système triangulaire supérieur $(MA)X = Mb$ *par l'algorithme de remontée.*

Remarque 6.3. *En pratique on ne calcule pas M d'une façon explicite, mais par des transformations équivalentes on ramène le système de départ en un système à matrice triangulaire supérieure. Autrement dit*

$$(A,b) \overset{transformation}{\longrightarrow} (A^{(n)}, b^{(n)}),$$

où $A^{(n)}$ est une matrice triangulaire supérieure, puis on résout le système triangulaire.

$$A^{(n)}X = b^{(n)}.$$

Algorithme d'élimination de Gauss :

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

$$\left(A^{(1)} \vdots\, b^{(1)}\right) = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & \vdots & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & \vdots & b_2^{(1)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & \vdots & b_n^{(1)} \end{pmatrix} \begin{matrix} L_1^{(1)} \\ L_2^{(1)} \\ \vdots \\ \\ L_n^{(1)} \end{matrix}$$

A la $1^{ère}$ étape : *Si $a_{11}^{(1)} \neq 0$ (sinon on fait une permutation de lignes) on fait les affectations suivantes :*

$$\begin{cases} L_1^{(2)} \longleftarrow L_1^{(1)} \\ \\ L_i^{(2)} \longleftarrow L_i^{(1)} - \alpha_{i1} L_1^{(1)} \quad où\ \alpha_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}},\ \ 2 \leq i \leq n. \end{cases}$$

On obtient donc

$$\left(A^{(2)} \vdots\, b^{(2)}\right) = \begin{pmatrix} a_{11}^{(2)} & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} & \vdots & b_1^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & \vdots & b_2^{(2)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & \vdots & b_n^{(2)} \end{pmatrix} \begin{matrix} L_1^{(2)} \\ L_2^{(2)} \\ \vdots \\ \\ L_n^{(2)} \end{matrix}$$

où

$$\begin{cases} a_{1j}^{(2)} = a_{1j}^{(1)},\ 1 \leq j \leq n;\ \ b_1^{(2)} = b_1^{(2)}; \\ \\ a_{i1}^{(2)} = 0,\ \ 2 \leq i \leq n; \\ \\ a_{ij}^{(2)} = a_{ij}^{(1)} - \alpha_{i1} a_{1j}^{(1)},\quad 2 \leq i,j \leq n; \\ \\ b_i^{(2)} = b_i^{(1)} - \alpha_{i1} b_1^{(1)},\quad 2 \leq i \leq n. \end{cases}$$

Remarque 6.4. *Si on pose*

$$E^{(1)} = \begin{pmatrix} 1 & & & \\ -\alpha_{21} & \ddots & \boldsymbol{0} & \\ -\alpha_{31} & & 1 & \\ \vdots & \boldsymbol{0} & \ddots & \\ -\alpha_{n1} & & & 1 \end{pmatrix}$$

on aura $A^{(2)} = E^{(1)}.A^{(1)}$.

A la $k^{ième}$ étape $(1 \leq k \leq n-1)$ **:** *Si $a_{kk}^{(k)} \neq 0$ (sinon on fait une permutation de lignes) on fait les affectations suivantes :*

$$\begin{cases} L_i^{(k+1)} \longleftarrow L_i^{(k)}, \ \ 1 \leq i \leq k; \\ \\ L_i^{(k+1)} \longleftarrow L_i^{(k)} - \alpha_{ik} L_k^{(k)} \quad où \ \alpha_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}, \ \ k+1 \leq i \leq n. \end{cases}$$

On obtient donc

$$\left(A^{(k+1)} \vdots\, b^{(k+1)}\right) = \begin{pmatrix} a_{11}^{(k)} & a_{12}^{(k)} & \cdots & \cdots & \cdots & \cdots & a_{1n}^{(k)} & \vdots & b_1^{(k)} \\ 0 & a_{22}^{(k)} & \cdots & \cdots & \cdots & \cdots & a_{2n}^{(k)} & \vdots & b_2^{(k)} \\ \vdots & 0 & \ddots & & & & & \vdots & \\ \vdots & & \ddots & a_{kk}^{(k)} & \cdots & \cdots & a_{kn}^{(k)} & \vdots & b_k^{(k)} \\ & & & & & & & & \\ \vdots & & & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} & \vdots & b_{k+1}^{(k+1)} \\ \vdots & & & \vdots & \vdots & & \vdots & \vdots & \\ 0 & & & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} & \vdots & b_n^{(k+1)} \end{pmatrix} \begin{matrix} L_1^{(k+1)} \\ L_2^{(k+1)} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ L_n^{(k+1)} \end{matrix}$$

où

$$\begin{cases} \begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)}, \ 1 \leq j \leq n; \\ b_i^{(k+1)} = b_i^{(k)}, \end{cases} & 1 \leq i \leq k; \\ \\ a_{ij}^{(k+1)} = 0, \ \ 1 \leq j \leq k, \ \ k+1 \leq i \leq n; & \\ \\ \begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)} - \alpha_{ik} a_{kj}^{(k)}, \quad k+1 \leq j \leq n; \\ b_i^{(k+1)} = b_i^{(k)} - \alpha_{ik} b_k^{(k)} \end{cases} & k+1 \leq i \leq n. \end{cases}$$

Remarque 6.5. *Si on pose*

$$E^{(k)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & \boldsymbol{0} & \\ & & 1 & & \\ & & -\alpha_{k+1,k} & & \\ & \boldsymbol{0} & -\alpha_{k+2,k} & \ddots & \\ & & \vdots & \boldsymbol{0} & \\ & & -\alpha_{nk} & & 1 \end{pmatrix}$$

on aura $A^{(k+1)} = E^{(k)}.A^{(k)}$, $1 \leq k \leq n-1$.

En réitérant $(n-1)$ fois l'opération on obtient :

$$AX = b \Leftrightarrow A^{(n)} X = b^{(n)}$$

avec

$$A^{(n)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} \\ & & \ddots & & \vdots \\ & \boldsymbol{0} & & \ddots & \vdots \\ & & & & a_{nn}^{(n)} \end{pmatrix}, \quad b^{(n)} = \begin{pmatrix} b_1^{(n)} \\ b_2^{(n)} \\ \vdots \\ \vdots \\ b_n^{(n)} \end{pmatrix}$$

$A^{(n)}$ étant une matrice triangulaire supérieure.

Remarque 6.6. *1. Pour $A^{(1)} = A$ on a trouvé que*

$$A^{(k+1)} = E^{(k)}.A^{(k)} = E^{(k)}.E^{(k-1)}.A^{(k-1)} = E^{(k)}.E^{(k-1)} \cdots E^{(1)}.A^{(1)}.$$

Alors

$$A^{(n)} = \underbrace{E^{(n-1)}.E^{(n-2)} \cdots E^{(1)}}.A = MA.$$

2. Les $a_{kk}^{(k)}$, $k = \overline{1,n}$ sont appelés "pivots de la méthode de Gauss".

3. La méthode de Gauss permet de calculer $\det(A)$ *par*

$$\det(A) = (-1)^j \prod_{i=1}^{n} a_{ii}^{(i)} \quad \text{où } j \text{ est le nombre de permutations .}$$

4. Si à la $k^{ième}$ étape $a_{kk}^{(k)} = 0$, alors il existe p, $k+1 \leq p \leq n$ tel que $a_{pk}^{(k)} \neq 0$, car

$$\det(A) = a_{11}^{(1)} \times a_{22}^{(2)} \times \cdots a_{k-1,k-1}^{(k-1)} \times \begin{vmatrix} a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ \vdots & & \vdots \\ a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{vmatrix} \neq 0.$$

5. Au cours de la triangularisation, si l'on trouve que l'un des pivots $a_{kk}^{(k)} = 0$, on permute la ligne du pivot avec une ligne supérieure L_p, $k+1 \leq p \leq n$ dont l'élément de la $k^{ième}$ colonne $a_{pk}^{(k)}$ est non nul.

6. La méthode de Gauss sans permutation de lignes s'appelle "Gauss ordinaire".

Exemple 6.1.

$$A = \begin{pmatrix} 2 & -5 & 1 \\ -1 & 3 & -1 \\ 3 & -4 & 2 \end{pmatrix} \hookrightarrow A^{(2)} = \begin{pmatrix} 2 & -5 & 3 \\ 0 & 0 & -1 \\ 0 & -6 & -12 \end{pmatrix} \hookrightarrow A^{(3)} = \tilde{A}^{(2)} = \begin{pmatrix} 2 & -5 & 3 \\ 0 & -6 & -12 \\ 0 & 0 & -1 \end{pmatrix}$$

Coût de la méthode d'élimination de Gauss ordinaire

Pour passer de $\left(A^{(k)} \vdots\, b^{(k)}\right)$ *à* $\left(A^{(k+1)} \vdots\, b^{(k+1)}\right)$ *on utilise*

$$\begin{cases} a_{ij}^{(k+1)} = a_{ij}^{(k)} - \frac{a_{ik}^{(1)}}{a_{kk}^{(k)}} a_{kj}^{(k)}, & k+1 \leq j \leq n; \\ & k+1 \leq i \leq n. \\ b_i^{(k+1)} = b_i^{(k)} - \frac{a_{ik}^{(1)}}{a_{kk}^{(k)}} b_k^{(k)} \end{cases}$$

Alors à chaque étape k, $1 \leq k \leq n-1$ on a $(n-k+1)(n-k)$ additions (soustractions), $(n-k+1)(n-k)$ multiplications et $(n-k)$ divisions, donc

$$\textit{Coût total } = 2\sum_{k=1}^{n-1}(n-k+1)(n-k) + \sum_{k=1}^{n-1}(n-k) = \tfrac{2}{3}n^3 + \tfrac{1}{2}n^2 - \tfrac{7}{6}n.$$

Rappel : $\sum_{i=1}^{n} k^2 = \frac{n}{6}(n+1)(2n+1)$. *D'où, le coût total de la méthode de Gauss (élimination+résolution d'un système triangulaire) est $\frac{2}{3}n^3 + \frac{1}{2}n^2 - \frac{7}{6}n + n^2 = O(\frac{2}{3}n^3)$ car pour n grand, n^2 est négligeable devant n^3.*

6.2.3 Problème posé par la (quasi) annulation des pivots

Exemple 6.2. *Soit ϵ un petit nombre réel et considérons le système linéaire :*

$$(S_\epsilon)\begin{cases} \epsilon x_1 + x_2 = 1, \\ x_1 + x_2 = 2. \end{cases}$$

Résolvons d'abord le système (S_ϵ) par la méthode de Gauss ordinaire (sans permutations).

$$\left(A^{(1)}\vdots\, b^{(1)}\right) = \left(\begin{array}{cc:c} \epsilon & 1 & 1 \\ 1 & 1 & 2 \end{array}\right) \underset{a_{11}^{(1)}=\epsilon}{\hookrightarrow} \left(A^{(2)}\vdots\, b^{(2)}\right) = \left(\begin{array}{cc:c} \epsilon & 1 & 1 \\ 0 & 1-\frac{1}{\epsilon} & 2-\frac{1}{\epsilon} \end{array}\right) \begin{array}{l} L_1^{(1)} \\ L_2^{(1)} - \frac{1}{\epsilon}L_1^{(1)}. \end{array}$$

On a donc immédiatement la valeur de x_2 : $x_2 = \frac{1-2\epsilon}{1-\epsilon}$.
Par substitution, on obtient x_1 : $x_1 = \frac{1}{1-\epsilon}$.
Pour $\epsilon = 0$, on trouve $x_1 = x_2 = 1$.
Pour $\epsilon = 10^{-9}$, où 10^{-9} est la précision de la machine, on trouve $1 - 10^9 \approx -10^9$ et $2 - 10^9 \approx -10^9$ sur la machine, d'où $x_2 = 1$ et $x_1 = 0$ qui n'est pas la solution du système donné!

Si par contre, on permute les lignes 1 et 2, on obtient :

$$\left(A^{(2)}\vdots\, b^{(2)}\right) = \left(\begin{array}{cc:c} 1 & 1 & 2 \\ 0 & 1-10^{-9} & 1-2.10^{-9} \end{array}\right),$$

or $1-10^{-9} \approx 1$ et $1-2.10^{-9} \approx 1$ sur la machine, d'où $x_2 = 1$ et $x_1 = 1$ qui est la solution du système donné.

Conclusion : *Il ne faut pas utiliser des pivots trop petits car les erreurs d'arrondi peuvent donner des solutions fausses. Un moyen de contourner le problème d'un pivot nul ou presque est d'utiliser la technique de pivotage. Il existe deux stratégies de pivotage :*

Pivotage partiel : *à la $k^{ième}$ étape $(1 \leq k \leq n-1)$ d'élimination de Gauss le pivot partiel est choisi parmi les coefficients $(a_{ik}^{(k)})_{k\leq i\leq n}$ tel que sa valeur absolue soit la plus grande, soit $a_{i_0k}^{(k)}$ cet élément $\left(|a_{i_0k}^{(k)}| = \max_{k\leq i\leq n}(\left|a_{ik}^{(k)}\right|)\right)$. On permute ensuite si $(i_0 \neq k)$ la $k^{ième}$ ligne et la ligne (i_0).*

Pivotage total : *à la $k^{ième}$ étape $(1 \leq k \leq n-1)$ d'élimination de Gauss le pivot total est choisi parmi les coefficients $(a_{ij}^{(k)})_{k\leq i,j\leq n}$ tel que sa valeur absolue soit la plus grande, soit $a_{i_0j_0}^{(k)}$ cet élément $\left(|a_{i_0j_0}^{(k)}| = \max_{k\leq i,j\leq n}(\left|a_{ij}^{(k)}\right|)\right)$ puis on fait les permutations des lignes et des colonnes correspondantes. Catte technique n'est utilisée que dans de rares cas pratiques.*
Attention *A chaque permutation de colonnes les inconnues changent de places.*

Remarque 6.7.

1. *La méthode de Gauss n'est jamais programmée sans au moins une stratégie de pivot partiel.*
2. *Les pivots petits ne disparaissent pas, ils sont juste rejetés à la fin de l'algorithme où leur influence est moins grande.*

6.2.4 Méthode de la décomposition LU

Principe de la méthode :

1. *Décomposition de la matrice A de façon à la mettre sous la forme $A = LU$ où L est une matrice triangulaire inférieure unitaire et U est une matrice triangulaire supérieure.*
2. *Résolution : Le système $AX = b$ devient*

$$AX = b \Longleftrightarrow L\underbrace{UX}_{Y} \Longleftrightarrow \begin{cases} LY = b \\ UX = Y \end{cases}$$

donc la résolution du système $AX = b$ revient à la résolution de deux systèmes triangulaires.

Théorème 6.1. *Soit A une matrice telle que les sous matrices principales $A_{[k]} = (a_{ij})_{1\leq i,j\leq k}$ de A soient inversibles pour tous $1 \leq k \leq n$, alors il existe une matrice $L = (l_{ij})_{1\leq i,j\leq n}$ triangulaire inférieure telle que $l_{ii} = 1$, $i = \overline{1,n}$ et une matrice triangulaire supérieure U telle que $A = LU$.*
De plus cette décomposition est unique.

Démonstration. **Existence :** Montrons que tous les pivots d'élimination de Gauss sont non nuls, c'est à dire $a_{kk}^{(k)} \neq 0$ pour $1 \leq k \leq n-1$. On le démontre par récurrence.
Le premier pivot $a_{11}^{(1)}$ est forcément non nul car $a_{11}^{(1)} = \det(A_{[1]}) \neq 0$.
Supposons que $a_{kk}^{(k)} \neq 0$ pour $1 \leq k \leq r-1$ et montrons que $a_{rr}^{(r)} \neq 0$.
On a $\det(A_{[r]}) = a_{11}^{(1)} a_{22}^{(2)} \dots a_{r-1,r-1}^{(r-1)} a_{rr}^{(r)}$. Or d'une part, par hypothèse $\det(A_{[r]})$ est différent de zéro et d'autre part, par hypothèse de récurrence $a_{kk}^{(k)} \neq 0$ pour $1 \leq k \leq r-1$. Donc $a_{rr}^{(r)}$ est aussi différent de zéro.

Unicité : Soit $A = L_1U_1 = L_2U_2$, d'où $U_1U_2^{-1} = L_1^{-1}L_2 = D$.
Comme $U_1U_2^{-1}$ est une matrice triangulaire supérieure et $L_1^{-1}L_2$ est une matrice triangulaire inférieure et D a des 1 sur la diagonale, alors $D = I_n$ ce qui implique que $U_1 = U_2$ et $L_1 = L_2$.
□

Détermination des matrices L et U

a) En utilisant l'algorithme d'élimination de Gauss ordinaire : *Au premier pas d'élimination de Gauss, on trouve*

$$A^{(2)} = E^{(1)}.A^{(1)} \quad où \quad E^{(1)} = \begin{pmatrix} 1 & & & \\ -\alpha_{21} & \ddots & \boldsymbol{0} & \\ -\alpha_{31} & & 1 & \\ \vdots & \boldsymbol{0} & & \ddots \\ -\alpha_{n1} & & & 1 \end{pmatrix}$$

Au deuxième pas, on trouve

$$A^{(3)} = E^{(2)}.A^{(2)} \quad où \quad E^{(2)} = \begin{pmatrix} 1 & & & & \\ 0 & 1 & & & \boldsymbol{0} \\ \vdots & -\alpha_{32} & \ddots & & \\ \vdots & -\alpha_{42} & \boldsymbol{0} & \ddots & \\ \vdots & \vdots & & & \\ 0 & -\alpha_{n2} & & & 1 \end{pmatrix}$$

de la même manière au k-ième pas d'élimination, on obtient

$$A^{(k+1)} = E^{(k)}.A^{(k)} \quad où \quad E^{(k)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \boldsymbol{0} \\ & & 1 & & \\ & & -\alpha_{k+1,k} & & \\ & \boldsymbol{0} & -\alpha_{k+2,k} & \ddots & \\ & & \vdots & \boldsymbol{0} & \\ & & -\alpha_{nk} & & 1 \end{pmatrix}$$

ce qui nous donne

$$\begin{aligned} A^{(n)} &= E^{(n-1)}.A^{(n-1)} \\ &= E^{(n-1)}.E^{(n-2)}.A^{(n-2)} \\ &= E^{(n-1)}.E^{(n-2)} \dots E^{(1)}.A. \end{aligned}$$

Posons $U = A^{(n)}$ *et* $L^{-1} = E^{(n-1)}.E^{(n-2)} \cdots E^{(1)}$
alors $U = L^{-1}A$ *d'où* $A = LU$ *où*

$$\begin{aligned} L &= \left(E^{(n-1)}.E^{(n-2)} \dots E^{(2)}.E^{(1)}\right)^{-1} \\ &= \left(E^{(1)}\right)^{-1}.\left(E^{(2)}\right)^{-1} \dots \left(E^{(n-1)}\right)^{-1} \\ &= \begin{pmatrix} 1 & & & & \\ \alpha_{21} & 1 & & \boldsymbol{0} & \\ \alpha_{31} & \alpha_{32} & \ddots & & \\ \vdots & \vdots & \ddots & \ddots & \\ \alpha_{n1} & \alpha_{n2} & \cdots & \alpha_{n,n-1} & 1 \end{pmatrix} \end{aligned}$$

et

$$U = A^{(n)} = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & \cdots & a_{1n}^{(1)} \\ & a_{22}^{(2)} & \cdots & \cdots & a_{2n}^{(2)} \\ & & \ddots & & \vdots \\ & \boldsymbol{0} & & \ddots & \vdots \\ & & & & a_{nn}^{(n)} \end{pmatrix}$$

b) En appliquant l'algorithme de la méthode :

En connaissant $A = (a_{ij})_{1 \le i,j \le n}$, *on écrit l'égalité* $A = LU$

$$\begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ \vdots & & \vdots & & \vdots \\ a_{i1} & \cdots & a_{ij} & \cdots & a_{in} \\ \vdots & & \vdots & & \vdots \\ a_{n1} & \cdots & a_{n2} & \cdots & a_{nn} \end{pmatrix} = \begin{pmatrix} 1 & & & & \\ l_{21} & 1 & & \boldsymbol{0} & \\ l_{31} & & \ddots & & \\ \vdots & & \ddots & \ddots & \\ l_{n1} & & \cdots & l_{n,n-1} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & \cdots & \cdots & u_{1n} \\ & u_{22} & \cdots & \cdots & u_{2n} \\ & & \ddots & & \vdots \\ & \boldsymbol{0} & & \ddots & \vdots \\ & & & & u_{nn} \end{pmatrix}$$

(U contient $\frac{n(n+1)}{2}$ éléments et L contient $\frac{n(n-1)}{2}$ éléments). Par identification on obtient un système linéaire de n^2 équations à n^2 inconnues. En résolvant le système obtenu dans des cas particuliers ($n = 2, 3, 4$), on constate que la détermination des éléments de L et U cherchés se fait suivant l'algorithme général :

$$\begin{cases} l_{ii} = 1, \quad 1 \leq i \leq n; \\ \begin{cases} u_{1j} = a_{1j}, \ 1 \leq j \leq n; \\ l_{i1} = \frac{a_{i1}}{u_{11}}, \ 2 \leq i \leq n; \end{cases} \\ \begin{cases} u_{mj} = a_{mj} - \sum_{k=1}^{m-1} l_{mk}.u_{kj}, \quad m \leq j \leq n; \\ l_{im} = \left(a_{im} - \sum_{k=1}^{m-1} l_{ik}.u_{km}\right)/u_{mm}, \quad m+1 \leq i \leq n; \end{cases} \end{cases} \quad 2 \leq m \leq n.$$

Coût de la méthode

- *Calcul de U : coût(×)=coût(+)=* $\sum_{m=1}^{n}(m-1)(n-m+1) = O(\frac{1}{6}n^3)$
- *Calcul de L : coût(×)=coût(+)=* $\sum_{m=1}^{n-1}(m-1)(n-m) = O(\frac{1}{6}n^3)$

  et coût(/)= $\sum_{m=1}^{n}(n-m) = \frac{n(n-1)}{2}$.

D'où, coût total $= O(\frac{2}{3}n^3)$*=coût de Gauss.*

Utilité de la détermination *LU*

Calcul de déterminant *Grâce à la factorisation LU, on peut calculer le déterminant d'une matrice carrée avec* $O(\frac{2}{3}n^3)$ *opérations, vu que*

$$det(A) = det(L) \times det(U) = det(U) = \prod_{k=1}^{n} u_{kk}.$$

Résolution : *Supposons qu'on veut résoudre le système $AX = b$. Décomposons A sous forme LU, alors $AX = b$ devient $(LU)X = b$ ou encore $L(UX) = b$. Posons $Y = UX$, on cherche alors Y tel que $LY = b$ est un système triangulaire inférieur qu'on résout par la méthode descendante. Y étant trouvé, on cherche X tel que $UX = Y$ est un système triangulaire supérieur qu'on résout par la méthode ascendante.*

Coût total : *coût(décomposition LU)+coût(résolution de deux systèmes triangulaires), c'est à dire, Coût total=* $O(\frac{2}{3}n^3) + 2n^2 = O(\frac{2}{3}n^3)$.

Calcul de l'inverse d'une matrice : *Soit A une matrice carrée inversible d'ordre n, notons par $v^{(1)}, \ldots, v^{(n)}$ les colonnes de sa matrice inverse A^{-1}, i.e. $A^{-1} = (v^{(1)}, \ldots, v^{(n)})$. La relation $A.A^{-1} = I_n$ se traduit par les n systèmes linéaires suivants :*

$$Av^{(k)} = e^{(k)}, \quad 1 \leq k \leq n, \tag{6.3}$$

où $e^{(k)}$ est le vecteur colonne ayant toutes les composantes nulles sauf la $k-$ième composonte qui est 1, $e^{(k)} = (0, \ldots, 1, \ldots, 0)^t$. Une fois connues les matrices L et U qui décomposent la matrice A, résoudre les n systèmes (6.3) gouvernés par la même matrice A.

Exemple 6.3. *Soit le système linéaire suivant*

$$\begin{pmatrix} 2 & -5 & 1 \\ -1 & 3 & -1 \\ 3 & -4 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 12 \\ -8 \\ 16 \end{pmatrix}$$

A l'aide de la décomposition LU de A :

1. *Résoudre le système donné en déterminant les matrices L et U :*
   - **a)** *En utilisant l'algorithme d'élimination de Gauss ordinaire.*
   - **b)** *En appliquant l'algorithme de la méthode.*
2. *Calculer l'inverse de A.*

6.2.5 Méthode de Cholesky

Dans le cas d'une matrice A symétrique définie positive, il est possible de résoudre le système $AX = b$ avec un nombre d'opérations égal presque à la moitie du nombre d'opérations utilisées dans la méthode de Gauss.

Définition 6.1. *Une matrice $A \in M_n(\mathbb{R})$ est dite définie positive si et ssi $X^tAX > 0,\ \forall X \in \mathbb{R}^n - \{0_{\mathbb{R}_n}\}$.*

Exemple 6.4. *Soient $A = \begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix}$ et $X = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}_2$. On a*

$$X^tAX = (x_1,\ x_2) \begin{pmatrix} 1 & 2 \\ 2 & 8 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = (x_1 + 2x_2)^2 + 4x_2^2 > 0\ \forall X \in \mathbb{R}^2 - \{(0,0)\}.$$

Définition 6.2. *$A \in M_n(\mathbb{R})$ est dite définie positive si et ssi toute sous-matrice principale $A_{[k]} = (a_{ij})_{1 \le i,j \le k},\ k = \overline{1,n}$ de A est de déterminant strictement positif.*

Théorème 6.2. *Si A est une matrice symétrique et définie positive alors, il existe (au moins) une matrice triangulaire inférieure R telle que $A = RR^t$.*
De plus, si on impose que les éléments diagonaux de R soient tous positifs, alors cette décomposition est unique.

Démonstration. On a A définie positive assure l'existence de la décomposition LU de A où $L = (l_{ij})_{1 \le i,j \le n}$ est une matrice triangulaire inférieure unitaire et $U = (u_{ij})_{1 \le i,j \le n}$ est une matrice triangulaire supérieure (car toute sous-matrice principale $A_{[k]}),\ k = \overline{1,n}$ de A est de déterminant non nul).

Remarquons que $\det(A_{[k]}) = \prod_{j=1}^{k} u_{jj},\ k = 1, 2, ..., n$ ce qui implique que $u_{ii} > 0,\ i = \overline{1,n}$.

Soit D et $D^{\frac{1}{2}}$ les matrices diagonales définies par

$$D = diag(u_{11}, u_{22}, ..., u_{nn}),\quad D^{\frac{1}{2}} = diag(\sqrt{u_{11}}, \sqrt{u_{22}}, ..., \sqrt{u_{nn}})$$

alors $\left(D^{\frac{1}{2}}\right)^{-1} = diag(\frac{1}{\sqrt{u_{11}}}, \frac{1}{\sqrt{u_{22}}}, ..., \frac{1}{\sqrt{u_{nn}}})$. D'une part, on a

$$A = LU = LI_nU = LD^{\frac{1}{2}}(D^{\frac{1}{2}})^{-1}U$$

Posons

$$\begin{array}{ll} R=(r_{ij})_{1\le i,j\le n}=LD^{\frac{1}{2}}, & \text{matrice triangulaire inférieure avec } r_{ii}=\sqrt{u_{ii}} \\ H=(h_{ij})_{1\le i,j\le n}=(D^{\frac{1}{2}})^{-1}U, & \text{matrice triangulaire supérieure avec } h_{ii}=\sqrt{u_{ii}}, \end{array}$$

alors

$$A=RH.$$

D'autre part, on a A est symétrique, alors

$$\begin{array}{ll} & A=A^t \\ \Longleftrightarrow & RH=H^tR^t \\ \Longleftrightarrow & H\,(R^t)^{-1}=(R)^{-1}\,H^t \ \text{(multiplions les deux côtés à gauche par } (R)^{-1} \text{ et à droite par } (R^t)^{-1}\ ) \\ \Longrightarrow & H\,(R^t)^{-1}=I_n \text{ et } (R)^{-1}\,H^t=I_n \ \ (H\,(R^t)^{-1} \text{ est M.T.S. et } (R)^{-1}\,H^t \text{ est M.T.I. }) \\ \Longrightarrow & H=R^t. \end{array}$$

□

Algorithme de Cholesky[2]

Son principe est le suivant : on écrit la matrice A sous la forme RR^t où $R=(r_{ij})_{1\le i,j\le n}$, $r_{ij}=0$ si $j>i$, ensuite on identifie colonne après colonne on aura

Première colonne : *($j=1$)*

$a_{11}=r_{11}^2\Longrightarrow r_{11}=\sqrt{a_{11}}$ *(si on impose que $r_{11}>0$)*

$a_{i1}=r_{i1}r_{11}\Longrightarrow r_{i1}=\frac{a_{i1}}{r_{11}},\quad 2\le i\le n.$

Deuxième colonne : *($j=2$)*

$a_{22}=r_{21}^2+r_{22}^2\Longrightarrow r_{22}=\sqrt{a_{22}-r_{21}^2}$ *(si on impose que $r_{22}>0$)*

$a_{i2}=r_{i1}r_{21}+r_{i2}r_{22}\Longrightarrow r_{i2}=\frac{1}{r_{22}}\,(a_{i2}-r_{i1}r_{i2}),\quad 3\le i\le n.$

Ainsi on a construit l'algorithme suivant :

$$\left\{\begin{array}{l} r_{11}=\sqrt{a_{11}}; \\ \\ r_{i1}=\frac{a_{i1}}{r_{11}},\quad 2\le i\le n; \\ \\ \textit{pour } 2\le j\le n \\ \\ r_{jj}=\sqrt{a_{jj}-\sum\limits_{k=1}^{j-1}r_{jk}^2}; \\ \\ r_{ij}=\frac{1}{r_{jj}}\left(a_{ij}-\sum\limits_{k=1}^{j-1}r_{ik}\,r_{jk}\right),\quad j+1\le i\le n. \end{array}\right.$$

Coût de la méthode

n extractions de racines carrés, $\sum\limits_{j=1}^{n}(n-j)=\frac{n(n-1)}{2}$ *(divisions),*

$2\sum\limits_{j=1}^{n}(j-1)(n-j)=O(\frac{1}{3}n^3)$ *(coût($\times$) et coût($+$))*

$\sum\limits_{j=1}^{n}(j-1)=O(\frac{1}{2}n^2)$ *(coût($\times$) et coût($+$) dans les racines carrés).*

Coût total$=O(\frac{1}{3}n^3)\approx$ *la moitié de celui de Gauss.*

[2] André Louis Cholesky, français, 1875-1918

Conclusion :

On peut utiliser la décomposition RR^t d'une matrice A symétrique définie positive pour résoudre le système linéaire $AX = b$ ou pour inverser A, de la même manière que pour la décomposition (LU).

Remarque 6.8. *1. Si $A = RR^t$ où $R = (r_{ij})_{1\leq i,j\leq n}$ alors, $\det(A) = (\det(R))^2 = \prod_{i=1}^{n} r_{ii}^2$.*

2. La décomposition $A = RR^t$ n'est pas unique si on n'impose pas que $r_{ii} > 0$, $i = 1, 2, \ldots, n$.

3. Pour une matrice définie positive on a : $r_{jj}^2 = a_{jj} - \sum_{k=1}^{j-1} r_{jk}^2 > 0$, donc, si a une étape de calcul on trouve que $r_{jj}^2 < 0$, la matrice A n'est pas définie positive.

6.2.6 Méthode de Gauss-Jordan

Principe de la méthode :

1. *Transformation de la matrice A en la matrice identité :*

$$(A, b) \overset{transformation}{\longrightarrow} (I_n, b^{(n)}),$$

où I_n est la matrice identité dans $\mathbb{M}_n(\mathbb{R})$.

2. *Résolution du système :*

$$AX = b \Longleftrightarrow I_n X = b^{(n)} \Longleftrightarrow X = b^{(n)}.$$

Algorithme de Gauss-Jordan :

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

$$\left(A^{(1)} \vdots\, b^{(1)}\right) = \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} & \vdots & b_1^{(1)} \\ a_{21}^{(1)} & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} & \vdots & b_2^{(1)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ a_{n1}^{(1)} & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} & \vdots & b_n^{(1)} \end{pmatrix} \begin{matrix} L_1^{(1)} \\ L_2^{(1)} \\ \vdots \\ \\ L_n^{(1)} \end{matrix}$$

A la $1^{ère}$ étape : *Si $a_{11}^{(1)} \neq 0$ (sinon on fait une permutation de lignes) on fait les affectations suivantes :*

$$\begin{cases} L_1^{(2)} & \longleftarrow & \frac{1}{a_{11}^{(1)}} L_1^{(1)} \\ \\ L_i^{(2)} & \longleftarrow & L_i^{(1)} - a_{i1}^{(1)} L_1^{(2)} \quad 2 \leq i \leq n. \end{cases}$$

On obtient

$$\left(A^{(2)} \vdots\, b^{(2)}\right) = \begin{pmatrix} 1 & a_{12}^{(2)} & \cdots & a_{1n}^{(2)} & \vdots & b_1^{(2)} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} & \vdots & b_2^{(2)} \\ \vdots & \vdots & & \vdots & \vdots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} & \vdots & b_n^{(2)} \end{pmatrix} \begin{matrix} L_1^{(2)} \\ L_2^{(2)} \\ \vdots \\ \\ L_n^{(2)} \end{matrix}$$

où

$$\begin{cases} a_{1j}^{(2)} = \frac{a_{1j}^{(1)}}{a_{11}^{(1)}},\ 1 \leq j \leq n; \quad b_1^{(2)} = \frac{b_1^{(2)}}{a_{11}^{(1)}}; \\ a_{i1}^{(2)} = 0, \quad 2 \leq i \leq n; \\ a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i1}^{(1)} a_{1j}^{(2)}, \quad 2 \leq i, j \leq n; \\ b_i^{(2)} = b_i^{(1)} - a_{i1}^{(1)} b_1^{(2)}, \quad 2 \leq i \leq n. \end{cases}$$

A la $k^{ième}$ étape $(1 \leq k \leq n)$ **:** *Si $a_{kk}^{(k)} \neq 0$ (sinon on fait une permutation de lignes) on fait les affectations suivantes :*

$$\begin{cases} L_i^{(k+1)} & \longleftarrow & \frac{1}{a_{kk}^{(k)}} L_i^{(k)}, \\ L_i^{(k+1)} & \longleftarrow & L_i^{(k)} - a_{ik}^{(k)} L_k^{(k+1)} \quad i = \overline{1,n},\ i \neq k. \end{cases}$$

On obtient donc

$$\left(A^{(k+1)} \vdots\, b^{(k+1)}\right) = \begin{pmatrix} 1 & 0 & \cdots & 0 & a_{1,k+1}^{(k+1)} & \cdots & a_{1,n}^{(k+1)} & \vdots & b_1^{(k+1)} \\ \vdots & 1 & \ddots & & \vdots & & \vdots & \vdots & b_2^{(k+1)} \\ \vdots & 0 & \ddots & 0 & a_{k-1,k+1}^{(k+1)} & \cdots & a_{k-1,n}^{(k+1)} & \vdots & \\ \vdots & & \ddots & 1 & a_{k,k+1}^{(k+1)} & \cdots & a_{kn}^{(k+1)} & \vdots & \vdots \\ \vdots & & & 0 & a_{k+1,k+1}^{(k+1)} & \cdots & a_{k+1,n}^{(k+1)} & \vdots & \vdots \\ \vdots & & & \vdots & \vdots & & \vdots & \vdots & \\ 0 & & & 0 & a_{n,k+1}^{(k+1)} & \cdots & a_{n,n}^{(k+1)} & \vdots & b_n^{(k+1)} \end{pmatrix} \begin{matrix} L_1^{(k+1)} \\ L_2^{(k+1)} \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ L_n^{(k+1)} \end{matrix}$$

où

$$\begin{cases} \begin{cases} a_{kj}^{(k+1)} = \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}}; \\ a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k+1)}, \quad i = \overline{1,n},\ i \neq k, \end{cases} & k+1 \leq j \leq n; \\ \begin{cases} b_k^{(k+1)} = \frac{b_k^{(k)}}{a_{kk}^{(k)}} \\ b_i^{(k+1)} = b_i^{(k)} - a_{ik}^{(k)} b_k^{(k+1)} \quad i = \overline{1,n},\ i \neq k. \end{cases} \end{cases}$$

Remarque 6.9. *1. Pour résoudre un système d'ordre n, la méthode de Gauss Jordan nécessite $O(n^3)$ opérations (moins rapide que celle de Gauss).*

2. Elle est conseillée pour inverser une matrice.

$$(A, I_n) \overset{transformation}{\longrightarrow} (I_n, A^{-1}).$$

6.3 Méthodes itératives

Quand n est assez grand les méthodes directes ne sont plus envisageables vu le nombre très grand d'opérations à effectuer qui engendre la propagation des erreurs d'arrondi. On a alors recours

aux méthodes itératives qui consistent à générer -à partir d'un vecteur initial $X^{(0)}$ choisi dans $\mathbb{R}^n$- une suite $(X^{(n)})_{n\in\mathbb{N}}$ telle que : $X^{(n+1)} = F(X^{(n)})$ et $\lim_{n\to+\infty} X^{(n)} = X$; où $F : \mathbb{R}^n \to \mathbb{R}^n$ est un opérateur lié à la méthode.

Principe des méthodes itératives :

Ecrivons d'abord la matrice A sous la forme $A = M - N$ où M est inversible, alors

$$\begin{aligned} AX = b &\iff (M-N)X = b \\ &\iff MX = NX + b. \end{aligned}$$

Multiplions les deux côtés par M^{-1}, on aura

$$X = M^{-1}NX + M^{-1}b.$$

Le principe de toutes les méthodes itératives est le suivant :
- *choisir un vecteur $X^{(0)} \in \mathbb{R}^n$,*
- *générer la suite $(X^{(k)})_{k\in\mathbb{N}}$ telle que $X^{(k+1)} = M^{-1}NX^{(k)} + M^{-1}b$,*
- *si la suite $(X^{(k)})_{k\in\mathbb{N}}$ converge vers X^*, alors X^* est la solution du système $AX = b$.*

Remarque 6.10. *1. Le critère d'arrêt se fait, en général, sur l'erreur relative de deux itérés successifs $X^{(k)}$ et $X^{(k+1)}$, c'est à dire*

$$Test\Big(\frac{\|X^{(k+1)} - X^{(k)}\|}{\|X^{(k+1)}\|} < \epsilon\Big)$$

où ϵ choisi petit, ou sur l'erreur absolue si $\|X^{(k+1)}\|$ est très petite.

2. La méthode itérative est dite convergente si : $\forall\, X^{(0)} \in \mathbb{R}^n$, $X^{(k)} \longrightarrow X^*$ quand $k \to +\infty$ où X^* est la solution du système $AX = b$.

6.3.1 Matrice d'itération et les conditions de convergence

On appelle, pour M et N choisies, matrice d'itération "la matrice $B = M^{-1}N$."

Condition nécessaire de convergence

Notons par $E^{(k)} = X^{(k)} - X^*$ le vecteur erreur à l'étape k ($k \in \mathbb{N}$), on a

$$\begin{aligned} X^* &= B\,X^* + M^{-1}b \qquad (1) \\ X^{(k)} &= B\,X^{(k-1)} + M^{-1}b, \qquad (2) \end{aligned}$$

en soustrayant (1) de (2), on aura

$$X^{(k)} - X^* = B(X^{(k-1)} - X^*) = B\,E^{(k-1)}.$$

Alors

$$E^{(k)} = B\,E^{(k-1)} = B^2\,E^{(k-2)} = \cdots = B^k\,E^{(0)} \quad où \;\; E^{(0)} = X^{(0)} - X^*,$$

ou encore

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

La méthode converge si $\forall\, X^{(0)}$, $\lim_{k\to+\infty} X^{(k)} = X^*$, ce qui est vraie si

$$\lim_{k\to+\infty} B^k = 0 \;\; (0 \textit{ au sens matriciel}).$$

Condition nécessaire et suffisante de convergence

Soit $\rho(B) = \max\{|\lambda| / \ \lambda$ *valeur propre de* $B\}$ *le rayon spectral de* B. *On a le théorème suivant :*

Théorème 6.3. *La suite* $(X^{(k)})_{k\in\mathbb{N}}$ *définie par*

$$\begin{cases} X^{(0)} \in \mathbb{R}^n \ quelconque \\ X^{(k+1)} = B\,X^{(k)} + M^{-1}b, \end{cases}$$

converge vers X^* *si et ssi* $\rho(B) < 1$.

Condition suffisante de convergence

Rappel

1. *Soit* $X = (x_1, \cdots, x_n)^t \in \mathbb{R}^n$, *les normes définies par :*
   - $\|X\|_1 = \sum_{i=1}^{n} |x_i|$
   - $\|X\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$
   - $\|X\|_\infty = \max_{1\le i\le n} |x_i|$

   sont équivalentes.
2. *Soit* $A = (a_{ij}) \in \mathbb{M}_n(\mathbb{R})$, *les normes définies par :*
   - $\|A\|_1 = \max_{1\le j\le n} \left(\sum_{i=1}^{n} |a_{ij}|\right)$
   - $\|A\|_2 = \sqrt{\rho({}^tA.A)}$
   - $\|A\|_\infty = \max_{1\le i\le n} \left(\sum_{j=1}^{n} |a_{ij}|\right)$

   sont équivalentes.

Lemme 6.1. *(***Relation entre** $\rho(B)$ **et** $\|B\|$*)*

$$\rho(B) \le \|B\|_i \quad (i = 1 \vee 2 \vee \infty).$$

Démonstration. Soit λ une valeur propre de B, alors $\exists\, X \in (\mathbb{R}^*)^n : BX = \lambda X$ et donc

$$\|BX\|_i = \|\lambda X\|_i \Longrightarrow |\lambda|\|X\|_i \le \|B\|_i\|X\|_i,\ X \in (\mathbb{R}^*)^n\ (\|BX\|_i \le \|B\|_i\|X\|_i).$$

D'où, $|\lambda| \le \|B\|_i \Longrightarrow \rho(B) \le \|B\|_i,\ i = 1 \vee 2 \vee \infty$. □

D'après le lemme 6.1, on tire que l'existence d'une norme $\|.\|_i$, $(i = 1 \vee 2 \vee \infty)$ *de* B *qui vérifie* $\|B\|_i < 1$ *est une condition suffisante pour la convergence de la méthode itérative.*

6.3.2 Principales méthodes itératives

On considère la décomposition suivante de la matrice A

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

où,

$$\begin{cases} D & : & \text{la diagonale de } A, \\ -E & : & \text{la partie au dessous de la diagonale de} A, \\ -F & : & \text{la partie au dessus de la diagonale de } A. \end{cases}$$

Exemple 6.5. *Soit* $A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$ *alors*

$$D = \begin{pmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}, \quad E = \begin{pmatrix} 0 & 0 & 0 \\ -a_{21} & 0 & 0 \\ -a_{31} & -a_{32} & 0 \end{pmatrix}, \quad F = \begin{pmatrix} 0 & -a_{12} & -a_{13} \\ 0 & 0 & -a_{23} \\ 0 & 0 & 0 \end{pmatrix}.$$

On suppose que $\forall\, i = 1, 2 \cdots, n,\ \ a_{ii} \neq 0_{\mathbb{R}}$ *(on peut s'y ramener si* A *est inversible).*

i) Méthode de Jacobi : $M = D, \quad N = E + F$

ii) Méthode de Gauss-Seidel : $M = D - E, \quad N = F$

iii) Méthode de relaxation : $M = (\frac{1}{\omega})D - E, \quad N = (\frac{1}{\omega} - 1)D + F,\ \omega \in \mathbb{R}^*.$

Remarque 6.11.

1. *Si* $\omega = 1$, *la méthode de relaxation coïncide avec celle de Gauss-Seidel.*
2. *La matrice* M *est inversible, car elle possède toujours les* a_{ii} *sur la diagonale.*
3. *la méthode de relaxation est une variante qui généralise la méthode de Gauss-Seidel. L'introduction du paramètre* ω *vise à accélérer la convergence de cette dernière.*

Présentation des algorithmes

On va, dans ce qui suit, expliciter les méthodes de Jacobi et Gauss-Seidel, et cela sur un système linéaire à trois équations $(n = 3)$.
Soit donc $AX = b$ *où* $A \in \mathbb{M}_3(\mathbb{R}),\ b \in \mathbb{R}^3$

$$AX = b \Longleftrightarrow \begin{cases} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \quad (1) \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \quad (2) \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \quad (3) \end{cases}$$

les $(a_{ii})_{i=1,2,\ldots,n}$ *étant supposés non nuls ; tirons* x_1 *de (1),* x_2 *de (2) et* x_3 *de (3), on obtient*

$$\begin{aligned} x_1 &= (b_1 - a_{12}x_2 + a_{13}x_3)/a_{11} = f(x_2, x_3) \quad (4) \\ x_2 &= (b_2 - a_{21}x_1 + a_{23}x_3)/a_{22} = g(x_1, x_3) \quad (5) \\ x_3 &= (b_3 - a_{31}x_1 + a_{32}x_2)/a_{33} = h(x_1, x_2) \quad (6) \end{aligned}$$

soit $X^{(0)} = (x_1^{(0)}, x_2^{(0)}, x_3^{(0)})^t \in \mathbb{R}^3$ *quelconque.*

La méthode de Jacobi revient au processus itératif suivant :

$1^{ère}$ **étape :**

$$\begin{aligned} x_1^{(1)} &= f(x_2^{(0)}, x_3^{(0)}) \\ x_2^{(1)} &= g(x_1^{(0)}, x_3^{(0)}) \\ x_3^{(1)} &= h(x_1^{(0)}, x_2^{(0)}) \end{aligned} \longrightarrow X^{(1)} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix}$$

$2^{ième}$ **étape :**

$$\begin{aligned} x_1^{(2)} &= f(x_2^{(1)}, x_3^{(1)}) \\ x_2^{(2)} &= g(x_1^{(1)}, x_3^{(1)}) \\ x_3^{(2)} &= h(x_1^{(1)}, x_2^{(1)}) \end{aligned} \longrightarrow X^{(2)} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix}$$

et ainsi de suite, jusqu'à satisfaction du critère d'arrêt. (Il est évident qu'on ne peut pas calculer le coût de la méthode).

La méthode de Gouss-Seidel[3] *revient au processus itératif suivant :*

$1^{ère}$ **étape :**

$$\begin{array}{rcl} x_1^{(1)} & = & f(x_2^{(0)}, x_3^{(0)}) \equiv \ Jacobi \\ x_2^{(1)} & = & g(x_1^{(1)}, x_3^{(0)}) \\ x_3^{(1)} & = & h(x_1^{(1)}, x_2^{(1)}) \end{array} \longrightarrow X^{(1)} = \begin{pmatrix} x_1^{(1)} \\ x_2^{(1)} \\ x_3^{(1)} \end{pmatrix}$$

$2^{ième}$ **étape :**

$$\begin{array}{rcl} x_1^{(2)} & = & f(x_2^{(1)}, x_3^{(1)}) \\ x_2^{(2)} & = & g(x_1^{(2)}, x_3^{(1)}) \\ x_3^{(2)} & = & h(x_1^{(2)}, x_2^{(2)}) \end{array} \longrightarrow X^{(2)} = \begin{pmatrix} x_1^{(2)} \\ x_2^{(2)} \\ x_3^{(2)} \end{pmatrix}$$

(c'est à dire toutes les composantes calculées $(x_1, x_2, ..., x_{i-1})$ sont utilisées pour calculer x_i), et ainsi de suite, jusqu'à satisfaction du critère d'arrêt.

Généralisation

Soit $AX = b$ un système linéaire d'ordre n où $a_{ii} \neq 0, \ \forall\, i = 1, 2, \cdots, n$. Les équations (4), (5) et (6) précédentes se généralisent comme suit :

$$x_i = \left(b_i - \sum_{j=1,\ j\neq i}^{n} a_{ij}\, x_j \right) / a_{ii}, \ \ \forall\, i = 1, 2, \cdots, n.$$

Etapes principales de la méthode de Jacobi :

1. *Etant données $A, b, X^{(0)} = (x_1^{(0)}, x_2^{(0)}, \cdots, x_n^{(0)})^t, \epsilon$(la précision) et/ou $KMax$(le nombre maximal d'itérations).*

2. $x_i^{(k+1)} = \left(b_i - \sum_{j=1,\ j\neq i}^{n} a_{ij}\, x_j^k \right) / a_{ii}, \ \ \forall\, i = 1, 2, \cdots, n.$

   A répéter pour $k = 0, ..., KMax$.

   Si

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

   arrêter les itérations, et $X^{(k+1)}$ est une solution approchée de X solution du système donné avec une précision relative ϵ.

Etapes principales de la méthode de Gauss-Seidel :

1. *Etant données $A, b, X^{(0)}, \cdots, x_n^{(0)})^t, \epsilon$ et/ou $KMax$.*

2. $x_i^{(k+1)} = \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}\, x_j^{(k)} \right) / a_{ii}, \ \ \forall\, i = 1, 2, \cdots, n.$

   A répéter pour $k = 0, ..., KMax$.

   Si

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

   arrêter les itérations, et $X^{(k+1)}$ est une solution approchée de X solution du système donné avec une précision relative ϵ.

Remarque 6.12. *1. S'il existe $i_0 \in [1,n]$ tel que $a_{i_0 i_0} = 0$, on procède à une permutation de ligne sur A (et sur b).*

2. En général, la convergence de l'une de ces méthodes n'implique pas la convergence de l'autre.

3. Plus que $\rho(B) << 1$, plus que la convergence du processus itératif

$$\begin{cases} X^{(0)} \in \mathbb{R}^n \ \ donné \\ X^{(k+1)} = BX^{(k)} + C \end{cases}$$

vers la solution exacte du système $AX = b$ est plus rapide.

Exemple 6.6. *Soient les quatre systèmes linéaires $A_i X = b_i, \ \ i = 1,2,3,4.$*

$$A_1 = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix} \quad A_2 = \begin{pmatrix} 2 & -1 & 1 \\ 2 & 2 & 2 \\ -1 & -1 & 2 \end{pmatrix}$$

$$A_3 = \begin{pmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{pmatrix} \quad A_4 = \begin{pmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{pmatrix}$$

Etudier la convergence des méthodes de Jacobi et Gauss-Seidel appliquées à chaque système, pour tout choix de $X^{(0)} \in \mathbb{R}^3$, en utilisant le rayon spectral des matrices d'itérations. Conclure.

Réponse *On trouve les résultats suivants :*

1.

$$J_1 = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix} \quad G_1 = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}$$

avec $\rho(J_1) = 0 < 1, \quad \rho(G_1) = 2 > 1$, d'où la méthode de Jacobi converge $\forall\, X^{(0)} \in \mathbb{R}^3$, alors que celle de Gauss-Seidel ne converge pas $\forall\, X^{(0)} \in \mathbb{R}^3$.

2.

$$J_2 = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ -1 & 0 & -1 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{pmatrix} \quad G_2 = \begin{pmatrix} 0 & \frac{1}{2} & -\frac{1}{2} \\ 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}$$

avec $\rho(J_2) = 1,118 > 1, \quad \rho(G_2) = \frac{1}{2} < 1$, d'où la méthode de Gauss-Seidel converge $\forall\, X^{(0)} \in \mathbb{R}^3$, alors que celle de Jacobi ne converge pas $\forall\, X^{(0)} \in \mathbb{R}^3$.

3.

$$J_3 = \begin{pmatrix} 0 & -\frac{1}{4} & -\frac{1}{4} \\ \frac{2}{9} & 0 & 0 \\ 0 & -\frac{4}{3} & 0 \end{pmatrix} \quad G_3 = \begin{pmatrix} 0 & -\frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{1}{8} & \frac{1}{8} \\ 0 & -\frac{1}{6} & -\frac{1}{6} \end{pmatrix}$$

avec $\rho(J_3) = 0,44 < 1, \quad \rho(G_3) = 0,018 < 1$, d'où les deux méthodes convergent $\forall\, X^{(0)} \in \mathbb{R}^3$, mais comme $\rho(G_3) < \rho(J_3)$ alors, la méthode qui converge plus rapidement est celle de Gauss-Seidel.

4.

$$J_4 = \begin{pmatrix} 0 & -\frac{6}{7} & -\frac{9}{7} \\ -\frac{4}{5} & 0 & \frac{4}{5} \\ \frac{7}{8} & \frac{3}{8} & 0 \end{pmatrix} \quad G_4 = \begin{pmatrix} 0 & -\frac{6}{7} & -\frac{9}{7} \\ 0 & \frac{24}{35} & \frac{64}{35} \\ 0 & -\frac{69}{140} & -\frac{123}{280} \end{pmatrix}$$

avec $\rho(J_4) = 0,64 < 1, \quad \rho(G_3) = 0,77 < 1$, d'où les deux méthodes convergent $\forall\, X^{(0)} \in \mathbb{R}^3$, mais comme $\rho(J_4) < \rho(G_4)$ alors, la méthode qui converge plus rapidement est celle de Jacobi.

D'autres conditions suffisantes de convergence des méthodes de Jacobi et Gauss-Seidel

Soit le système linéaire $AX = b$ où $A \in \mathbb{M}_n(\mathbb{R})$ inversible, J la matrice d'itération de Jacobi définie par :

$$J = (\alpha_{ij})_{1\leq i,j\leq n} \quad où \quad \alpha_{ij} = \begin{cases} -\frac{a_{ij}}{a_{ii}}, & si\ i \neq j; \\ 0, & si\ i = j. \end{cases}$$

On a vu qu'une condition suffisante pour que le processus itératif converge (vers la solution du système (S)) est que $\|J\|_N$ où N est l'une des trois normes matricielles présentées.

Prenons la norme $\|.\|_\infty$: $\|J\|_\infty = \max\limits_{1\leq i\leq n} \left(\sum\limits_{j=1}^{n} |\alpha_{ij}| \right)$, *alors*

$$\|J\|_\infty < 1 \Longleftrightarrow \sum_{j=1}^{n} |\alpha_{ij}| < 1, \ \ \forall i = 1, 2, \cdots, n$$

et pour i fixé, on a

$$(|\alpha_{i1}| + |\alpha_{i2}| + \cdots + |\alpha_{in}|) < 1$$

ou encore

$$\frac{1}{a_{ii}} (|a_{i1}| + |a_{i2}| + \cdots + |a_{in}|) < 1 \Longrightarrow \sum_{j=1,\, j\neq i}^{n} |a_{ij}| < |a_{ii}|$$

auquel cas on dit que la matrice A est à diagonale dominante stricte.

Remarque 6.13. *On aboutit à la même conclusion pour la méthode de Gauss-Seidel (exercice).*

D'où le résultat suivant :

Théorème 6.4. *Pour que les processus itératifs de Jacobi et de Gauss-Seidel convergent ($\forall X^{(0)} \in \mathbb{R}^n$), il suffit que la matrice A du système $AX = b$ soit à diagonale dominante stricte (D.D.S).*

Proposition 6.1. *Si A est symétrique définie positive, alors le processus itératif de Gauss-Seidel converge $\forall X^{(0)} \in \mathbb{R}^n$.*

Méthode de relaxation

La matrice d'itération est dans ce cas $B = M^{-1}N$ où $M = (\frac{1}{\omega})D - E$, $N = (\frac{1-\omega}{\omega})D + F$, avec ω un paramètre dans $\mathbb{R}^*$.
On remarque que si $\omega = 1$, cette méthode coïncide avec celle de Gauss-seidel; on l'utilise pour accélérer la convergence de Gauss-Seidel :

$$\begin{array}{ll} \omega > 1 & \textit{procédé de sur-relaxation,} \\ \omega = 1 & \textit{méthode de Gauss-Seidel,} \\ \omega < 1 & \textit{procédé de sous-relaxation.} \end{array}$$

En écrivant le processus itératif :

$$X^{(k+1)} = (M^{-1}N)X^{(k)} + M^{-1}b,$$

alors la méthode de relaxation consiste en le schéma suivant :
lors du passage de $X^{(k)}$ à $X^{(k+1)}$ on ne retient pas $X^{(k+1)}$ pour la suite, mais $\omega X^{(k+1)} + (1 - \omega)X^{(k)}$. Autrement dit

$$X^{(k+1)} \longleftarrow X^{(k)} + \omega(\underbrace{X}_{G-S}^{(k+1)} - X^{(k)}),$$

et en injectant cette expression dans l'algorithme de Gauss-Seidel, on trouve l'algorithme suivant :

$$x_i^{(k+1)} = x_i^{(k)} + \omega \left(b_i - \sum_{j=1}^{i-1} a_{ij}\, x_j^{(k+1)} - \sum_{j=i}^{n} a_{ij}\, x_j^{(k)} \right) / a_{ii}, \quad \forall\, i = 1, 2, \cdots, n.$$

Convergence de la méthode de relaxation

Théorème 6.5. *(conditions sur le paramètre ω)*

Cas d'une matrice A quelconque (mais inversible), *une condition nécessaire pour que la méthode de relaxation converge $\forall\, X^{(0)} \in \mathbb{R}^n$) est que $\omega \in]0, 2[$.*

Si A est à diagonale dominante sticte, *alors la condition suffisante de convergence est que $\omega \in]0, 1]$.*

Si A est symétrique définie positive, *alors la condition nécessaire et suffisante de convergence est que $\omega \in]0, 2[$.*

6.4 Exercices

Méthodes directes

Exercice 6.1. *Soient les systèmes d'équations linéaires suivants :*

$$a)\begin{cases} x+y-z=3 \\ 2x-y+3z=0 \\ -x-2y+z=-5 \end{cases} \qquad b)\begin{cases} 3x+y-5z=14 \\ 2x+y-3z=5 \\ -x-y-z=4 \end{cases} \qquad c)\begin{cases} 3x-y+2z=-3 \\ x-y+z=-4 \\ 2x+y-z=-3 \end{cases}$$

$$d)\begin{cases} 2x+6y+2z+8t=16 \\ x+4y+2z+2t=5 \\ x+y+2z+4t=9 \\ x+y+z+t=2 \end{cases} \qquad e)\begin{cases} 10x-y+2z=6 \\ -x+11y-z+3t=25 \\ 2x-y+10z-t=-11 \\ 3y+8t-z=15 \end{cases} \qquad f)\begin{cases} 9x-z+3t=5 \\ 2x+4y+9z+t=1 \\ x+2y+4z-8t=3 \\ x+t-z+4y=7 \end{cases}$$

1. *Résoudre les systèmes linéaires donnés par la méthode d'élimination de Gauss (lorsque cela est possible).*
2. *Résoudre les systèmes linéaires donnés avec la technique du pivot total.*
3. **a)** *Donner la décomposition L.U des matrices des systèmes a) et e).*
   b) *Retrouver les solutions de ces systèmes.*

Exercice 6.2. *Soient α_i, $i = 1, 2, 3, 4$, quatre nombres réels non nuls et distincts deux à deux et soit $A \in \mathbf{M}_4(\mathbb{R})$ définie par :*

$$A = \begin{pmatrix} 1 & \alpha_2/\alpha_1 & \alpha_3/\alpha_1 & \alpha_4/\alpha_1 \\ 1 & 1 & \alpha_3/\alpha_2 & \alpha_4/\alpha_2 \\ 1 & 1 & 1 & \alpha_4/\alpha_3 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

On note par Δ la matrice diagonale $\Delta = diag(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$.

1. *Trianguläriser le système $AX = b$ par la méthode de Gauss ordinaire en justifiant pourquoi ceci est possible.*
2. *Donner alors la décomposition LU de A et en déduire la décomposition $A = LDV$ où $D = diag(U)$.*
3. **a)** *Calculer l'inverse de la matrice L par deux méthodes différentes.*
   b) *Vérifier que $V = \Delta^{-1} L^t \Delta$ et en déduire V^{-1}.*
   c) *Pour $\alpha_i = ia$, $i = 1, 2, 3, 4$ et $a \in \mathbf{R}^*$, écrire A et calculer A^{-1}.*

Exercice 6.3. *On désire résoudre le système linéaire $HX = K$ où K est un vecteur donné et H est une matrice tridiagonale inversible.*

$$H = \begin{pmatrix} a_1 & b_1 & 0 & 0 & \dots & 0 \\ c_2 & a_2 & b_2 & 0 & \dots & 0 \\ 0 & c_3 & a_3 & b_3 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & c_{n-1} & a_{n-1} & b_{n-1} \\ 0 & 0 & \dots & 0 & c_n & a_n \end{pmatrix}, \qquad K = \begin{pmatrix} k_1 \\ k_2 \\ \vdots \\ \vdots \\ \vdots \\ k_n \end{pmatrix}$$

1. *En appliquant la méthode d'élimination de Gauss ordinaire (les pivots étant supposées tous non nuls), écrire la forme générale de la matrice triangulaire obtenue, ainsi que le vecteur constant.*
2. *Résoudre ce dernier système.*

Exercice 6.4. *On considère le système $AX = b$ défini par :*

$$A = \begin{pmatrix} \alpha & 1 & 2 \\ 1 & 3 & 1 \\ 2 & 1 & 3 \end{pmatrix} \quad \alpha \in \mathbb{R} \quad et \quad b \in \mathbb{R}^3.$$

1. *Pour quelles valeurs du paramètre α, la matrice A est-elle symétrique définie positive ?*
2. *On pose $\alpha = 0$. Quelle méthode directe peut-on utiliser pour résoudre le système $AX = b$? justifier votre réponse.*
3. *On considère $\alpha = 1$.*
   - **i)** *Résoudre les systèmes linéaires $AY^{[i]} = e_i$ $(i = 1, 2, 3)$ où e_i est le ième vecteur de la base canonique de $\mathbb{R}^3$.*
   - **ii)** *Donner la décomposition LU de A, en justifiant son existence.*
   - **iii)** *En déduire A^{-1} puis U^{-1}.*

Exercice 6.5. *Soit $A = (a_{ij})_{1 \le i,j \le 3}$ la matrice définie par :*

$$a_{ij} = \begin{cases} i(4i+5), & si \ \ i = j, \\ 3\min(i,j), & si \ \ i \neq j. \end{cases}$$

1. *Donner la décomposition $R.R^t$ de A à l'aide de l'algorithme de Cholesky en justifiant son existence.*
2. *En déduire la solution du système $AX = b$ où $b = (1,\ 1,\ 0)^t$.*
3. *Comment peut-on résoudre, avec un minimum d'opérations, le système $A^2X = b$? Justifier votre réponse.*

Méthodes itératives

Exercice 6.6. *On considère la matrice A et le vecteur b définis par :*

$$A = \begin{pmatrix} \frac{1}{8} & 0 & \frac{1}{4} \\ 0 & \frac{1}{3} & 0 \\ \frac{1}{2} & 0 & \frac{1}{8} \end{pmatrix}, \qquad b = \begin{pmatrix} \frac{3}{8} \\ \frac{1}{3} \\ \frac{5}{8} \end{pmatrix}$$

1. **a)** *Résoudre le système $AX = b$ par la méthode d'élimination de Gauss.*
   b) *En déduire le déterminant de A.*
   c) *Peut-on utiliser les méthodes LU et Cholesky pour retrouver la solution du système $AX = b$?*
2. **a)** *Ecrire le processus itératif de Jacobi correspondant au système $AX = b$.*
   b) *Soit $X^{(0)} = (\frac{1}{2}, 0, 1)^t$, calculer $X^{(1)}$ et $X^{(2)}$. Donner l'erreur absolue commise dans chacun des cas en norme $\|.\|_\infty$.*
   c) *Etudier la convergence de cette méthode.*
3. *Peut-on réécrire le système $AX = b$ de manière à assurer la convergence de la méthode de Jacobi $\forall\, X^{(0)} \in \mathbb{R}^3$?*

Exercice 6.7. *Soit le système linéaire $AX = b$ où* $A = \begin{pmatrix} 2 & \frac{2}{3} & \frac{4}{3} \\ 1 & 2 & \frac{2}{3} \\ 3 & 1 & 4 \end{pmatrix}$ *et* $b = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$

1. *On note par $a_{kk}^{(k)}$ le pivot de la $k^{ième}$ étape d'élimination de Gauss et $A_{[k]}$ la sous matrice d'ordre k de A .*
   i) *Montrer l'équivalence :* $\det(A_{[k]}) \neq 0,\ k = 1, ..., n \Longleftrightarrow a_{kk}^{(k)} \neq 0,\ k = 1, ..., n$
   ii) *Peut-on utiliser les méthodes LU et Cholesky pour résoudre le système $AX = b$?*
2. *On considère le processus itératif de Jacobi associé au système $AX = b$:*

$$X^{(k+1)} = JX^{(k)} + C.$$

   i) *Déterminer J et C.*
   ii) *Calculer le polynôme caractéristique P de J.*
   iii) *Séparer les racines de P.*
   vi) *Que peut-on dire sur la convergence de la méthode de Jacobi.*
3. **i)** *Montrer que si tout coefficient (α_{ij}) de la matrice de Jacobi J associée à un système linéaire d'ordre n vérifie : $|\alpha_{ij}| < \frac{1}{n}$, alors la méthode de Jacobi converge pour tout vecteur initial $X^{(0)} \in \mathbb{R}^n$.*
   ii) *En déduire que la méthode de Jacobi associée à la matrice $B = (b_{ij})_{0 \leq i,j \leq 3}$ où*

$$b_{ij} = \begin{cases} a_{ij}, & si\ i \neq j; \\ 8a_{ii}, & si\ i = j, \end{cases} \quad converge\ \ \forall X^{(0)} \in \mathbb{R}^3.$$

Exercice 6.8.
I) Soit le système linéaire $AX = b$ où $A \in M_n(\mathbb{R})$ et $b \in \mathbb{R}^n$. Montrer que si le processus itératif associé à ce système

$$\begin{cases} X^{(0)} \in \mathbb{R}^n, \\ X^{(k+1)} = BX^{(k)} + C,\ k \geq 0. \end{cases}$$

converge vers $\overline{X}$ la solution exacte du système $AX = b$, alors

i) $\|\overline{X} - X^{(k)}\|_* \leq \frac{\|B\|_*}{1-\|B\|_*}\|X^{(k)} - X^{(k-1)}\|_*$ *(estimation pas à pas)*

ii) $\|\overline{X} - X^{(k)}\|_* \leq \frac{\|B\|_*^k}{1-\|B\|_*}\|X^{(1)} - X^{(0)}\|_*$ *(estimation à priori)*

II) Soit le système d'équations linéaires $(1) : Ax = b$ *défini par*

$$A = \begin{pmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{pmatrix}, \qquad et \quad b = (4,\ 4,\ 2)^t.$$

1. *Vérifier que les processus itératifs de Jacobi et Gauss-Seidel associés au système linéaire (1) convergent, quelque soit le vecteur initial de $\mathbb{R}^3$. On désigne par $\overline{X}$ la solution exacte du système* (1).
2. *Dans cette partie, on considère la méthode de Jacobi associée au système (1)*

$$\begin{cases} X^{(0)} \in \mathbb{R}^3, \\ X^{(k+1)} = JX^{(k)} + f,\ k \geq 0, \end{cases} \qquad où\ \ J\ et\ f\ sont\ à\ déterminer$$

On pose $X^{(0)} = (0\ \ 2\ \ 0)^t$.

a) *Calculer $X^{(2)}$ et estimer l'erreur $\|X^{(2)} - \overline{X}\|_\infty$.*

b) *Quel est le nombre suffisant d'itérations n_1 à partir duquel on a : $\|X^{(n)} - \overline{X}\|_\infty \leq 10^{-2}$?*

c) *Montrer que : $\forall\, n \in \mathbb{N},\ \ x_1^{(n)} + x_2^{(n)} = 2$ et $x_3^{(n)} = 0$.*

d) *En déduire la solution exacte $\overline{X}$ du système (1).*

3. *Dans cette partie, on considère la méthode de Gauss-Seidel associée au système (1)*

$$\begin{cases} Y^{(0)} \in \mathbb{R}^3, \\ Y^{(k+1)} = GY^{(k)} + g,\ k \geq 0, \end{cases} \qquad où\ \ G\ et\ g\ sont\ à\ déterminer$$

On pose $Y^{(0)} = (0,\ \ 2,\ \ 0)^t$.

a) *Calculer $Y^{(2)}$ et estimer l'erreur $\|Y^{(2)} - \overline{X}\|_\infty$.*

b) *Quel est le nombre suffisant d'itérations n_1 à partir duquel on a $\|Y^{(n)} - \overline{X}\|_\infty \leq 10^{-2}$?*

4. *Commenter les résultats obtenus.*

Exercice 6.9. *On considère la matrice A et le vecteur b définis par :*

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 2 & 1 \\ 1 & 1 & -\frac{1}{2} \end{pmatrix}, \qquad b = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$$

1. *Résoudre le système $AX = b$ par la méthode d'élimination de Gauss.*

2. *Soit la matrice* $B = \begin{pmatrix} 0 & -1 & -1 \\ 0 & 0 & -\frac{1}{2} \\ 2 & 2 & 0 \end{pmatrix}$

   a) *Quelle relation y-a-t-il entre les matrice* A *et* B*? Ecrire le processus itératif correspondant.*

   b) *Montrer que* $P(\lambda) = \det(B - \lambda I_3) = -\lambda^3 - 3\lambda + 1$.

   c) *Séparer graphiquement les racines de* P *et vérifier que* P *admet une racine séparée* $\overline{\lambda} \in [\frac{1}{10}, \frac{1}{3}]$. *Conclure.*

   d) *Comment peut-on déterminer l'approximation initiale* $X^{(0)}$ *pour retrouver la solution exacte du système* $AX = b$ *(de la question 1) après une étape ?*

Exercice 6.10. *On considère le système d'équations linéaires* $(S) : Ax = b$ *défini par*

$$A = \begin{pmatrix} 2 & \alpha & 1 \\ \alpha & 1 & \beta \\ 1 & \beta & 2 \end{pmatrix}, \qquad \text{où } \alpha, \beta \in \mathbb{R}^*, \quad \text{vérifient } \alpha\beta = 1 \text{ et } b \in \mathbb{R}^3.$$

1. **a)** *Calculer* $P(\lambda) = det(J - \lambda I_3)$ *où* J *est la matrice d'itération de Jacobi associée à* A.

   b) *Montrer que* $P(-1) < 0$ *et étudier la convergence de la méthode de Jacobi appliquée à (S) pour tout choix de* $x^{(0)} \in \mathbb{R}^3$.

   c) *La matrice* A *peut-elle être une matrice D.D.S ?*

2. *On considère la méthode de Gauss-Seidel associée au système (S)*

$$\begin{cases} X^{(0)} \in \mathbb{R}^3, \\ X^{(k+1)} = GX^{(k)} + g. \end{cases}$$

   Calculer G *et* g *; en déduire l'ensemble des valeurs de* α, β *pour lesquelles la méthode de Gauss-Seidel appliquée à (S) converge pour tout choix de* $x^{(0)} \in \mathbb{R}^3$.

Exercice 6.11. *Soit à résoudre le système d'équations linéaires* $Ax = b$ *défini par*

$$A = \begin{pmatrix} \beta & -1 & 0 & 0 \\ -1 & \beta & -1 & 0 \\ 0 & -1 & \beta & -1 \\ 0 & 0 & -1 & \beta \end{pmatrix}, \qquad \beta \in \mathbb{R}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{pmatrix}$$

1. *Que peut-on dire de la convergence des méthodes de Jacobi et Gauss-Seidel sans calculer les matrices d'itération si*

   (a) $\beta = 3$;

   (b) $\beta = 2$.

2. *Dans le cas où* $\beta = 3$, *définissons* $b =^t (1, 0, 2, 28)$.

   i) *Calculer la solution exacte du système* $Ax = b$, *en utilisant la méthode de Gauss.*

   ii) *Calculer les trois premiers itérés des méthodes de Jacobi et Gauss-Seidel en partant de* $x^{(0)} =^t (0, 0, 0, 0)$.

   iii) *Comparer ces itérés à la solution exacte. les résultats sont-ils cohérents avec ceux de la question 1-(a).*

iv) *S'il y a convergence des deux méthodes, déterminer celle qui converge le plus vite. Pourquoi ?*

Exercice 6.12. *On considère la matrice A et le vecteur b définis par :*

$$A = \begin{pmatrix} 1 & \frac{1}{3} & 0 \\ \frac{1}{3} & 1 & \frac{1}{3} \\ 0 & \frac{1}{3} & 1 \end{pmatrix}, \qquad b = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$$

1. *Vérifier que les processus itératifs de Jacobi et de Gauss-Seidel associés au système linéaire $AX = b$ convergent, pour tout vecteur initial de $\mathbb{R}^3$.*
2. *On considère le processus itératif de Jacobi associé au système (1) :*

$$\begin{cases} X^{(0)} \in \mathbb{R}^3, \\ X^{(k+1)} = JX^{(k)} + f, \ k \geq 0, \end{cases} \qquad \text{où } J \text{ et } f \text{ sont à déterminer .}$$

On pose $X^{(0)} = 0_{\mathbb{R}^3}$.

a) *Calculer $X^{(2)}$ et estimer l'erreur $\|X^{(2)} - \overline{X}\|_1$, où $\overline{X}$ est la solution exacte du système (1).*

b) *Calculer $\rho(J)$ le rayon spectral de J.*

3. *On considère le processus itératif de Gauss-Seidel associé au système $AX = b$:*

$$\begin{cases} Y^{(0)} \in \mathbb{R}^3, \\ Y^{(k+1)} = GY^{(k)} + g, \ k \geq 0, \end{cases} \qquad \text{où } G \text{ et } g \text{ sont à déterminer .}$$

a) *Calculer $\rho(G)$ le rayon spectral de G. En déduire que $\rho(G) = (\rho(J))^2$.*

b) *On pose $Y^{(0)} = 0_{\mathbb{R}^3}$. Montrer que : $\forall\, k \in \mathbb{N}, \ 3y_3^{(k)} + y_2^{(k)} = 0$.*

c) *En déduire la solution exacte $\overline{X}$ du système $AX = b$.*

Exercice 6.13. 1. *Pour une matrice carrée quelconque A montrer que $\rho(A) \leq \|A\|$ pour toute norme matricielle.*

2. *On suppose A symétrique définie positive. Montrer que $\rho(A) \leq \|A\|_2$.*
3. *Soit*

$$A = \begin{pmatrix} 1 & \alpha & \alpha \\ \alpha & 1 & \alpha \\ \alpha & \alpha & 1 \end{pmatrix}, \quad \alpha \in \mathbb{R}.$$

a) *Pour quelles valeurs de α la matrice A est-elle définie positive ?*

b) *Ecrire la matrice J de l'itération de Jacobi.*

c) *Pour quelles valeurs de α la matrice de Jacobi converge-t-elle ?*

d) *Ecrire la matrice J de l'itération de Gauss-Seidel.*

e) *Calculer $\rho(G)$. pour quelles valeurs de α cette méthode converge-t-elle plus vite que celle de Jacobi ?*

Chapitre 7

Calcul des valeurs et vecteurs propres d'une matrice

7.1 Position du problème

Soit A une matrice carrée à coefficients dans $\mathbb{K}$ ($\mathbb{K} = \mathbb{R}$ ou $\mathbb{C}$). On cherche les valeurs et vecteurs propres $\lambda \in \mathbb{C}$, $X \in (\mathbb{K}^n)^*$ (respectivement) de A tels que :

$$\lambda \text{ valeur propre de } A \Longleftrightarrow \exists\, X \in (\mathbb{K}^n)^* :\ AX = \lambda X.$$

$$X \text{ vecteur propre de } A \Longleftrightarrow \exists \lambda \in \mathbb{C} :\ AX = \lambda X.$$

Soit P le polynôme caractéristique de A. Alors

$$\lambda \text{ valeur propre de } A \Longleftrightarrow P(\lambda) = \det(A - \lambda I) = 0,$$

c'est à dire λ est une racine du polynôme caractéristique P.
Les méthodes de recherche des valeurs et vecteurs propres sont classées en deux type :

Méthodes directes : *elles procèdent, généralement, en trois étapes principales :*

i) *recherche des coefficients de P,*

ii) *calcul des racines de P,*

iii) *obtention de vecteurs propres.*

Méthodes itératives : *ce sont des méthodes qui ne consistent pas en la recherche du polynôme caractéristique P.*

7.2 Méthodes directes

1. Calcul direct de $P(\lambda) = \det(A - \lambda I)$: Pour $n = 2, 3, 4$, le déterminant d'une matrice est, relativement, facile à calculer manuellement. Pour n (assez) grand, il faut utiliser un ordinateur, qui fait l'évaluation du déterminant d'une matrice d'ordre n en coût de $O(n!)$ (complexité exponentielle). Par exemple, si $n = 20$, un ordinateur très puissant mettra des années pour calculer $det(A)$.
2. Méthode de Krylov : *Rappelons d'abord le théorème important suivant :*

Théorème 7.1. *(Théorème de Cayley-Hamiltion) Si l'on note par $P(\lambda)$ le polynôme caractéristique d'une martice carrée d'ordre n A. Alors $P(A) = 0_{M_n(\mathbb{K})}$.*

On a

$$\begin{aligned} P(\lambda) &= (-1)^n\left(\lambda^n - Tr(A)\lambda^{n-1} + \alpha_{n-2}\lambda^{n-2} + \alpha_{n-3}\lambda^{n-3}\ldots + \det(A)\right) \\ &= (-1)^n\left(\lambda^n + \sum_{k=0}^{n-1}\alpha_k\,\lambda^k\right) \quad o\grave{u}\ \alpha_0 = \det(A),\ \alpha_{n-1} = -Tr(A). \end{aligned}$$

Cette méthode permet le calcul des coefficients du polynôme P comme suit :
D'une part, d'après le théorème précédent, on a

$$P(A) = (-1)^n[A^n + \sum_{k=0}^{n-1}\alpha_k\,A^k] = 0_{M_n(\mathbb{K})} \Longrightarrow A^n = \sum_{k=0}^{n-1}\beta_k\,A^k, \quad o\grave{u}\ \beta_k = -\alpha_k.$$

D'autre part, soient Y_0 un vecteur non nul de $\mathbb{R}^n$, B une matrice carrée d'ordre n définie par : $B = [Y_0, AY_0, A^2Y_0, \ldots, A^{n-1}Y_0]$, *et* $X = \begin{pmatrix} \beta_0 \\ \beta_2 \\ \vdots \\ \beta_{n-1} \end{pmatrix}$. *On trouve,*

$$BX = \alpha_0\,Y_0 + \beta_1\,AY_0 + \ldots + \beta_{n-1}\,A^{n-1}\,Y_0 = \left(\sum_{k=0}^{n-1}\beta_k\,A^k\right)Y_0 = A^nY_0.$$

C'est à dire le vecteur X qui contient les symétriques des coefficients du polynôme P vérifie le système linéaire $BX = b$ où $b = A^nY_0$.

Remarque 7.1. *1. Il faut choisir Y_0 de sorte que B soit régulière.*

2. La résolution du système linéaire obtenu peut se faire par l'une des méthodes étudiées dans le chapitre précédent.

Exemple 7.1. *Soient* $A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$, $Y_0 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$. *Alors,* $B = [Y_0, AY_0] = \left[\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 3 \\ 7 \end{pmatrix}\right]$ *et* $b = A^2Y_0 = \begin{pmatrix} 17 \\ 37 \end{pmatrix}$. *Donc le vecteur $X = (\beta_0, \beta_1)^t$ vérifiant le sysème linéaire :*

$$\begin{pmatrix} 1 & 3 \\ 1 & 7 \end{pmatrix}\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} = \begin{pmatrix} 17 \\ 37 \end{pmatrix} \Longrightarrow \begin{cases} \beta_0 = 2 \\ \beta_1 = 5 \end{cases} \Longrightarrow \begin{cases} \alpha_0 = -2 \\ \alpha_1 = -5 \end{cases}$$

d'où $P(\lambda) = \lambda^2 - 5\lambda - 2$.

Remarque 7.2. *D'autres méthodes de recherche des coefficients de P existent(Faddeev, Leverrier,...), mais l'inconvénient des méthodes directes réside dans le fait qu'il faut chercher les racines de P, ce qui nécessite l'emploi de méthodes de résolution d'équations non linéaires.*

7.3 Méthodes itératives

7.3.1 Méthode de la puissance itérée

Cette méthode permet d'obtenir par itération la valeur propre de plus grand module d'une matrice A ainsi que le vecteur propre correspondant.
Soit A une matrice réelle d'ordre n. On suppose, pour simplifier, que toute les valeurs propres $\lambda_1, \ldots, \lambda_n$ de A sont distinctes, de sorte que les vecteurs propres correspondants $V_1, \ldots, V_n$ forment une base de $\mathbb{R}^n$. Posons

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| \quad (i.e.\ \lambda_1 \in \mathbb{R}).$$

Soit $X^{(0)}$ un vecteur quelconque dans $\mathbb{R}^n_*$, alors

$$\exists\, \alpha_i \in \mathbb{R},\ i = \overline{1,n} \text{ tel que } X^{(0)} = \sum_{i=1}^{n} \alpha_i V_i.$$

Appliquons la matrice A au vecteur $X^{(0)}$, on obtient

$$AX^{(0)} = \sum_{i=1}^{n} \alpha_i A V_i = \sum_{i=1}^{n} \alpha_i\, \lambda_i\, V_i.$$

Appliquons de nouveau A

$$A^2X^{(0)} = \sum_{i=1}^{n} \alpha_i A^2 V_i = \sum_{i=1}^{n} \alpha_i\, \lambda_i^2\, V_i,$$

et ainsi de suite, on aura

$$\begin{aligned} A^kX^{(0)} = \sum_{i=1}^{n} \alpha_i \lambda_i^k V_i &= \alpha_1 \lambda_1^k V_1 + \sum_{i=2}^{n} \alpha_i\, \lambda_i^k\, V_i \\ &= \lambda_1^k \left(\alpha_1 V_1 + \sum_{i=2}^{n} \alpha_i \left(\tfrac{\lambda_i}{\lambda_1}\right)^k V_i \right). \end{aligned}$$

On voit que lorsque k tend vers $+\infty$, $\left(\alpha_1 V_1 + \sum_{i=2}^{n} \alpha_i \left(\tfrac{\lambda_i}{\lambda_1}\right)^k V_i \right)$ tend vers $\alpha_1\, V_1$.
Par conséquent, au bout d'un nombre suffisamment grand d'itérations, on peut écrire

$$X^{(k)} = A^kX^{(0)} \simeq \lambda_1^k \alpha_1 V_1,$$

d'où à l'itération suivante :

$$X^{(k+1)} = AX^{(k)} = A^{k+1}X^{(0)} \simeq \lambda_1^{k+1}\alpha_1 V_1 \simeq \lambda_1 X^{(k)}.$$

Par conséquent,

$$|\lambda_1| \simeq \frac{\|X^{(k+1)}\|}{\|X^{(k)}\|}.$$

***En pratique,** on utilise cette procédure comme suit :*
Dans le but d'éviter la manipulation (éventuelle) de nombres assez grands, on a intérêt à normaliser chaque vecteur $X^{(k)}$ pour $k \geq 1$, c'est à dire qu'on n'applique pas la matrice A au vecteur $X^{(k)}$, mais au vecteur $Y^{(k)} = \frac{X^{(k)}}{\|X^{(k)}\|}$. Donc si on suppose que $|x_p^{(k)}| = \max_i |x_i^{(k)}|$,

- *la $p^{ième}$ composante de $Y^{(k)}$ vaut 1,*
- *la valeur propre λ_1 estimée à la $k^{ième}$ itérations est la $p^{ième}$ composante du vecteur $X^{(k)}$.*
  ***En effet ;** quand k est assez grand et la $p^{ième}$ composante de V_1 est non nulle, on aura*

$$\frac{x_p^{(k+1)}}{x_p^{(k)}} \simeq \frac{\lambda_1^{k+1}\alpha_1 v_{1,p}}{\lambda_1^k \alpha_1 v_{1,p}} = \lambda_1.$$

L'algorithme relatif à la méthode de la puissance itérée se présente comme suit :

1. *Etant données A, $X^{(0)} = (x_1^{(0)}, \ldots, x_n^{(0)})$ vecteur arbitraire dans $\mathbb{R}_*^n$, ϵ une précision sur λ.*
2. *On pose $Y^{(0)} = X^{(0)}$, pour $(k \geq 1)$, on calcule le vecteur $X^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})$ tel que*

$$X^{(k)} = A\,Y^{(k-1)},$$

puis le vecteur

$$Y^{(k)} = \left(\frac{x_1^{(k)}}{x_p^{(k)}}, \ldots, \frac{x_n^{(k)}}{x_p^{(k)}}\right), \quad où \;\; |x_p^{(k)}| = \max_i |x_i^{(k)}|.$$

Alors,

$$\lambda_1^{(k)} = x_p^{(k)} \quad et \quad V_1^{(k)} = Y^{(k)}.$$

3. *Répéter l'étape (2) tant que*

$$|\lambda_1^{(k)} - \lambda_1^{(k-1)}| \geq \epsilon.$$

4.

$$\lambda_1 \simeq \tilde{x}_p^{(k)} \quad et \quad V_1 \simeq \tilde{Y}^{(k)}.$$

Remarque 7.3. *Lorsque l'algorithme de la méthode de la puissance itérée est appliqué à la matrice A^{-1}, il détermine la plus petite valeur propre de A, c'est la méthode de la puissance itérée inverse.*

Exemple 7.2. *Soit*

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

Les valeurs propres de la matrice A sont $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1$ et les vecteurs propres correspondants sont $V_1 = (0,0,1)$, $V_2 = (1,1,1)$ et $V_3 = (1,0,0)$. Montrons que la méthode de la puissance permet de recalculer λ_1 et V_1. Posons $X^{(0)} = (0,1,0) \Longrightarrow Y^{(0)} = (0,1,0)$ alors

$$X^{(1)} = AY^{(0)} = (1,2,-1) \Longrightarrow p = 2,\ \lambda^{(1)} = 2 \ et \ Y^{(1)} = (\frac{1}{2}, 1, \frac{-1}{2})$$

$$X^{(2)} = AY^{(1)} = (\frac{3}{2}, 2, \frac{-5}{2}) \Longrightarrow p = 3,\ \lambda^{(2)} = \frac{-5}{2} \ et \ Y^{(2)} = (\frac{-3}{5}, \frac{-4}{5}, 1)$$

$$X^{(3)} = AY^{(2)} = (\frac{-7}{5}, \frac{-8}{5}, \frac{19}{5}) \Longrightarrow p = 3,\ \lambda^{(3)} = \frac{19}{5} \ et \ Y^{(3)} = (\frac{-7}{19}, \frac{-8}{19}, 1)$$

$$X^{(4)} = AY^{(3)} = (\frac{-15}{19}, \frac{-16}{19}, \frac{65}{19}) \Longrightarrow p = 3,\ \lambda^{(4)} = \frac{65}{19} \ et \ Y^{(4)} = (\frac{-15}{65}, \frac{-16}{65}, 1)$$

$$X^{(5)} = AY^{(4)} = (\frac{-31}{65}, \frac{-32}{65}, \frac{211}{65}) \Longrightarrow p = 3,\ \lambda^{(5)} = \frac{211}{65} \ et \ Y^{(4)} = (\frac{-31}{211}, \frac{-32}{211}, 1)$$

$$X^{(6)} = AY^{(5)} = (\frac{-63}{211}, \frac{-64}{211}, \frac{665}{211}) \Longrightarrow p = 3,\ \lambda^{(6)} = \frac{665}{211} \ et \ Y^{(4)} = (\frac{-63}{665}, \frac{-64}{665}, 1).$$

Remarquons que la suite $(\lambda^{(k)})$ converge vers $\lambda_1 = 3$ et $(Y^{(k)})$ converge vers $V_1 = (0,0,1)$.

Remarque 7.4. *Si on choisit un vecteur propre de la matrice A comme vecteur initial $X^{(0)}$ de la méthode, on risque de ne pas avoir la plus grande des valeurs propres.*

7.3.2 Méthode de Rutishauser

Principe de la méthode : *Construction d'une suite de matrices semblables à A, et qui converge vers une matrice triangulaire (ou diagonale si A est symétrique).*

Rappels :

1. *Une matrice B est dite semblable à une matrice A, si elle existe une matrice inversible P telle que : $B = P^{-1}AP$.*
2. *Deux matrices semblables ont mêmes valeurs propres.*
3. *Les valeurs propres d'une matrice triangulaire (ou diagonale) sont situées sur sa diagonale.*
4. *Une matrice A est dite diagonalisable si toutes les valeurs propres de cette matrice sont simples.*
5. *Si toutes les sous matrices principales d'une matrice A sont inversibles, alors elle se décompose sous la forme $A = LU$ où L est triangulaire inférieure unitaire et U est triangulaire supérieure.*

La méthode de Rutishauser utilise cette décomposition plusieurs fois de suite. On supposera, dans le cadre de cette méthode, que toutes les décompositions LU sont possibles.
Soit $A_1 = A = LU = L_1U_1$,
A_2 est calculée comme suit : $A_2 = U_1L_1$, qu'on décompose encore en $A_2 = L_2U_2$,
de même : $A_3 = U_2L_2$, qu'on décompose encore en $A_3 = L_3U_3$,
et ainsi de suite, on obtient une suite de matrices $(A_k)_{k\in\mathbb{N}^*}$ définie par :

$$\begin{cases} A_1 = A = L_1\,U_1, \\ A_{k+1} = U_kL_k, \quad \text{qu'on décompose encore en } A_{k+1} = L_{k+1}\,U_{k+1}. \end{cases}$$

Réécrivons

$$A_2 = U_1L_1 = (L_1^{-1}L_1)U_1L_1 = L_1^{-1}(L_1U_1)L_1 = L_1^{-1}A_1L_1,$$

ce qui implique que A_2 est semblable à A. De plus, on a

$$A_{k+1} = U_kL_k = L_k^{-1}(L_kU_k)L_k = L_k^{-1}A_kL_k,$$

et

$$A_{k+1} = L_k^{-1}(L_kU_k)L_k = L_k^{-1}(U_{k-1}L_{k-1})L_k = L_k^{-1}L_{k-1}^{-1}A_{k-1}L_{k-1}L_k,$$

donc de proche en proche, on obtient

$$A_{k+1} = L_k^{-1}L_{k-1}^{-1}\dots L_2^{-1}L_1^{-1}\,A\,L_1L_2\dots L_{k-1}L_k.$$

Posons $\mathcal{L}_k = L_1L_2\dots L_k$, alors on a $\mathcal{L}_k$ et $\mathcal{L}_k^{-1}$ sont des matrices triangulaires inférieures unitaires pour tout $k \in \mathbb{N}^*$ et $A_{k+1} = \mathcal{L}_k^{-1}A\mathcal{L}_k$.
Donc d'une part, on a A_k est semblable à A pour tout $k \geq 2$,
et d'autre part,

$$\lim_{k\to+\infty} A_k = \lim_{k\to+\infty} \mathcal{L}^{-1}A\mathcal{L}, \tag{7.1}$$

où $\mathcal{L}$ et $\mathcal{L}^{-1}$ sont des matrices triangulaires inférieures unitaires.
Or si A est diagonalisable, elle existe une matrice P inversible telle que $P^{-1}AP = D$ où D est une matrice diagonale, alors on a A_k est semblable à A pour tout $k \in \mathbb{N}^*$ et A est semblable à D, ce qui entraîne que A_k est semblable à D, $\forall\, k \in \mathbb{N}^*$. Faisons tendre k vers $+\infty$, elle existe une matrice Q inversible telle que

$$\lim_{k\to+\infty} A_k = Q^{-1}DQ = Q^{-1}P^{-1}APQ = (PQ)^{-1}APQ.$$

D'après (7.1), on tire que les matrices P et Q sont triangulaires, ce qui implique que la suite $(A_k)_{k\in\mathbb{N}^*}$ converge vers une matrice triangulaire, elle a donc les mêmes valeurs propres que D et donc que A.

Dans la pratique, on se donne ϵ (très petit>0) puis on construit la suite (A_k) jusqu'à ce que $\exists\, k_0 \in \mathbb{N}^*$ tel que A_{k_0} soit approximativement triangulaire à ϵ près.

Chapitre 8

Résolution des équations non linéaires

8.1 Position du problème

Soit $f : [a, b] \subseteq \mathbb{R} \longrightarrow \mathbb{R}$ une fonction au moins continue sur $[a, b]$. On cherche les racines de f c'est à dire les points $x \in [a, b]$ tels que $f(x) = 0$.
Ce problème est important en particulier en optimisation où l'on cherche à minimiser (ou maximiser) une fonction, car on cherche les points où la dérivée s'annule.

Si $f(a).f(b) < 0$, alors par le théorème des valeurs intermédiaires, il existe au moins une racine α de f dans l'intervalle $[a, b]$. Si de plus, f est strictement monotone alors f est une bijection et cette racine α est unique dans $[a, b]$.

Nous supposerons que α est l'unique racine de f dans l'intervalle $[a, b]$.
En général, les méthodes de résolution numérique de $f(x) = 0$ sont des méthodes itératives. Elles consistent à construire une suite $(x_n)_n$ convergente (le plus rapidement possible) vers α.

8.1.1 Séparation des racines

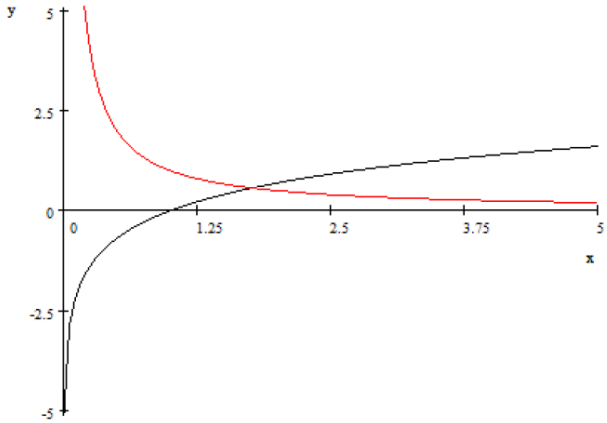
La plupart des méthodes de résolution nécessitent la séparation des racines, c'est à dire celles qui consistent en la détermination d'un (ou des) intervalle(s) fermé(s) et borné(s) de $[a, b]$ dans le(s)quel(s) f admet une et une seule racine.
La séparation des racines s'effectue, en général, en utilisant deux types de méthodes :

i.Méthode graphique : ***Soit*** *on trace (expérimentalement ou par étude de 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 f.*

Remarque 8.1. *On choisit souvent f_1 et f_2 de façon que leurs courbes soient des courbes connues.*

Exemple 8.1. *La fonction f définie par : $f(x) = \ln x - \frac{1}{x},\ x \in]0, +\infty[$ a une racine unique dans l'intervalle $[\frac{5}{4}, \frac{5}{2}]$. En effet ; posons $f_1(x) = \ln x$ et $f_2(x) = \frac{1}{x}$. Alors*
$f(x) = 0 \Leftrightarrow f_1(x) = f_2(x)$ et d'après le graphe

$$G(f_1) \cap G(f_2) = \left\{ (\alpha, f(\alpha)),\ \alpha \in [\frac{5}{4}, \frac{5}{2}] \right\}.$$



ii. Méthode de balayage : *On considère une suite croissante finie* $\{x_i,\ i = 0, 1, \ldots, n\}$ *de valeurs de* x *réparties sur l'intervalle* $[a, b]$. *Si* f *est continue et* $f(x_i)f(x_{i+1}) < 0$, *alors il existe entre* x_i *et* x_{i+1} *au moins une racine de* f *(c'est le théorème des valeurs intermédiaires).*

Exemple 8.2. *Le polynôme* P *défini par :* $P(x) = x^4 - x^3 - x^2 + 7x - 6$ *a au moins deux racines réelles* $\alpha_1 \in]-3, -1[$, $\alpha_2 \in]0, 2[$ *car :* $P(-4) = 270, p(-3) = 72, P(-1) = -12$, $P(0) = -6$, $P(2) = 12$ *et* $P(3) = 60, \ldots,$ *ect.*

8.2 Méthode de dichotomie (ou de la bissection)

Cette méthode est utilisée pour approcher les racines d'une fonction continue $f : \mathbb{R} \to \mathbb{R}$. *S'il existe* $a, b\,(a < b)$ *avec* $f(a)f(b) < 0$, *on sait alors qu'il existe au moins une racine* α *de* f *dans l'intervalle* $]a, b[$. *Posons* $a_0 = a$, $b_0 = b$, $I_0 =]a_0, b_0[$ *et* $x_0 = \frac{a_0+b_0}{2}$.
Pour $n \geq 1$, *on choisit le sous-intervalle* $I_n =]a_n, b_n[$ *de l'intervalle* $]a_{n-1}, b_{n-1}[$ *de la façon suivante :*

a) *posons* $x_{n-1} = \frac{a_{n-1}+b_{n-1}}{2}$,

b) *si* $f(x_{n-1}) = 0$, *alors* $\alpha = x_{n-1}$ *et la méthode est terminée,*

c) *si* $f(x_{n-1}) \neq 0$, *alors*

- **i)** *si* $f(a_{n-1}).f(x_{n-1}) < 0$, *alors* $\alpha \in]a_{n-1}, x_{n-1}[$ *et on pose* $a_n = a_{n-1}$, $b_n = x_{n-1}$;
- **ii)** *si* $f(x_{n-1}).f(b_{n-1}) < 0$, *alors* $\alpha \in]x_{n-1}, b_{n-1}[$ *et on pose* $a_n = x_{n-1}$, $b_n = b_{n-1}$;

d) *on définit l'intervalle* $I_n =]a_n, b_n[$; *on augmente* n *de 1 et on recommence du point a).*

On obtient donc une suite de valeurs approchées de α :

$$x_0 = \frac{a_0 + b_0}{2}, x_1 = \frac{a_1 + b_1}{2}, \ldots, x_n = \frac{a_n + b_n}{2}, \ldots$$

telle que

$$|x_n - \alpha| \leq \frac{b - a}{2^{n+1}}. \tag{8.1}$$

Exercice 8.1. *Retrouver l'inégalité (8.1).*

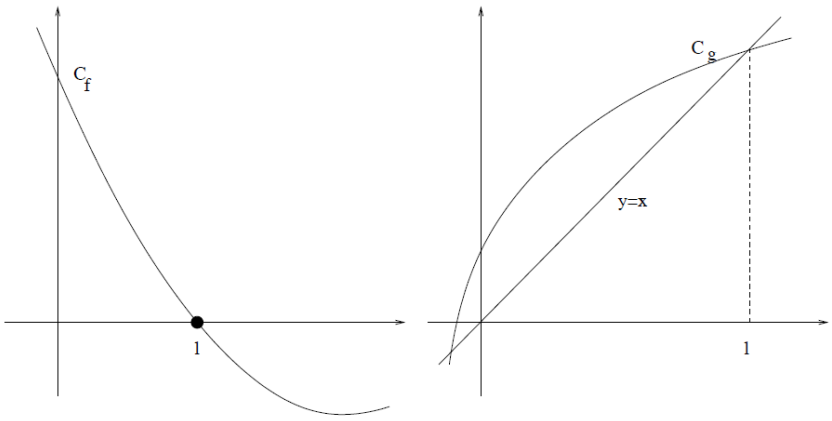
Remarque 8.2. *L'inégalité (8.1) nous permet d'estimer le nombre suffisant d'itérations* n *pour approcher* α *avec une précision donnée* ϵ, *il suffira de résoudre l'inégalité :* $\frac{b-a}{2^{n+1}} \leq \epsilon$ *par rapport à* n.

8.3 Méthode du point fixe (des approximations successives)

On suppose que f ne possède qu'une seule racine, notée α, dans $[a,b]$.
Principe de la méthode des approximations successives :
On remplace l'équation $f(x)=0$ par une équation ***équivalente*** *$x=\phi(x)$ où ϕ est continue, et on construit la suite récurrente définie par*

$$\begin{cases} x_0 \quad \text{fixé dans } [a,b]; \\ x_{n+1}=\phi(x_n), \quad n\geq 0, \end{cases} \tag{8.2}$$

avec l'espoir que la suite (x_n) converge vers la solution α du problème posé.
Interprétation géométrique :
- *Chercher α tel que $f(\alpha)=0$ revient à chercher l'intersection du graphe de f avec l'axe des abscisses.*
- *Chercher α tel que $\phi(\alpha)=\alpha$ revient à chercher l'intersection du graphe de f avec la droite $y=x$, c'est à dire à trouver le point fixe de ϕ, d'où le nom de la méthode.*



Exemples : *On peut remplacer l'équation $f(x)=0$ par l'équation $x=x-f(x)=\phi(x)$ ou par $x=x-\frac{f(x)}{a}=\phi(x), a\neq 0$.*
Questions :
- *La suite $(x_n)_n$ converge-t-elle ?*
- *Si oui, converge-t-elle vers α?*
- *Question d'analyse numérique : s'il y a convergence vers α, à quelle vitesse converge-t-on ? Peut-on estimer l'erreur commise sur l'approximation de α?*

La réponse à la deuxième question est oui, si $x_n\in[a,b]$, $\forall\, n\in\mathbb{N}$. En effét ;
Soit $x=\lim\limits_{n\to+\infty} x_n$. Alors comme ϕ est continue on a

$$\begin{array}{ccc} x_{n+1} & = & \phi(x_n) \\ \downarrow & & \downarrow \\ x & = & \phi(x) \end{array}$$

donc la limite de $(x_n)_n$ vérifie $x=\phi(x)$, et si $x\in[a,b]$ on obtient $x=\alpha$.
Donc pour que $(x_n)_n$ converge vers α, il suffit que (x_n) converge et que $x_n\in[a,b]$, $\forall\, n\in\mathbb{N}$.

Définition 8.1. *Soit I un sous-intervalle de $\mathbb{R}$ et $\phi: I\longrightarrow I$ une application continue.*

1. *On dit que $\alpha \in I$ est un point fixe de ϕ dans I si $\phi(\alpha) = \alpha$.*
2. *On dit que ϕ est contractante si ϕ est Lipschitzienne du rapport K avec $0 < K < 1$, c'est à dire s'il existe $K \in]0,1[$ tel que*

$$\forall\, x, y \in I,\ |\phi(x) - \phi(y)| \leq K|x - y|.$$

Théorème 8.1. *Soit ϕ : **[a,b]** $\subset \mathbb{R} \longrightarrow$ **[a,b]** une application **contractante**. Alors*

1. *ϕ admet un point fixe unique $\alpha \in [a,b]$.*
2. *La suite définie par : $x_{n+1} = \phi(x_n)$, $n \geq 0$ converge vers α, $\forall\, x_0 \in [a,b]$, et on a l'estimation suivante de l'erreur*

$$|\alpha - x_n| \leq \frac{K^n}{1-K}|x_1 - x_0| \quad \text{où } K \text{ est la constante de Lipschitz.} \tag{8.3}$$

Démonstration. **i. Existence :** on a $|x_{n+1} - x_n| \leq K^n|x_1 - x_0|, \forall\, n \in \mathbb{N}$. En effet ; puisque $x_{i+1} - x_i = \phi(x_i) - \phi(x_{i-1})$, $i = 1, \ldots, n$ et ϕ est $K-$*Lipschitzienne*, alors

$$|x_{n+1} - x_n| \leq K|x_n - x_{n-1}| \leq K^2|x_{n-1} - x_{n-2}| \leq \ldots \leq K^n|x_1 - x_0|.$$

D'où, $\forall\, n, p \in \mathbb{N}$ on aura

$$\begin{array}{rcl} |x_{n+p} - x_n| & \leq & |x_{n+p} - x_{n+p-1}| + |x_{n+p-1} - x_{n+p-2}| + \ldots + |x_{n+1} - x_n| \\ & \leq & \left(K^{n+p-1} + K^{n+p-2} + \ldots + K^n\right)|x_1 - x_0| \\ & = & K^n\left(1 + K + K^2 + \ldots + K^{p-1}\right)|x_1 - x_0| \\ & = & K^n \frac{1-K^p}{1-K}|x_1 - x_0| \ (\text{ car : } 1 - K^p < 1) \\ & \leq & \frac{K^n}{1-K}|x_1 - x_0| \longrightarrow 0 \ \text{ quand } n \to +\infty \ (\text{ car : } 0 < K < 1). \end{array}$$

La suite $(x_n)_n$ est donc de Cauchy dans $\mathbb{R}$, donc convergente dans $\mathbb{R}$ (car $\mathbb{R}$ est complet). Soit α la limite de cette suite ; par continuité de ϕ, $\phi(\alpha) = \alpha$, d'où l'existence du point fixe. De plus, si on fixe n et en faisant tendre p vers $+\infty$ on obtient

$$|\alpha - x_n| \leq \frac{K^n}{1-K}|x_1 - x_0|.$$

ii. Unicité : soient α_1 et α_2, deux points fixes de ϕ. On a :

$$|\alpha_1 - \alpha_2| = |\phi(\alpha_1) - \phi(\alpha_2)| \leq K|\alpha_1 - \alpha_2|,$$

et par suite, $K < 1 \Longrightarrow \alpha_1 = \alpha_2$.

□

Remarque 8.3. *1. ϕ est contractante sur $[a,b] \Longrightarrow \phi$ est continue sur $[a,b]$.*

2. *Plus la constante de Lipschitz K est petite, plus la convergence est rapide.*
3. *Dans la pratique, il faut d'abord déterminer l'intervalle sur lequel la fonction ϕ est contractante (si elle l'est !) puis éventuellement le réduire pour que $\phi([a,b]) \subset [a,b]$.*
4. *L'inégalité (8.3) nous permet d'estimer le nombre suffisant d'itérations n pour approcher α avec une précision ϵ donnée, il suffira de prendre*

$$n = \left[\frac{\ln\left(\frac{\epsilon(1-K)}{|x_1 - x_0|}\right)}{\ln K}\right] + 1.$$

5. *En pratique, on arrête le processus* $x_{n+1} = \phi(x_n)$ *si l'erreur relative* $\frac{|x_{n+1}-x_n|}{|x_{n+1}|} \leq \epsilon'$ *où* ϵ' *est la précision relative, (ou si l'erreur absolue* $|x_{n+1} - x_n| \leq \epsilon$ *lorsque* $x_{n+1} \simeq 0$*).*

Il est important d'avoir un critère pratique assurant qu'une fonction ϕ *est contractante (Lipschitzienne, avec* $0 < K < 1$*).*

Proposition 8.1. *Soit* ϕ *une fonction dérivable sur* $[a, b] \subset \mathbb{R}$.
Si $|\phi'(x)| \leq K \; \forall\, x \in [a, b]$, *alors* ϕ *est* $K-$*Lipschitzienne sur* $[a, b]$.

Démonstration. Soient $x, y \in [a, b]$. Puisque $[x, y] \subseteq [a, b]$, on a

$$\phi(y) - \phi(x) = \int_x^y \phi'(s)ds = \int_0^1 \phi'(x + t(y - x))(y - x)dt \quad (\text{ posons } s = x + t(y - x))$$

donc,

$$\begin{aligned} |\phi(y) - \phi(x)| &= \left|\int_0^1 \phi'(x + t(y - x))(y - x)dt\right| \\ &\leq \int_0^1 |\phi'(x + t(y - x))||(y - x)|dt \leq K|y - x|. \end{aligned}$$

Ou bien, on utilise la formule des accroissements finis :

$$\phi(y) - \phi(x) = \phi'(\xi)(y - x), \quad \text{où} \quad \xi \in]x, y[,$$

donc

$$|\phi(y) - \phi(x)| = |\phi'(\xi)||y - x| \leq K|y - x|.$$

□

Remarque 8.4. *1. Si* $f \in \mathcal{C}^1([a, b])$, *alors*

$$\max_{x\in[a,b]} |\phi'(x)| < 1 \Longrightarrow \phi \text{ est contractante sur } [a, b] \text{ avec } K = \max_{x\in[a,b]} |\phi'(x)|.$$

2. *Cette proposition est bien utile car il est beaucoup plus facile de regarder le max de* ϕ' *que de vérifier que* ϕ *est contractante.*
3. ***Attention :*** *soit la fonction* $\phi(x) = x + \frac{1}{x}$ *pour* $x \geq 1$. *On a* $\phi'(x) = 1 - \frac{1}{x^2}$, *alors* $|\phi'(x)| < 1, \forall\, x \in [1, +\infty[$ *mais* $\sup\limits_{x\geq 1} |\phi'(x)| = 1$ *et*

$$|\phi(x) - \phi(y)| < |x - y|.$$

Pourtant ϕ *n'a pas de point fixe, en effet* $x = \phi(x) \Leftrightarrow \frac{1}{x} = 0$!

Proposition 8.2. *L'image d'un intervalle fermé borné* $[a, b]$ *par une fonction continue est un intervalle fermé borné. Autrement dit : si* ϕ *est continue sur* $[a, b]$ *alors*

$$\phi([a, b]) = [m, M] \text{ où } m = \min_{x\in[a,b]} \phi(x) \text{ et } M = \max_{x\in[a,b]} \phi(x).$$

De plus,

1. ϕ *est croissante* $\Longrightarrow \phi([a, b]) = [\phi(a), \phi(b)]$.
2. ϕ *est décroissante* $\Longrightarrow \phi([a, b]) = [\phi(b), \phi(a)]$.

Proposition 8.3. *Soit* ϕ *une fonction continue au voisinage de* α *tel que* $\phi(\alpha) = \alpha$.

1. *Si* $|\phi'(\alpha)| < 1$, *alors il existe un intervalle* $[a, b]$ *contenant* α *pour lequel la suite* $(x_n)_n$ *définie par (8.2) converge vers* α, $\forall\, x_0 \in [a, b]$
2. *Si* $|\phi'(\alpha)| > 1$, *alors pour tout intervalle réel* $[a, b]$ *contenant* α,
soit $\forall\, x_0 \in [a, b]$ $(x_0 \neq \alpha)$, *la suite* $(x_n)_n$ *définie par (8.2) ne converge pas vers* α
(on dit que la suite (x_n) *diverge),* ***soit*** *elle est stationnaire en* α *à partir d'un certain rang.*
On dit alors de α *qu'il est* ***un point fixe répulsif.***

Etude de quelques cas (suivant ϕ)

Les exemples qui suivent montrent que les transformations en $x = \phi(x)$ aboutissent à des comportements différents :

1. *Si $f(x) = x^2 + 3e^x - 12 = 0$, on peut l'écrire sous les différentes formes :*
   - ♣ $x = x + f(x) = \phi_1(x),\ x \in \mathbb{R}$.
   - ♣ $x = x - f(x) = \phi_2(x),\ x \in \mathbb{R}$.
   - ♣ $x = \sqrt{12 - 3e^x} = \phi_3(x),\ x \geq \ln 4$.
   - ♣ $x = \ln(\frac{12-x^2}{3}) = \phi_4(x),\ -2\sqrt{3} < x < 2\sqrt{3}$.
2. *Soit $f(x) = \tan x - x$. Cette fonction admet une seule racine α dans $[\pi, \frac{3\pi}{2}]$. On peut écrire l'équation $f(x) = 0$ sous les deux formes suivantes :*
   a) $x = \tan x = \phi_1(x)$ *et* *b)* $x = \arctan x = \phi_2(x)$.

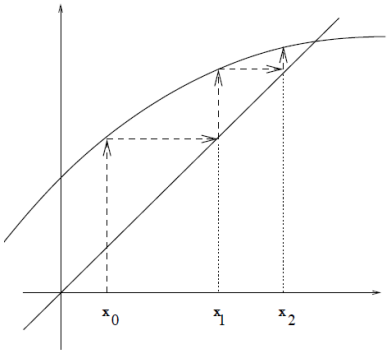
   Cas a) : $|\phi_1'(x)| = 1 + \tan^2 x > 1,\ \forall\, x \in [\pi, \frac{3\pi}{2}]$, *alors*
   la méthode du point fixe correspondante diverge.

   Cas b) : $|\phi_2'(x)| = \frac{1}{1+x^2} < 1,\ \forall\, x \in [\pi, \frac{3\pi}{2}]$, *alors*
   la fonction ϕ_2 est contractante sur $[\pi, \frac{3\pi}{2}]$.
3. *Soit $f(x) = x^2 - 6x + 8 = 0$. Cette fonction admet deux racines : $\alpha_1 = 2$ et $\alpha_2 = 4$.*
   On peut écrire l'équation $f(x) = 0$ sous les deux formes suivantes :
   i) $x = \frac{x^2+8}{6} = \phi_1(x)$ *et* *ii)* $x = \sqrt{6x - 8} = \phi_2(x)$.

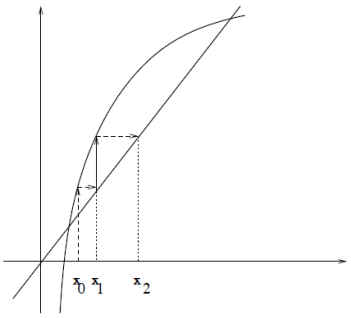
   Cas i) : $|\phi_1'(x)| = \frac{|x|}{3} < 1 \Longrightarrow |x| < 3,$ *alors*
   la méthode du point fixe correspondante converge vers $\alpha_1 = 2$ si on prend, par exemple $[a, b] = [1, \frac{5}{2}]$.

   Cas ii) : $|\phi_2'(x)| = |\frac{3}{\sqrt{6x-8}}| < 1 \Longrightarrow x > \frac{17}{6} \simeq 2.82$, *alors*
   la méthode du point fixe correspondante converge vers $\alpha_2 = 4$ si on prend, par exemple $[a, b] = [3, 5]$.

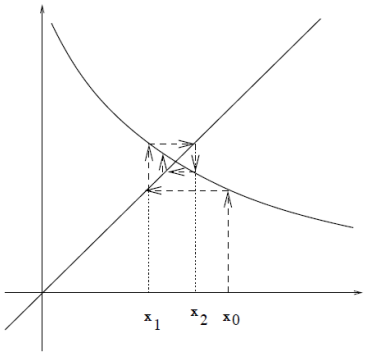
Interprétation géométrique



$$\phi'(\alpha) < 1$$



$$\phi'(\alpha) > 1$$



$$-1 < \phi'(\alpha) < 0$$

8.3.1 Ordre de convergence d'une méthode itérative

En plus de la convergence d'une méthode itérative $x_{n+1} = \phi(x_n)$, x_0 donné; il est important de connaître la rapidité de cette convergence. On souhaite savoir la relation entre l'erreur d'approximation d'une étape n et celle de l'étape $(n+1)$, ceci nous conduit à la définition suivante :

Définition 8.2. *Une méthode itérative définie par $x_{n+1} = \phi(x_n)$ est d'ordre p ssi $\exists\, k \in \mathbb{R}_+^*$ telles que*

$$\lim_{n\to+\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = \lim_{n\to+\infty} \frac{|e_{n+1}|}{|e_n|^p} = k,$$

où le terme $e_n = x_n - \alpha$ est l'erreur à l'étape n.

Dans le cas où ϕ est p fois dérivables sur $[a, b]$, le développement de Taylor autour de α donne :

$$\begin{aligned} x_{n+1} - \alpha &= \phi(x_n) - \phi(\alpha) \\ &= (x_n - \alpha)\phi'(\alpha) + \tfrac{(x_n-\alpha)^2}{2!}\phi''(\alpha) + \ldots + \tfrac{(x_n-\alpha)^p}{p!}\phi^{(p)}(\alpha) + (x_n - \alpha)^p \epsilon_n(x_n - \alpha), \end{aligned}$$

avec $\lim_{n\to+\infty} \epsilon_n(x_n - \alpha) = 0.$

- *Si* $\phi'(\alpha) \neq 0$, *on obtient* $x_{n+1}-\alpha \approx \phi'(\alpha)(x_n-\alpha)$ ($(x_n-\alpha)^a$, $a \geq 2$ *est assez petit devant* $(x_n-\alpha)$ *quad* n *grand*), *d'où* $\lim\limits_{n\to+\infty} \frac{|e_{n+1}|}{|e_n|} = |\phi'(\alpha)| \neq 0 \Longrightarrow p = 1$.
- *Si* $\phi'(\alpha) = 0$ *et* $\phi''(\alpha) \neq 0$, *on obtient* $x_{n+1} - \alpha \approx \frac{(x_n-\alpha)^2}{2}\phi''(\alpha)$, *d'où*

$$\lim_{n\to+\infty} \frac{|e_{n+1}|}{|e_n|^2} = \frac{1}{2}|\phi''(\alpha)| \neq 0 \Longrightarrow p = 2.$$

- *La méthode est dite d'ordre* p *si et seulement si* $\phi'(\alpha) = \phi''(\alpha) = \ldots = \phi^{(p-1)}(\alpha) = 0$ *et* $\phi^{(p)}(\alpha) \neq 0$.

Remarque 8.5.

Si $p = 1$, *la convergence est dite linéaire. Ce qui est le cas général des méthodes d'aproximations successives.*

- *Si* $p = 2$, *la convergence est dite quadratique.*
- *Si* $p = 3$, *la convergence est dite cubique.*

Ordre et rapidité de convergence. *Supposons qu'on a deux méthodes telles que :*
Méthode (1) : $\frac{|x_{n+1}-\alpha|}{|x_n-\alpha|} = 0.75$.
Méthode (2) : $\frac{|x_{n+1}-\alpha|}{|x_n-\alpha|^2} = 0.75$.

Méthode (1) : $|x_{n+1} - \alpha| = 0.75|x_n - \alpha| = (0.75)^2|x_{n-1} - \alpha| = ... = (0.75)^n|x_0 - \alpha|$.

Méthode (2) :

$$\begin{aligned} |x_{n+1} - \alpha| = 0.75|x_n - \alpha|^2 &= (0.75)(0.75|x_{n-1} - \alpha|^2)^2 \\ &= (0.75)^3|x_{n-1} - \alpha|^4 \\ &= (0.75)^{2^{n+1}-1}|x_0 - \alpha|^{2^{n+1}}. \end{aligned}$$

Question : *Quel est pour chacune des deux méthodes, le nombre minimal d'itérations pour avoir une erreur* $\leq 10^{-8}$, *en supposant que* $|x_0 - \alpha| = 0.5$ ***?***
méthode (1) : $|x_{n+1} - \alpha| = (0.75)^n|x_0 - \alpha| \leq 10^{-8} \Rightarrow\ n \geq 62$.
méthode (2) : $|x_{n+1} - \alpha| = (0.75)^{2^{n+1}-1}|x_0 - \alpha|^{2^{n+1}} \leq 10^{-8} \Rightarrow\ n \geq 4$.
Donc la deuxième méthode converge plus rapidement que la première.
D'où, plus l'ordre p *est grand, plus vite l'erreur décroît.*

8.4 Méthodes de type $x_{n+1} = \phi(x_n) = x_n - \dfrac{f(x_n)}{g(x_n)}$

Dans cette catégorie de méthodes, ϕ a la forme particulière $\phi(x) = x - \frac{f(x)}{g(x)}$. Il est clair que, sous certaines conditions sur g, un point fixe α de ϕ est une racine de f.

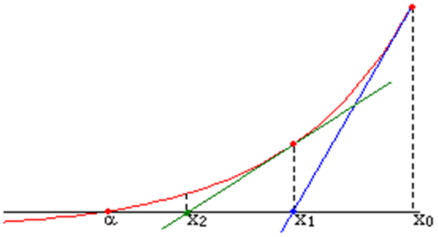
8.4.1 Méthode de Newton-Raphson (méthode de la tangente)

La méthode de Newton-Raphson est une méthode de type point fixe pour la fonction

$$\phi(x) = x - \frac{f(x)}{f'(x)}.$$

Supposons qu'on a déterminé un intervalle $[a, b]$ dans lequel f admet une racine séparée α.

Interprétation géométrique



Graphiquement la méthode de Newton-Raphson fonctionne comme suit :
à partir d'un point x_0 bien choisi dans $[a, b]$, x_1 est l'abscisse du point d'intersection de la tangente de graphe de f au point $(x_0, f(x_0))$ avec l'axe des abscisses. D'où

$$f'(x_0) = \frac{0 - f(x_0)}{x_1 - x_0} = \frac{-f(x_0)}{x_1 - x_0} \Longrightarrow x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \ (\ si\ f'(x_0) \neq 0),$$

en répétant le processus sur x_1 on obtient un point x_2, et ainsi de suite.
Les points $(x_n)_{n\in\mathbb{N}}$ vérifient donc la relation de récurrence :

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \quad n \geq 0,$$

c'est la formule de Newton-Raphson[1] *la plus utilisée dans la recherche des racines de f.*

Théorème 8.2. *Soit $f \in C^2([a, b])$ vérifiant*

1. $f(a).f(b) < 0$;
2. $f'(x) \neq 0 \ \forall x \in [a, b]$, *(c.à.d : f' ne change pas de signe dans $[a, b]$) ;*
3. $f''(x) \neq 0 \ \forall x \in [a, b]$, *(c.à.d :$f''$ ne change pas de signe dans $[a, b]$).*

[1] Joseph Raphson 1648-1715

Alors

a) *f admet une racine unique dans $[a,b]$,*

b) *la suite $(x_n)_n$ converge vers α $\forall x_0 \in [a,b]$ vérifiant $f(x_0)f''(x_0) > 0$. De plus, cette convergence est quadratique tel que*

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

Démonstration.
a) D'après le théorème des valeurs intermédiaires, les conditions (1) et (2) impliquent l'existence et l'unicité de $\alpha \in [a,b]$ telle que $f(\alpha) = 0$.
b) Montrons la convergence de $(x_n)_n$ vers α.

i) Etudions la bornitude de la suite $(x_n)_n$ On a $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$, alors

$$\begin{aligned} x_{n+1}-\alpha &= x_n - \alpha - \tfrac{f(x_n)}{f'(x_n)} \\ &= x_n - \alpha + \tfrac{f(\alpha)-f(x_n)}{f'(x_n)} \quad (f(\alpha)=0). \end{aligned} \tag{1}$$

Le développement de Taylor de f au voisinage de x_n à l'ordre 2 donne

$$f(\alpha) = f(x_n) + (\alpha - x_n)f'(x_n) + \frac{(\alpha-x_n)^2}{2!}f''(\xi_n),\ \min(x_n,\alpha) \le \xi_n \le \max(x_n,\alpha).$$

En substituant l'expression de $f(\alpha)$ dans (1), on obtient

$$\begin{aligned} x_{n+1}-\alpha &= x_n - \alpha + \tfrac{f(x_n)+(\alpha-x_n)f'(x_n)+\frac{(\alpha-x_n)^2}{2!}f''(\xi_n)-f(x_n)}{f'(x_n)} \\ &= \tfrac{(\alpha-x_n)^2}{2!}\ \tfrac{f''(\xi_n)}{f'(x_n)}. \end{aligned}$$

D'où les deux cas suivants :

1. f' et f'' sont de même signe $\Longrightarrow$ $x_{n+1} > \alpha$, $\forall n \in \mathbb{N}$, c.à.d. (x_n) est minorée par α.
2. f' et f'' sont de signes différents $\Longrightarrow$ $x_{n+1} < \alpha$, $\forall n \in \mathbb{N}$, c.à.d. (x_n) est majorée par α.

ii) Etudions la monotonie de la suite $(x_n)_{n\in\mathbb{N}}$. On distingue quatre cas :

Cas (1) $f'(x) > 0$ et $f''(x) > 0$ $\quad \forall x \in [a,b]$.

Cas (2) $f'(x) > 0$ et $f''(x) < 0$ $\quad \forall x \in [a,b]$.

Cas (3) $f'(x) < 0$ et $f''(x) < 0$ $\quad \forall x \in [a,b]$.

Cas (4) $f'(x) < 0$ et $f''(x) > 0$ $\quad \forall x \in [a,b]$.

Cas (1) Soit $f'(x) > 0$ et $f''(x) > 0, \forall x \in [a,b]$, alors $x_n > \alpha$, $\forall n \in \mathbb{N}$.
Supposons que x_0 est choisi de sorte que $f(x_0).f''(x_0) > 0$ et montrons que $(x_n)_n$ est décroissante. On raisonne par récurrence :
Pour $n = 0$: $\quad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \Rightarrow x_1 - x_0 = -\frac{f(x_0)}{f'(x_0)} < 0$; en effet on choisit x_0 tel que $f(x_0) > 0$ car $f''(x_0) > 0$. Donc $x_1 - x_0 < 0 \Rightarrow x_1 < x_0$.
Pour $n \in \mathbb{N}^*$: $x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} < 0$; en effet

$$\left\{ \begin{array}{l} x_n > \alpha,\ \forall n \in \mathbb{N}^* \\ f \text{ croissante sur } [a,b] \end{array} \right. \Rightarrow f(x_n) > f(\alpha) = 0,\ \forall n \in \mathbb{N}^*,$$

donc $x_{n+1} - x_n < 0 \Rightarrow x_{n+1} < x_n \; \forall n \in \mathbb{N}^*$.
Conclusion $\forall n \in \mathbb{N}$, $x_{n+1} < x_n$, donc la suite $(x_n)_{n\in\mathbb{N}}$ est décroissante minorée, d'où elle est convergente.

Cas (2) Soit $f'(x) > 0$ et $f''(x) < 0, \forall x \in [a,b]$, alors $x_n < \alpha$, $\forall n \in \mathbb{N}$. Supposons que x_0 est choisi de sorte que $f(x_0).f''(x_0) > 0$ et montrons que $(x_n)_n$ est croissante.
Pour $n = 0$: $x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} \Rightarrow x_1 - x_0 = -\frac{f(x_0)}{f'(x_0)} >$; en effet, on choisit x_0 tel que $f(x_0) < 0$ car $f''(x_0) < 0$. Donc $x_1 - x_0 > 0 \Rightarrow x_1 > x_0$.
Pour $n \in \mathbb{N}^*$: $x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} > 0$; en effet,

$$\begin{cases} x_n < \alpha, \; \forall n \in \mathbb{N}^* \\ f \text{ croissante sur } [a,b] \end{cases} \Rightarrow f(x_n) < f(\alpha) = 0, \; \forall n \in \mathbb{N}^*,$$

donc $x_{n+1} - x_n > 0 \Rightarrow x_{n+1} > x_n \; \forall n \in \mathbb{N}^*$.
Conclusion $\forall n \in \mathbb{N}$, $x_{n+1} > x_n$, donc la suite $(x_n)_{n\in\mathbb{N}}$ est croissante majorée, d'où elle est convergente.
De la même manière on traite les deux autres cas.

□

Commentaires : *Il est très important de choisir x_0 aussi proche que possible de α, sinon, non seulement la suite $(x_n)_{n\in\mathbb{N}}$ peut converger vers une autre racine, mais peut aussi diverger.*

Conclusion sur la méthode de Newton-Raphson

1. ***Avantage principal :*** *en règle générale, la convergence de cette méthode (si elle a lieu) est quadratique (d'ordre 2) telle que $|x_{n+1} - \alpha| = k|x_n - \alpha|^2$, $k \in \mathbb{R}$, (celle du point fixe quelconque est, en général, d'ordre 1 $|x_{n+1} - \alpha| = k|x_n - \alpha|$, $k \in \mathbb{R}$).*
2. ***Inconvénients :***
   a) *choix du point de départ x_0 pour avoir la convergence,*
   b) *calcul, à chaque étape, de $f'(x_n)$ en plus de $f(x_n)$.*

8.4.2 Méthode de la sécante

Dans certaines situations, la dérivée de f' est très compliquée ou même impossible à expliciter. On ne peut pas alors utiliser telle qu'elle la méthode de Newton-Raphson. L'idée est de remplacer f' par le taux d'accroissement de f sur un petit intervalle $[x_{n-1}, x_n]$:

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

Remarque 8.6. *On notera que l'algorithme itératif ne peut démarrer que si on dispose déjà de deux valeurs approchées x_0, x_1 de α.*

8.4.3 Méthode de la corde

Cette méthode est obtenue en remplaçant $f'(x_n)$ par une constante Q fixée dans la formule de Newton-Raphson :

$$x_{n+1} = x_n - \frac{f(x_n)}{Q}, \; n \geq 0.$$

On peut prendre par exemple $Q = f'(x_0)$, ou bien $Q = \frac{f(b)-f(a)}{b-a}$, dans le cas où l'on cherche une racine dans l'intervalle $[a,b]$.

8.5 Conclusion

1. *L'avantage de la méthode du point fixe est que si elle converge, elle convergera quelle que soit la valeur initiale choisie dans l'intervalle de convergence. Son inconvénient principal est sa lenteur de convergence, on peut néanmoins l'accélérer.*
2. *La méthode de Newton-Raphson est plus rapide car la vitesse de convergence est d'ordre au moins 2; la difficulté réside dans le choix de la valeur initiale!*
3. *Les différentes méthodes proposées ont leurs avantages et leurs inconvénients, on choisira celle qui convient le mieux au problème posé! On voit qu'il est donc nécessaire d'étudier correctement les algorithmes avant de se lancer dans la programmation! C'est la démarche obligatoire en analyse numérique.*

8.6 Exercices

Exercice 8.2. *On considère la fonction f définie par $f(x) = x^3 - e^{-x}$, $x \in \mathbb{R}$.*

1. *Séparer les racines de f, puis montrer que $f(x) = 0$ admet une solution réelle unique α et que $\alpha \in [\frac{1}{2}, 1]$.*
2. *Déterminer, par la dichotomie, une approximation de α à 5.10^{-2} près en utilisant le test d'arrêt $|x_{n+1} - x_n| \leq \varepsilon$. Déterminer le nombre d'itérations suffisant pour avoir cette précision en utilisant la formule de l'erreur. Commenter.*

Exercice 8.3. *On considère l'équation :*

$$x = f(x) = \frac{1}{\lambda + 1}(e^{-\lambda x} + \lambda x),\ x \in [0, 1],\quad \lambda \in \mathbb{R}_+^*. \tag{8.4}$$

1. **a)** *Montrer que les hypothèses du théorème du point fixe sont vérifiées par f dans l'intervalle $[0, 1]$.*

   b) *En déduire que l'équation (8.4) admet une solution séparée $\alpha \in [\frac{1}{\lambda+1}, 1]$.*
2. *Soit l'équation :*

$$F(x) = e^{-x} - x = 0. \tag{8.5}$$

   a) *Quelle relation y a t-il entre les équations (8.4) et (8.5) ?*

   b) *Les hypothèses du théorème de Newton sont elles vérifiées dans l'intervalle [0,1] ?*

   c) *Trouver un sous intervalle I fermé de $[0, 1]$ dans lequel l'algorithme de Newton appliqué à l'équation (8.5) converge $\forall x_0 \in I$.*

Exercice 8.4. *Soit l'équation $f(x) = 0$ où $f(x) = x^2 e^x + 2x - 1$, $x \in \mathbb{R}$.*

1. *Montrer que l'équation $f(x) = 0$ admet une seule solution réelle positive s. Vérifier que $s \in [\frac{3}{10}, \frac{1}{2}]$.*
2. *Montrer que la méthode de Newton est applicable sur $[\frac{3}{10}, \frac{1}{2}]$.*
3. *Approcher s à $0,5.10^{-3}$ près, par cette méthode.*

Exercice 8.5.
On considère la matrice A définis par :

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

a) *Montrer que $P(\lambda) = \det(A - \lambda I_3) = -\lambda^3 - 3\lambda + 1$.*

b) *Séparer graphiquement les racines de P et vérifier que P admet une racine séparée $\overline{\lambda} \in [\frac{1}{10}, \frac{1}{3}]$.*

c) *Déterminer une fonction ϕ telle que la suite définie par : $\lambda_{n+1} = \phi(\lambda_n)$, $n \geq 0$ converge vers $\overline{\lambda}$, $\forall\, \lambda_0 \in [\frac{1}{10}, \frac{1}{3}]$.*

d) *Pour $\lambda_0 = \frac{1}{3}$, calculer λ_1 et λ_2 avec 5 chiffres significatifs, puis estimer l'erreur $|\lambda_2 - \overline{\lambda}|$.*

a) *Montrer que l'algorithme de Newton associé à l'équation $P(\lambda) = 0$ s'écrit :*

$$\lambda_{n+1} = \frac{1}{3}\left(\frac{2\lambda_n^3 + 1}{\lambda_n^2 + 1}\right), n \geq 0.$$

b) *Montrer que l'algorithme de Newton converge vers $\overline{\lambda}$, $\forall\, \lambda_0 \in [\frac{1}{10}, \frac{1}{3}]$ vérifiant une certaine condition à préciser.*

c) *En utilisant le théorème des accroissement finis à la fonction P sur l'intervalle $[\overline{\lambda}, \lambda_{n+1}]$ ou $([\lambda_{n+1}, \overline{\lambda}])$ et le développement de Taylor de la fonction P à l'ordre 2 en λ_n suivant :*

$$f(\lambda_{n+1}) = f(\lambda_n) + (\lambda_{n+1} - \lambda_n)f'(\lambda_n) + \frac{(\lambda_{n+1} - \lambda_n)^2}{2!}f'(\xi_n)$$

où $\xi_n \in]\lambda_{n+1}, \lambda_n[$ (ou $]\lambda_n, \lambda_{n+1}[$). Montrer que

$$|\lambda_{n+1} - \overline{\lambda}| \leq \frac{M}{2m}|\lambda_{n+1} - \lambda_n|^2$$

où $M = \max\limits_{x\in[\frac{1}{10},\frac{1}{3}]} |P''(\lambda)|, \quad m = \min\limits_{x\in[\frac{1}{10},\frac{1}{3}]} |P'(\lambda)|.$

d) *Pour $\lambda_0 = \frac{1}{3}$, calculer λ_1 et λ_2 avec 5 chiffres significatifs, puis estimer l'erreur $|\lambda_2 - \overline{\lambda}|$.*

Comparer la rapidité de convergence des deux méthodes (Newton et point fixe).

Exercice 8.6. *Soit la fonction f définie par : $f(x) = -e^{-1}x + \frac{3}{4} - e^{-x}$, $x \in \mathbb{R}$.*

1. *Montrer que l'équation $f(x) = 0$ admet deux racines réelles positives α et β où $\alpha \in [0, 1]$ et $\beta \in [1, e]$.*
2. **a)** *Montrer que l'équation $f(x) = 0$ est équivalente sur $[1, e]$ à l'équation $\varphi(x) = x$ où $\varphi(x) = \frac{3}{4}e - e^{1-x}$.*

   b) *La fonction φ vérifie-t-elle les conditions du théorème du point fixe sur l'intervalle $[1, e]$?*

   c) *Montrer que la suite itérative $x_{n+1} = \varphi(x_n)$, $n \geq 0$ converge vers la racine β $\forall\, x_0 \in [1 + \frac{1}{10}, e]$.*

   d) *En prenant $x_0 = 1, 1$; calculer x_2 et estimer l'erreur $|x_2 - \beta|$.*

(Indication : $e^{-e} = 0, 0660$ et $e^{1-e} = 0, 1794$)

Exercice 8.7. *Soit la fonction f définie par :*

$$f(x) = e^{-x} - x + 1, \ x \in \mathbb{R}.$$

1. *Montrer que l'équation $f(x) = 0$ admet une solution séparée α dans l'intervalle $[1, 2]$.*
2. *Déterminer un intervalle I et une fonction φ tels que la méthode du point fixe converge vers la solution α, en prenant $x_0 = 1$.*

Exercice 8.8. *On considère la fonction f définie par : $f(x) = x^2 - ln(x + 1)$, $x \in I = [\frac{1}{2}, 1]$.*

1. *Montrer graphiquement que $f(x) = 0$ admet une seule solution réelle strictement positive α et que $\alpha \in [\frac{1}{2}, 1]$.*
2. *Montrer que l'équation $f(x) = 0$ est équivalente aux équations $x = \phi_i(x)$, $i = 1, 2$ dans I.*

   où $\phi_1(x) = \sqrt{ln(x + 1)}, \quad \phi_2(x) = e^{x^2} - 1, \ x \in I.$
3. *Montrer que ϕ_1 vérifie les hypothèses du théorème du point fixe dans I, conclure.*
4. *En prenant $x_0 = 0$, déterminer le nombre d'itérations suffisant pour approcher α à 10^{-1} près, par la méthode des approximations successives, calculer cette approximation.*
5. *La suite itérative* $\begin{cases} x_0 \in I, \\ x_{n+1} = \phi_2(x_n), \ n \geq 0, \end{cases}$ *converge-t-elle, $\forall\, x_0 \in I$?*

Exercice 8.9. *Montrer que la suite $(x_n)_n$ converge vers $\sqrt{2}$ pour tout $x_0 \in [1, +\infty[$,*

$$\text{où} \qquad \forall\, n \in \mathbb{N}^*, \quad x_n = \frac{1}{2}\left(x_{n-1} + \frac{2}{x_{n-1}}\right).$$

Exercice 8.10. *Soit f définie par $f(x) = x^3 - e^{-x}$.*

1. *Montrer que l'équation $f(x) = 0$ admet une seule solution réelle α puis vérifier que $\alpha \in [\frac{1}{2}, 1]$.*
2. **a)** *Montrer que la suite $x_0 \in [\frac{1}{2}, 1]$, $x_{n+1} = e^{-\frac{x_n}{3}}$ converge vers α.*
   b) *Pour $x_0 = 1$, déterminer n le nombre d'itérations minimal pour que $|x_n - \alpha| \leq 5.10^{-2}$.*
3. **a)** *Peut-on appliquer la méthode de Newton dans $[\frac{1}{2}, 1]$? Donner l'expression de $(x_n)_n$ en précisant le choix de x_0.*
   b) *Calculer α à 5.10^{-2} prés, partant de $x_0 = 1$. Exprimer le résultat au dernier c.s.e.*
4. *Conclure.*

Exercice 8.11. *Soit f la fonction définie par : $f(x) = x + \sin x - 1$, $x \in \mathbb{R}$.*

1. *Montrer que la fonction f admet une racine séparée α dans l'intervalle $[0, \pi]$.*
2. *Montrer que l'équation $f(x) = 0$ est équivalente à l'équation $x = \phi(x)$ où $\phi(x) = 1 - \sin x$*
3. **a)** *Montrer que la fonction ϕ vérifie les hypothèses du théorème du point fixe dans l'intervalle $[0,39, 0,65]$*
   b) *Quel est le nombre d'itérations suffisant pour approcher α à 10^{-2} prés en partant de $x_0 = 0,39$.*
4. *Montrer que l'équation $f(x) = 0$ est aussi équivalente à l'équation $x = \psi(x)$ où $\psi(x) = \alpha\phi(x) + (1-\alpha)x$, α est un paramètre réel et ϕ est la fonction donnée en deuxième question.*
5. *Dans cette question on prend $\alpha = 0,6$.*
   a) *Montrer que la fonction ψ vérifie les hypothèses du théorème du point fixe dans l'intervalle $[0,39, 0,65]$*
   b) *Quel est le nombre d'itérations suffisant pour avoir une valeur approchée de α à 10^{-2} près en partant de $x_0 = 0,39$.*
6. *Avec quelle fonction est-il préférable de travailler, justifier votre réponse*
7. **a)** *Montrer que la méthode de Newton est applicable dans l'intervalle $[0,39, 0,65]$*
   b) *Calculer une valeur approchée de α à 10^{-2} près par la méthode de Newton.*
8. *Quelle méthode est-il préférable d'utiliser.*

Exercice 8.12. ***I.*** *(Méthode de la sécante) : Soit $f : [a, b] \to \mathbb{R}$ une fonction de classe C^2 dans $[a, b]$. On fait de plus les hypothèses suivantes*

$$\begin{array}{ll} (H1): & f(a)f(b) < 0, \\ (H2): & f'(x) > 0,\ \forall\, x \in [a, b], \\ (H3): & f''(x) > 0,\ \forall\, x \in [a, b]. \end{array}$$

On considère le processus itératif

$$x_{n+1} = x_n - \frac{b - x_n}{f(b) - f(x_n)} f(x_n) \qquad et \quad x_0 = a.$$

1. *Donner une interprétation géométrique de ce processus.*
2. *Montrer que $(x_n)_n$ converge vers $\overline{x}$ solution exacte de $f(x) = 0$.*

II. *On considère le processus itératif $x_{n+1} = \varphi(x_n)$, $x_0 \in [1, \alpha]$ où*

$$\varphi(x) = -\frac{1}{2\alpha}x^2 + x + \frac{1}{2}, \quad \alpha > 1.$$

1. *Montrer que φ est contractante dans $[1,\alpha]$ et que $[1,\alpha]$ est stable par φ.*
2. *En déduire que le processus converge vers ξ à calculer.*
3. *Montrer que si $x_0 = 1$, alors*

$$|x_n - \xi| \leq \frac{1}{2}\alpha(1 - \frac{1}{\alpha})^{n+1}.$$

III. Application numérique : *Soient*

$$f(x) = \frac{1}{4}x^2 - \frac{1}{2}, \quad [a,b] = [1,2], \ \ \alpha = 2, \ \ x_0 = 1.$$

1. *Montrer que l'équation $\varphi(x) = x$ est équivalente à l'équation $f(x) = 0$.*
2. *Déterminer, par la méthode de la sécante, une approximation de la solution exacte ξ_1 de l'équation $f(x) = 0$ à $\varepsilon = 10^{-1}$ près en utilisant le test d'arrêt $|x_{n+1} - x_n| \leq \varepsilon$.*
3. *Déterminer le nombre d'itérations permettant de calculer une valeur approchée de ξ_1 à 10^{-1} près, par la méthode du point fixe. Calculer cette approximation.*
4. *Quelle est la méthode la plus avantageuse ?*

Exercice 8.13.
On considère la fonction g définie par : $g(x) = \frac{x}{2} + \frac{1}{x}, \ \ x > 0$.

1. *Montrer que l'équation $g(x) = x$ admet une solution unique α dans $[1,3]$.*

2. *Calculer α à 5×10^{-3} près en arrondissant au nombre de chiffres significatifs exacts. Justifier le choix de la méthode d'approximation.*

3. *Montrer que la fonction g est la fonction de Newton-Raphson d'une certaine fonction f que l'on explicitera.*
4. *En déduire que la méthode du point fixe pour la fonction ϕ définie par $\phi(x) = \frac{x}{2} + \frac{A}{2x}$, $x > 0$ $(A > 0)$, est un moyen de déterminer $\sqrt{A}$.*

Chapitre 9

Résolution des équations différentielles ordinaires

9.1 Position du problème

On appelle équation différentielle une équation reliant une fonction et ses dérivées successives. Si l'équation ne fait intervenir que la fonction et sa dérivée, on parle d'équation du premier ordre. Nous prenons comme point de départ, une équation différentielle du premier ordre avec condition initiale.
Soient $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ *une fonction suffisamment différentiable et* $y_0 \in \mathbb{R}$.
La tâche consiste à déterminer une fonction $y : [t_0, T] \to \mathbb{R}$ *solution du* ***problème de Cauchy :***

$$\begin{cases} y'(t) &= f(t, y(t)), \quad t \in [t_0, T] \\ y(t_0) &= y_0 (\ \textbf{\textit{condition initiale ou condition de Cauchy}}\ ). \end{cases} \tag{9.1}$$

La résolution numérique des équations différentielles est probablement le domaine de l'analyse numérique où les applications sont les plus nombreuses. Que ce soit en mécanique des fluides, en transfert de chaleur ou en analyse de structures, on aboutit souvent à la résolution d'équations différentielles, de systèmes d'équations différentielles ou plus généralement d'équations aux dérivées partielles.
La précision des diverses méthodes de résolution proposées dans ce cours est proportionnelle à l'ordre de ces méthodes. Nous commençons le cours par des méthodes relativement simples ayant une interprétation géométrique. Elles nous conduiront progressivement à des méthodes plus complexes telles les méthodes de Runge-Kutta d'ordre 4, qui permettent d'obtenir des résultats d'une grande précision.

Remarque 9.1. *La variable indépendante* t *représente très souvent (mais pas toujours) le temps. La condition initiale* $y(t_0) = y_0$ *est en quelque sorte l'état de la solution au moment où on commence à s'y intéresser. Il s'agit d'obtenir* $y(t)$ *pour* $t \in [t_0, T]$ *si on cherche une solution analytique, ou une approximation de* $y(t)$*, si on utilise une méthode numérique.*

Commençons par présenter quelques exemples des problèmes de Cauchy qu'on peut résoudre analytiquement :

Exemple 9.1. *Soit le problème de Cauchy suivant :*

$$\begin{cases} y'(t) &= t, \quad t \in [0, 1] \\ y(0) &= 1. \end{cases}$$

Cet exemple est l'un des exemples les plus simples que l'on puisse imaginer. En intégrant de chaque côté, on aura $y(t) = \frac{t^2}{2} + c$ *où* c *est une constante réelle, pour déterminer cette constante, il suffit d'imposer la codition initiale,* $y(0) = 1$*. On obtient*

$$y(t) = \frac{t^2}{2} + 1, \quad t \in [0, 1].$$

Exemple 9.2. *Soit le problème de Cauchy :*

$$\begin{cases} y'(t) &= \sqrt{y}, \quad t \in [0, 1] \\ y(0) &= 0. \end{cases}$$

Il suffit pas dans ce cas d'intégrer les deux côtes de l'équation pour obtenir la solution. On doit d'abord séparer les variables en écrivant par exemple :

$$\frac{dy}{\sqrt{y}} = dt, \quad y \neq 0$$

on peut maintenant faire l'intégration de deux côtes on aura

$$2\sqrt{y} = t + c \Rightarrow y(t) = \left(\frac{t + c}{2}\right)^2$$

puis, $y(0) = 0 \Rightarrow c = 0$*. Donc* $y(t) = \frac{t^2}{4}$ *est une solution du problème donné.*
Remarquons aussi que $y \equiv 0$ *est une solution de notre problème.*

Exemple 9.3. *Soit le problème de Cauchy :*

$$\begin{cases} y' &= -\frac{y}{t \ln t} + \frac{1}{\ln t}, \quad t \in [e, 5] \\ y(e) &= e. \end{cases}$$

L'équation différentielle de ce problème est linéaire du premier ordre. En appliquant la méthode élémentaire pour résoudre ce type des équations, on obtient

$$y(t) = \frac{t + c}{\ln t} \quad \text{où} \quad c \text{ est une constante réelle arbitraire,}$$

puis la codition initiale donne $c = 1$*. La solution du problème donné est alors* $y(t) = \frac{t}{\ln t}$*.*

Le théorème suivant nous donne des conditions suffisantes qui assurent l'existence et l'unicité de la solution (théorique) du problème (9.1).

Théorème 9.1. *Soit* $f : [t_0, T] \times \mathbb{R} \to \mathbb{R}$ *une fonction telle que*

1. f *est continue sur* $[t_0, T] \times \mathbb{R}$*.*
2. f *est Lipschitzienne par rapport à la deuxième variable, c'est à dire il existe une constante* $L > 0$ *telle que pour tout* $t \in [t_0, T]$ *et* $y_1, y_2 \in \mathbb{R}$*, on ait :*

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|.$$

Alors, le problème de Cauchy[1] *(9.1) admet une unique solution* $y \in C^1([t_0, T], \mathbb{R})$*.*

[1] Augustin Louis Cauchy, français, 1789-1857

Remarque 9.2. *1. Remarquons que la fonction* $f(t,y) = -\frac{y}{t\ln t} + \frac{1}{\ln t}$ *est continue sur* $]0,+\infty[\times\mathbb{R}$, *donc elle est continue sur l'intervalle* $[e,5]\times\mathbb{R}$. *De plus, elle est Lipschitzienne par rapport à* y *sur* $[e,5]$ *avec la constante de Lipschitz* $L = \frac{1}{e}$. *En effet; pour tout* $t \in [e,5]$ *et* $y_1, y_2 \in \mathbb{R}$ *on a*

$$\begin{aligned} |f(t,y_1) - f(t,y_2)| &= |-\tfrac{y_1}{t\ln t} + \tfrac{1}{\ln t} + \tfrac{y_2}{t\ln t} - \tfrac{1}{\ln t}| \\ &= \tfrac{1}{t\ln t}|y_1 - y_2| \\ &\leq \tfrac{1}{e}|y_1 - y_2|, \end{aligned}$$

ce qui justifie l'unicité de la solution du troisième problème qu'on a trouvé analytiquement.

2. Par contre le second membre du deuxième problème $f(t,y) = \sqrt{y}$ *n'est pas Lipschitzien par rapport à* y. *En effet; Pour* $y_1 \in \mathbb{R}$ *quelconque et* $y_2 = 0$ *on a*

$$|f(t,y_1) - f(t,y_2)| = |\sqrt{y_1} - 0| = \sqrt{y_1} \geq y_1.$$

D'où, la possibilité d'avoir plus d'une solution de ce problème, (analytiquement on a trouvé deux solutions $y_1(t) = 0$ *et* $y_2(t) = \frac{t^2}{4}$ *sur* $[0,1]$*).*

On peut résoudre analytiquement que quelques types d' équations différentielles (par exemple les équations à variables séparables, linéaire, Bernoulli, Riccati...) mais il y a une large classe d'équations différentielles non résolubles analytiquement. La résolution numérique est ici essentielle. *Dans la suite, nous nous placerons dans les conditions du théorème 9.1.*

Remarque 9.3. *1. Avec les outils numériques de résolution d'équations différentielles il n'est pas possible d'obtenir une solution pour toutes les valeurs de la variable* t. *On obtient plutôt une approximation de la solution analytique seulement pour certaines valeurs de* t *notées* t_i *et distancées d'une valeur* $h_i = t_{i+1} - t_i$. *Dans les méthodes présentées, cette distance est constante pour tout* i *et est notée* h. *On l'appelle le pas de temps.*

2. On note $y(t_i)$ *la solution analytique de l'équation différentielle (9.1) en* $t = t_i$, *et* y_i *la solution approximative en* $t = t_i$ *à l'aide d'une méthode numérique.*

Méthodes numériques à un pas

9.2 Méthode d'Euler

Reprenons l'équation différentielle de (9.1) et la condition initiale $y(t_0) = y_0$. *Le but est d'obtenir une approximation de la solution en* $t = t_1 = t_0 + h$. *Avant d'effectuer la première itération, il faut déterminer dans quelle direction on doit avancer à partir du point* (t_0, y_0) *pour obtenir le point* (t_1, y_1), *qui est une approximation du point* $(t_1, y(t_1))$.
L'équation différentielle (9.1) assure que :

$$y'(t_0) = f(t_0, y(t_0)) = f(t_0, y_0).$$

On peut donc suivre la droite passant par (t_0, y_0) *et de pente* $f(t_0, y_0)$. *L'équation de cette droite, notée* $d_0(t)$, *est :*

$$d_0(t) = f(t_0, y_0)(t - t_0) + y_0$$

et est illustrée par la figure (A)

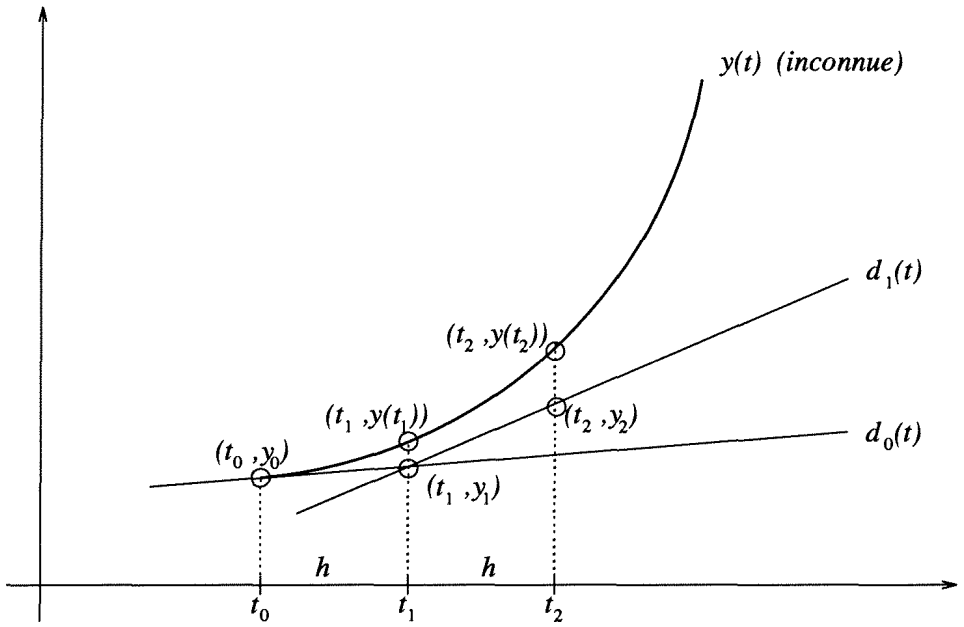


Figure 7.1: Méthode d'Euler

En $t = t_1$, *on a :*

$$d_0(t_1) = f(t_0, y_0)(t_1 - t_0) + y_0 = y_0 + hf(t_0, y_0)) = y_1.$$

En d'autres termes, $d_0(t_1)$ *est proche de la solution analytique* $y(t_1)$, *c'est à dire*

$$y(t_1) \simeq y_1 = d_0(t_1) = y_0 + hf(t_0, y_0).$$

Il est important de noter que, le plus souvent, $y_1 \neq y(t_1)$. *Donc si on souhaite faire une deuxième itération et obtenir une approximation de* $y(t_2)$, *on peut refaire l'analyse précédente à partir du point* (t_1, y_1). *on remarque cependant que la pente de la solution analytique en* $t = t_1$ *est :*

$$y'(t_1) = f(t_1, y(t_1)).$$

On ne connaît pas exactement $y(t_1)$, *mais nous possédons l'approximation* y_1 *de* $y(t_1)$. *On doit alors utiliser l'expression :*

$$y\prime(t_1) = f(t_1, y(t_1)) \simeq f(t_1, y_1)$$

et construire la droite

$$d_1(t) = f(t_1, y_1)(t - t_1) + y_1,$$

qui permettra d'estimer $y(t_2)$. *On a alors*

$$y(t_2) \simeq y_2 = y_1 + hf(t_1, y_1).$$

On remarque que l'erreur commise à la première itération est réintroduite dans les calculs de la deuxième itération.

Algorithme d'Euler[2]

(1) Etant donné un pas h, une condition initiale (t_0, y_0), et un nombre maximal d'itérations N.
(2)

$$\begin{array}{lll} \textit{Pour} \quad 0 \leq n \leq N & & \\ & y_{n+1} = y_n + hf(t_n, y_n) & \\ \textit{et} & t_{n+1} = t_n + h & \\ \textit{écrire} & t_{n+1} \quad \textit{et} \quad y_{n+1}. & \end{array}$$

(3) Arrêt.

Remarque 9.4. *La méthode d'Euler est de loin la méthode la plus simple de résolution numérique d'équations différentielles ordinaires. Elle possède une belle interprétation géométrique et son emploi est facile. Toutefois, elle est relativement peu utilisée en raison de sa précision.*

Exemple 9.4. *Soit l'équation différentielle*

$$\begin{cases} y'(t) = -y(t) + t + 1 \\ y(0) = 1. \end{cases}$$

On prend $h = 0,1$ et $f(t, y) = -y + t + 1$. Le tableau suivant rassemble les résultats des dix premières itérations. On peut montrer que la solution analytique de cette équation est :

$$y(t) = e^{-t} + t,$$

ce qui permet de comparer les solutions numérique et analytique et de constater la croissance de l'erreur

t_i	$y(t_i)$	y_i	$\lvert y(t_i) - y_i \rvert$
$0,0$	$1,000000$	$1,000000$	$0,000000$
$0,1$	$1,004837$	$1,000000$	$0,004837$
$0,2$	$1,018731$	$1,010000$	$0,008731$
$0,3$	$1,040818$	$1,029000$	$0,011818$
$0,4$	$1,070302$	$1,056100$	$0,014220$
$0,5$	$1,106531$	$1,090490$	$0,016041$
$0,6$	$1,148812$	$1,131441$	$0,017371$
$0,7$	$1,196580$	$1,178297$	$0,018288$
$0,8$	$1,249329$	$1,230467$	$0,018862$
$0,9$	$1,306570$	$1,287420$	$0,019150$
$1,0$	$1,367879$	$1,348678$	$0,019201$

9.3 Méthodes de Taylor

Le développement de Taylor autorise une généralisation immédiate de la méthode d'Euler, qui permet de diminuer l'erreur d'approximation. Nous nous limitons cependant à la méthode de Taylor du second ordre. On cherche, au temps $t = t_n$, une approximation de la solution en $t = t_{n+1}$. On a immédiatement :

$$\begin{aligned} y(t_{n+1}) &= y(t_n + h) \\ &= y(t_n) + y'(t_n)h + \tfrac{y''(t_n)}{2}h^2 + o(h^2). \end{aligned}$$

[2]Leonhard Euler, 1707-1783

En se servant de l'équation différentielle (9.1), on trouve :

$$y(t_{n+1}) = y(t_n) + f(t_n, y(t_n))h + f'(t_n, y(t_n))h^2 + o(h^2)$$

et on a :

$$f\,'(t, y(t)) = \frac{\partial f(t, y(t))}{\partial t} + \frac{\partial f(t, y(t))}{\partial y}\, y'(t),$$

donc

$$f\,'(t, y(t)) = \frac{\partial f(t, y(t))}{\partial t} + \frac{\partial f(t, y(t))}{\partial y} f(t, y(t)).$$

Ainsi, on obtient

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \frac{h^2}{2}(\frac{\partial f(t_n, y(t_n))}{\partial t} + \frac{\partial f(t_n, y(t_n))}{\partial y} f(t_n, y(t_n))) + o(h^2), \quad (9.2)$$

en négligeant les termes d'ordres supérieurs ou égaux à 3. D'où

$$y(t_{n+1}) \simeq y(t_n) + hf(t_n, y(t_n)) + \frac{h^2}{2}(\frac{\partial f(t_n, y(t_n))}{\partial t} + \frac{\partial f(t_n, y(t_n))}{\partial y} f(t_n, y(t_n))). \quad (9.3)$$

Cette relation sera la base de la méthode de Taylor.

Commentaire

En fait, la méthode de Taylor consiste à approcher la solution de l'équation (9.1) par des arcs de paraboles au lieu des segments de droites (des tangentes) utilisés dans la méthode d'Euler.

Algorithme de Taylor d'ordre 2

(1) Etant donné un pas de temps h, une condition initiale (t_0, y_0), et un nombre maximal d'itérations N.

(2)

Pour $0 \leq n \leq N$

$$y_{n+1} = y_n + hf(t_n, y_n) + \tfrac{h^2}{2}(\tfrac{\partial f(t_n, y_n)}{\partial t} + \tfrac{\partial f(t_n, y_n)}{\partial y} f(t_n, y_n))$$

et

$$t_{n+1} = t_n + h$$

écrire t_{n+1} *et* y_{n+1}.

(3) Arrêt.

Remarque 9.5.

Dans cet algorithme, on a remplacé la solution analytique $y(t_n)$ par son approximation y_n dans la relation (9.3). On en conclut que les erreurs se propagent d'une itération à une autre.

Exemple 9.5. *Soit l'équation différentielle déjà résolue par la méthode d'Euler*

$$\begin{cases} y'(t) = -y(t) + t + 1 \\ y(0) = 1. \end{cases}$$

On prend $h = 0,1$. Dans ce cas $f(t, y) = -y + t + 1$ et $\frac{\partial f}{\partial t}(t, y) = 1$, $\frac{\partial f}{\partial ty}(t, y) = -1$. Le tableau suivant rassemble les résultats des dix premières itérations ce qui permet de comparer les

solutions numérique et analytique et de constater la croissance de l'erreur et la comparer avec la méthode d'Euler

t_i	$y(t_i)$	y_i	$\|y(t_i) - y_i\|$
$0,0$	$1,000000$	$1,000000$	$0,000000$
$0,1$	$1,004837$	$1,005000$	$0,000163$
$0,2$	$1,018731$	$1,019025$	$0,000400$
$0,3$	$1,040818$	$1,041218$	$0,000482$
$0,4$	$1,070302$	$1,070802$	$0,000482$
$0,5$	$1,106531$	$1,107075$	$0,000544$
$0,6$	$1,148812$	$1,149404$	$0,000592$
$0,7$	$1,196580$	$1,197210$	$0,000625$
$0,8$	$1,249329$	$1,249975$	$0,000646$
$0,9$	$1,306570$	$1,307228$	$0,000658$
$1,0$	$1,367879$	$1,368541$	$0,000662$

Remarque 9.6. *On remarque que l'erreur est plus petite avec la méthode de Taylor d'ordre 2 qu'avec la méthode d'Euler.*

Remarque 9.7. *Si l'on veut encore réduire la marge d'erreur, on poursuive le développement de Taylor dans (9.2) jusqu'à des termes d'ordre élevé. On doit alors évaluer les dérivées de la fonction $f(t, y(t))$ d'ordre de plus en plus élevé, ce qui nécessite le calcul supplémentaire de :*

$$\frac{\partial^2 f}{\partial t^2}, \frac{\partial^2 f}{\partial y^2}, \frac{\partial^2 f}{\partial t \partial y}, ..., \frac{\partial^{i+j} f}{\partial t^i \partial^j}.$$

Pour cette raison, les méthodes obtenues sont difficiles à utiliser. pour contourner cette difficulté on développe les méthodes de Runge-Kutta.

9.4 Méthodes de Runge-Kutta

Il serait avantageux de disposer de méthodes d'ordres de plus en plus élevées tout en évitant les inconvénients des méthodes de Taylor, qui nécessitent l'évaluation des dérivées partielles de la fonction $f(t, y(t))$. Une voie est tracée par les méthodes de Runge[3]-Kutta[4], qui sont calquées sur les méthodes de Taylor du même ordre.

9.4.1 Méthodes de Runge-Kutta d'ordre 2

On a vu que le développement de la méthode de Taylor passe par la relation :

$$y(t_{n+1}) = y(t_n) + hf(t_n, y(t_n)) + \frac{h^2}{2}(\frac{\partial f(t_n, y(t_n))}{\partial t} + \frac{\partial f(t_n, y(t_n))}{\partial y} f(t_n, y(t_n))) + o(h^2) \quad (9.4)$$

Le but est de remplacer cette dernière relation par une expression équivalente possédant le même ordre de précision ($o(h^2)$). On propose la forme :

$$y(t_{n+1}) = y(t_n) + a_1 hf(t_n, y(t_n)) + a_2 hf(t_n + a_3 h, y(t_n) + a_4 h) \quad (9.5)$$

[3] Carle Runge, 1856-1927
[4] Martin Kutta, 1867-1944

où on doit déterminer les paramètres a_1, a_2, a_3 et a_4 de telle sorte que les expressions (9.4) et (9.5) aient toutes les deux une erreur en $o(h^2)$. Pour y arriver, on doit utiliser le développement de Taylor en deux variables autour du point $(t_n, y(t_n))$. On a ainsi :

$$f(t_n + a_3 h, y(t_n) + a_4 h) = f(t_n, y(t_n)) + a_3 h \frac{\partial f(t_n, y(t_n))}{\partial t} + a_4 h \frac{\partial f(t_n, y(t_n))}{\partial y} + o(h)$$

La relation (9.5) devient alors :

$$\begin{aligned} y(t_{n+1}) &= y(t_n) + (a_1 + a_2) h f(t_n, y(t_n)) + a_2 a_3 h^2 \tfrac{\partial f(t_n, y(t_n))}{\partial t} \\ &+ a_2 a_4 h^2 \tfrac{\partial f(t_n, y(t_n))}{\partial y} + o(h^2) \end{aligned} \tag{9.6}$$

On voit immédiatement que les expressions (9.4) et (9.6) sont de même ordre. Pour déterminer les coefficients a_i, $i = 1, 2, 3, 4$, il suffit de comparer ces deux expressions terme à terme. On obtient un système non linéaire de 3 équations à 4 inconnues :

$$\begin{cases} a_1 + a_2 = 1 \\ a_2 a_3 = \frac{1}{2} \\ a_2 a_4 = \frac{f(t_n, y(t_n))}{2} \end{cases} \tag{9.7}$$

Le système (9.7) a moins d'équations que d'inconnues et donc n'a pas de solution unique. Cela offre une marge de manoeuvre qui favorise la mise au point de plusieurs variantes de la méthode de Runge-Kutta. Voici le choix le plus couramment utilisé.

9.4.2 Méthode d'Euler modifiée

Cette méthode correspond au choix suivant des coefficients a_i :

$$a_1 = a_2 = \frac{1}{2}, a_3 = 1, et\ a_4 = f(t_n, y(t_n))$$

Il suffit ensuite de remplacer ces valeurs dans l'équation (9.5). Pour ce faire, on néglige le terme en $o(h^2)$ et on remplace $y(t_n)$ par son approximation y_n. On obtient alors l'algorithme suivant :

Algorithme d'Euler modifié

(1) Etant donné un pas de temps h, une condition initiale (t_0, y_0) , et un nombre maximal d'itérations N.

(2)

Pour $0 \leq n \leq N$

$$\begin{aligned} \widehat{y} &= y_n + h f(t_n, y_n) \\ y_{n+1} &= y_n + \tfrac{h}{2}[f(t_n, y_n) + f(t_{n+1}, \widehat{y})] \end{aligned}$$

et $t_{n+1} = t_n + h$

écrire t_{n+1} *et* y_{n+1}.

(3) Arrêt.

Remarque 9.8.
Pour faciliter les calculs, l'évaluation de y_{n+1} est faite en deux étapes. La variable temporaire $\widehat{y}$ correspond tout simplement à une itération de la méthode d'Euler. On fait ainsi une prédiction $\widehat{y}$ de la solution en t_{n+1} qui est corrigée (et améliorée) à la deuxième étape de l'algorithme.

Exemple 9.6.
Soit le problème de cauchy :

$$\begin{cases} y'(t) &= -y(t) + t + 1 \\ y(0) &= 1. \end{cases}$$

On choisit le pas de temps $h = 0,1$.

Itération 1 : $\widehat{y} = hf(t_0, y_0) = 1$
qui est le résultat obtenu à l'aide de la méthode d'Euler. La deuxième étape donne :

$$y_1 = y_0 + \frac{h}{2}[f(t_0, y_0) + f(t_1, \widehat{y})] = 1,005.$$

Itération 2 : $\widehat{y} = hf(t_1, y_1) = 1,0145$
La correction conduit à son tour à :

$$y_2 = y_1 + \frac{h}{2}[f(t_1, y_1) + f(t_2, \widehat{y})] = 1,019025.$$

9.4.3 Méthode du point milieu

Une autre méthode de Runge-Kutta d'ordre 2 qui est très utilisée est la méthode du point milieu, qui correspond au choix suivant des coefficients a_i.

$$a_1 = 0, a_2 = 1, a_3 = \frac{1}{2}, et\ a_4 = \frac{f(t_n, y(t_n)}{2}.$$

En remplaçant ces valeurs des coefficients a_i dans l'équation (9.6), on obtient alors l'algorithme suivant :

Algorithme du point milieu

(1) Etant donné un pas de temps h, une condition initiale (t_0, y_0) , et un nombre maximal d'itérations N.
(2)

$$\begin{aligned} &Pour\ 0 \leq n \leq N \\ &\qquad K &&= hf(t_n, y_n) \\ &\qquad y_{n+1} &&= y_n + h[f(t_n + \tfrac{h}{2}, y_n + \tfrac{K}{2})] \\ &et \qquad t_{n+1} &&= t_n + h \\ &écrire \qquad t_{n+1} \quad et \quad y_{n+1}. \end{aligned}$$

(3) Arrêt.

Remarque 9.9. *La méthode est dite du point milieu car la fonction $f(t, y)$ est évaluée au point milieu de l'intervalle $[t_n, t_{n+1}]$.*

Remarque 9.10. *Les méthodes d'Euler modifiée et du point milieu étant du même ordre de troncature locale, leur précision est semblable. D'autre choix sont possibles pour les coefficients a_i, mais nous nous limitons aux deux méthodes précédentes.*

9.4.4 Méthode de Runge-Kutta d'ordre 4

En reprenant le développement de Taylor de la fonction f, mais cette fois à l'ordre 5, un raisonnement similaire à celui qui a mené aux méthodes de Runge-Kutta d'ordre 2 aboutit à un système de 8 équations non linéaires comprenant 10 inconnues. Le résultat final est la méthode de Runge-Kutta d'ordre 4, qui représente un outil d'une grande utilité.

Algorithme de Runge-Kutta d'ordre 4

(1) Etant donné un pas de temps h, une condition initiale (t_0, y_0) , et un nombre maximal d'itération N
(2)

$$
\begin{array}{lll}
\textit{Pour } 0 \leq n \leq N & & \\
& K_1 = hf(t_n, y_n) & \\
& K_2 = hf(t_n + \frac{h}{2}, y_n + \frac{K_1}{2}) & \\
& K_3 = hf(t_n + \frac{h}{2}, y_n + \frac{K_2}{2}) & \\
& K_4 = hf(t_n + h, y_n + K_3) & \\
& y_{n+1} = y_n + \frac{1}{6}[K_1 + 2K_2 + 2K_3 + K_4] & \\
\textit{et} & t_{n+1} = t_n + h & \\
\textit{écrire} & t_{n+1} \quad \textit{et} \quad y_{n+1}. &
\end{array}
$$

(3) Arrêt.

Remarque 9.11. *La méthode de Rung-Kutta d'ordre 4 est très fréquemment utilisée en raison de sa grande précision qui est mise en évidence dans l'exemple suivant :*

Exemple 9.7. *Soit l'équation différentielle déjà résolue par la méthode d'Euler*

$$\begin{cases} y'(t) = -y(t) + t + 1 \\ y(0) = 1. \end{cases}$$

On prend $h = 0,1$ et $f(t, y) = -y + t + 1$.

Itération 1 :

$$
\begin{array}{rcl}
K_1 & = & hf(t_0, y_0) = 0 \\
K_2 & = & hf(t_0 + \frac{h}{2}, y_0 + \frac{K1}{2}) = 0,005 \\
K_3 & = & hf(t_0 + \frac{h}{2}, y_0 + \frac{K_2}{2}) = 0,00475 \\
K_4 & = & hf(t_0 + h, y_0 + K_3) = 0,009525
\end{array}
$$

ce qui entraîne que : $y_1 = y_0 + \frac{1}{6}[K_1 + 2K_2 + 2K_3 + K_4] = 1,0048375.$

Itération 2 :

$$
\begin{array}{rcl}
K_1 & = & hf(t_1, y_1) = 0,00951625 \\
K_2 & = & hf(t_1 + \frac{h}{2}, y_1 + \frac{K_1}{2}) = 0,014040438 \\
K_3 & = & hf(t_1 + \frac{h}{2}, y_1 + \frac{K_2}{2}) = 0,0138142281 \\
K_4 & = & hf(t_1 + h, y_1 + K_3) = 0,0187309014
\end{array}
$$

ce qui entraîne que : $y_2 = y_1 + \frac{1}{6}[K_1 + 2K_2 + 2K_3 + K_4] = 1,0187309014.$

Le tableau suivant rassemble les résultats des dix premières itérations ce qui permet de comparer les solutions numérique et analytique et de constater la croissance de l'erreur, et la comparer avec les méthodes (Euler, Taylor).

t_i	$y(t_i)$	y_i	$\lvert y(t_i) - y_i\rvert$
$0,0$	$1,0000000000$	$1,000000$	0
$0,1$	$1,0048374180$	$1,0048375000$	$0,819 \times 10^{-7}$
$0,2$	$1,0187307798$	$1,0187309014$	$0,148 \times 10^{-6}$
$0,3$	$1,0408182207$	$1,0408184220$	$0,210 \times 10^{-6}$
$0,4$	$1,0703200460$	$1,0703202889$	$0,242 \times 10^{-6}$
$0,5$	$1,1065306597$	$1,1065309344$	$0,274 \times 10^{-6}$
$0,6$	$1,1488116361$	$1,1488119343$	$0,298 \times 10^{-6}$
$0,7$	$1,1965853034$	$1,1965856186$	$0,314 \times 10^{-6}$
$0,8$	$1,2493289641$	$1,2493292897$	$0,325 \times 10^{-6}$
$0,9$	$1,3065696598$	$1,3065799912$	$0,331 \times 10^{-6}$
$1,0$	$1,3678794412$	$1,3678797744$	$0,333 \times 10^{-6}$

Remarque 9.12.
On constate que l'erreur se situe autour de 10^{-6}, ce qui se compare avantageusement avec les erreurs obtenues à l'aide de méthodes d'ordre moins élevé (Euler, Taylor, Runge-Kutta d'ordre 2).

9.5 Méthodes à un pas générique

Une méthode de résolution numérique d'équations différentielles est dite à un pas si elle est de la forme :

$$\begin{cases} y_{i+1} = y_i + h\phi(t_i, y_i, h), \\ t_{i+1} = t_i + h,\ i = 0, 1, \ldots, n, \\ y_0 = y(0), \end{cases} \tag{9.8}$$

où n est le nombre de subdivisions de l'intervalle $[t_0, T]$, $\phi : [a, b] \times \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ une fonction que l'on supposera continue par rapport aux trois variables.
Choisir une méthode revient à choisir la fonction ϕ. Quelles conditions imposer à ϕ pour que la méthode fonctionne ?

Exemple 9.8. *1. Méthode d'Euler :$\phi(t, y, h) = f(t, y)$.*

2. Méthode du point milieu : $\phi(t, y, h) = f(t + \frac{h}{2}, y + \frac{h}{2} f(t, y))$.

Définition 9.1. *(**La convergence**) Une méthode à un pas est dite convergente sur $[t_0, T]$ si quelle que soit y_0 condition initiale on a*

$$\lim_{h \to 0} \max_{0 \le i \le n} |y(t_i) - y_i| = 0.$$

C'est à dire, pour tout $i = 0, \ldots, n$ on a convergence de la solution approchée y_i vers la solution exacte au point t_i, $y(t_i)$.

Définition 9.2. *(**La consistance**) Une méthode à un pas est dite consistante si pour tout y solution de $y' = f(t, y)$ on a*

$$\lim_{h \to 0} \max_{0 \le i \le n} \left| \frac{1}{h}(y(t_{i+1}) - y(t_i)) - \phi(t_i, y(t_i), h) \right| = 0.$$

C'est à dire, la méthode approche bien l'équation différentielle.

Définition 9.3. ***(La stabilité)*** *Une méthode à un pas est dite stable s'il existe une constante K indépendante de h telle que, pour tout $y_0, \hat{y}_0$ et $y_{i+1} = y_i + \phi(t_i, y(t_i), h)$ on a*

$$\hat{y}_{i+1} = \hat{y}_i + \phi(t_i, y(t_i), h) + \epsilon_i \quad \textit{vérifiant } |y_i - \hat{y}_i| \leq K|y_0 - \hat{y}_0| + \sum_{i=0}^{i-1} |\epsilon_i| \; \forall i = 0, \ldots, n.$$

C'est à dire, on peut contrôler la répercussion des erreurs d'arrondi.

Proposition 9.1. *Pour une méthode à un pas, la convergence implique la consistance.*

Proposition 9.2. *Si une méthode à un pas est consistante et stable, alors elle est convergente.*

9.5.1 Ordre d'une méthode à un pas

Il est intéressant de disposer d'un critère, valable dans le cas général, qui nous indiquera si telle méthode est meilleure qu'une autre (au sens qu'elle approche mieux la solution exacte). Pour cela, on introduit les définitions suivantes :

Définition 9.4. *(Erreur de Troncature locale)*
Soit M une méthode de résolution à un pas d'un problème de Cauchy, son algorithme est donc de la forme :

$$y_{i+1} = y_i + h\phi(t_i, y_i, h), \quad 0 \leq i \leq n.$$

On appelle " erreur de troncature locale au point $t = t_i$ " de M la quantité

$$e_i = \frac{y(t_{i+1}) - y(t_i)}{h} - \phi(t_i, y_i, h).$$

Soit f de classe C^p sur $[t_0, T] \times \mathbb{R}$. La méthode M est d'ordre p si $e_i = O(h^p)$ c.à.d

$$\exists C \in \mathbb{R}_+^* / \quad \left| \frac{y(i+1) - y(i)}{h} - \phi(t_i, y_i, h) \right| \leq C.h^p \; \forall \, i \in (0, n).$$

La méthode M convergera ($y_n \to y(n)$ quand $h \to 0$ $\forall \, i \in (0, n)$), donc d'autant plus vite que p sera grand.

Condition nécessaire et suffisante pour que M soit d'ordre $p (p \geq 1)$?

On suppose que $f \in C^p([t_0, T], \mathbb{R})$ et ϕ de classe C^p par rapport à h. D'après ce qui précède, le pas doit être au voisinage de 0.

Théorème 9.2. *M est d'ordre $p \geq 1 \Leftrightarrow \phi$ est choisie telle que :*

$$\frac{\partial^k}{\partial h^k} \phi(t, y, h)_{|h=0} = \frac{1}{k+1} f^{(k)}(x, y), \quad 0 \leq k \leq p - 1.$$

Ordres respectifs des méthodes d'Euler et du point milieu

Méthode d'Euler : *Dans ce cas $\phi(t, y, h) = f(t, y)$ (f supposée de classe C^1)*
Déterminons $p \geq 1$ tel que $\frac{\partial^k}{\partial h^k} \phi(t, y, h)_{|h=0} = \frac{1}{k+1} f^{(k)}(x, y)$, $0 \leq k \leq p - 1$.

1. *$k = 0$: $\phi(t, y, 0) = f(t, y)$ vérifiée*
2. *$k = 1$: $\frac{\partial}{\partial h} \phi(t, y, h)_{|h=0} = 0 \neq \frac{1}{2} f'(t, y) = \frac{1}{2}(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f)$.*

On en tire que $p = 1$, donc la méthode d'Euler est d'ordre 1.

Méthode du point milieu : *Dans ce cas $\phi(t, y, h) = f(t + \frac{h}{2}, y + \frac{h}{2} f(t, y))$ (f supposée de classe C^2). Déterminons $p \geq 1$ tel que*

$$\frac{\partial^k}{\partial h^k}\phi(t, y, h)_{|h=0} = \frac{1}{k+1} f^{(k)}(x, y), \quad 0 \leq k \leq p-1.$$

1. $k = 0$: $\phi(t, y, 0) = f(t, y)$ *vérifiée*
2. $k = 1$: $\frac{\partial}{\partial h}\phi(t, y, h)_{|h=0} = \frac{1}{2} f'(t, y) = \frac{1}{2}(\frac{\partial f}{\partial t} + \frac{\partial f}{\partial y} f)$. *En effet ;*

$$\frac{\partial}{\partial h}\phi(t, y, h) = \frac{\partial f}{\partial t}\frac{\partial(t + \frac{h}{2})}{\partial h} + \frac{\partial f}{\partial y}\frac{\partial(y + \frac{h}{2} f(t, y))}{\partial h}$$

$$\Rightarrow \frac{\partial}{\partial h}\phi(t, y, h)_{|h=0} = \frac{1}{2}\left(\frac{\partial f(t, y)}{\partial t} + \frac{1}{2}\frac{\partial f(t, y)}{\partial y} f(t, y)\right).$$

On montrera de même que, pour $k = 2$, l'égalité n'a pas lieu. On en tire que $p = 2$, donc la méthode d'Euler est d'ordre 2.

9.6 Exercices

Exercice 9.1. *Faire trois itérations avec le pas $h = 0,1$ des méthodes d'Euler, d'Euler modifiée, du point milieu et de Runge-Kutta d'ordre 4 pour les équations différentielles suivantes :*

a) $y'(t) = tsin(y(t))$ *avec* $y(0) = 2$.

b) $y'(t) = t^2 + (y(t))^2 + 1$ *avec* $y(1) = 0$.

c) $y'(t) = (y(t)e^t$ *avec* $y(0) = 2$.

Exercice 9.2. *Soit le problème de Cauchy suivant :*

$$\begin{cases} y'(t) = 2y(t) \;\; t \in [0,1] \\ y(0) = 1. \end{cases}$$

1. *Trouver la solution exacte de ce problème.*
2. *Prendre cinq subdivisions sur l'intervalle $[0,1]$ et appliquer les méthodes : d'Euler, Euler modifiée, point milieu.*
   Ecrire les valeurs obtenues, dans un tableau et comparer chacune avec la valeur exacte (en calculant l'erreur relative correspondante).
3. *Quelle est la méthode qui donne de meilleures valeurs approchées de la solution exacte ?*

Exercice 9.3. *Soit le problème de Cauchy suivant :*

$$\begin{cases} y'(t) = t + y(t) \;\; t \in [0,2] \\ y(0) = 1,24, \end{cases}$$

qui possède la solution analytique :

$$y(t) = 2,24e^t - (t-1)$$

I. Résoudre numériquement le même problème à l'aide de la méthode :

1. *d'Euler, avec un pas $h = 0,2$.*
2. *de Runge-Kutta d'ordre 4, avec le pas $h = 1$.*

II. Comparer au point $t = 1$, les valeurs numériques à la valeur analytique et donner (en %) l'erreur relative commise par chacune des deux méthodes.

Exercice 9.4. *Soit le problème de Cauchy suivant :*

$$(P) \quad \begin{cases} t^2y'(t) - ty(t) + 1 = 0, \quad t \in [2,3] \\ \\ y(2) = 0. \end{cases}$$

1. *Résoudre numériquement le problème (P) à l'aide de la méthode :*
   i) *du point milieu, avec un pas $h = \frac{1}{3}$.*
   ii) *de Runge-Kutta d'ordre 4, avec un pas $h = 1$.*
   Ecrire les valeurs obtenues (de la question (i)) dans un tableau et comparer chacune d'elles avec la valeur exacte, sachant que la solution analytique de (P) est $y(t) = -\frac{t}{8} + \frac{1}{2t}$, $t \in [2,3]$.
2. *Quelle est la méthode qui donne la meilleure valeur approchée de la solution exacte au point $t = 3$?*

Exercice 9.5. *Soit l'équation différentielle :* $y'(t) = t(y(t))$ *avec* $y(1) = 2$, *dont on connaît la solution exacte :*

$$y(t) = 2e^{(t^2-1)/2}.$$

1. *En prenant successivement* $h = 0,5$, $h = 0,25$, $h = 0,125$ *et* $h = 0,0625$.
   Approcher dans chaque cas $y(2)$ *en appliquant la méthode de Taylor d'ordre 2 et calculer l'erreur absolue commise dans chaque cas en comparant les résultats obtenus avec la valeur exacte* $y(2)$.
2. *Conclure*

Exercice 9.6. *Etant données trois méthodes de résolution numérique* M_1, M_2 *et* M_3 *d'un problème de Cauchy du premier ordre, de même pas de discrétisation* h.

1. *Comparer la précision de ces trois méthodes sachant que les erreurs absolues respectives sont de la forme :*
   $$E_1 = o(h), \quad E_2 = o(h^2) \quad E_3 = o(h^4).$$
2. *Proposer trois méthodes de résolution dont les erreurs absolues sont* E_1, E_2 *et* E_3, *en donnant leurs algorithmes.*
3. *Soit le problème de Cauchy :*
   $$(P) \quad \begin{cases} (1+e^t)yy' = e^t, & t \in [0,1] \\ \\ y(0) = 1. \end{cases}$$
   a. *On pose* $h = \frac{1}{2}$. *Calculer deux approximations de la solution exacte du problème* (P) *en* $t = 1$ *avec quatre décimales.*
   b. *Comparer chaque approximation avec la solution exacte que l'on déterminera.*