Université Moulay ISMAÏL
Faculté des Sciences et Techniques d'Errachidia
Département de Mathématiques

Cours d'Analyse Numérique 1

Filière MIP-S4 (M148)

Pr. Samir KHALLOUQ

Année universitaire 2020-2021

Table des matières

Analyse Numérique C'est Quoi ? **1**

1 Notions sur les erreurs numériques **3**
1.1 Les différentes sources d'erreurs . . . 3
1.2 Traitement des nombres réels par ordinateur . . . 4
1.2.1 Comment les représentons-nous? . . . 4
1.2.2 Arithmétique flottante . . . 6
1.3 Instabilité numérique . . . 8

2 Interpolation numérique **11**
2.1 Motivation pratique . . . 11
2.2 Approximation polynomiale . . . 11
2.3 Interpolation polynomiale . . . 12
2.3.1 Représenter-evaluer un polynôme en un point . . . 12
2.3.2 Position du problème . . . 13
2.3.3 Polynômes de Lagrange . . . 14
2.4 Le polynôme d'interpolation de Newton . . . 15
2.4.1 Forme de Newton . . . 16
2.4.2 Existence et unicité . . . 16
2.4.3 Différences divisées . . . 17
2.4.4 Différence divisées et polynôme d'interpolation de Newton . . . 18
2.4.5 Algorithme de calcul des c_i . . . 19
2.5 Erreur d'interpolation . . . 20
2.5.1 Analyse de l'erreur d'interpolation linéaire . . . 21
2.5.2 Meilleurs points d'interpolation - phénomène de Runge . . . 21

3 Dérivation numérique **25**
3.1 Principe . . . 25
3.2 Dérivée première . . . 25
3.3 Formule générale en trois points . . . 27
3.4 Dérivées d'ordre supérieur . . . 27
3.5 Étude de l'erreur commise . . . 28

4 Intégration numérique **31**
4.1 Principe . . . 31
4.1.1 Intervalle de référence . . . 31
4.2 Exemples de formules de quadrature élémentaires . . . 31
4.3 Étude de l'erreur pour les formules de quadrature élémentaires . . . 32
4.3.1 Degré de précision d'une méthode d'intégration numérique . . . 32

4.3.2 Majorations d'erreur pour les méthodes de quadrature élémentaires . 34
4.3.3 Quelques outils/rappels 34
4.4 Méthodes de quadrature composites 38
4.4.1 Principe 38
4.4.2 Méthodes composites couramment utilisées 39

5 Résolution numérique des équations non linéaires $f(x) = 0$ 41
5.1 Introduction 41
5.2 Existence de solutions et localisation des solutions 41
5.3 Premiers résultats théoriques 41
5.4 Construction de solutions approchées 42
5.4.1 Méthodes de la dichotomie et de Lagrange 43
5.4.2 Méthodes de points fixe 45
5.4.3 La méthode de Newton 47
5.4.4 Méthode de la corde et méthode de la sécante 49

6 Initiation à la résolution des systèmes linéaires : méthode de Gauss et LU 51
6.1 Introduction 51
6.2 Systèmes linéaires 51
6.3 Opérations élémentaires sur les lignes 53
6.3.1 Multiplier une ligne par un scalaire 53
6.3.2 Permuter deux lignes 54
6.3.3 Opération $(L_i \leftarrow L_i + \lambda L_j)$ 55
6.4 Triangularisation 55
6.5 Méthode d'élimination de Gauss 56
6.6 Décomposition LU 57
6.7 Gauss avec pivot partiel 60
6.8 Gauss avec pivot total 60
6.9 Effet sur la décomposition LU 61
6.10 Cas des matrices bandes 63
6.11 Calcul de l'inverse d'une matrice 63

Bibliographie 67

Analyse Numérique C'est Quoi ?

Pour aborder le calcul numérique (à l'aide d'un outil informatique) des solutions d'un problème "réel", on passe par les étapes suivantes :

1. **Description qualitative des phénomènes physiques :** Cette étape est effectuée par des spécialistes des phénomènes que l'on veut quantifier (ingénieurs, chimistes, biologistes etc.....)

2. **Modélisation :** Il s'agit, à partir de la description qualitative précédente, d'écrire un modèle mathématique. Dans la plupart des cas, on ne saura pas calculer une solution analytique, explicite, du modèle ; on devra faire appel à des techniques de résolution approchée.

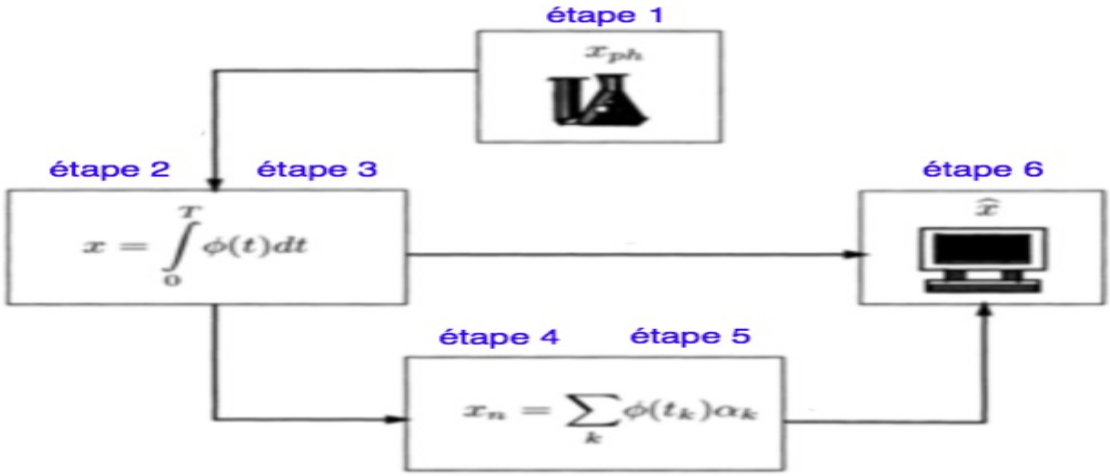
3. **Analyse mathématique :** Même si l'on ne sait pas trouver une solution explicite du modèle, il est important d'en étudier les propriétés mathématiques, dans la mesure du possible. Il est bon de se poser les questions suivantes :
   - Le problème est-il bien posé ? c'est-à-dire y-a-t'il existence et unicité de la solution ?
   - Les propriétés physiques auxquelles on s'attend sont elles satisfaites par les solutions du modèle mathématique ?

4. **Discrétisation et résolution numérique :** Un problème posé sur un domaine continu (espace - temps) n'est pas résoluble tel quel par un ordinateur, qui ne peut traiter qu'un nombre fini d'inconnues. Pour se ramener à un problème en dimension finie, on discrétise l'espace et/ou le temps.

5. **Analyse numérique :** Une fois le problème discret obtenu, il est raisonnable de se demander si la solution de ce problème est proche, et en quel sens, du problème continu.

6. **Mise en oeuvre, programmation et analyse des résultats**

La figure suivante donne les différentes étapes pour trouver une solution d'un problème réel :

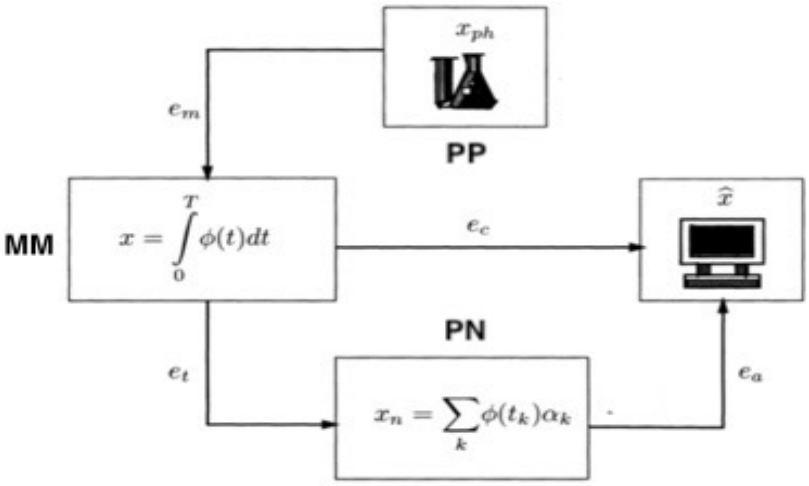


Chapitre 1

Notions sur les erreurs numériques

1.1 Les différentes sources d'erreurs

Dans le calcul numérique, l'erreur est inévitable. Le simple fait d'utiliser un ordinateur pour représenter des nombres réels introduit des erreurs. Ce qui est donc important, c'est de ne pas chercher à éliminer les erreurs. Mais plutôt pour pouvoir contrôler leur effet. De manière générale, nous pouvons identifier plusieurs niveaux d'erreurs qui surviennent lors de l'approximation et de la résolution d'un problème physique (voir Figure).



Au niveau le plus élevé se trouve l'erreur $\mathbf{e_m}$ qui se produit lorsque l'on force la réalité physique (**PP** signifie problème physique et $\mathbf{x_{ph}}$ désigne ses solutions) à obéir à un modèle mathématique (**MM** dont la solution est $\mathbf{x}$). De telles erreurs limiteront l'applicabilité du modèle mathématique à certaines situations et échappent au contrôle du calcul scientifique. D'autre part, il est souvent nécessaire d'introduire d'autres erreurs car toute procédure du modèle mathématique impliquant une suite infinie d'opérations arithmétiques ne peut être effectuée par l'ordinateur que de manière approximative. Par exemple, le calcul de la somme d'une série se fera nécessairement de manière approximative en considérant une troncature appropriée.
Il faudra donc introduire un problème numérique, **PN**, dont la solution $\mathbf{x_n}$ diffère de $\mathbf{x}$ par erreur et qui est appelée erreur de troncature. De telles erreurs ne sont pas exactes

uniquement dans les modèles mathématiques qui sont déjà définis en dimension finie (par exemple, lors de la résolution d'un système linéaire). La somme des erreurs $\mathbf{e_a}$ et $\mathbf{e_t}$ constitue l'erreur de calcul $\mathbf{e_c}$.

1.2 Traitement des nombres réels par ordinateur

Tout le monde connaît l'ensemble $\mathbb{R}$ des nombres réels, mais, la manière dont les ordinateurs les traitent est peut-être moins connue. D'une part, puisque les machines ont des ressources limitées, seul un sous-ensemble $\mathbb{F}$ de dimension finie de $\mathbb{R}$ peut être représenté. Ce sont des nombres à virgule flottante. D'autre part, comme nous le verrons dans la section 1.2.2, $\mathbb{F}$ est caractérisé par des propriétés différentes de celles de $\mathbb{R}$. La raison est que tout nombre réel x est en principe tronqué par la machine, donnant naissance à un nouveau nombre (appelé nombre à virgule flottante), noté $fl(x)$, qui ne coïncide pas nécessairement avec le nombre d'origine x.

1.2.1 Comment les représentons-nous?

Pour nous familiariser avec les différences entre $\mathbb{R}$ et $\mathbb{F}$, faisons quelques expériences utilisant MATLAB qui illustrent la manière dont un ordinateur (par exemple un PC) traite les nombres réels. Que nous utilisions MATLAB plutôt qu'un autre langage est juste une question de commodité. Les résultats de notre calcul dépendent en effet principalement du langage de programmation. Considérons le nombre rationnel $x = \frac{1}{7}$, dont la représentation décimale est 0.142857. Il s'agit d'une représentation infinie, car le nombre de chiffres décimaux est infini. Pour obtenir sa représentation informatique, introduisons après l'invite le rapport $\frac{1}{7}$ et obtenons

» $\frac{1}{7}$
ans = 0.1429

qui est un nombre avec seulement quatre chiffres décimaux, le dernier étant différent du quatrième chiffre du numéro d'origine. Si nous considérions maintenant $\frac{1}{3}$, nous trouverions 0.3333, de sorte que le quatrième chiffre décimal serait maintenant exact. Ce comportement est dû au fait que les nombres réels sont arrondis sur l'ordinateur. Cela signifie, tout d'abord, que seul un nombre fixe de chiffres décimaux est retourné, et de plus le dernier chiffre décimal qui apparaît est augmenté d'une unité chaque fois que le premier chiffre décimal ignoré est supérieur ou égal à 5. La première remarque à faire est que l'utilisation de seulement quatre chiffres décimaux pour représenter des nombres réels est discutable. En effet, la représentation interne du nombre est faite avec autant de 16 chiffres décimaux, et ce que nous avons vu est simplement l'un des nombreux formats de sortie possibles de MATLAB. Le même nombre peut prendre une expression différente selon la déclaration de format spécifique qui est faite.

***Exemple* 1.** *Pour le nombre $\frac{1}{7}$, certains formats de sortie possibles sont :*

format short	*rendements*	$0.1429,$
format short e	$''$	$1.4286e-01,$
format short g	$''$	$0.14286,$
format long	$''$	$0.142857142857143.$
format long e	$''$	$1.428571428571428e-01,$
format long	$''$	$0.142857142857143,$

En général, un ordinateur stocke un nombre réel de la manière suivante

$$x = (-1)^s.(0.a_1a_2a_3\ldots a_t).\beta^e = (-1)^s.m.\beta^{e-t}, \quad a_1 \neq 0, \tag{1.1}$$

- ***s*** *est 0 ou 1,*
- β *(un entier positif supérieur ou égal à 2) est* ***la base*** *adoptée par l'ordinateur spécifique concerné,*
- ***m*** *est un entier appelé* ***mantisse*** *dont la longueur* ***t*** *est le nombre maximum de chiffres* a_i *(allant entre 0 et* $\beta - 1$*) qui sont stockés, on l'appel nombre de* ***bits*** *ou de* ***digits****,*
- ***e*** *est un nombre entier appelé exposant, borné dans un intervalle* (L, U) *(avec* $L < 0$ *et* $U > 0$*)*

Les nombres dont la forme est donnée en (1.1) sont appelés ***nombres à virgule flottante "n.v.f"****, car la position du point décimal n'est pas fixe. Les chiffres* $a_1, a_2, \ldots, a_p$ *(avec* $p \leq t$*) sont souvent appelés les p premiers* ***digits*** *significatifs de* x*.*
La condition $a_1 \neq 0$ *garantit qu'un nombre ne peut pas avoir plusieurs représentations.*

Remarque 1.1. *Sans la restriction* $a_1 \neq 0$*, le nombre* $\dfrac{1}{10}$ *pourrait être représenté (en décimal) comme* $0, 1.10^0$*, mais aussi comme* $0, 01.10^1$*,... etc.*

L'ensemble $\mathbb{F}$ *est donc complètement caractérisé par la base* β*, le nombre des digits significatifs* t*,* L *et* U*. Ainsi, il est noté* $\mathbb{F}(\beta, t, L, U)$*.*

Exemple 2.

- *Pour* $\beta = 2$ *(****système binaire****), les* $a_k = 0$ *ou* 1 *appelés bits. C'est le système numérique des machines.*

$$(1001)_2 = 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 = (9)_{10}.$$

- *Pour* $\beta = 8$ *(****système octal****), les digits sont* $a_k = 0, 1, \ldots, 7$*.*

$$(31)_8 = 3 \times 8^1 + 1 \times 8^0 = (25)_{10}.$$

- *Pour* $\beta = 10$ *(****système décimal****), les digits sont* $a_k = 0, 1, \ldots, 9$*.*
- *Pour* $\beta = 16$ *(****système hexadécimal****), les digits sont* $a_k = 0, 1, \ldots, 15$*. Par convention, on adopte la notation suivante pour* $a_k = 0, 1, \ldots, 15$*.*

10	*11*	*12*	*13*	*14*	*15*
A	*B*	*C*	*D*	*E*	*F*

***Exemple* 3.** *Dans MATLAB nous avons* $\mathbb{F} = \mathbb{F}(2, 53, -1021, 1024)$ *(en effet, 53 digits significatifs en base 2 correspondent aux 15 digits significatifs qui sont montrés par MATLAB en base 10 avec le format long, voir Exemple 1).*

Heureusement, l'erreur d'arrondi qui est inévitablement générée chaque fois qu'un nombre réel $x \neq 0$ *est remplacé par son représentant* $fl(x)$ *dans* $\mathbb{F}$*, est faible, car*

$$\frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\epsilon_M, \text{ où } \epsilon_M = \beta^{1-t}, \tag{1.2}$$

Exemple* 4.** *Dans MATLAB,* ϵ_M *peut être obtenu via la commande* ***eps*, et nous obtenons* $\epsilon_M = 2^{-52} \backsimeq 2.2204.10^{-16}$.

Soulignons que dans (1.2) nous estimons ***l'erreur relative*** *sur* x*, qui est plus significative que* ***l'erreur absolue*** $|x - fl(x)|$*, car cette derniere ne tient pas compte de l'ordre de grandeur de* x*, contrairement au premiere.*
Le nombre 0 *n'appartient pas à* $\mathbb{F}$*, dans ce cas on aurait* $a_1 = 0$ *dans (1.1), c'est pour cela qu'on le traite séparément. De plus,* L *et* U *étant finis, on ne peut pas représenter des nombres dont la valeur absolue est soit arbitrairement grande soit arbitrairement petite. Précisément, le plus petit et le plus grand nombre réel positif de* $\mathbb{F}$ *sont donnés respectivement par*

$$x_{min} = \beta^{L-1}, \quad x_{max} = \beta^U(1 - \beta^{-t}).$$

Dans Matlab, ces valeurs peuvent être obtenues via les commandes ***realmin*** *et* ***realmax****, ce qui donne*

$$x_{min} = 2.225073858507201.10^{-308}, \quad x_{max} = 1.7976931348623158.10^{+308}.$$

***Remarque* 1.2.**

- $x < x_{min}$ *produit un message de sous-dépassement (****underflow****) et il est traité comme* 0.
- $x > x_{max}$ *donne un message de débordement (****overflow****) et il est stocké dans la variable* ***Inf*** *(qui est la représentation informatique de* $+\infty$*).*
- *Les éléments dans* $\mathbb{F}$ *sont plus denses près de* x_{min}*, et moins denses près de* x_{max}*. En fait, le nombre dans* $\mathbb{F}$ *le plus proche de* x_{min} *(à sa droite) et celui le plus proche de* x_{max} *(à sa gauche) sont respectivement*

$$x^+_{min} = 2,225073858507202,10^{-308}, \quad x^-_{max} = 1,7976931348623157,10^{+308}.$$

- $x^+_{min} - x_{min} \backsimeq 10^{-323}$*, tandis que* $x_{max} - x^-_{max} \backsimeq 10^{292}$*. Cependant, la distance relative est faible dans les deux cas, comme on peut le déduire de (1.2)*

1.2.2 Arithmétique flottante

Les erreurs de représentations ont des conséquences plus ou moins graves sur les opérations classiques sur les nombres (Arithmétique flottante). Ces opérations sont effectuées et notées de la manière suivante : Soient x*,* y *deux nombres réels, on a*

$x \oplus y$	$fl(fl(x) + fl(y))$	addition
$x \ominus y$	$fl(fl(x) - fl(y))$	soustraction
$x \otimes y$	$fl(fl(x) \times fl(y))$	multiplication
$x \oslash y$	$fl(fl(x) \div fl(y))$	division

***Exemple* 5.** *Application sur une machine à $t = 4bits$*

- ***La multiplication***

$$\begin{aligned} 1.429 \otimes 13 &= fl(fl(1.429) \times fl(13)), \\ &= fl(0.1429 \times 10^1 \times 0.1300 \times 10^2), \\ &= fl(0.018577 \times 10^3), \\ &= 0.185\underline{8} \times 10^2. \end{aligned}$$

Si on fait la multiplication directement on trouve 18.577. Donc

$$\begin{aligned} Erreur \;\; absolue &= |18.577 - 0.1858.10^2| = 0.003, \\ Erreur \;\; relative &= \frac{0.003}{18.577} \backsimeq 0.000165 = 0.0165\%_{\diagup}. \end{aligned}$$

- ***La division*** *se fait de la même manière*

$$\begin{aligned} 0.56789 \times 10^3 \oslash 0.1234321 \times 10^{-3} &= fl(fl(0.56789 \times 10^3) \div fl(0.1234321 \times 10^{-3})), \\ &= fl(0.5679 \times 10^3 \div 0.1234 \times 10^{-3}), \\ &= fl(4.602106969 \times 10^6), \\ &= 0.4602 \times 10^7. \end{aligned}$$

- ***Addition et Soustraction****, la règle est d'ajouter des* 0 *dans la mantisse du nombre qui a le plus petit exposant (aligner avec le plus grand exposant)*

$$\begin{aligned} 4.0351 \times 10^5 \oplus 0.1978 \times 10^4 &= fl(0.4035 \times 10^6 + 0.1978 \times 10^4), \\ &= fl(0.4035 \times 10^6 + 0.\underline{00}1978 \times 10^6), \\ &= fl(0.405478 \times 10^6), \\ &= 0.4055 \times 10^6. \end{aligned}$$

Les opérations algébriques élémentaires sur les nombres à virgule flottante ne vérifient pas toutes les propriétés des opérations sur $\mathbb{R}$. Précisément, la commutativité tient toujours pour l'addition (c'est-à-dire $fl(x + y) = fl(y + x)$) ainsi que pour la multiplication ($fl(xy) = fl(yx)$), mais d'autres propriétés telles que l'associativité et la distributivité sont violées.

***Exemple* 6.** *Ona :* $122 \times (333 + 695) = 122 \times 333 + 122 \times 695 = 125416$
Maintenant on fait les calculs sur une machine à $t = 3bits$

$$\begin{aligned} 122 \otimes (333 \oplus 695) &= fl(0.122 \times 10^3 \times fl(0.333 \times 10^3 + 0.695 \times 10^3)), \\ &= fl(0.122 \times 10^3 \times fl(1.028 \times 10^3)), \\ &= fl(0.122 \times 10^3 \times 0.103 \times 10^4), \\ &= fl(0.012566 \times 10^7), \\ &= 0.126 \times 10^6. \end{aligned}$$

$$\begin{aligned}122 \otimes 333 \oplus 122 \otimes 695 &= fl(fl(0.122 \times 10^3 \times 0.333 \times 10^3),\\ &\quad + fl(0.122 \times 10^3 \times 0.695 \times 10^3)),\\ &= fl(fl(0.040626 \times 10^6) + fl(0.08479 \times 10^6)),\\ &= fl(0.406 \times 10^5 + 0.848 \times 10^5),\\ &= fl(1.254 \times 10^5),\\ &= 0.125 \times 10^6.\end{aligned}$$

Donc $122 \otimes (333 \oplus 695) \neq 122 \otimes 333 \oplus 122 \otimes 695$.
Un autre problème qu'il faut faire aussi attention et qu'on appelle ***phénomène d'absorption****. On le trouve lorsqu'on effectue l'addition ou la soustraction de deux nombres* x *et* y $(x \pm y)$*, où l'ordre de grandeur de* y *est beaucoup plus petit que celui de* x*.*

***Exemple* 7.** *Soient* $x = 0.7654 \times 10^2$ *et* $y = 0.4856 \times 10^{-2}$*, on cherche à calculer* $x \oplus y$ *sur une machine à* t *= 4bits*

$$\begin{aligned}x \oplus y &= fl(fl(x) + fl(y)),\\ &= fl((0.7654 + 0.00004856) \times 10^2),\\ &= fl(0.76544856 \times 10^2),\\ &= 0.7654 \times 10^2 = x.\end{aligned}$$

Donc y *est complètement absorbée (négligée). On peut affirmer aussi que la propriété dans* $\mathbb{R}$ *:* $x + y = x \Leftrightarrow y = 0$ *n'est pas vérifier en arithmétique de l'ordinateur (c.à.d. dans* $\mathbb{F}$*).*

1.3 Instabilité numérique

Définition 1.1. *L'agorithme est une démarche qui décrit, à l'aide des opérations élémentaires, toutes les étapes nécessaires à la résolution d'un problème spécifique.*
Lorsqu'une erreur de donnée, de représentation ou de calcul est commise dans un algorithme, elle est transmise dans les calculs suivants : on parle de ***propagation*** *ou* ***accumulation d'erreurs****. Celà peut aboutir à des résultats peu précis, voir aberrants. La méthode ou l'algorithme qui cause ce type d'erreurs est dit* ***numériquement instable****.*

***Exemple* 8.** *On cherche à calculer pour* $n \geq 0$ *donné l'intégrale*

$$I_n = \int_0^1 \frac{x^n}{10 + x} dx.$$

On peut montrer que $I_n = \dfrac{1}{n} - 10I_{n-1},\ n \geq 1,$ *avec* $I_0 = ln(\dfrac{11}{10})$ *(en exercice).* I_0 *ne peut être calculée exactement, mais avec une petite erreur* ϵ_0*. Cette erreur a des conséquences sur le calcul de* I_n*.*
En effet, supposons que I_{n-1} *est calculée avec une erreur* ϵ_{n-1} *, alors*

$$I_n + \epsilon_n = \frac{1}{n} - 10(I_{n-1} + \epsilon_{n-1}), n = 1, 2, \ldots$$

Les erreurs successives vérifient alors la relation $\epsilon_n = -10\epsilon_{n-1}$*. Donc* $\epsilon_n = (-10)^n\epsilon_0$*.*
On remarque que l'erreur augmente de manière exponentielle même si ϵ_0 *est petit. Le problème provient de la puissance croissante de* 10*. Par exemple, si* $\epsilon_0 = 10^{-10}$*, on aura*

$\epsilon_{30} = (-10)^{30} \times 10^{-10} \backsimeq 10^{+20}$, *ce qui est énorme.*
Pour résoudre ce problème, on inverse la récurrence pour $k \geq n$, *c.à.d*

$$I_{p-1} = \frac{1}{10p} - \frac{1}{10} I_p, \ p = k, k-1, \ldots, n.$$

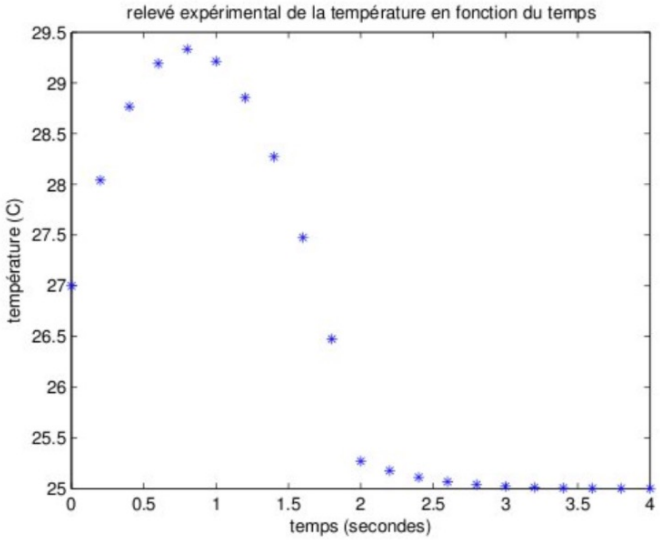
Par exemple, on part d'une valeur approchée grossière de I_{40} *avec une erreur* ϵ. *En descendant à* I_{30}, *l'erreur est* $\epsilon_{30} = (\frac{-1}{10})^{10}\epsilon$. *Donc si* $\epsilon \backsimeq 1$, *alors* $\epsilon_{30} \backsimeq 10^{-10}$.

Chapitre 2

Interpolation numérique

2.1 Motivation pratique

Considérons le relevé expérimental de la température d'une solution chimique au cours du temps : à des temps discrets, notés $t_1, t_2, ..., t_N$ on mesure les températures T_i et on reporte les résultats sur un graphe en prenant le temps en abscisse et la température en ordonnée, comme ci-dessous



La théorie chimique prévoit que la température T dépend du temps suivant la loi

$$T = f(t)$$

Pour savoir si l'expérience a été éffectuée dans de bonnes conditions, il faut pouvoir comparer T_i à $f(t_i)$. Parfois f n'est pas connue explicitement mais seulement tabulée et il est important de disposer de valeurs intermédiaires. L'approximation de f vue comme une fonction, en utilisant les données expérimentales (t_i, T_i) s'impose naturellement.
Comment approcher le plus simplement possible une fonction par un polynôme ?

2.2 Approximation polynomiale

L'approximation d'une fonction f (connue ou non) par un polynôme est une démarche naturelle que l'on rencontre dans divers contexte en analyse : lorsque f est assez régulière,

elle permet d'analyser le comportement local (développements de Taylor) mais aussi dans certain cas de décrire globalement la fonction comme somme infinie de monones (fonctions analytiques). Dans ces deux situations la précision avec laquelle on peut approcher f par un polynôme dépend de la régularité de la fonction. A l'inverse, avec une hypothèse de régularité relativement faible, le théorème de Stone-Weierstrass nous assure que l'on peut approcher uniformément toute fonction continue sur un intervalle compact (fermé borné), d'aussi près que l'on veut, par un polynôme.

Théorème 2.2.1. *Soit f une fonction de $C([a,b])$. Alors pour tout $\epsilon > 0$, il existe un polynôme P tel que*

$$\max_{x\in[a,b]} |f(x) - P(x)| < \epsilon$$

2.3 Interpolation polynomiale

2.3.1 Représenter-evaluer un polynôme en un point

La façon la plus simple de représenter un polynôme p de degré inférieur ou égal à n est de l'exprimer dans la base canonique $\{1, x, x^2, ..., x^n\}$ comme

$$p(x) = a_0 + a_1 x + ... + a_{n-1}x^{n-1} + a_n x^n$$

Pour autant cette écriture ne suggère pas la méthode la plus efficace pour calculer $p(x)$ en un point x donné. Faisant le décompte des opérations.

Pour calculer $a_j x^j$, on doit effectuer j multiplications. Le nombre totale de multiplications est donc

$$N_{mult} = \sum_{k=1}^{n} k = \frac{n(n+1)}{2}$$

il faut ajouter à cela n additions. Si une addition a le même cout qu'une multiplications, alors le nombre total d'opérations est $N_T = n + \frac{n(n+1)}{2} \simeq \frac{n^2}{2} = O(n^2)$.

Horner propose de factoriser $p(x)$ sous la forme :

$$p(x) = a_0 + x(a_1 + x(a_2 + ... + x(a_{n-1} + xa_n)...)) \tag{2.1}$$

et d'évaluer successivement les quantités

$$\begin{aligned} b_n &= a_n \\ b_{n-1} &= a_{n-1} + xb_n \\ &... \\ b_1 &= a_1 + xb_2 \\ b_0 &= a_0 + xb_1 \end{aligned}$$

Faisons maintenant le décompte des opérations. A chaque ligne on effectue exactement une addition et une multiplication. Avec n lignes, nous obtenons

$$N_T^{Horner} = 2n = O(n)$$

ce qui est asymptotiquement une réduction trés importante du nombre d'opérations. En effet si $n = 1000$, avec la méthode classique nous devrons effectuer environ 500000 opérations tandis que celle d'Horner n'en recquiert que 2000. Les opérations sont réduites, le calcul est plus rapide mais aussi plus sûr car sur ordinateur, en précision finie, chaque opération génère des erreurs, d'arrondis notamment.

2.3.2 Position du problème

Problème *Étant donnés un nuage de point $\mathcal{X} = (x_i, f_i)_{0\leq i\leq n}$ ou les x_i sont distincts, déterminer un polynôme p qui vérifi*

$$(\mathcal{P}) \begin{cases} 1.\ deg(p) \leq n, \\ 2.\ p(x_i) = f_i,\ i = 0, 1, ..., n. \end{cases}$$

Il est naturel de considérer les questions suivants :

- *Se donner des critères d'existence et d'unicité pour p.*
- *Disposer d'un procédé systématique (un algorithme) permettant de construire p dans la pratique.*
- *Étudier l'erreur comise en remplçant f par p.*

Dans la suite de ce cours, nous répondrons totalement ou partiellement à ces questions.

Théorème 2.3.1. *Le problème $(\mathcal{P})$ admet une solution unique si et seulemnt si les réels $(x_i)_{0\leq i\leq n}$ sont distincts deux à deux.*

Preuve 2.3.1. *Si on exprime p dans la base canonique, on obtient*

$$p(t) = a_0 + a_1 t + a_2 t^2 + ... + a_n t^n,$$

le résultat est le système linéaire

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f_0 \\ f_1 \\ \vdots \\ f_n \end{pmatrix}. \tag{2.2}$$

*La matrice de ce système qu'on la note V est appelée matrice de **Vandermonde** associée aux points $x_0, x_1, x_2, \ldots, x_n$.*

On montre par récurrence que

$$\det(V) = \prod_{0\leq i\prec j\leq n} (x_j - x_i)$$

Du fait que les x_i sont distincts deux a deux, le déterminant de V est non nul, alors V est inversible et par suite il existe un unique vecteur $\vec{a}$ vérifiant (2.2).

La résolution de ce système linéaire, par inversion de la matrice V ou par substitution dans les équations, lorsque n est grand, est une tâche très pénible (matrice pleine), d'une part. D'autre part cette approche peut ramener à un système mal conditionner (amplification des erreurs d'arrondi).
Dans la suite nous allons présenter une méthode plus astucieuse pour construire l'unique polynôme p.

2.3.3 Polynômes de Lagrange

Existence

L'existence du polynômes d'interpolation p peut être établi de la façon suivante. Étant donnée n polynômes $l_j(t)$ qui vérifient les conditions suivantes :

$$l_j(x_i) = \begin{cases} 0 \ si & i \neq j, \\ 1 \ si & i = j, \end{cases} \tag{2.3}$$

Donc le polynôme d'interpolation p s'écrit sous la forme

$$p(t) = f_0 l_0(t) + f_1 l_1(t) + ... + f_n l_n(t). \tag{2.4}$$

Notons que lorsque le seconde membre de (2.4) est évalué en x_i. tous les termes sauf le i ème s'annulent, puisque $l_j(x_i)$ s'annule pour $j \neq i$. Nous obtenons

$$p(x_i) = \sum_{j=0}^{n} f_j l_j(x_i) = f_i l_i(x_i) = f_i.$$

Nous devons donc montrer que les polynômes l_j qui ont la propriété (2.3) existent. Pour $n = 2$, nous avons

$$l_0(t) = \frac{(t - x_1)(t - x_2)}{(x_0 - x_1)(x_0 - x_2)}, \ l_1(t) = \frac{(t - x_0)(t - x_2)}{(x_1 - x_0)(x_1 - x_2)},$$

$$l_2(t) = \frac{(t - x_0)(t - x_1)}{(x_2 - x_0)(x_2 - x_1)}.$$

Il est facile de vérifier que ces polynômes satisfont la propriété (2.3).
Généralisons ce cas quadratique, nous remarquons que les polynômes suivants satisfont (2.3)

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

Ces polynômes sont appelés ***polynômes de Lagrange****.*
Une conséquence de ce résultat d'existence est que le système linéaire (2.2) admet une solution, autrement dit, la matrice de Vandermonde associées aux points distincts $x_0, x_1, ..., x_n$ est inversible.

Unicité

Pour établir l'unicité du polynôme d'interpolation, nous utilisons le resultat suivant

Proposition 2.3.1. *Si un polynôme de degré n s'annule en $n + 1$ points distincts, alors le polynôme est identiquement null.*

Maintenant supposons que le problème $(\mathcal{P})$ admet deux solutions p et q. Alors $r(t) = p(t) - q(t)$ est de degré au plus égale à n. Mais puisque

$$r(x_i) = p(x_i) - q(x_i) = f_i - f_i = 0, \tag{2.5}$$

le polynôme r s'annule en $n + 1$ points, donc $r = 0$ ou équivalent à $p = q$. D'ou l'unicité.

***Remarque* 2.1.** *La conditions $deg(p) \leq n$ dans la position du problème d'interpolation apparait pour certains gens non naturel. Pour quoi pas le degré du polynôme est exactement n ? L'unicité du polynôme d'interpolation répond à cette question.*
Supposons, par exemple qu'on cherche à interpoler trois points appartennant au même droite. Le polynôme d'interpolation est de degré au plus égale à deux. Puisque une droite est un polynôme linéaire interpolant ces points, par l'unicité, le polynôme d'interpolation quadratique coincide avec le polynôme d'interpolation linéaire. C-à-d le coefficient de t^2 doit être égale à zéros.

***Exemple* 9.** *Prenons $n = 2,\ x_0 = -1,\ x_1 = 0,\ x_2 = 1$.*
Les polynômes de Lagrange associée aux points $-1,\ 0$ et 1 sont donnés par

$$l_0(t) = \frac{(t-x_1)(t-x_2)}{(x_0-x_1)(x_0-x_2)} = \frac{1}{2}(t-1)t = \frac{1}{2}t^2 - \frac{1}{2}t \tag{1.5}$$

$$l_1(t) = \frac{(t-x_0)(t-x_2)}{(x_1-x_0)(x_1-x_2)} = -(t+1)(t-1) = 1 - t^2 \tag{1.6}$$

$$l_2(t) = \frac{(t-x_0)(t-x_1)}{(x_2-x_0)(x_2-x_1)} = \frac{1}{2}(t+1)t = \frac{1}{2}t^2 + \frac{1}{2}t \tag{1.7}$$

***Exemple* 10.** *Trouver un polynôme de degré deux qui en $x_0 = -1$ vaut $f_0 = 8$, en $x_1 = 0$ vaut $f_1 = 3$ et en $x_2 = 1$ vaut $f_2 = 6$.*
D'après ce qui précède, nous avons $p(t) = 8l_0(t) + 3l_1(t) + 6l_2(t)$ où $l_0,\ l_1$ et l_2 sont donnés par (1.5)-(1.7). Nous obtenons donc

$$\begin{aligned} p(t) &= 8(\frac{1}{2}t^2 - \frac{1}{2}t) + 3(1-t^2) + 6(\frac{1}{2}t^2 + \frac{1}{2}t) \\ p(t) &= 4t^2 - t + 3 \end{aligned}$$

Exercice 1
1) Déterminer le polynôme d'interpolation de Lagrange associé au tableau suivant

k	1	2	3
x_k	-1	0	1
y_k	1	3	2

où f est une fonction continue sur $\mathbb{R}$ et $y_k = f(x_k)$.
2) Retrouver les coefficient du polynôme d'interpolation cherché en résolavant un système linéaire.

2.4 Le polynôme d'interpolation de Newton

La méthode de Vandermonde pour déterminer le polynôme d'interpolation est une méthode trés couteuse mais une fois les coefficients sont déterminés, l'évaluation de p est facile. Pour la forme de Lagrange, la détermination de p est facile mais l'évaluation est trés couteuse. La question naturel qu'on se pose est la suivante : y a-t-il une solution pour remédier à ce problème ? la réponse est oui et c'est la forme de Newton.

2.4.1 Forme de Newton

La forme de Newton consiste à considérer comme base

$$1,\ (t-x_0),\ (t-x_0)(t-x_1),\ ...,\ (t-x_0)(t-x_1)...(t-x_{n-1}),$$

ou équivalent à écrire p sous la forme

$$p(t) = c_0 + c_1(t-x_0) + c_2(t-x_0)(t-x_1) + ... + c_n(t-x_0)(t-x_1)...(t-x_{n-1}). \tag{2.6}$$

Evaluation *Pour dériver l'algorithme permettant d'evaluer p en un point, on écrit (2.6) sous la forme*

$$p(t) = c_0 + (t-x_0)(c_1 + (t-x_1)(c_2 + (t-x_2)(c_3 + ... + (t-x_{n-2})(c_{n-1} + c_n(t-x_{n-1})))). \tag{2.7}$$

On obtient l'algorithme suivant

$p = c_n$
pour $i = n-1 : (-1) : 0$ *faire*
  $p = p * (t - x_i) + c_i$
fin pour i

Cet algorithme nécessite $2n$ additions et n multiplications.

2.4.2 Existence et unicité

Dans cette section, nous établissons l'existence de la forme de Newton avant de procéder à la façons de la calculer. Commençons par évaluer $p(x_i)$.

$$p(x_0) = c_0$$

$$p(x_1) = c_0 + c_1(x_1 - x_0)$$

plus généralement, $p(x_i)$ contient seulement $i+1$ termes non nulls. La condition d'interpolation $f_i = p(x_i)$ donne

$$\begin{aligned} f_0 &= c_0 \\ f_1 &= c_0 + c_1(x_1 - x_0), \\ &\cdots \\ f_n &= c_0 + c_1(x_n - x_0) + ... + c_n(x_n - x_0)(x_n - x_1)...(x_n - x_{n-1}) \end{aligned}$$

La matrice de ce système linéaire est une matrice triangulaire inférieur de la forme

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & x_1 - x_0 & 0 & \cdots & 0 \\ 1 & (x_2 - x_0) & (x_2 - x_0)(x_2 - x_1) & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & (x_n - x_0) & (x_n - x_0)(x_n - x_1) & \cdot & (x_n - x_0)(x_n - x_1)...(x_n - x_{n-1}) \end{pmatrix} \tag{2.8}$$

Les éléments diagonaux sont non nulls, la matrice est donc inversible, d'ou l'existence et l'unicité des c_i.

Une conséquence importante de la triangularitée de la matrice est que l'addition d'un nouveau point au problème d'interpolation n'affect pas les coefficients qui sont déja calculé. Autrement dit

$$\begin{aligned} &c_0 \text{ interpole } (x_0, f_0) \\ &c_0 + c_1(t - x_1) \text{ interpole } (x_0, f_0), (x_1, f_1) \\ &c_0 + c_1(t - x_1) + c_2(t - x_0)(t - x_1) \text{ interpole } (x_0, f_0), (x_1, f_1), (x_2, f_2) \end{aligned} \tag{2.9}$$

ainsi de suite.

2.4.3 Différences divisées

En principe, le système triangulaire peut être résolu en $O(n^2)$ opérations pour déterminer les coefficients c_i. Cependant les coefficients de ce système peuvent facilement poser un problème d'éxcés de mémoire. Pour remédier à ce problème nous allons construire un algorithme de même complexité permetant de calculer les c_i en $O(n^2)$ opérationss. Nous commencons par définir les différences divisées.

Soit $f(x)$ une fonction numérique définie sur une partie D de $\mathbb{R}$. Soit x_0 et x_1 deux éléments de D tel que :

$$x_0 < x_1 \quad et \quad f(x_0) = f_0 \quad , \; f(x_1) = f_1$$

$f[x_0] = f(x_0)$
La différence divisée première entre x_0 et x_1 est égale à :

$$f[x_0, x_1] = \frac{f[x_1] - f[x_0]}{x_1 - x_0}$$

Plus généralement si $x_0 < x_1 < x_2 < \ldots < x_n$ sont des points de D alors la différence divisée première entre x_i et $x_j = x_{i+1}$ est :

$$f[x_i, x_j] = \frac{f[x_j] - f[x_i]}{x_j - x_i} \qquad (j = i + 1)$$

Une différence seconde est définie par :

$$\begin{aligned} f[x_0, x_1, x_2] &= \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} \\ f[x_2, x_3, x_4] &= \frac{f[x_3, x_4] - f[x_2, x_3]}{x_4 - x_2} \end{aligned}$$

$$j = i + 1 \; , \; k = i + 2 \; , \; f[x_i, x_j, x_k] = \frac{f[x_j, x_k] - f[x_i, x_j]}{x_k - x_i}$$

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}$$

est une différence d'ordre n.

Une table de différences est une représentation trés commode de celle-ci. Par exemple pour une suite $(x_0, x_1, x_2, x_3, x_4)$ on a la table diagonale suivante :

$$\begin{array}{llllll}
x_k & f_k & & & & \\
x_0 & f_0 & & & & \\
 & & f[x_0,x_1] & & & \\
x_1 & f_1 & & f[x_0,x_1,x_2] & & \\
 & & f[x_1,x_2] & & f[x_0,x_1,x_2,x_3] & \\
x_2 & f_2 & & f[x_1,x_2,x_3] & & f[x_0,x_1,x_2,x_3,x_4] \\
 & & f[x_2,x_3] & & f[x_1,x_2,x_3,x_4] & \\
x_3 & f_3 & & f[x_2,x_3,x_4] & & \\
 & & f[x_3,x_4] & & & \\
x_4 & f_4 & & & &
\end{array}$$

Exemple : *Calculons les différences divisées relatives aux valeurs des f_k données par les tableaux :*

x_k	f_k
0	1
1	1
2	2
4	5

La solution est :

$$\begin{array}{llllll}
x_k & f_k & & & & \\
0 & 1 & & & & \\
 & & 0 & & & \\
1 & 1 & & \frac{1}{2} & & \\
 & & 1 & & -\frac{1}{12} & \\
2 & 2 & & \frac{1}{6} & & \\
 & & \frac{3}{2} & & & \\
4 & 5 & & & &
\end{array}$$

2.4.4 Différence divisées et polynôme d'interpolation de Newton

Théorème 2.4.1. *Soit f une fonction numérique définie sur une partie D de $\mathbb{R}$. Soient $x_0, x_1, \ldots, x_n$ des points de D tels que : $f(x_k) = f_{k.}$. Alors le polynôme d'interpolation de Newton est de la forme :*

$$\begin{aligned} p(x) &= f_0 + c_1(x-x_0) + c_2(x-x_0)(x-x_1) \\ &\quad + c_3(x-x_0)(x-x_1)(x-x_2) + \cdots \\ &\quad + c_n(x-x_0)(x-x_1)\ldots(x-x_{n-1}), \end{aligned}$$

avec

$$c_k = f[x_0, x_1, ..., x_k].$$

Preuve 2.4.1.

• *Pour $k = 0$ il est évident que $p(x_0) = f_0$*
Par définition des différences divisées et la propriété de symétrie, on a

$$\begin{aligned} f_k &= f_0 + (x_k - x_0) f[x_0, x_k], \\ f[x_0, x_k] &= f[x_0, x_1] + (x_k - x_1) f[x_0, x_1, x_k], \\ f[x_0, x_1, x_k] &= f[x_0, x_1, x_2] + (x_k - x_2) f[x_0, x_1, x_2, x_k], \\ \vdots &= \quad \vdots \qquad\qquad \vdots \\ f[x_0, x_1, \cdots, x_{n-2}, x_k] &= f[x_0, x_1, \cdots, x_{n-1}] + (x_k - x_{n-1}) f[x_0, x_1, \cdots, x_{n-1}, x_k]. \end{aligned}$$

Par exemple, la première ligne découle du fait que

$$f[x_0, x_k] = \frac{f_k - f_0}{x_k - x_0}$$

La seconde ligne se déduit de

$$f[x_0, x_1, x_k] = f[x_1, x_0, x_k] = \frac{f[x_0, x_k] - f[x_1, x_0]}{x_k - x_1}$$

• *Pour $k = 1$, d'après la première ligne, on a : $f_1 = f_0 + (x_1 - x_0)f[x_0, x_1]$. D'autre part*

$$p(x_1) = f_0 + (x_1 - x_0)f[x_0, x_1] = f_1.$$

En substituant la seconde ligne dans la première, on obtient

$$f_k = f_0 + (x_k - x_0)f[x_0, x_1] + (x_k - x_0)(x_k - x_1)f[x_0, x_1, x_k]$$

• *Pour $k = 2$, on obtient alors*

$$f_2 = f_0 + (x_2 - x_0)f[x_0, x_1] + (x_2 - x_0)(x_2 - x_1)f[x_0, x_1, x_2]$$

et par définition du polynôme $p(x)$, on a

$$p(x_2) = f_0 + (x_2 - x_0)f[x_0, x_1] + (x_2 - x_0)(x_2 - x_1)f[x_0, x_1, x_2],$$

on en déduit que $p(x_2) = f_2$.
Par substitution successives, nous vérifions que $p(x_k) = f_k$ pour chaque x_k jusqu'à ce que nous arrivons à :

$$\begin{aligned} f_n &= f_0 + (x_n - x_0)f[x_0, x_1] + (x_n - x_0)(x_n - x_1)f[x_0, x_1, x_2] \\ &\quad + \cdots + (x_n - x_0)(x_n - x_1)\cdots(x_n - x_{n-1})f[x_0, x_1, \cdots, x_{n-1}, x_n] \end{aligned}$$

qui établit que $f_n = p(x_n)$.

2.4.5 Algorithme de calcul des c_i

Par définition de différences divisées, on a

$$c_k = f[x_0, x_1, ..., x_k] = \frac{f[x_1, x_2, ..., x_k] - f[x_0, x_1, ..., x_{k-1}]}{x_k - x_0}$$

Une conséquence de cela est qu'on peut calculer les différents quotient d'une manière récursif

$$\begin{array}{lllll} f_0 = f[x_0] & & & & \\ f_1 = f[x_1] & f[x_0, x_1] & & & \\ f_2 = f[x_2] & f[x_1, x_2] & f[x_0, x_1, x_2] & & \\ f_3 = f[x_3] & f[x_2, x_3] & f[x_1, x_2, x_3] & f[x_0, x_1, x_2, x_3] \end{array}$$

On calcul $d_{i0} = f[x_i] = f_i$ pour $i = 0, ..., n$, puis

$$d_{ij} = \frac{d_{i,j-1} - d_{i-1,j-1}}{x_i - x_{i-j}},\ j = 1, 2, .., n,\ \text{et } i = j, j+1, ..., n$$

***Exemple* 11.** *Construire le polynôme d'interpolation de Newton pour les données $(0, 1)$, $(1, 2)$, $(2, 9)$, $(3, 28)$.*
Le polynôme d'interpolation de Newton p_3 est de degré ≤ 3.

Table des DD :

x_i	Ordre 0	Ordre 1	Ordre 2	Ordre 3
0	[1]			
1	2	[1]		
2	9	7	[3]	
3	28	19	6	[1]

$$p_3(x) = 1 + 1 \times (x-0) + 3 \times (x-0)(x-1) + 1 \times (x-0)(x-1)(x-2)$$

2.5 Erreur d'interpolation

Soit p_n le polynôme d'interpolation de f aux point $x_0, x_1, ..., x_n$. On pose $e_n(f)(t) = f(t) - p_n(t)$

Théorème 2.5.1. *Si la fonction f est $(n+1)$ fois continûment dérivable sur $[a,b]$, $\{x_i / i = 0, 1, \cdots, n\} \subset [a,b]$, pour tout $t \in [a,b]$, il existe ζ_t contenu dans le plus petit fermé contenant $x_0, x_1, \cdots, x_{n-1}, x_n$ tel que*

$$e_n(f)(t) = \frac{f^{(n+1)}(\zeta_t)}{(n+1)!}\omega(t) \text{ où } \omega(t) = \prod_{i=0}^{n}(t - x_i).$$

Preuve 2.5.1.
On a l'égalité i.e $f(x) = p_n(x)\ \forall x \in [a,b] \,/\, x = x_i,\ i = 0, 1, \cdots, n$.
Soit $t \in [a,b]$, t fixé / $t \neq x_i\ \forall\ i = 0, 1, \cdots, n$. On définit la fonction

$$F(u) = f(u) - p_n(u) - \frac{f(t) - p_n(t)}{\omega(t)}\omega(u).$$

La fonction F est $(n+1)$ fois continûment dérivable (même régularité que f)

$$\text{si } u = x_i \quad F(x_i) = \underbrace{f(x_i) - p_n(x_i)}_{=0} - \frac{f(t) - p_n(t)}{\omega(t)}\underbrace{\omega(x_i)}_{=0} = 0,\ \forall\ i = 0, 1, \cdots, n.$$

Si $u = t$ alors $F(t) = f(t) - p_n(t) - \dfrac{f(t) - p_n(t)}{\omega(t)}\omega(t) = 0$. La fonction F admet au moins $(n+2)$ racines réelles distinctes $x_0, x_1, \cdots, x_{n-1}, x_n$ et t. Puisque la fonction F est continue sur $[x_0, x_1]$, dérivable sur $]x_0, x_1[$ et $F(x_0) = F(x_1) = 0$ alors le théorème de Rôlle nous fournit l'existence d'un point $\lambda_0 \in]x_0, x_1[$ tel que $F'(\lambda_0) = 0$. Par suite, F' admet au moins $(n+1)$ racines réelles distinctes ainsi, on montre que $F^{(n+1)}$ admet au moins une racine notée ζ_t i.e $F^{(n+1)}(\zeta_t) = 0$. Nous avons

$$F^{(n+1)}(u) = f^{(n+1)}(u) - 0 - \frac{f(t) - p_n(t)}{\omega(t)}(n+1)!$$

En remplaçant u par ζ_t dans la relation ci dessus nous concluons que

$$e_n(f)(t) = f(t) - p_n(t) = \frac{f^{(n+1)}(\zeta_t)}{(n+1)!}\omega(t)$$

Corollaire 2.5.1. *Si f est $(n+1)$ fois continûment dérivable alors*

$$|e_n(f)(t)| \leq \frac{|\omega(t)|}{(n+1)!}M_{n+1} \leq \frac{(b-a)^{n+1}}{(n+1)!}M_{n+1} \text{ avec } M_{n+1} = \sup_{[a,b]}\left|f^{(n+1)}(\zeta_t)\right| \qquad (2.10)$$

2.5.1 Analyse de l'erreur d'interpolation linéaire

Nous montrons maintenant comment utiliser le théorème précedent pour obtenir une borne sup de l'erreur d'interpolation dans un cas simple. Soit $p(t)$ le polynôme d'interpolation linéaire de f aux points x_0 et x_1 et on suppose que

$$|f''(t)| \leq M,$$

dans un certain interval I. Alors

$$|f(t) - p(t)| = \frac{|f''(\xi)|}{2}|(t-x_0)(t-x_1)| \leq \frac{M}{2}|(t-x_0)(t-x_1)|$$

Le traitement de cette borne dépend de choix de t si à l'intérieur ou à l'extérieur de l'intervale $[x_0, x_1]$.

Si t est à l'extérieur de $[x_0, x_1]$. On dit qu'on ***extrapole*** *le polynôme d'approximation de f. Puisque $|(t-x_0)(t-x_1)|$ devient large quand t s'éloigne de l'intervale $[x_0, x_1]$, donc le problème devient difficille.*

Si t est à l'intérieur de $[x_0, x_1]$, on parle de l'interpolation de f. Dans ce cas, on peut obtenir une borne sup de l'erreur. La fonction $|(t-x_0)(t-x_1)|$ atteint son maximum dans $[x_0, x_1]$ au point $t = \frac{x_0+x_1}{2}$, et le maximum est $(x_1-x_0)^2/4$. Donc

$$t \in [x_0, x_1] \;\Rightarrow\; |f(t) - p(t)| \leq \frac{M}{8}(x_1 - x_0)^2.$$

Application *$f(t) = \sin t$, la question est comment choisir la distance h entre les deux points x_0 et x_1 pour prescrire une précision donnée. supposons qu'on veut approcher f avec une précision de 10^{-4}. Comme $|f''(t)| = |sin(t)| \leq 1$, on a*

$$|sin(t) - p(t)| \leq h^2/8.$$

On doit donc choisir h tel que $h^2/8 \leq 10^{-4}$ soit $h \leq 0.02\sqrt{2}$.

***Remarque* 2.2.** *Le comportement de l'erreur d'interpolation dépend de f et du choix des noeuds x_i.*

2.5.2 Meilleurs points d'interpolation - phénomène de Runge

Avant de chercher les meilleurs point d'interpolation, on doit se rappeler de quelques résultats :

- *Soit $f : [a,b] \subset \mathbb{R}$, continue sur $[a,b]$. on rappelle que la norme infinie de f notée $||f||_\infty$ est la quantité définie par $max_{x\in[a,b]}|f(x)|$.*
- *Une suite de fonctions f_n, $n \geq 0$ continues sur $[a,b]$ converge uniformément vers une fonction f ssi $||f_n - f||_\infty \to 0$ lorsque $n \to +\infty$.*
- *D'après la relation (2.10), on a*

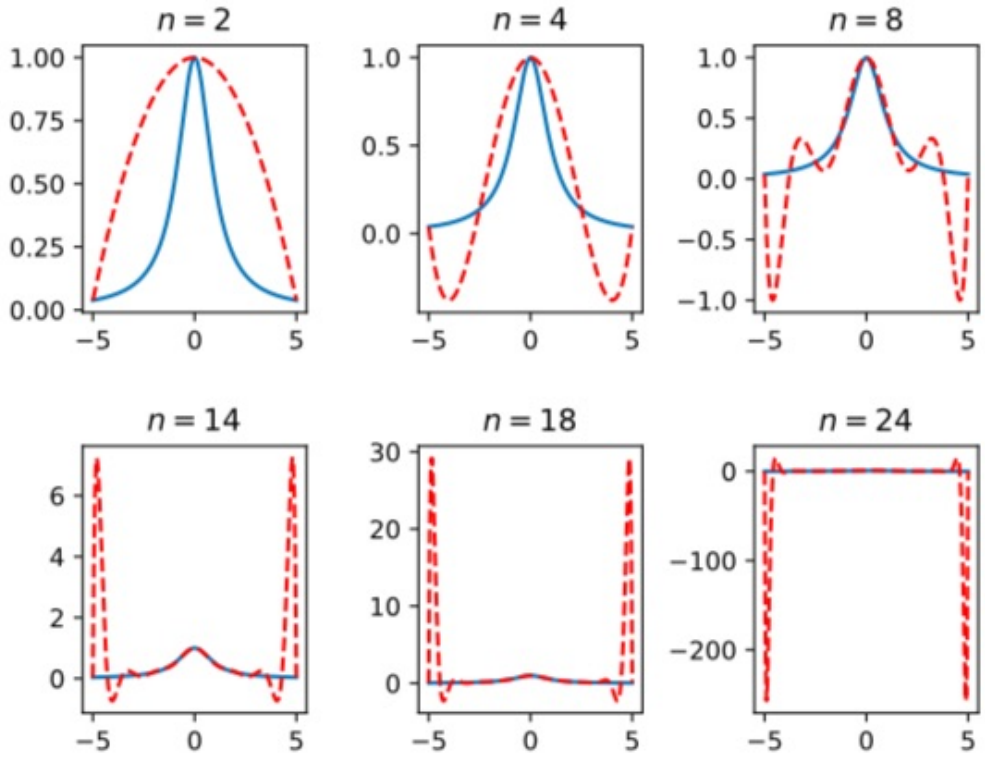
$$||e_n(f)||_\infty \leq \frac{||\omega||_\infty}{(n+1)!}M_{n+1}. \tag{2.11}$$

D'après ces résultats on remarque que la convergence uniforme de p_n vers f quand n croit, c.à.d. $||f - p_n||_\infty = ||e_n(f)||_\infty \to 0$ est réalisée si $\lim_{n\to+\infty} \frac{||\omega||_\infty}{(n+1)!} M_{n+1} = 0$. ***Malheureusement ceci n'est pas toujours vraie en général****, car le produit $||\omega||_\infty M_{n+1}$ risque de grandir plus rapidement que $(n+1)!$ pour certaines fonctions f, ce qui confirme la remarque 2.2.*

Exemple* 12.** *Un contre exemple célèbre a été proposé par Runge. On cherche à interpoler la fonction $f(x) = \frac{1}{1+x^2}$ en subdivisant $[-5, 5]$ en n* ***parties égales *: $x_{i+1} - x_i = h$ où $h = \frac{10}{n}$ est le pas de la subdivision et $x_i = -5 + i \times h$, $0 \leq i \leq n$ sont les $n+1$ noeuds de l'interpolation.*
Le tableau suivant montre l'évolution de l'erreur d'interpolation en fonction du degré n du polynôme d'interpolation pour les noeuds equidistants :

n	2	4	8	14	18	24
$\|\|e_n(f)\|\|_\infty$	0.64623	0.43836	1.04518	7.19488	29.19058	257.21305

Les figuures suivantes montrent graphiquement le phénomène de Runge (la courbe du polynôme est en pointilles) :



***Remarque* 2.3.** *Losqu'on choisit les x_i équidistants, on observe que :*

- *Plus n est grand plus les effets de bords de l'intervalle s'aggravent.*
- *Par conséquent, la norme de l'erreur croit de manière exponentielle lorsque n devient grand.*

***Question* 1.** *Quel est le meilleur choix de la distribution des x_i pour minimiser l'erreur ? Pour répondre à cette question, on doit donner quelques définitions et résultats*

Définition 2.5.1. *Soit $T_n(x) = \cos(n \arccos(x)), x \in [-1, 1], n \geq 0$, on montre (en exercice) que $T_n(x)$ est un polynôme de degré n appelé polynôme de Tchebychev et qui vérifie la relation de recurrence :*

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

Déterminons les racines de T_n.
Posons $\theta = \arccos(x)$ donc $T_n(x) = \cos(n\theta) = 0$ ssi $n\theta = \frac{\pi}{2} + k\pi, \quad 0 \leq k \leq n-1 \Longrightarrow \theta_k = \frac{2k+1}{n}\pi$ donc les racines de $T_n(x)$ sont $x_k = \cos(\theta_k), \quad k = 0, \ldots, n-1$.

Définition 2.5.2. *On appelle points d'interpolation de Tchebychev d'ordre n les racines de T_{n+1} donnés par :*

$$x_i = \cos(\frac{2i+1}{2n+2}\pi), \quad i = 0, \ldots, n$$

On se ramène à l'intervalle $[a, b]$ par le changement de variable $x = \frac{a+b}{2} + \frac{b-a}{2}t, t \in [-1, 1]$. Donc les points de Tchebychev d'ordre n sur $[a, b]$ sont :

$$x_i = \frac{a+b}{2} + \frac{b-a}{2}\cos(\frac{2i+1}{2n+2}\pi), \quad i = 0, \ldots, n.$$

On peut démontrer le résultat suivant :

Théorème 2.5.2. *La quantité $max_{x\in[a,b]}|\omega_{n+1}(x)|$ est minimale lorsqu'on prend pour noeuds d'interpolation $x_i, 0 \leq i \leq n$ les points de Tchebychev d'ordre n définis sur $[a, b]$.*

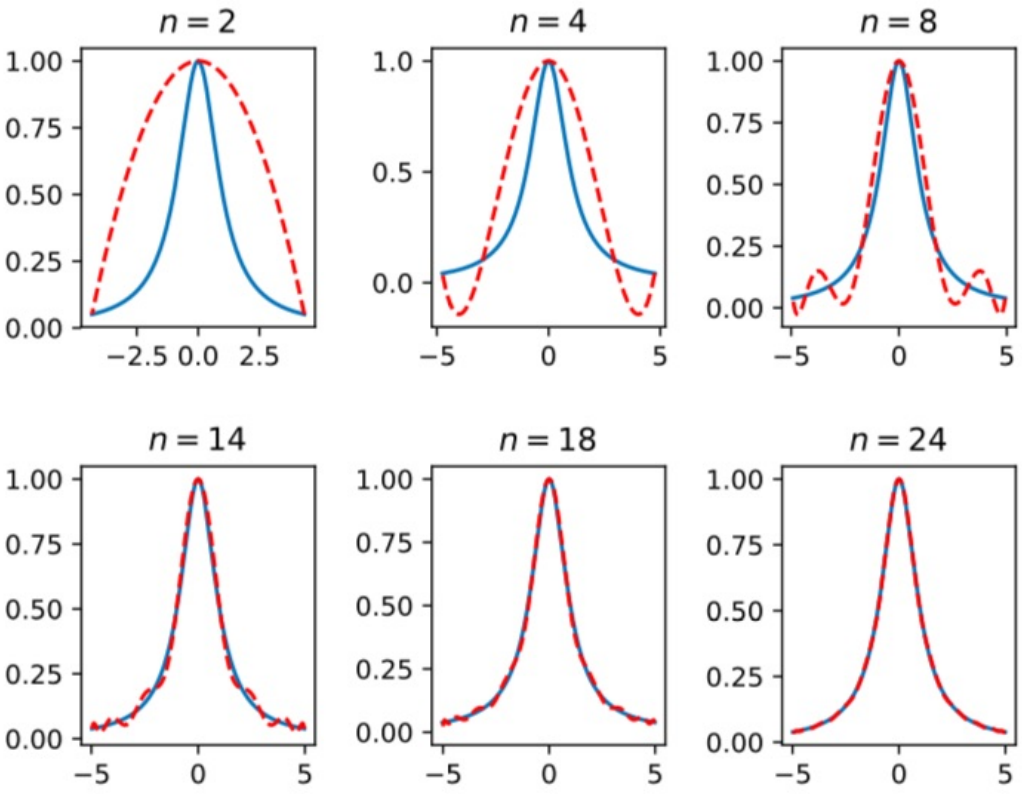
Pour la même fonction $f(x) = \frac{1}{1+x^2}$ en utilisant les noeuds de Tchebytchev sur $[-5, 5]$

$$x_i = 5\cos(\frac{2i+1}{2n+2}\pi), \quad i = 0, \ldots, n.$$

Les résultats sont nettement améliorés aux bords de $[-5, 5]$ (voir le tableau suivant)

n	2	4	8	14	18	24
$\|\|e_n(f)\|\|_\infty$	0.75021	0.55582	0.26909	0.08305	0.03755	0.01139

Les figuures suivantes montrent graphiquement la disparition du pénomène de Runge (la courbe du polynôme est en pointilles) :


$n = 2$
$n = 4$
$n = 8$
$n = 14$
$n = 18$
$n = 24$

Chapitre 3

Dérivation numérique

3.1 Principe

La dérivation numérique consiste à dériver de façon approchée une fonction sur un intervalle borné $[a, b]$, c'est-à-dire calculer la pente de la courbe représentant la fonction, à partir d'un calcul ou d'une mesure en un nombre fini de points.

- *La répartition des points en abscisse est généralement uniforme (pas d'échantillonnage constant h) mais il existe des méthodes à pas variable, ou encore à pas adaptatif.*
- *La dérivation des polynômes étant très simple, l'opération consiste généralement à construire une interpolation polynomiale par morceaux (de degré plus ou moins élevé) puis de dériver le polynôme sur chaque morceau. On obtient alors des formules dites de différences finies.*
- *La précision de la dérivation numérique peut alors s'améliorer en augmentant le nombre de points n (en diminuant le pas d'échantillonnage h) ou en augmentant le degré de l'interpolation polynomiale (sous réserve de bonnes propriétés de continuité de la courbe).*

3.2 Dérivée première

Soit f une fonction connue seulement par sa valeur en $(n+1)$ points donnés $x_i \quad i = 0; 1; \cdots; n$ distincts. Les formules de différence les plus simples basées sur l'utilisation de la ligne droite pour interpoler les données utilisent deux points pour estimer la dérivée. On suppose connue la valeur de la fonction en x_{i-1}, x_i et x_{i+1}; on pose $f(x_{i-1}) = y_{i-1}, f(x_i) = y_i$, et $f(x_{i+1}) = y_{i+1}$ Si on suppose que l'espace entre deux points successifs est constant, donc on pose $h = x_i - x_{i-1} = x_{i+1} - x_i$ Alors les formules standards en deux points sont :

- *Formule de différence progressive :*

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = \frac{y_{i+1} - y_i}{x_{i+1} - x_i}.$$

- *Formule de différence régressive :*

$$f'(x_i) \simeq \frac{f(x_i) - f(x_{i-1})}{x_i - x_{i-1}} = \frac{y_i - y_{i-1}}{x_i - x_{i-1}}.$$

- *Formule de différence centrale :*

$$f'(x_i) \simeq \frac{f(x_{i+1}) - f(x_{i-1})}{x_{i+1} - x_{i-1}} = \frac{y_{i+1} - y_{i-1}}{x_{i+1} - x_{i-1}}.$$

***Remarque* 3.1.**

Les formules de différences classiques peuvent être trouvées en utilisant la formule de Taylor.

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\eta)$$
$$x \leq \eta \leq x+h$$

-Formule progressive :

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_i)}{h} - \frac{h}{2}f''(\eta)$$
$$x_i \leq \eta \leq x_i + h$$

l'erreur est $\frac{h}{2}f''(\eta)$ donc en $O(h)$: Cette formule peut être trouvée aussi en utilisant le polynôme d'interpolation de Lagrange pour les points $(x_i, f(x_i))$ et $(x_{i+1}, f(x_{i+1}))$.
- Formule régressive

$$h = x_i - x_{i-1}$$
$$f'(x_i) = \frac{f(x_i) - f(x_{i-1})}{h} + \frac{h}{2}f''(\eta)$$
$$x_{i-1} \leq \eta \leq x_i$$

- La formule de différence centrale de la dérivée en x_i peut être trouvée en utilisant la formule de Taylor d'ordre 3 avec $h = x_{i+1} - x_i = x_i - x_{i-1}$

$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2}f''(x_i) + \frac{h^3}{3!}f^{(3)}(\eta_1)$$
$$f(x_{i-1}) = f(x_i) - hf'(x_i) + \frac{h^2}{2}f''(x_i) - \frac{h^3}{3!}f^{(3)}(\eta_2)$$
$$x_i \leq \eta_1 \leq x_{i+1}, x_{i-1} \leq \eta_2 \leq x_i$$

si on suppose que $f^{(3)}$ est continue sur $[x_{i-1}, x_{i+1}]$ on peut écrire la formule suivante :

$$f'(x_i) = \frac{f(x_{i+1}) - f(x_{i-1})}{2h} + \frac{h^2}{6}f^{(3)}(\eta)$$

l'erreur est $\frac{h^2}{6}f^{(3)}(\eta)$ donc en $O(h^2)$.
-La formule de différence centrale peut aussi être trouvée à partir du polynôme d'interpolation de Lagrange en 3 points. On peut interpoler les données par un polynôme au lieu d'utiliser la droite, nous obtenons alors les formules de différence qui utilisent plus de deux points. On suppose que le pas h est constant.
- Formule de différence progressive utilisant trois points :

$$f'(x_i) \simeq \frac{-f(x_{i+2}) + 4f(x_{i+1}) - 3f(x_i)}{x_{i+2} - x_i}$$

Formule de différence régressive utilisant trois points :

$$f'(x_i) \simeq \frac{3f(x_i) - 4f(x_{i-1}) + f(x_{i-2})}{x_i - x_{i-2}}$$

3.3 Formule générale en trois points

La formule d'approximation en 3 points de la dérivée première, basée sur le polynôme d'interpolation de Lagrange, n'utilise pas des points équidistants. Étant donné trois points $(x_1, y_1), (x_2, y_2)$ *et* (x_3, y_3) *avec* $x_1 < x_2 < x_3$, *la formule suivante permet d'approcher la dérivée en un point* $x \in [x_1, x_3]$. *Les dérivées aux points* x_i *sont les suivantes :*

$$\begin{aligned}
f'(x_1) &= \frac{2x_1 - x_2 - x_3}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{x_1 - x_3}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{x_1 - x_2}{(x_3 - x_1)(x_3 - x_2)} y_3, \\
f'(x_2) &= \frac{x_2 - x_3}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{2x_2 - x_1 - x_3}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{x_2 - x_1}{(x_3 - x_1)(x_3 - x_2)} y_3, \\
f'(x_3) &= \frac{x_3 - x_2}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{x_3 - x_1}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{2x_3 - x_2 - x_1}{(x_3 - x_2)(x_3 - x_1)} y_3.
\end{aligned}$$

Le polynôme de Lagrange est donnée par

$$P(x) = L_1(x)y_1 + L_2(x)y_2 + L_3(x)y_3,$$

où

$$\begin{aligned}
L_1(x) &= \frac{(x - x_2)(x - x_3)}{(x_1 - x_2)(x_1 - x_3)}, \\
L_2(x) &= \frac{(x - x_1)(x - x_3)}{(x_2 - x_1)(x_2 - x_3)}, \\
L_3(x) &= \frac{(x - x_1)(x - x_2)}{(x_3 - x_1)(x_3 - x_2)}.
\end{aligned}$$

L'approximation de la dérivée première est donnée par $f'(x) \simeq P'(x)$, *qui peut s'écrire*

$$P'(x) = L_1'(x)y_1 + L_2'(x)y_2 + L_3'(x)y_3,$$

où

$$\begin{aligned}
L_1'(x) &= \frac{2x - x_2 - x_3}{(x_1 - x_2)(x_1 - x_3)}, \\
L_2'(x) &= \frac{2x - x_1 - x_3}{(x_2 - x_1)(x_2 - x_3)}, \\
L_3'(x) &= \frac{2x - x_2 - x_1}{(x_3 - x_2)(x_3 - x_1)}.
\end{aligned}$$

Donc

$$f'(x) \simeq \frac{2x - x_2 - x_3}{(x_1 - x_2)(x_1 - x_3)} y_1 + \frac{2x - x_1 - x_3}{(x_2 - x_1)(x_2 - x_3)} y_2 + \frac{2x - x_2 - x_1}{(x_3 - x_2)(x_3 - x_1)} y_3.$$

3.4 Dérivées d'ordre supérieur

Les formules de dérivées d'ordre supérieur, peuvent être trouvées à partir des dérivées du polynôme de Lagrange ou en utilisant les formules de Taylor. Par exemple, étant donné 3 points $x_{i-1}; x_i; x_{i+1}$ *équidistants, la formule de la dérivée seconde est donnée par :*

$$f''(x_i) \simeq \frac{1}{h^2} \left[f(x_{i+1}) - 2f(x_i) + f(x_{i-1}) \right],$$

l'erreur est en $O(h^2)$:
- Dérivée seconde à partir du polynôme de Taylor

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2!}f''(x) + \frac{h^3}{3!}f^{(3)}(x) + \frac{h^4}{4!}f^{(4)}(\eta_1),$$
$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2!}f''(x) - \frac{h^3}{3!}f^{(3)}(x) + \frac{h^4}{4!}f^{(4)}(\eta_2),$$
$$x \leq \eta_1 \leq x+h \text{ et } x-h \leq \eta_2 \leq x,$$
$$f''(x) \simeq \frac{f(x+h) - 2f(x) + f(x-h)}{h^2},$$

l'erreur est en $O(h^2)$: Pour obtenir les formules de la troisième et la quatrième dérivée, on prend une combinaison linéaire des développement de Taylor, pour $f(x+2h)$, $f(x+h)$, $f(x-h)$ et $f(x-2h)$.
La table suivante donne différentes formules centrales toutes en $O(h^2)$:

$$f'(x_i) \simeq \frac{1}{2h}\left[f(x_{i+1}) - f(x_{i-1})\right].$$

$$f''(x_i) \simeq \frac{1}{h^2}\left[f(x_{i+1}) - 2f(x_i) + f(x_{i-1})\right]$$
$$f^{(3)}(x_i) \simeq \frac{1}{2h^3}\left[f(x_{i+2}) - 2f(x_{i+1}) + 2f(x_{i-1}) - f(x_{i-2})\right]$$
$$f^{(4)}(x_i) \simeq \frac{1}{h^4}\left[f(x_{i+2}) - 4f(x_{i+1}) + 6f(x_i) - 4f(x_{i-1}) + f(x_{i-2})\right]$$

En utilisant les polynômes d'interpolation de Lagrange les dérivées d'ordre p sont calculées par :

$$f^{(p)}(\alpha) \sim \sum_{i=0}^{n} A_i(\alpha) f(x_i)$$

où $A_i(\alpha) = L_i^{(p)}(\alpha) \quad p \leq n$.

***Remarque* 3.2.** *La formule est exacte pour les polynômes de degrés n : Le système linéaire donnant les $A_i(\alpha)$ a un déterminant de type Vandermonde différent de zéro si les x_i sont distincts. Les $A_i(\alpha)$ sont indépendants de f et peuvent être calculés une fois pour toutes.*

3.5 Étude de l'erreur commise

D'après le chapitre interpolation polynômiale, si f est connue en $(n+1)$ points $x_i; i = 0; \ldots; n$ alors $f(x) = P_n(x) + e(x)$; où $e(x)$ est l'erreur d'interpolation. En dérivant on obtient

$$f'(x) = P_n'(x) + e'(x) = \sum_{i=0}^{n} A_i(x) f(x_i) + e'(x).$$

Avec :

$$e'(x) = \frac{d}{dx}\left[\frac{1}{(n+1)!}\omega(x) f^{(n+1)}(\xi_x)\right] = \frac{1}{(n+1)!}\omega'(x) f^{(n+1)}(\xi_x) + \frac{1}{(n+1)!}\omega(x)\frac{d}{dx}\left[f^{(n+1)}(\xi_x)\right].$$

On remarque tout de suite que l'erreur de dérivation est nulle si f est un polynôme de degré inférieur ou égale à n. Si on prend pour x un point x_i; le second terme de la dernière somme

s'annule, sinon il faut connaitre $\frac{d}{dx}\left[f^{(n+1)}(\xi_x)\right]$ *; ce qui est difficile car la fonction* $x \to \xi_x$ *étant inconnue. On constate qu'on devra se contenter d'une estimation :*

$$|e'(x)| \leq \frac{1}{(n+1)!}\omega'(x)M_{n+1} + \frac{1}{(n+1)!}\omega(x)M_{n+2},$$

avec $M_{n+1} = \max_{[a,b]} |f^{(n+1)}(x)|$ *et* $M_{n+2} = \max_{[a,b]} |f^{(n+2)}(x)|$.

Chapitre 4

Intégration numérique

4.1 Principe

Pour les méthodes élémentaires/simples (à opposer aux méthodes composites, voir dernière section du chapitre) de type interpolatrice, on approche $I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x$ *par l'intégrale sur* $[a,b]$ *du polynôme d'interpolation de Lagrange de* f *aux points* $(x_i)_{i=0}^n$. *On note* p *le polynôme d'interpolation, dans la base des polynômes de Lagrange* p *s'écrit*

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

En intégrant cette expression sur l'intervalle $[a,b]$ *on obtient l'expression générale des formules de quadrature élémentaires de type interpolatrice*

$$I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x \simeq \tilde{I}_{[a,b]}(f) = \int_a^b p(x)\mathrm{d}x = \sum_{i=0}^{n} \alpha_i f(x_i) \qquad avec \qquad \alpha_i = \int_a^b l_i(x)\mathrm{d}x$$

4.1.1 Intervalle de référence

Il est souvent plus facile de faire les calculs sur l'intervalle de référence $[-1,1]$. *Une fois ces calculs réalisés, on revient à l'intervalle* $[a,b]$ *grâce au changement de variable* φ.

$$\begin{aligned} \varphi : [-1,1] &\to [a,b] \\ t &\mapsto \frac{a+b}{2} + t\frac{b-a}{2}. \end{aligned}$$

Pour f *continue sur* $[a,b]$, *on pose* $\tilde{f}(t) = f(\varphi(t)) = f(x)$. *Alors* $\tilde{f}$ *est continue sur* $[-1,1]$ *et on obtient*

$$\int_a^b f(x)\mathrm{d}x = \frac{b-a}{2}\int_{-1}^{1} \tilde{f}(t)\mathrm{d}t.$$

***Remarque* 4.1.** Dans le cas où l'intervalle de référence est $[0,1]$, on utilise le changement de variable suivant : $\psi(t) = (b-a)t + a$.

4.2 Exemples de formules de quadrature élémentaires

Une catégorie remarquable de formules de quadrature par interpolation est celle des formules de Newton-Cotes. Ce sont des formules obtenues avec des points d'interpolation équidistants

$$x_i = a + i\frac{b-a}{n}, \quad n \in \mathbb{N}^*.$$

• ***Interpolation à un point : Formule du rectangle à gauche***
Sur l'intervalle de référence $[-1,1]$: $\tilde{I}_{[-1,1]}(\tilde{f}) = 2\tilde{f}(-1)$.
Sur l'intervalle $[a,b]$ *:* $\tilde{I}_{[a,b]}(f) = (b-a)f(a)$
• ***Formule du rectangle à droite :***
Sur l'intervalle de référence $[-1,1]$ *:* $\tilde{I}_{[-1,1]}(\tilde{f}) = 2\tilde{f}(1)$.
Sur l'intervalle $[a,b]$: $\tilde{I}_{[a,b]}(f) = (b-a)f(b)$.

• ***Formule du point milieu :***
sur l'intervalle de référence $[-1,1]$: $\tilde{I}_{[-1,1]}(\tilde{f}) = 2\tilde{f}(0)$.
Sur l'intervalle $[a,b]$: $\tilde{I}_{[a,b]}(f) = (b-a)f\left(\frac{a+b}{2}\right)$.

• ***Interpolation à deux points : formule de trapèze***
Sur l'intervalle de référence $[-1,1]$: $\tilde{I}_{[-1,1]}(\tilde{f}) = \tilde{f}(-1) + \tilde{f}(1)$.
Sur l'intervalle $[a,b]$: $\tilde{I}_{[a,b]}(f) = (b-a)\frac{f(a)+f(b)}{2}$.
• ***Interpolation à trois points : formule de Simpson***
Sur l'intervalle de référence $[-1,1]$: $\tilde{I}_{[-1,1]}(\tilde{f}) = \frac{\tilde{f}(-1)+4\tilde{f}(0)+\tilde{f}(1)}{3}$.
Sur l'intervalle $[a,b]$ $\tilde{I}_{[a,b]}(f) = \frac{b-a}{6}\left(f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right)$.

4.3 Étude de l'erreur pour les formules de quadrature élémentaires

4.3.1 Degré de précision d'une méthode d'intégration numérique

Définition 4.1. *On définit le degré de précision (ou degré d'exactitude) d'une formule d'intégration numérique comme le plus grand entier* $r \geq 0$ *pour lequel on assure*

$$\forall q \in \mathbb{R}_r[X], \quad I_{[a,b]}(q) = \tilde{I}_{[a,b]}(q).$$

Autrement dit, une formule de quadrature est dite d'ordre r *si elle est exacte sur* $\mathbb{R}_r[X]$ *et inexacte pour au moins un polynôme de degré* $r+1$.

***Remarque* 4.2.**

- Pour vérifier qu'une formule est au moins d'ordre r, il suffit de vérifier qu'elle est exacte sur la base canonique $(1, X, \ldots, X^r)$.
- Pour les formule de quadrature considérées dans ce cours, on a

$$I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x \simeq \int_a^b p(x)\mathrm{d}x,$$

où p est le polynôme d'interpolation de Lagrange de f aux points $(x_i)_{i=0}^n$. On en déduit que le degré de précision de la méthode de quadrature est au moins n, le degré du polynôme d'interpolation. En effet si $f \in \mathbb{R}_n[x]$ alors p coïncide avec son polynôme d'interpolation sur tout $[a,b]$ et donc la formule de quadrature est exacte.

• ***Méthode du rectangle à gauche :***
Cette méthode est au moins de degré 0 (exacte pour les constantes). Pour savoir si elle est au moins de degré 1, *on teste la méthode avec* $f(x) = x$, *on a*

$$\int_a^b f(x)\mathrm{d}x = \int_a^b x\mathrm{d}x = \frac{b^2-a^2}{2} \neq (b-a)a = (b-a)f(a).$$

On en déduit que la méthode du rectangle à gauche est de degré de précision 0.

• ***Méthode du rectangle à droite :***

Cette méthode est au moins de degré 0 (exacte pour les constantes). Pour savoir si elle est au moins de degré 1*, on teste la méthode avec* $f(x) = x$*, on a*

$$\int_a^b f(x)\mathrm{d}x = \int_a^b x\mathrm{d}x = \frac{b^2 - a^2}{2} \neq (b-a)b = (b-a)f(b).$$

On en déduit que la méthode du rectangle à droite est de degré de précision 0.

• ***Méthode du point milieu :***

Cette méthode est au moins de degré 0 (exacte pour les constantes). Pour savoir si elle est au moins de degré 1, on teste la méthode avec $f(x) = x$*, on a*

$$\int_a^b f(x)\mathrm{d}x = \int_a^b x\mathrm{d}x = \frac{b^2 - a^2}{2} = (b-a)\frac{a+b}{2} = (b-a)f\left(\frac{a+b}{2}\right).$$

Donc la méthode est au moins d'ordre 1. On teste alors $f(x) = x^2$

$$\int_a^b f(x)\mathrm{d}x = \int_a^b x^2\mathrm{d}x = \frac{b^3 - a^3}{3} \neq (b-a)\left(\frac{a+b}{2}\right)^2 = (b-a)f\left(\frac{a+b}{2}\right).$$

On en déduit que la méthode du point milieu est de degré de précision 1.

• ***Méthode du trapèze :***

Cette méthode est moins de degré 1 car le polynôme d'interpolation considéré est d'ordre 1. On teste

$$f(x) = x^2, \int_a^b f(x)\mathrm{d}x = \int_a^b x^2\mathrm{d}x = \frac{b^3 - a^3}{3} \neq (b-a)\frac{a^2+b^2}{2} = (b-a)\frac{f(a)+f(b)}{2}.$$

On en déduit que la méthode du trapèze est de degré de précision 1.

• ***Méthode du Simpson :***

Cette méthode est moins de degré 2 car le polynôme d'interpolation considéré est d'ordre 2*. Pour simplifier, on se place sur* $[-1, 1]$*, si la formule est exacte sur* $[-1, 1]$ *alors par le changement de variable qui permet de se ramener à* $[a, b]$*, la formule sera aussi exacte sur* $[a, b]$*. On teste* $\tilde{f}(t) = t^3$*. On rappelle que la formule de quadrature de Simpson sur* $[-1, 1]$ *est donnée par*

$$\tilde{I}_{[-1,1]}(\tilde{f}) = \frac{\tilde{f}(-1) + 4\tilde{f}(0) + \tilde{f}(1)}{3}$$

On vérifie alors que

$$\int_{-1}^1 \tilde{f}(t)\mathrm{d}t = \int_{-1}^1 t^3\mathrm{d}t = 0 = \frac{(-1)^3 + 4\times 0^3 + 1^3}{3}$$

mais

$$\int_{-1}^1 t^4\mathrm{d}t = \frac{2}{5} \neq \frac{2}{3} = \frac{(-1)^4 + 4\times 0^4 + 1^4}{3}$$

ce qui montre que la méthode de Simpson est de degré de précision 3.

4.3.2 Majorations d'erreur pour les méthodes de quadrature élémentaires

Maintenant qu'on a une première idée de la précision des méthodes simples grâce au degré de précision, nous allons essayer de caractériser un peu mieux l'erreur commise quand on utilise une formule de quadrature élémentaire. On s'intéresse donc à l'erreur $E(f)$ et notre but est de trouver pour chacune des méthodes introduites précédemment $M(f)$ (qui dépendra de l'intervalle $[a,b]$, de f et de ses dérivées), le plus petit possible, tel que

$$|E(f)| = \left|I_{[a,b]}(f) - \tilde{I}_{[a,b]}(f)\right| \leq M(f)$$

4.3.3 Quelques outils/rappels

Théorème 4.1. *(Théorème des accroissements finis TAF) Soit $f : [a,b] \to \mathbb{R}$ une fonction continue sur $\mathbb{R}$ dérivable sur $]a,b[$. Il existe $c \in]a,b[$ tel que*

$$f(b) - f(a) = f'(c)(b-a).$$

Théorème 4.2. *(Formule de Taylor-Lagrange) Soient $k \in \mathbb{N}$, $f : [a,b] \to \mathbb{R}$ de classe $\mathcal{C}^k$ sur $[a,b]$ dérivable $k+1$, sur $]a,b[$. Il existe $c \in]a,b[$ tel que*

$$f(b) = f(a) + (b-a)f'(a) + \frac{(b-a)^2}{2!}f''(a) + \ldots + \frac{(b-a)^k}{k!}f^{(k)}(a) + \frac{(b-a)^{k+1}}{(k+1)!}f^{(k+1)}(c).$$

Théorème 4.3. *(Erreur interpolation de Lagrange) Soient $n \in \mathbb{N}^*$, $f : [a,b] \to \mathbb{R}$ une application continue sur $[a,b]$ de classe $\mathcal{C}^{n+1}$ sur $]a,b[$.*
On prend $x_0 < \ldots < x_n$, $(n+1)$ réels distincts dans $[a,b]$, tels que :

$$x_i = a + ih \quad \text{avec } h = \frac{b-a}{n}.$$

Soient p le polynôme d'interpolation de Lagrange de f aux points x_i et e l'erreur qu'on définit par :

$$e(x) = f(x) - p(x).$$

Alors
$$\forall x \in [a,b], \quad |e(x)| \leq \tfrac{M_{n+1}}{(n+1)!}|\omega(x)|.$$

Avec $M_{n+1} = \max_{x\in[a,b]} \left|f^{(n+1)}(\xi_x)\right|$ et $\omega(x) = \Pi_{i=0}^{n}(x - x_i)$.

•***Méthode du rectangle à gauche :***

Soit f de classe $\mathcal{C}^1$ sur $[a,b]$, on a

$$\begin{aligned}
|E(f)| = \left|I_{[a,b]}(f) - \tilde{I}_{[a,b]}(f)\right| &= \left|\int_a^b f(x)dx - (b-a)f(a)\right| \\
&= \left|\int_a^b (f(x) - f(a))dx\right| \\
&\leq \int_a^b |f(x) - f(a)|dx \quad \textit{(Par inégalité triangulaire)} \\
&\leq \int_a^b \max_{c\in[a,b]} |f'(c)|\,|x-a|dx \quad \textit{(Par TAF)} \\
&= M_1 \int_a^b (x-a)dx \\
&= \frac{(b-a)^2}{2} M_1.
\end{aligned}$$

Est-ce que $M(f) = \frac{(b-a)^2}{2} M_1$ est le plus petit majorant possible ?

La réponse et oui ; si on prend la fonction $f(x) = x$, on a $M_1(f) = \max_{[a,b]} |f'(x)| = 1$ et donc

$$\begin{aligned}
E(f) &= \int_a^b f(x)\mathrm{d}x - (b-a)f(a) \\
&= \frac{b^2-a^2}{2} - (b-a)a \\
&= \frac{(b-a)^2}{2} = \frac{(b-a)^2}{2} M_1.
\end{aligned}$$

On en déduit que cette fonction f particulière réalise l'égalité dans l'estimation d'erreur précédente et donc que la majoration trouvée est optimale (on ne peut pas trouver de plus petit majorant pour cette méthode).

Propriété 4.1. *(Erreur pour la méthode du rectangle à gauche)*

$$|E(f)| \leq \frac{M_1}{2}(b-a)^2.$$

• ***Méthode du rectangle à droite :***
Le même raisonnement (en remplaçant $f(a)$ par $f(b)$) conduit à la même estimation. L'erreur pour la méthode du rectangle à droite est donnée par :

$$|E(f)| \leq \frac{M_1}{2}(b-a)^2.$$

• ***Méthode du point milieu :***

Soit f une fonction de classe $\mathcal{C}^2$ sur $[a,b]$, on a

$$\begin{aligned}|E(f)| = \left|I_{[a,b]}(f) - \tilde{I}_{[a,b]}(f)\right| &= \left|\int_a^b f(x)\mathrm{d}x - (b-a)f\left(\frac{a+b}{2}\right)\right| \\ &= \left|\int_a^b \left(f(x) - f\left(\frac{a+b}{2}\right)\right)\mathrm{d}x\right| \\ &\leq \int_a^b \left|f(x) - f\left(\frac{a+b}{2}\right)\right|\mathrm{d}x \\ &\leq \int_a^b \max_{c\in[a,b]}|f'(c)|\left|x - \frac{a+b}{2}\right| dx \\ &= \frac{(b-a)^2}{4}M_1.\end{aligned}$$

Est-ce que $M(f) = \frac{(b-a)^2}{2}M_1$ est le plus petit majorant possible ? Non, car on a vu précédemment que la méthode était exacte pour les polynômes de degré 1, donc que l'erreur pour des polynômes de degré 1 était nulle. Cette information pour l'instant n'apparaît pas dans notre majoration d'erreur.
Comment obtenir une majoration optimale pour la méthode du point milieu ?
Pour simplifier les calculs, on se place sur l'intervalle de référence $[-1,1]$ et on s'intéresse à l'erreur

$$|E(\tilde{f})| = \left|\int_{-1}^1 \tilde{f}(t)\mathrm{d}t - \tilde{I}_{[-1,1](\tilde{f})}\right|,$$

pour revenir à l'erreur $|E(f)|$ sur $[a,b]$, on appliquera le changement de variable φ. Sur l'intervalle $[-1,1]$ l'erreur pour la méthode du point milieu s'écrit :

$$|E(\tilde{f})| = \left|\int_{-1}^1 \tilde{f}(t)\mathrm{d}t - 2\tilde{f}(0)\right|.$$

D'après la formule de Taylor-Lagrange à l'ordre 2 , il existe $c \in]a,b[$ tel que

$$\tilde{f}(t) = \tilde{f}(0) + t\tilde{f}'(0) + \frac{t^2}{2}\tilde{f}''(c).$$

et on pose $\tilde{q}(t) = \tilde{f}(0) + t\tilde{f}'(0)$. Comme $\tilde{q}$ est un polynôme de degré 1, la formule du point milieu est exacte pour $\tilde{q}$

$$\int_{-1}^1 \tilde{q}(t)\mathrm{d}t = \tilde{I}_{[-1,1]}(\tilde{q}) = 2\tilde{q}(0) = 2\tilde{f}(0)$$

Ainsi

$$\begin{aligned}|E(\tilde{f})| &\leq \left|\int_{-1}^1 \tilde{f}(t)\mathrm{d}t - \int_{-1}^1 \tilde{q}(t)\mathrm{d}t\right| \\ &\leq \int_{-1}^1 |\tilde{f}(t) - \tilde{q}(t)|\mathrm{d}t \\ &= \int_{-1}^1 \left|\tilde{f}(t) - \tilde{f}(0) - \tilde{f}'(0)t\right|\mathrm{d}t \\ &\leq \int_{-1}^1 \max_{[-1,1]}\left|\tilde{f}''(c)\right|\frac{t^2}{2}\mathrm{d}t \quad \textit{Par Taylor Lagrange} \\ &= \frac{\tilde{M}_2}{2}\times\frac{2}{3} = \frac{\tilde{M}_2}{3}.\end{aligned}$$

où on a noté $\tilde{M}_k = \max_{c\in[-1,1]} \left|\tilde{f}^{(k)}(c)\right|$. *On vérifie que cette majoration est optimale en testant* $\tilde{f}(t) = t^2$, *on*

$$\int_{-1}^{1} \tilde{f}(t)\mathrm{d}t - 2\tilde{f}(0) = \int_{-1}^{1} t^2\mathrm{d}t - 2\times 0 = \frac{2}{3} = \frac{\tilde{M}_2}{3}$$

Si on revient à l'intervalle de départ $[a,b]$ *en appliquant* φ, *on a*

$$|E(f)| = \left|\int_a^b f(x)\mathrm{d}x - \tilde{I}_{[a,b]}(f)\right| = \frac{b-a}{2}\left|\int_{-1}^{1}\tilde{f}(t)\mathrm{d}t - \tilde{I}_{[-1,1]}(\tilde{f})\right| = \frac{b-a}{2}|E(\tilde{f})|$$

et donc

$$|E(f)| \leq \frac{b-a}{2}\times\frac{\tilde{M}_2}{3}$$

On conclut en revenant aux dérivées de f. *Pour cela, on rappelle que*

$$\tilde{f}(t) = f(\varphi(t))$$

et donc par différentiation d'une fonction composée

$$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}t}(t) = \frac{\mathrm{d}f}{\mathrm{d}\varphi(t)}(\varphi(t))\times\frac{\mathrm{d}\varphi}{\mathrm{d}t}(t) = \frac{\mathrm{d}f}{\mathrm{d}x}(x)\times\frac{(b-a)}{2}$$

Par récurrence sur k, *l'ordre de dérivation, on montre que*

$$\tilde{f}^{(k)}(t) = \left(\frac{b-a}{2}\right)^k f^{(k)}(x)$$

ce qui, avec $k=2$ *et en passant au max dans l'égalité, donne*

$$\tilde{M}_2 = \left(\frac{b-a}{2}\right)^2 M_2 = \frac{(b-a)^2}{4}M_2.$$

Propriété 4.2. *(Erreur pour la méthode du point milieu)*

$$|E(f)| \leq \frac{M_2}{24}(b-a)^3.$$

- ***Méthode de trapèze :***

Comme pour la méthode du point milieu, on prend f *de classe* $\mathcal{C}^2$ *et on se ramène à la fonction* $\tilde{f}$ *définie sur* $[-1,1]$. *Soit* $\tilde{p}$ *le polynôme d'interpolation de Lagrange de* $\tilde{f}$ *aux points* $-1, 1$. *On sait que l'erreur d'interpolation est majorée par*

$$|e(t)| = |\tilde{f}(t) - \tilde{p}(t)| \leq \frac{|t^2-1|}{2}\tilde{M}_2.$$

On en déduit que

$$\begin{aligned}
|E(\tilde{f})| &= \left|I_{[-1,1]}(\tilde{f}) - I_{[-1,1]}(\tilde{p})\right| \\
&= |\int_{-1}^{1}\tilde{f}(t) - \tilde{p}(t)\Big)\,\mathrm{d}t| \\
&\leq \int_{-1}^{1}|\tilde{f}(t) - \tilde{p}(t)|\mathrm{d}t \\
&\leq \frac{\tilde{M}_2}{2}\int_{-1}^{1}\left(1-t^2\right)\mathrm{d}t = \frac{2\tilde{M}_2}{3}.
\end{aligned}$$

Cette majoration est optimale car on atteint l'égalité (en valeur absolue) avec la fonction $\tilde{f}(t) = t^2$

$$\int_{-1}^{1} \tilde{f}(t)\mathrm{d}t - [f(-1) + f(1)] = \int_{-1}^{1} t^2\mathrm{d}t - 2 = -\frac{4}{3} = -\frac{2\tilde{M}_2}{3}$$

On revient à l'intervalle $[a, b]$ *de la même manière que pour la méthode du point milieu.*

Propriété 4.3. *(Erreur pour la méthode du trapèze)*

$$|E(f)| \leq \frac{M_2}{12}(b-a)^3.$$

***Remarque* 4.3.** La méthode des trapèzes, bien que construite à partir d'un polynôme d'interpolation de degré 1 est moins précise que celle du point milieu construite elle à partir d'un polynôme de degré 0.

- ***Méthode de Simpson :***

Soit f *de classe* $\mathcal{C}^4$ *sur* $[a, b]$ *et* $\tilde{f}(t) = f(\varphi(t))$ *définie sur* $[-1, 1]$. *Si, comme pour la méthode du trapèze, on considère* $\tilde{p}$ *le polynôme d'interpolation de* $\tilde{f}$, *on obtient une majoration d'erreur dépendant de* $\tilde{M}_3$ *et donc non optimale car on sait que la méthode de Simpson est d'ordre de précision 3 (* $cf.$ *première majoration faite pour la méthode du point milieu). Essayons de dériver directement la majoration optimale en généralisant l'idée utilisée pour la méthode du point milieu. Soit* $\tilde{q}$ *l'unique polynôme de degré 3 tel que*

$$\tilde{q}(0) = \tilde{f}(0), \quad \tilde{q}'(0) = \tilde{f}'(0), \quad \tilde{q}(-1) = \tilde{f}(-1), \quad \tilde{q}(1) = \tilde{f}(1).$$

Lemme 4.1. *Lorsqu'on interpole* $\tilde{f}$ *par* $\tilde{q}$ *on commet une erreur* $\tilde{e}$ *majorée par*

$$|\tilde{e}(t)| = |\tilde{f}(t) - \tilde{q}(t)| \leq \frac{\tilde{M}_4}{24} t^2 \left(1 - t^2\right) \quad \forall t \in [-1, 1].$$

Propriété 4.4. *(Erreur pour la méthode de Simpson)*

$$|E(f)| \leq \frac{M_4}{2880}(b-a)^5.$$

4.4 Méthodes de quadrature composites

4.4.1 Principe

Pour approcher $I_{[a,b]}(f)$, *les méthodes dites composites consistent à partitionner l'intervalle* $[a, b]$ *en petits sous-intervalles* $[a_{j-1}, a_j]$ *sur lesquels on applique une méthode élémentaire du type de celles présentées en section précédente. On ne considérera ici que le cas où les sous-intervalles sont de même taille (subdivision uniforme).*

Plus précisément, soient $J \in \mathbb{N}^*$ *et* $h = \frac{b-a}{J}$, *on pose* $a_j = a + jh$ *pour tout* $j = 0, \ldots, J$. *Par la relation de Chasles*

$$\int_a^b f(x)\mathrm{d}x = \sum_{j=1}^{J} \int_{a_{j-1}}^{a_j} f(x)\mathrm{d}x.$$

Puis par une formule de quadrature élémentaire, sur $[a_{j-1}, a_j]$

$$\int_{a_{j-1}}^{a_j} f(x)\mathrm{d}x \simeq \tilde{I}_j(f)$$

Ce qui nous donne la formule de quadrature composite

$$I_{[a,b]}(f) = \sum_{i=1}^{J} \tilde{I}_j(f).$$

***Remarque* 4.4.**

- Dans la suite on va supposer qu'on applique la même formule de quadrature élémentaire sur tous les sous-intervalles.
- Pour obtenir une majoration de l'erreur totale commise, on écrit

$$\begin{aligned} E_J(f) &= \left| \int_a^b f(x)\mathrm{d}x - \sum_{j=1}^{J} \tilde{I}_j(f) \right| \\ &= \left| \sum_{j=1}^{J} \int_{a_{j-1}}^{a_j} f(x)\mathrm{d}x - \tilde{I}_j(f) \right| \\ &\leq \sum_{j=1}^{J} \left| \int_{a_{j-1}}^{a_j} f(x)\mathrm{d}x - \tilde{I}_j(f) \right| = \sum_{j=1}^{J} E_j(f). \end{aligned}$$

4.4.2 Méthodes composites couramment utilisées

- ***Méthode des rectangles à gauche :***

$$\begin{gathered} I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x \simeq h \sum_{j=1}^{J} f\left(a_{j-1}\right) \\ E_J(f) = \left| \int_a^b f(x)\mathrm{d}x - h \sum_{j=1}^{J} f\left(a_{j-1}\right) \right| \leq \frac{Jh^2}{2} M_1 \leq \frac{(b-a)^2}{2J} M_1. \end{gathered}$$

- ***Méthode des rectangles à droite :***

$$I_{[a,b](f)} = \int_a^b f(x)\mathrm{d}x \simeq h \sum_{j=1}^{J} f\left(a_j\right).$$

$$E_J(f) = \left| \int_a^b f(x)\mathrm{d}x - h \sum_{j=1}^{J} f\left(a_j\right) \right| \leq \frac{Jh^2}{2} M_1 \leq \frac{(b-a)^2}{2J} M_1.$$

- ***Méthodes des points milieux :***

$$I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x \simeq h \sum_{j=1}^{J} f\left(\frac{a_{j-1}+a_j}{2}\right).$$

$$E_J(f) = \left| \int_a^b f(x)\mathrm{d}x - h \sum_{j=1}^{J} f\left(\frac{a_{j-1}+a_j}{2}\right) \right| \leq \frac{(b-a)^3}{24J^2} M_2.$$

- ***Méthode des trapèzes :***

$$\begin{gathered} I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x \simeq h \sum_{j=1}^{J} \frac{f(a_{j-1})+f(a_j)}{2} \\ E_J(f) = \left| \int_a^b f(x)\mathrm{d}x - h \sum_{j=1}^{J} \frac{f(a_{j-1})+f(a_j)}{2} \right| \leq \frac{(b-a)^3}{12J^2} M_2. \end{gathered}$$

***Remarque* 4.5.** En pratique on n'implémente pas directement cette formule car on voit qu'elle nécessite de calculer deux fois chaque $f(a_j), j = 1, \ldots, J-1$. On réécrit la méthode sous la forme

$$\begin{aligned} I_{[a,b]}(f) \simeq & h \sum_{j=1}^{J} \frac{f(a_{j-1}) + f(a_j)}{2} \\ &= \frac{h}{2} \left(\sum_{j=1}^{J} f(a_{j-1}) + \sum_{j=1}^{J} f(a_j) \right) \\ &= \frac{h}{2} \left(\sum_{j=0}^{J-1} f(a_j) + \sum_{j=1}^{J} f(a_j) \right) \\ &= h \frac{f(a) + f(b)}{2} + h \sum_{j=1}^{J-1} f(a_j). \end{aligned}$$

- ***Méthode de Simpson composite :***

$$I_{[a,b]}(f) = \int_a^b f(x)\mathrm{d}x \simeq h \sum_{j=1}^{J} \frac{f(a_{j-1}) + 4f\left(a_{j-1/2}\right) + f(a_j)}{6}$$

avec

$$a_{j-1/2} = \frac{a_{j-1} + a_j}{2} = a + (j - 1/2)h$$

ou encore (forme plus économe en calculs)

$$I_{[a,b]}(f) = \frac{h}{3} \left(\frac{f(a) + f(b)}{2} + \sum_{j=1}^{J-1} f(a_j) + 2 \sum_{j=1}^{J} f\left(a_{j-1/2}\right) \right).$$

$$E_J(f) = \left| \int_a^b f(x)\mathrm{d}x - h \sum_{j=1}^{J} \frac{f(a_{j-1}) + 4f\left(a_{j-1/2}\right) + f(a_j)}{6} \right| \leq \frac{(b-a)^5}{2880 J^4} M_4.$$

***Remarque* 4.6.** La méthode de Simpson est très utilisée, elle constitue un compromis entre précision (erreur en h^4) et quantité de calculs ($2J + 1$ évaluations) considéré généralement comme satisfaisant.

Chapitre 5

Résolution numérique des équations non linéaires $f(x) = 0$

5.1 Introduction

Ce chapitre a pour but de rechercher des solutions de l'équation non linéaire $f(x) = 0$ où f est une fonction donnée. Les méthodes numériques pour approcher une solution consistent à localiser grossièrement un zéro de f en procédant le plus souvent par des considérations graphiques ; la solution grossière est noté x_0, et aussi à construire à partir de x_0 une suite $x_1, x_2, x_3, \ldots$ telle que

$$\lim_{n\to\infty} x_n = \bar{x} \qquad \text{où } \bar{x} \text{ vérifie } f(\bar{x}) = 0.$$

5.2 Existence de solutions et localisation des solutions

On se donne une application f continue d'un ouvert I de $\mathbb{R}$ dans $\mathbb{R}$ et on cherche à approcher numériquement une solution de l'équation $f(x) = 0$. Est ce qu'un tel x existe ?

Définition 5.1. *Soit $f : I \longrightarrow \mathbb{R}$, on appelle zéro de f tout $\bar{x} \in I$ qui satisfait*

$$f(\bar{x}) = 0.$$

On dit aussi que $\bar{x}$ est une racine de f.

Définition 5.2. *On appelle point fixe de f tout $\bar{x}$ qui satisfait*

$$f(\bar{x}) = \bar{x}.$$

***Remarque* 5.1.** Si $\bar{x}$ est un zéro de f alors $\bar{x}$ est un point fixe de $g : x \mapsto f(x) + x$.

5.3 Premiers résultats théoriques

Théorème 5.1. *(Théorème des valeurs intermédiaires) Soit f une fonction continue sur $I = [a, b]$. Alors f atteint toutes les valeurs entre $f(a)$ et $f(b)$. Par exemple, pour une fonction croissante on a :*

$$\forall d \in [f(a), f(b)] \text{ il existe } c \in I \text{ tel que } f(c) = d.$$

Corollaire 5.3.1. *Soit $f : I = [a, b] \to \mathbb{R}$ une application continue telle que $f(a)f(b) < 0$, c'est-à-dire que $f(a)$ et $f(b)$ sont non nuls et de signes opposés. Alors il existe $\bar{x} \in]a, b[$ tel que $f(\bar{x}) = 0$. Si de plus f est strictement monotone, alors $\bar{x}$ est unique.*

***Exemple* 13.**

- *Une fonction polynomiale à coefficients réels de degré impair admet au moins un zéro sur $\mathbb{R}$.*
- *L'équation $x(1 + 2x) = e^x$ admet une unique solution dans l'intervalle $]0, 1[$.*

Corollaire 5.3.2. *(Théorème de point fixe)*
Soit $g : [a, b] \to [a, b]$ continue sur $[a, b]$. Alors g admet un point fixe $\bar{x}$ dans l'intervalle $[a, b]$

Preuve 5.3.1. *Supposons par l'absurde que g n'admet pas de point fixe sur $[a, b]$. Alors en particulier*

$$\begin{cases} g(a) > a \\ g(b) < b \end{cases}$$

Posons $f(x) = g(x) - x$, f est continue puisque g l'est. De plus

$$\begin{cases} f(a) > 0 \\ f(b) < 0 \end{cases}$$

Le TVI nous donne alors l'existence d'un zéro de f dans $[a, b], \bar{x}$, qui est donc par définition de f un point fixe de g.

5.4 Construction de solutions approchées

Définition 5.3. *On appelle méthode itérative un procédé de calcul de la forme*

$$x_{n+1} = F(x_n), \quad n = 0, 1, 2, \ldots$$

qui part d'une valeur donnée x_0 pour calculer x_1, puis à l'aide de x_1 on calcule x_2, etc. La formule qui donne x_{n+1} est dite formule de récurrence. Le procédé est dit convergent si x_n tend vers un nombre fini lorsque n tend vers l'infini.

Définition 5.4. *Soit p un entier positif. On dit qu'une méthode convergente est d'ordre p s'il existe une constante C telle que si $\bar{x} = \lim_n x_n$ alors*

$$|\bar{x} - x_{n+1}| \leq C\,|\bar{x} - x_n|^p$$

ou encore

$$\lim_{n\to\infty} \frac{|\bar{x} - x_{n+1}|}{|\bar{x} - x_n|^p} = C.$$

Si $p = 1$ on parle de convergence linéaire, si $p = 2$ la convergence est dite quadratique.

***Remarque* 5.2.** Dans le cas où $p = 1$, il est nécessaire que $C < 1$ pour que (x_n) converge vers $\bar{x}$.

5.4.1 Méthodes de la dichotomie et de Lagrange

Principe des deux méthodes

Ces méthodes s'appuient sur le théorème des valeurs intermédiaires. On considère la fonction f continue sur $[a,b]$ à valeurs dans $\mathbb{R}$ avec $f(a)f(b)<0$. On sait donc qu'il existe un zéro de f dans $I_0=]a,b[$ qu'on note $\bar{x}$. Pour localiser $\bar{x}$ on va calculer à chaque itération un sous-intervalle $I_n=[a_n,b_n]$ de I_{n-1} dans lequel $\bar{x}$ est localisé.

- ***Algorithme de dichotomie :***

La méthode de dichotomie consiste à découper l'intervalle I_n en deux intervalles de même longueur. Concrètement, supposons par exemple que $f(a_n)<0, f(b_n)>0$ et notons $x_n=\frac{a_n+b_n}{2}$. On étudie le signe de $f(x_n)$.

- *Si $f(x_n)=0$ alors $\bar{x}=x_n$.*
- *Si $f(x_n)<0$, d'après de TVI, il existe un zéro de f sur l'intervalle $]x_n,b_n[$ On pose donc*
$$a_{n+1}=x_n \text{ et } b_{n+1}=b_n.$$
- *Si $f(x_n)>0$ d'après de TVI, il existe un zéro de f sur l'intervalle $]a_n,x_n[$. On pose donc*
$$a_{n+1}=a_n \text{ et } b_{n+1}=x_n.$$

On poursuit alors la construction jusqu'à obtenir la précision souhaitée.
La méthode de dichotomie consiste donc à construire deux suites adjacentes (a_n) et (b_n) (qui sont les extrémités des intervalles successifs dans lesquels $\bar{x}$ est localisé) convergeant vers $\bar{x}$.

Théorème 5.2. *Soient a et b dans $\mathbb{R}$ tels que $a<b$, et $f:[a,b]\mapsto\mathbb{R}$ une application continue possédant un unique zéro noté $\bar{x}\in]a,b[$. On suppose de plus que $f(a)f(b)<0$. Alors les deux suites (a_n) et (b_n) convergent vers $\bar{x}$ et on a les majorations d'erreur suivantes :*

$$\forall n\geq 0,\quad 0\leq \bar{x}-a_n\leq\frac{b-a}{2^n},\quad 0\leq b_n-\bar{x}\leq\frac{b-a}{2^n}.$$

Preuve 5.4.1. *On va montrer que les deux suites (a_n) et (b_n) sont adjacentes. Plus précisément, on va montrer par récurrence sur n que*

$$\forall n\geq 1,\quad a_{n-1}\leq a_n,\quad b_n\leq b_{n-1},\quad b_n-a_n=\frac{b-a}{2^n}\quad (\text{ ou bien } a_n=b_n=\bar{x})\,(1)$$

On rappelle que : (a_n) et (b_n) sont dites adjacentes si l'une est croissante, l'autre décroissante et $|a_n-b_n|\to 0$ quand $n\to+\infty$.
D'autre part ont sait que : si (a_n) et (b_n) sont adjacentes alors ces deux suites sont convergentes et ont la même limite $l\in\mathbb{R}, \forall n, a_n\leq l\leq b_n$.
- Initialisation. $n=1$. Si $f(x_0)$ est du signe de $f(a)$ alors $a_1=x_0=\frac{a_0+b_0}{2}$ et $b_1=b_0=b$. Ainsi on vérifie que :

$$a_0\leq\tfrac{a+b}{2}=a_1\leq b_1=b_0$$
$$b_1-a_1=b-\tfrac{a+b}{2}=\tfrac{b-a}{2}$$

On fait le raisonnement si $f(x_0) f(a) < 0$.
-Hérédité : On suppose la propriété (1) est vraie au rang n. On étudie le signe de $f(x_n) = f\left(\frac{a_n+b_n}{2}\right)$. Si $f(x_n) f(a) > 0$ (l'opposé se traitant de manière complètement analogue). Alors

$$a_{n+1} = x_n \quad et \qquad a_n \leq \tfrac{a_n+b_n}{2} = a_{n+1},$$

et par hypothèse de récurrence on obtient

$$b_{n+1} - a_{n+1} = b_n - \frac{a_n + b_n}{2} = \frac{b_n - a_n}{2} = \frac{1}{2}\frac{b-a}{2^n} = \frac{b-a}{2^{n+1}},$$

ce qui achève la démonstration de (1).
On a ainsi montré que (a_n) et (b_n) sont adjacentes, on note l la limite commune. Afin de conclure la démonstration du théorème, il reste à montrer que $l = \bar{x}$. Pour cela on va prouver que $f(l) = 0$. Pour tout n on a

$$\begin{cases} f(a_n) f(a) \geq 0, \\ f(b_n) f(b) \geq 0, \end{cases}$$

ce qui donne en passant à la limite par continuité de f :

$$\begin{cases} f(l) f(a) \geq 0, \\ f(l) f(b) \geq 0. \end{cases}$$

Comme par ailleurs on a supposé que $f(a)$ et $f(b)$ étaient non nuls et de signes opposés on a nécessairement

$$\begin{cases} f(l) \leq 0 \\ f(l) \geq 0 \end{cases} \quad \Longrightarrow \quad f(l) = 0$$

On conclut enfin en disant que

$$\begin{aligned} 0 \leq x_n - a_n \leq b_n - a_n \leq \tfrac{b-a}{2^n} \\ 0 \leq b_n - x_n \leq b_n - a_n \leq \tfrac{b-a}{2^n}. \end{aligned}$$

***Remarque* 5.3.** Les itérations s'achèvent à la m-ième étape quand

$$|\bar{x} - x_m| \leq |I_m| < \varepsilon.$$

où ε est une tolérance fixée. On a $|I_m| = \frac{b-a}{2^m}$ donc pour avoir une erreur inférieure à ε, on doit prendre le plus petit m tel que

$$m \geq \frac{\log\left(\frac{b-a}{\varepsilon}\right)}{\log(2)} = \log_2\left(\frac{b-a}{\varepsilon}\right).$$

La méthode de dichotomie ne garantit pas la réduction monotone de l'erreur absolue d'une itération à l'autre, c'est-à-dire qu'on n'a pas

$$|\bar{x} - x_{n+1}| \leq C_n |\bar{x} - x_n| \qquad \text{pour tout } n \geq 0,$$

avec $C_n < 1$. La méthode de dichotomie n'est pas une méthode d'ordre 1.
Son avantage est qu'elle est facile à implémenter, une fois un zéro isolé on a convergence à coup sûr.
Ses inconvénients : Convergence lente, méthode pas généralisable en dimension supérieure, ne s'applique pas par exemple pour chercher les extremas, par ex. pour $x \mapsto x^2$.

• ***Algorithme de Lagrange :***
Plutôt que de couper l'intervalle en deux intervalles de même longueur, on découpe $I_n = [a_n, b_n]$ *en* $[a_n, x_n]$ *et* $[x_n, b_n]$ *où* x_n *est le point d'intersection de la droite passant par* $(a_n, f(a_n))$ *et* $(b_n, f(b_n))$ *et l'axe des abscisses. Autrement dit* x_n *satisfait les équations suivantes*

$$\frac{f(b_n) - f(a_n)}{b_n - a_n}(x_n - b_n) + f(b_n) = 0$$

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

On est alors ramené comme précédemment à étudier le signe de $f(x_n)$.

Propriété 5.1. *Si* f *(régulière sur* $[a,b]$ *) avec* f'' *de signe constant sur* $[a,b]$ *(càd* f *convexe ou concave sur* $[a,b]$ *alors soit il existe un* n *tel que* $f(x_n) = 0$*, soit* x_n *est bien défini pour tout* n *et* x_n *converge à l'ordre 1 vers* $\bar{x}$ *où* $\bar{x}$ *est l'unique zéro de* f *dans* $[a,b]$.

5.4.2 Méthodes de points fixe

Théorème 5.3. *(Théorème de point fixe contractant)*
Soient I *un intervalle fermé non vide de* $\mathbb{R}$ *et* $g : I \to I$ *une application strictement contractante càd qu'il existe une constante* $0 < k < 1$ *telle que*

$$\forall (x,y) \in I^2 \quad |g(x) - g(y)| < k|x - y|.$$

Alors il existe un unique $\bar{x} \in I$ *tel que* $g(\bar{x}) = \bar{x}$ *la suite définie par*

$$\begin{cases} x_0 \in I, \\ \forall n \in \mathbb{N} \quad x_{n+1} = g(x_n), \end{cases}$$

converge vers $\bar{x}$*. De plus, on a la majoration d'erreur*

$$|x_n - \bar{x}| \leq k^n |x_0 - \bar{x}| \quad \forall n \in \mathbb{N}.$$

***Remarque* 5.4.** I peut être de la forme $I = \mathbb{R}$, $I =]-\infty, a]$, $[a, +\infty[$ ou $I = [a,b]$.

Preuve 5.4.2. *On se place dans le cas plus simple où* $I = [a,b]$.
- Existence d'un point fixe : facile via le Corollaire 5.3.2 (on remarquera pour cela que la propriété de k*-contractivité entraîne la continuité de* g*).*
- Unicité du point fixe : soient $\bar{x}$ *et* $\tilde{x}$ *deux points fixes de* g *sur* I*. Comme* g *est strictement contractante on a :*

$$|\bar{x} - \tilde{x}| = |g(\bar{x}) - g(\tilde{x})| \leq k|\bar{x} - \tilde{x}|.$$

La condition $k < 1$ *impose donc* $|\bar{x} - \tilde{x}| = 0$ *càd* $\bar{x} = \tilde{x}$.
- Convergence et majoration d'erreur : On va la démontrer par récurrence sur n *que*

$$|x_n - \bar{x}| \leq k^n |x_0 - \bar{x}|.$$

- *Initialisation : on a bien* $|x_0 - \bar{x}| \leq k^0 |x_0 - \bar{x}|$.
- *Hérédité : on suppose l'inégalité vérifiée au rang* n*. Au rang* $n+1$ *on a*

$$|x_{n+1} - \bar{x}| = |g(x_n) - g(\bar{x})| \leq k|x_n - \bar{x}| \leq k^{n+1}|x_0 - \bar{x}|.$$

L'inégalité est bien satisfaite au rang $n+1$ *ce qui conclut la preuve par récurrence.*

Comme $k < 1$, on a $k^n \to 0$ et cette inégalité permet ainsi de montrer la convergence de x_n vers $\bar{x}$.
Dans le cas plus général où I est un fermé de $\mathbb{R}$, l'existence de $\bar{x}$ et la convergence de la suite s'obtiennent simultanément en montrant que (x_n) est une suite de Cauchy. En effet comme g est k-contractante on peut écrire pour $n \geq 1$:

$$\begin{aligned} |x_{n+1} - x_n| &= |g(x_n) - g(x_{n-1})| \\ &\leq k\,|x_n - x_{n-1}| \\ &\leq k^n\,|x_1 - x_0|\,. \end{aligned}$$

Soit maintenant $p > n \geq 0$, on a

$$\sum_{j=n+1}^{p} (x_j - x_{j-1}) = x_p - x_n,$$

et l'inégalité précédente donne :

$$\begin{aligned} |x_p - x_n| &\leq \sum_{j=n+1}^{p} k^{j-1}\,|x_1 - x_0| \\ &= \sum_{j=0}^{p-(n+1)} k^{n+j}\,|x_1 - x_0| \\ &= k^n\,|x_1 - x_0| \sum_{j=0}^{p-(n+1)} k^j \\ &\leq k^n\,|x_1 - x_0| \sum_{j=0}^{+\infty} k^j \\ &= \frac{k^n\,|x_1 - x_0|}{1-k} \to 0 \qquad \textit{quand} \quad n \to \infty. \end{aligned}$$

La suite (x_n) est donc une suite de Cauchy dans I qui est un fermé de $\mathbb{R}$. Elle converge donc dans I, on note l sa limite ; par continuité de g, en passant à la limite dans l'égalité $x_{n+1} = g(x_n)$, on obtient $l = g(l)$.
L'unicité se démontre de la même manière que pour le cas d'un segment.

Corollaire 5.4.1. *Supposons $g : \mathbb{R} \to \mathbb{R}$ de classe $\mathcal{C}^1$ et soit $\bar{x}$ un point fixe de g. Si $|g'(\bar{x})| < 1$ alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\bar{x} - x_0| \leq \varepsilon$, alors la suite donnée par $x_{n+1} = g(x_n)$ converge vers $\bar{x}$ lorsque n tend vers l'infini.*

Définition 5.5. *Soit $\bar{x}$ un point fixe d'une application g, on dit que $\bar{x}$ est un point fixe attractif si $|g'(\bar{x})| < 1$, répulsif si $|g'(\bar{x})| > 1$.*

Le corollaire précédent nous donne donc la convergence locale autour des point fixes attractifs. A contrario le résultat suivant donne la non-convergence pour les points fixes répulsifs.

Propriété 5.2. *Soient $g : I \to \mathbb{R}$ une application de classe $\mathcal{C}^1$ et $\bar{x}$ un point fixe répulsif de g. On pose (x_n) la suite définie par l'approximation de point fixe $x_{n+1} = g(x_n)$. Alors soit la suite (x_n) est stationnaire, égale à $\bar{x}$, soit (x_n) ne converge pas vers $\bar{x}$.*

5.4.3 La méthode de Newton

Soit f une fonction définie sur un intervalle $[a,b]$, continument dérivable sur $[a,b]$ (i.e. de classe $\mathcal{C}^1$ sur $[a,b]$). Soit $\bar{x}$ un zéro de f dans $[a,b]$ tel que $f'(\bar{x}) \neq 0$. Par relaxation on peut écrire :

$$\{f(\bar{x}) = 0\} \Longleftrightarrow \{\forall \lambda \neq 0, \quad \bar{x} = g_\lambda(\bar{x})\} \quad g_\lambda(x) = x + \lambda f(x)$$

Alors le meilleur choix de constante λ est celui pour lequel la méthode est d'ordre 2, soit celui pour lequel $g'_\lambda(\bar{x}) = 0$ donc :

$$\lambda = -\frac{1}{f'(\bar{x})}$$

En pratique il n'est pas possible (sauf cas particulier) de calculer $f'(\bar{x})$ puisqu'on ne connait pas $\bar{x}$. Une solution consiste alors à approximer à chaque itération $f'(\bar{x})$ par $f'(x_n)$. C'est la méthode de Newton.

Théorème 5.4. *(Théorème de convergence globale) On reprend les hypothèses précédentes. et on suppose de plus que f est de classe $\mathcal{C}^2$ sur I et que f' et f'' ne s'annulent pas sur I. Soit $x_0 \in I$ tel que $f(x_0)$ soit du même signe que f'' (on suppose qu'il existe au moins un tel x_0). Alors la suite définie par la méthode de Newton*

$$\begin{cases} x_0 \text{ donné} \\ x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = g(x_n) \end{cases}$$

est bien définie et converge de manière monotone vers $\bar{x}$ unique zéro de f (également point fixe de g) sur I. De plus $g'(\bar{x}) = 0$.

Preuve 5.4.3. *Comme f' est de signe constant sur I, f est strictement monotone sur I et l'équation $f(x) = 0$ admet au plus une solution. Le point $\bar{x}$ est donc l'unique zéro de f sur I. L'égalité $g'(\bar{x}) = 0$ découle directement de la définition de la méthode de Newton (on a choisit λ dans la méthode de relaxation pour avoir cette égalité). Reste à prouver la convergence de la suite (x_n) vers $\bar{x}$. On remarque que quitte à changer f en $-f$, on peut supposer sans perte de généralité que $f'' > 0$ (le cas $f'' < 0$ se traitant de manière identique). On doit maintenant distinguer deux cas $f' > 0$ et $f' < 0$.*
• $f' > 0$: f est donc strictement croissante sur I et comme $f(x_0) > 0 = f(\bar{x})$ ceci implique que $x_0 > \bar{x}$. Faire le tableau de signe de g', g : on observe que g est décroissante pour $x < \bar{x}$ et croissante pour $x > \bar{x}$. Comme $x_1 = g(x_0) > g(\bar{x})$, on a $x_1 > \bar{x}$. D'autre part

$$f(x_0) > 0, f'(x_0) > 0 \Longrightarrow x_1 = x_0 - \frac{f(x_0)}{f'(x_0)} < x_0$$

donc on a $\bar{x} < x_1 < x_0$. Par récurrence on montre que la suite (x_n) décroit. Comme elle est minorée par $\bar{x}$ elle converge et sa limite est l'unique point fixe de g sur I : $\bar{x}$.
• $f' < 0$: les variations sont échangées, on montre que $x_0 < x_1 < \bar{x}$ et on obtient une suite (x_n) croissante majorée qui doit converger vers $\bar{x}$.

***Remarque* 5.5.** Si on enlève l'hypothèse $f(x_0)$ de même signe que f'' : si $x_1 \in I$ alors on peut montrer que $f(x_1)$ est de même signe que f'' et donc on commence l'itération à partir de x_1.

- **Interprétation géométrique** : La définition de x_{n+1} peut se réécrire sous la forme

$$0 = f'(x_n)(x_{n+1} - x_n) + f(x_n).$$

Ceci signifie que $(x_{n+1}, 0)$ est le point d'intersection de la tangente à f en x_n avec l'axe des abscisses.

Théorème 5.5. *(Convergence quadratique locale) Soit $I = [a, b]$, on considère la fonction f définie sur I de classe $\mathcal{C}^2$ et $\bar{x} \in I$ tel que $f(\bar{x}) = 0$ mais $f'(\bar{x}) \neq 0$. Soient m et $M > 0$ tels que*

$$|f'(x)| \geq m, \quad |f''(x)| \leq M \quad \forall x \in I.$$

On pose $c = \frac{M}{2m}(b-a)$, Soit $x_0 \in I$. On suppose que la suite définie par la méthode de Newton à partir de x_0 est bien définie, est à valeurs dans I et converge vers $\bar{x}$. Alors on a :

$$\begin{array}{l} - |x_{n+1} - \bar{x}| \leq \frac{M}{m} \frac{(\bar{x}-x_n)^2}{2} \text{ pour tout } n \geq 0, \\ - |x_n - \bar{x}| \leq \frac{2m}{M} c^{2^n} \text{ pour tout } n \geq 0. \end{array}$$

- Si de plus $c < 1$ alors

$$\forall \varepsilon > 0, \quad n \geq \ln\left(\ln\left(\frac{M\varepsilon}{2m}\right) / \ln c\right) / \ln 2 \Rightarrow |x_n - \bar{x}| \leq \varepsilon.$$

Preuve 5.4.4. *La formule de Taylor-Lagrange à l'ordre 2 appliquée à f au point x_n s'écrit*

$$f(\bar{x}) = f(x_n) + (\bar{x} - x_n) f'(x_n) + \frac{(\bar{x} - x_n)^2}{2} f''(c_n),$$

avec c_n dans l'intervalle entre x_n et $\bar{x}$. Comme $f(\bar{x}) = 0$, on obtient en divisant par $f'(x_n)$:

$$-\frac{f(x_n)}{f'(x_n)} - \bar{x} + x_n = \frac{(\bar{x} - x_n)^2}{2} \frac{f''(c_n)}{f'(x_n)},$$

c'est-à-dire

$$x_{n+1} - \bar{x} = \frac{(\bar{x} - x_n)^2}{2} \frac{f''(c_n)}{f'(x_n)},$$

Il en découle l'inégalité

$$|x_{n+1} - \bar{x}| \leq \frac{(\bar{x} - x_n)^2}{2} \frac{M}{m}.$$

Le deuxième point s'en déduit alors par simple récurrence :
-Initialisation : pour $n = 0$ on a bien l'inégalité suivante

$$|x_0 - \bar{x}| \leq (b - a) = \frac{2m}{M} c.$$

- Hérédité : on suppose qu'au rang n l'inégalité est vérifiée i.e.

$$|x_n - \bar{x}| \leq \frac{2m}{M} c^{2^n}.$$

D'après ce qui précède,

$$\begin{aligned} |x_{n+1} - \bar{x}| &\leq (\bar{x} - x_n)^2 \frac{M}{m} \\ &\leq \left(\frac{2m}{M} c^{2^n}\right)^2 \frac{M}{m} \\ &\leq \frac{2m}{M} c^{2^{n+1}}. \end{aligned}$$

Ceci permet de conclure la démonstration du second point du théorème. Supposons que $c < 1$ et soit $\varepsilon > 0$. D'après les estimations précédentes on a $|\bar{x} - x_n| < \varepsilon$ dès que $\frac{2m}{M} c^{2^n} < \varepsilon$ càd

$$n \ln 2 \geq \ln\left(\left(\frac{M\varepsilon}{2m}\right) / \ln c\right).$$

5.4.4 Méthode de la corde et méthode de la sécante

• ***Méthode de la corde :*** *Cette méthode permet d'éviter qu'à chaque itération on ait à évaluer $f'(x_n)$. La méthode de la corde consiste à remplacer $f'(x_n)$ par $f'(x_0)$ ce qui donne*

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_0)}.$$

- Interprétation géométrique : Le calcul des itérés s'effectue en prenant toujours la même pente $f'(x_0)$.

Théorème 5.6. *Supposons f continûment dérivable $(\mathcal{C}^1)$ qui admet un zéro $\bar{x}$ tel que $f'(\bar{x}) \neq 0$. Alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\bar{x} - x_0| \leq \varepsilon$, la suite (x_n) donné par la méthode de la corde converge vers $\bar{x}$. La convergence est linéaire.*

• ***Méthode de la sécante :***
Toujours dans la même idée d'éviter le calcul de la dérivée de f, on peut faire l'approximation

$$f'(x_n) \simeq \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}},$$

ce qui nous définit la méthode de la sécante

$$\begin{cases} x_0, x_1 \text{ donnés dans } I, x_0 \neq x_1, \\ \forall n > 0, \quad x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n). \end{cases}$$

Théorème 5.7. *Soit f de classe $\mathcal{C}^2$ qui admet un zéro $\bar{x}$ tel que $f'(\bar{x}) \neq 0$. Alors il existe $\varepsilon > 0$ tel que si x_0 satisfait $|\bar{x} - x_0| \leq \varepsilon$, la suite (x_n) définie par la méthode de la sécante converge vers $\bar{x}$ et on*

$$|\bar{x} - x_{n+1}| \leq C |\bar{x} - x_n|^{\varphi},$$

où φ est le nombre d'or $\left(\varphi = \frac{1+\sqrt{5}}{2} \sim 1,62\right)$.

Chapitre 6

Initiation à la résolution des systèmes linéaires : méthode de Gauss et LU

6.1 Introduction

Les systèmes d'équations algébriques jouent un rôle très important en ingénierie. Ces systèmes peuvent être classé en deux grandes familles : les systèmes linéaires et les systèmes non linéaires. Les progrés de l'informatique et de l'analyse numérique permettent d'aborder des problèmes de taille important. On résout aujourd'hui des systèmes de plusieurs centaines de milliers d'inconnues. On rencontre ces applications en mécanique de fluide, traitement d'images, etc.

On distingue deux grandes familles de méthodes :

- *Les méthodes directes : on obtient la solution en un nombre fini d'opérations.*
- *Les méthodes itératives : on construit une suite qui tend vers la solution.*

Dans ce chapitre, nous allons parler des méthodes directes pour la résolution des systèmes linéaires.

6.2 Systèmes linéaires

De façon générale, la résolution d'un système linéaire consiste à trouver $x = (x_1\, x_2 \cdots x_n)^T$ *solution de*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + \cdots + a_{3n}x_n &= b_3 \\ \vdots &= \vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{6.1}$$

ou sous forme matricielle :

$$Ax = b \tag{6.2}$$

où $A = (a_{ij})_{1\leq i,j\leq n}$ *et* $b = (b_1\ b_2\ \cdots\ b_n)^T$. *Bien entendu, la matrice* A *et le vecteur* b *sont connus. Il reste à déterminer* x.

Nous supposons que la matrice A est non singulières ou inversible, c'est-à-dire la matrice inverse A^{-1} existe. Ainsi la solution de l'équation (6.2) peut s'écrire :

$$x = A^{-1}b$$

Nous verrons cependant que le calcul de la matrice A^{-1} est plus difficile et plus long que la résolution du système de départ.

***Exemple* 14.** *Considérons le système linéaire suivant :*

$$\begin{aligned} 2x_1 + 3x_2 &= 8 \\ 3x_1 + 4x_2 &= 11 \end{aligned}$$

Pour le résoudre, on peut utiliser la méthode classique qui consiste à éliminer les équations une à une par substitution successive. On isole x_1 de la première équation :

$$x_1 = \frac{8 - 3x_2}{2}$$

que l'on substitue dans la deuxième équation :

$$3\frac{8 - 3x_2}{2} + 4x_2 = 11$$

On déduit facilement que $x_2 = 2$ et $x_1 = 1$.

Théoriquement, on peut étendre cette méthode à des systèmes de grande taille, mais difficile de transcrire cette façon de faire sous forme d'algorithme programmable dans un langage informatique quelconque.

Avant de procéder aux méthodes de résolution, on peut se demander quels types de systèmes linéaire facile à résoudre. Le cas le plus simple est sans doute lorsque la matrice A est diagonale, dans ce cas la solution est :

$$x_i = \frac{b_i}{a_{ii}}, \ \textit{pour } i = 1, 2, ..., n.$$

On remarque aussi que la solution existe et unique si et seulement si les termes diagonaux sont non nuls.

Le deuxième type de système simple est le système triangulaire inférieur ou supérieur

Définition 6.1. *Une matrice est dite triangulaire inférieure (ou supérieure) si tous les a_{ij} (ou tous les a_{ji}) sont nuls pour $i < j$. Une matrice triangulaire inférieure est de la forme :*

$$\begin{pmatrix} a_{11} & 0 & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & 0 & 0 & \cdots & 0 \\ a_{31} & a_{32} & a_{33} & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ a_{n-1\,1} & a_{n-1\,2} & a_{n-1\,3} & \vdots & a_{n-1\,n} & 0 \\ a_{n1} & a_{n2} & a_{n3} & \vdots & a_{n\,n-1} & a_{nn} \end{pmatrix}$$

Si la matrice A est triangulaire, le système linéaire est facile à résoudre. On commence par la première équation si A est triangulaire inférieure et la dernière si A est triangulaire supérieure et de résoudre une à une les équations. On parle de descente triangulaire ou de remontée triangulaire.

Le cas général pour la descente triangulaire :

$$x_1 = \frac{b_1}{a_{11}}$$
$$x_i = \frac{\left(b_i - \sum_{k=1}^{i-1} a_{ik}x_k\right)}{a_{ii}} \quad pour\ i = 2, 3, ..., n.$$

Pour la remontée triangulaire, on a :

$$x_n = \frac{b_n}{a_{nn}}$$
$$x_i = \frac{\left(b_i - \sum_{k=i+1}^{n} a_{ik}x_k\right)}{a_{ii}} \quad pour\ i = n-1, n-2, ..., 2, 1.$$

***Exercice* 1.** *Calculer le nombre d'opérations nécessaire pour les deux algorithmes (descente et remontée).*

Dans les sections qui suivent, nous voyons comment ramener un système linéaire quelconque à un ou plusieurs systèmes triangulaires.

Définition 6.2. *Une méthode de résolution d'un système linéaire est dite directe si la solution du système peut être obtenue en un nombre prédéterminé d'opérations.*

On peut alors déduire le temps de calcul nécessaire à la résolution qui peut être très long si n est grand.

6.3 Opérations élémentaires sur les lignes

On peut multiplier à gauche les termes de la relation (6.2) par une matrice P inversible ; la solution n'est pas modifiée puisque l'on peut remultiplier par P^{-1} pour revenir au système de départ. Ainsi :

$$PAx = Pb$$

possède la même solution que le système (6.2).

***Remarque* 6.1.** *Ce résultat n'est plus vrai si la matrice P n'est pas inversible.*

Pour transformer un système quelconque en système triangulaire, il suffit d'utiliser trois opérations élémentaires sur les lignes de la matrice. Ces opérations correspondent à trois types de matrices P différents.

6.3.1 Multiplier une ligne par un scalaire

Remplacer la ligne i par un multiple d'elle-même ($L_i \leftarrow \lambda L_i$) revient à multiplier le système linéaire (6.2) par une matrice diagonale inversible $P = M(L_i \leftarrow \lambda L_i)$, dont tous les éléments diagonaux sont 1, sauf a_{ii}, qui vaut λ, les autres termes sont nuls.

Le déterminant de M est λ. M est donc inversible si $\lambda \neq 0$. La matrice inverse est

$$M^{-1} = M(L_i \leftarrow \frac{1}{\lambda} L_i)$$

Il suffit de remplacer λ par $\frac{1}{\lambda}$.

***Exemple* 15.** *Soit le système linéaire :*

$$\begin{pmatrix} 2 & 1 & 3 \\ 3 & 1 & 4 \\ 4 & 5 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 8 \\ 10 \end{pmatrix}$$

dont la solution est $x = (1\ 1\ 1)^T$. Si on souhaite multiplier la ligne 2 par un facteur 3, cela revient á multiplier le système par la matrice :

$$M(L_2 \leftarrow 3L_2) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

La solution du nouveau système reste la même.

6.3.2 Permuter deux lignes

L'opération élémentaire qui consiste à intervertir deux lignes $(L_i \leftrightarrow L_j)$ est également connue sous le nom de permutation de lignes. Cette opération est équivalente à multiplié le système (6.2) par une matrice inversible $P(L_i \leftrightarrow L_j)$ qui contient 1 sur la diagonale sauf à la ligne i, où le 1 est dans la colonne j, et à la ligne j, où le 1 est dans la colonne i. Les autres termes sont nuls.

***Exemple* 16.** *Intervertir la ligne 2 et la ligne 3 du système de l'exemple précédent. Il suffit de le multiplier le système (6.2) par la matrice :*

$$P(L_2 \leftrightarrow L_3) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

et on obtient :

$$\begin{pmatrix} 2 & 1 & 3 \\ 4 & 5 & 1 \\ 3 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ 10 \\ 8 \end{pmatrix}$$

La matrice P est inversible, la matrice inverse est :

$$P^{-1}(L_i \leftrightarrow L_j) = P(L_i \leftrightarrow L_j).$$

***Remarque* 6.2.** *On vérifie facilement que le déterminant de P est -1. Lorsque on permute deux lignes, le déterminant du système de départ change de signe.*

6.3.3 Opération $(L_i \leftarrow L_i + \lambda L_j)$

La dérnière opération élémentaire consiste à remplacer la ligne i par elle même plus un multiple de la ligne j. Cela est équivalent à multiplier le système originale par une matrice inversible $T(L_i \leftarrow L_i + \lambda L_j)$ vaut 1 sur la diagonale et 0 ailleurs sauf a_{ij}, qui vaut λ.

***Exemple* 17.** *Dans le système de l'exemple précédent, on souhaite remplacer la deuxième ligne par elle même moins deux fois la première ligne. On multiplie alors le système par*

$$T(L_2 \leftarrow L_2 - 2L_1) = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

ce qui donne

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

***Remarque* 6.3.** *La matrice $T(L_i \leftarrow L_i + \lambda L_j)$ est inversible. La matrice inverse est*

$$T^{-1}(L_i \leftarrow L_i + \lambda L_j) = T(L_i \leftarrow L_i - \lambda L_j)$$

Pour revenir en arrière, il suffit de soustraire la ligne qu'on vient d'ajouter

Nous allons donc utilisé ces opérations pour ramener un système d'équations linéaires quelconque à un système triangulaire. On distingue deux cas :

- *On souhaite résoudre une seule équation $Ax = b$. On travaille sur la matrice $[A, b]$, c'est la méthode de Gauss. Pour résoudre le système, il faut*
  - *une triangularisation,*
  - *une remontée.*
- *On doit résoudre plusieurs systèmes avec la même matrice*

  $$Ax = b_1, \cdots, Ax = b_k.$$

  On décompose A en produit de deux matrices triangulaires U (supérieure) et L (inférieure), puis on résout les deux systèmes

  $$Ly_k = b_k,\ Ux_k = y_k,$$

  c'est la décomposition LU.

6.4 Triangularisation

Soit A une matrice n lignes m colonnes $(m \geq n)$. Il y a $n-1$ étapes :

À l'étapes k, on annule sous la diagonale les coefficients de la colonne k. On appelle K^e pivot $(p^{(k)})$ le coefficient de la diagonale $p^{(k)} = a_{kk}$.

À chaque ligne $i > k$ on soustrait la ligne k multipliée par $\dfrac{q}{p^{(k)}}$ avec $q = a_{ik}$

$$\forall j, k \leq j \leq m,\ a_{ij} = a_{ij} - a_{kj}.\frac{q}{p^{(k)}}$$

lorsque $j = k$ on fait l'opération

$$\begin{aligned} a_{ik} &= a_{ik} - a_{ik} \\ &= 0 \end{aligned}$$

Mais dans la pratique il ne faut pas calculer ces coefficients pour éviter les erreurs de calcul.

Cet algorithme ne fonctionne pas si l'un des pivots est nul.

Algorithm 1 Algorithme de triangularisation

Require: $A = (A(i,j)), n, m$
1: **for** $k = 1$ to $n-1$ **do**
2: $p \leftarrow A(k,k)$;
3: **if** $p = 0$ **then**
4: Error(" pivot null ")
5: **end if**
6: **for** $i = k+1$ to n **do**
7: $q \leftarrow A(i,k)$;
8: $A(i,k) \leftarrow 0$;
9: $piv = \frac{q}{p}$;
10: **for** $j = k+1$ to m **do**
11: $A(i,j) \leftarrow A(i,j) - A(k,j) * piv$;
12: **end for**
13: **end for**
14: **end for**
15: **return** A matrice triangulaire ;

6.5 Méthode d'élimination de Gauss

La méthode d'élimination de Gauss repose sur le fait que les opérations élémentaires consistent à multiplier le système de départ par une matrice inversible. Il consiste à annuler tous les termes sous la diagonale de la matrice A autrement dit trianguläriser la matrice A . Puisque les opérations élémentaires doivent être effectuer à la fois sur les lignes de la matrice A et sur le vecteur b, introduisons la matrice augmentée

Définition 6.3. *La matrice augmentée du système linéaire (6.2) est la matrice de dimension n sur $n+1$ que l'on obtient en ajoutant le membre de droite à la matrice A, c'est-à-dire :*

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} & b_2 \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} & b_3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nn} & b_n \end{pmatrix}$$

L'algorithme de Gauss consiste donc à

- *construire la matrice augmentée $[A, b]$,*
- *appliquer l'algorithme de triangularisation sur $[A, b]$,*
- *appliquer l'algorithme de remontée.*

Calculons le nombre total d'opérations (additions, multiplication et divisions) nécessaires pour la résolution du système (6.2) par la méthode d'élimination de Gauss. Notons N_T ce nombre, pour triangulariser $[A, b]$, il faut

$$\begin{aligned}
N_1 &= \sum_{k=1}^{n-1}\left(\sum_{i=k+1}^{n}\left(1div + \sum_{j=k+1}^{n+1}(1add + 1mult)\right)\right)\\
&= \sum_{k=1}^{n-1}(n-k)div + \sum_{k=1}^{n-1}(n-k+1)(n-k)add + \sum_{k=1}^{n-1}(n-k+1)(n-k)mult\\
&= \frac{n(n-1)}{2}div + \left(\sum_{k=1}^{n-1}(n-k)^2 + \sum_{k=1}^{n-1}(n-k)\right)add + \left(\sum_{k=1}^{n-1}(n-k)^2 + \sum_{k=1}^{n-1}(n-k)\right)mult\\
&= \frac{n(n-1)}{2}div + \left(\frac{n(n-1)}{2} + \frac{n(n-1)(2n-1)}{6}\right)add + \left(\frac{n(n-1)}{2} + \frac{n(n-1)(2n-1)}{6}\right)mult\\
&= O(\frac{2}{3}n^3)
\end{aligned}$$

Pour la remontée triangulaire, le nombre d'opérations est $O(n^2)$, le nombre total d'opérations est $N_T = 0(\frac{2}{3}n^3)$.

6.6 Décomposition LU

On souhaite factoriser la matrice A en deux matrices triangulaires

$$A = LU$$

avec L à diagonale unité.

Pour construire L et U, on utilise la méthode d'élimination de Gauss en se souvenant des opérations faites. En effet à la fin de la triangularisation, on obtient la matrice U.

Une étape de l'élimination de Gauss revient à multiplier A par une matrice $M^{(k)}$ de la forme :

$$M^{(k)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & m_{k+1}^k & \ddots & \\ & & \vdots & & \\ & & m_n^k & & 1 \end{pmatrix}$$

où $m_i^k = -\frac{a_{ik}}{a_{kk}}$.

La première étape consiste à multiplier A par $M^{(1)} = (m_{ij})_{1\leq i,j\leq n}$, avec

$$m_{ii} = 1 \text{ et } m_{i1} = -\frac{a_{i1}}{a_{11}},\ i = 2, \cdots, n,$$

on obtient

$$M^{(1)}\,A = \begin{pmatrix} p^{(1)} & a_{12}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(1)} & \cdots & a_{2n}^{(1)} \\ \vdots & \vdots & & \vdots \\ 0 & a_{n2}^{(1)} & \cdots & a_{nn}^{(1)} \end{pmatrix}$$

Dans la deuxième étape, on multiplie par $M^{(2)}$

$$m_{ii} = 1 \text{ et } m_{i1} = -\frac{a_{i2}^{(1)}}{a_{22}^{(1)}},\ i = 3, \cdots, n,$$

On a alors

$$M^{(2)}\, M^{(1)}\, A = \begin{pmatrix} p^{(1)} & a_{12}^{(2)} & \cdots & & a_{1n}^{(2)} \\ 0 & p^{(2)} & a_{23}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(2)} & \cdots & a_{3n}^{(2)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & a_{n3}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}$$

Finalement, on obtient

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

Nous avons donc

$$M.A = U.$$

Puisque le produit de deux matrice triangulaires inférieures est une matrice triangulaire inférieure, et lorsque la décomposition existe, elle est unique. La matrice U *ainsi obtenue est celle de la décomposition* LU *et que* M *est l'inverse de* L*.*

La matrice $M^{(k)}$ *est inversible d'inverse* $L^{(k)}$ *donnée par*

$$L^{(k)} = \begin{pmatrix} 1 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & -m_{k+1}^k & \ddots & \\ & & \vdots & & \\ & & -m_n^k & & 1 \end{pmatrix}$$

La matrice L *de la décomposition est*

$$L = L^{(1)}.L^{(2)}.\cdots.L^{(n-1)}$$

$$L = \begin{pmatrix} 1 & & & & & & \\ -m_2^1 & 1 & & & & & \\ \vdots & \ddots & 0 & & & & \\ & & 1 & & & & \\ & & -m_{k+1}^k & 1 & & & \\ & & \vdots & & \ddots & & \\ -m_n^1 & & -m_n^k & \cdots & -m_n^{n-1} & 1 & \end{pmatrix}$$

Les m_i^k *sont les coefficients de l'élimination de Gauss* $m_i^k = -\frac{a_{ik}}{a_{kk}}$*.*

Le nombre d'opérations nécessaire pour la construction de L *et* U *est* $O(\frac{2}{3}n^3)$*.*

Algorithm 2 Algorithme de décomposition LU

Require: $A = (A(i,j))$, n ;
1: $U \leftarrow A$;
2: $L \leftarrow I_n$;
3: **for** $k = 1$ to $n-1$ **do**
4: $p \leftarrow U(k,k)$;
5: **if** $p = 0$ **then**
6: Error(" pivot null")
7: **end if**
8: **for** $i = k+1$ to n **do**
9: $q \leftarrow U(i,k)$;
10: $U(i,k) \leftarrow 0$;
11: $L(i,k) \leftarrow \frac{q}{p}$;
12: **for** $j = k+1$ to n **do**
13: $U(i,j) \leftarrow U(i,j) - U(k,j) * L(i,k)$;
14: **end for**
15: **end for**
16: **end for**
17: **return** L et U matrices triangulaires

***Exemple* 18.** *Faire la décomposition LU de la matrice*

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

***Remarque* 6.4.** *Les deux algorithmes précédents fonctionnent si tous les pivots $p^{(k)}$ sont non nuls.*

Le théorème suivant donne des conditions suffisantes pour que ces pivots ne seront pas nuls.

Théorème 6.1. *L'élimination de Gauss fonctionne sur une matrice A ssi toutes ses matrices principales $A_k = (a_{ij})_{1 \leq i,j \leq k}$ sont inversibles.*

***Remarque* 6.5.** *Le fait qu'une matrice principale A_k ne soit pas inversible ne signifie pas que la matrice A n'est pas inversible.*

***Exemple* 19.**

$$M_1 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}, \; M_2 = \begin{pmatrix} 1 & 1 & 0 \\ -1 & -1 & 1 \\ 0 & 1 & 1 \end{pmatrix}$$

ces matrices sont inversibles et peuvent être triangularisées.

D'une manière générale si un pivot est nul, alors on permute deux lignes.

Théorème 6.2. *Si tous les pivots possibles sont nuls, alors la matrice est singulière.*

Pourquoi cela n'est-il pas satisfaisant ? On considère le système linéaire

$$\begin{cases} \epsilon x_1 + x_2 = 1 \\ x_1 + x_2 = 2 \end{cases}$$

avec $\epsilon = 10^{-8}$*, la solution est* $(1,1)$*. Si on résout ce système par l'élimination de Gauss, on obtient* $(x_1, x_2) = (0,1)$ *qui est complètement différent de la solution exacte. En permutant les lignes et les colonnes, on obtient*

$$\begin{pmatrix} 1 & 1 \\ 1 & \epsilon \end{pmatrix} \begin{pmatrix} x_2 \\ x_1 \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

La solution de ce système par élimination de Gauss est $(1,1)$*. Cet exemple montre que si pivot non nul est petit peut conduire à des résultats complètement faux.*

6.7 Gauss avec pivot partiel

On cherche le pivot maximal parmi les éléments de la colonne k situés sous la diagonale

$$p^{(k)} = a_{sk}, \ \textit{avec} \ |a_{sk}| = \max_{i=k,\ldots,n} |a_{ik}|.$$

Si tous les pivots de cette colonne sont nuls la matrice est singulière.

Algorithm 3 Algorithme de Gauss avec pivot partiel

Require: $A = (A(i,j))$, n, m
1: **for** $k = 1$ to $n-1$ **do**
2: $p \leftarrow A(k,k)$; $l \leftarrow k$;
3: **for** $i = k$ to n **do**
4: **if** $|A(i,k)| > |p|$ **then**
5: $p \leftarrow A(i,k)$; $l \leftarrow i$;
6: **end if**
7: **end for**
8: **if** $l \neq k$ **then**
9: **for** $j = k$ to m **do**
10: $T \leftarrow A(k,j)$; $A(k,j) \leftarrow A(l,j)$;
11: $A(l,j) \leftarrow T$;
12: **end for**
13: **end if**
14: **for** $i = k+1$ to n **do**
15: $q \leftarrow A(i,k)$;
16: $A(i,k) \leftarrow 0$;
17: $piv = \frac{q}{p}$;
18: **for** $j = k+1$ to m **do**
19: $A(i,j) \leftarrow A(i,j) - A(k,j) * piv$;
20: **end for**
21: **end for**
22: **end for**
23: **return** A matrice triangulaire

6.8 Gauss avec pivot total

On recheche le pivot maximal parmi les éléments de la sous-matrice $[a_{ij}]$ $(i \geq k, j \geq k)$ *:*

$$p^{(k)} = a_{st} \ \textit{avec} \ |a_{st}| = \max_{i,j=k,\ldots,n} |a_{ij}|.$$

On permute les lignes s et k puis les colonnes t et k. C'est une méthode stable mais l'inconvénient est qu'on modifier l'ordre des inconnues.

6.9 Effet sur la décomposition LU

Permuter deux lignes d'une matrice correspond à la multiplier à gauche par une matrice P qui contient 1 sur la diagonale sauf la ligne i, où le 1 est dans la colonne j, et à la ligne j, où le 1 est dans la colonne i. Les autres termes sont nuls.

L'élimination de Gauss avec permutation revient à obtenir :

$$M^{(1)}.P_{1,l_1}.A = A^{(1)},$$

puis

$$M^{(2)}.P_{2,l_2}.M^{(1)}.P_{1,l1}.A = A^{(2)},$$

À la fin on obtient

$$M.A = A^{(n-1))} = U,$$

avec $M = M^{(n-1)}.P_{n-1,l_{n-1}}.\cdots.M^{(2)}.P_{2,l2}.M^{(1)}.P_{1,l1}$. La matrice U est triangulaire supérieure, mais la matrice M n'est plus triangulaire inférieure. Par contre il est possible de commuter les matrices de permutation et les matrices $M^{(k)}$. Si $k < i < j$:

$$P_{ij} \times M^{(k)} = \bar{M}^{(k)} \times P_{ij}$$

où $\bar{M}^{(k)}$ est la matrice $M^{(k)}$ dont les coefficients i et j ont été échangés, elle est de la même forme que $M^{(k)}$, elle est donc triangulaire inférieure et elle est inversée de la même façon.

Finalement,

$$M.P.A = U$$

où $M = \tilde{M}^{(n-1)}.\tilde{M}^{(n-2)}.\cdots.\tilde{M}^{(1)}$ et $P = P_{n-1,l_{n-1}}.\cdots.P_{2,l_2}.P_{1,l_1}$.

On obtient :

$$P.A = L.U$$

avec $L = M^{-1}$.

L'algorithme de décomposition PLU se résume aux étapes suivantes

- *On applique l'algorithme avec recherche du pivot partiel, en calculant U et L.*
- *À chaque fois qu'on permute une ligne de U, on permute aussi les lignes de L en dessous de la diagonale.*
- *On conserve la matrice de permutation P.*
- *On résout $L.U.x = Pb$.*

***Exemple* 20.** *Faire la décomposition PLU de la matrice*

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

Algorithm 4 Algorithme de décomposition PLU

Require: $A = (A(i,j))$, n

$U \leftarrow A$, $L \leftarrow I_n$, $P \leftarrow I_n$
for $k = 1$ to $n-1$ **do**
  $p \leftarrow U(k,k)$
  for $i = k$ to n **do**
    if $|U(i,k)| > |p|$ **then**
      $p \leftarrow U(i,k)$, $l \leftarrow i$
    end if
  end for
  if $l \neq k$ **then**
    for $j = 1$ to n **do**
      $T \leftarrow U(k,j)$; $U(k,j) \leftarrow U(l,j)$
      $U(l,j) \leftarrow T$
      if $j < k$ **then**
        $T \leftarrow L(k,j)$; $L(k,j) \leftarrow L(l,j)$
        $L(l,j) \leftarrow T$
      end if
      $T \leftarrow P(k,j)$; $P(k,j) \leftarrow P(l,j)$
      $P(l,j) \leftarrow T$
    end for
  end if
  if $p = 0$ **then**
    Error(" matrice singulière")
  end if
  for $i = k+1$ to n **do**
    $q \leftarrow U(i,k)$
    $U(i,k) \leftarrow 0$
    $L(i,k) \leftarrow \frac{q}{p}$
    for $j = k+1$ to n **do**
      $U(i,j) \leftarrow U(i,j) - U(k,j)\,\frac{q}{p}$
    end for
  end for
end for
return L, U matrices triangulaires et P

6.10 Cas des matrices bandes

Définition 6.4. *Soit n un entier strictement positif on dit qu'une matrice A est une matrice bande s'il existe des entiers positifs p et q strictement inférieurs à n tels que $a_{ij} = 0$ pour tous les couples d'entiers $(i, j) \in \{1, ..., n\}^2$ tels que $i - j > p$ ou $j - i > q$. La largeur de bande de la matrice vaut $p+q+1$ avec p éléments a priori non nuls à gauche de la diagonale et q éléments à droite sur chaque ligne.*

Proposition 6.1. *La factorisation LU conserve de la structure bande des matrices.*

6.11 Calcul de l'inverse d'une matrice

On note e_i le ième élément de la base canonique de $\mathbb{R}^n$, la iéme colonne de A est $c_i = Ae_i$. Si d_i designe la ième colonne de A^{-1}, alors $A^{-1}e_i = d_i$ ou encore $Ad_i = e_i$ pour $i = 1, 2, \cdots, n$, donc pour calculer A^{-1}, on décompose A sous forme $A = LU$, puis on résout $2n$ systèmes triangulaires $LY_i = e_i$ et $Ud_i = Y_i$ pour $i = 1, \cdots, n$.

Nombre d'opérations *: $\frac{2}{3}n^3$ opérations pour la décomposition LU et n^3 opérations pour la résolution de $2n$ systèmes triangulaires. le nombre total d'opérations est $\frac{5}{3}n^3$.*

Sur machine cependant, la propagation des erreurs d'arrondis (plus particulièrement pour les problèmes de grande taille) fait que les vecteurs q_j calculés ne sont pas linéairement indpendants, ce qui empêche la matrice Q d'être exactement orthogonale. Ces instabilités numériques sont dues au fait que la procédure d'orthonormalisation produit des valeurs très petites.

Exercice

1. Résoudre par élimination de Gauss le système linéaire suivant :
$$\begin{aligned} 2x_1 + x_2 + 4x_4 &= 2 \\ -4x_1 - 2x_2 + 3x_3 - 7x_4 &= -9 \\ 4x_1 + x_2 - 2x_3 + 8x_4 &= 2 \\ -3x_2 - 12x_3 - x_4 &= 2 \end{aligned} \quad (1)$$
(Utiliser la matrice augmentée du système et indiquer bien les opérations effectuées sur les lignes de celle-ci.)
2. Déterminer une matrice M telle que MA soit triangulaire supérieure.
3. En déduire une décomposition LU de PA où L est une matrice triangulaire inférieure à diagonale unité, U est une matrice triangulaire supérieure et P est une certaine matrice de permutation. la décomposition obtenue est-elle unique ?

1. On utilisant les éliminations successive de Gauss on a :

$$A = A^{(1)} = \left(\begin{array}{cccc|c} 2 & 1 & 0 & 4 & 2 \\ -4 & -2 & 3 & -7 & -9 \\ 4 & 1 & -2 & 8 & 2 \\ 0 & -3 & -12 & -1 & 2 \end{array}\right) \begin{array}{l} \\ \to l_2 + 2l_1 \\ \to l_3 - 2l_1 \\ \\ \end{array}$$

$$A^{(2)} = \left(\begin{array}{cccc|c} 2 & 1 & 0 & 4 & 2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & -3 & -12 & -1 & 2 \end{array}\right)$$

On remarque que le pivot est nul, donc il faut permuter la ligne 2 avec la ligne 3 ou la ligne 4.

Permutation $l_2 \leftrightarrow l_3$

$$\tilde{A}^{(2)} = \left(\begin{array}{cccc|c} 2 & 1 & 0 & 4 & 2 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & -3 & -12 & -1 & 2 \end{array}\right) \begin{array}{l} \\ \\ \\ \to l_4 - 3l_2 \end{array}$$

$$A^{(3)} = \left(\begin{array}{cccc|c} 2 & 1 & 0 & 4 & 2 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & 0 & -6 & -1 & 8 \end{array}\right) \begin{array}{l} \\ \\ \\ \to l_4 + 2l_3 \end{array}$$

$$A^{(4)} = \left(\begin{array}{cccc|c} 2 & 1 & 0 & 4 & 2 \\ 0 & -1 & -2 & 0 & -2 \\ 0 & 0 & 3 & 1 & -5 \\ 0 & 0 & 0 & 1 & -2 \end{array}\right)$$

On obtient le système triangulaire supérieur suivant :

$$\begin{aligned} 2x_1 + x_2 + 4x_4 &= +2 \\ -x_2 - 2x_3 &= -2 \\ 3x_3 + x_4 &= -5 \\ x_4 &= -2 \end{aligned}$$

La méthode de remontée donne :

$$x = \begin{pmatrix} +3 \\ +4 \\ -1 \\ -2 \end{pmatrix}$$

- On a le passage de $A^{(1)}$ à $A^{(2)}$ donne :

$$A^{(2)} = M^{(1)}A^{(1)} \quad \text{avec} \quad M^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

- Le passage de $A^{(2)}$ à $\tilde{A}^{(2)}$ est une permutation avec :

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

- Le passage de $\tilde{A}^{(2)}$ à $A^{(3)}$ donne :

$$A^{(3)} = M^{(2)}\tilde{A}^{(2)} \quad \text{avec} \quad M^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & -3 & 0 & 1 \end{pmatrix},$$

- Le passage de $A^{(3)}$ à $A^{(4)}$ donne :

$$A^{(4)} = M^{(3)}A^{(3)} \quad \text{avec} \quad M^{(3)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 2 & 1 \end{pmatrix},$$

- On pose $U = A^{(4)}$, $M = M^{(3)}M^{(2)}PM^{(1)}$ et on a $U = MA$. Le calcul donne :

$$M = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 10 & 2 & -3 & 1 \end{pmatrix}$$

3. Il est clair que M est inversible (produit de matrices inversibles). Donc $A = M^{-1}U$.
Mais M n'est pas triangulaire inférieure et

$$M^{-1} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ -2 & 0 & 1 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 3 & -2 & 1 \end{pmatrix}$$

Donc A n'est pas décomposable en LU à cause de la permutation. Mais sachant que $P^2 = I_4$, en écrivant

$$U = MP^2A = (MP)(PA),$$

il vient que

$$PA = (MP)^{-1}U = (PM^{-1})U.$$

On pose $L = PM^{-1}$.

Ce qui revient à permuter les lignes 2 et 3 de M^{-1}, on obtient $PA = LU$ avec

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

triangulaire inférieure à diagonale unité et U triangulaire supérieure. La décomposition est unique car si

$$PA = LU = L'U'$$

alors

$$(L')^{-1} L = U'U^{-1}$$

qui est à la fois triangulaire inférieur à diagonale 1 et triangulaire supérieur, donc

$$(L')^{-1} L = U'U^{-1} = I_4.$$

Bibliographie

[1] , Etapes vers l'Analyse, Exercices et Problèmes Résolus avec Rappels de Cours, Première année du premier cycle universitaire PC, Société d'Edition et de Diffusion Al Madariss.

[2] Z. Hammouch, Cours d'Analyse numérique 1, Module M148 Filières MIP S4, A.U. 2019-2020, Faculté des Sciences et Téchniques, Errachidia.

[3] F. Jedrzejewski, Introduction aux méthodes numériques, Deuxième édition, Springer-Verlag France, Paris 2005.

[4] M. L-Lahlou, Cours d'Analyse numérique 1, Filières SMA-SMI S4, A.U. 2018-2019, Faculté des Sciences, Marrakèch.

[5] A. Taakili, Cours d'Analyse numérique, Module M1113 Filières MIP, A.U. 2011-2012, Faculté des Sciences et Téchniques, Errachidia.