DÉPARTEMENT STPI
2ÈME ANNÉE IC, FILIÈRE FAS, 2026-2027.

OUTILS MATHÉMATIQUES POUR L'INGÉNIEUR

Florent CHAZEL, Aude RONDEPIERRE, Adeline ROUCHON

Mails :
florent.chazel@insa-toulouse.fr
aude.rondepierre@insa-toulouse.fr
rouchon@insa-toulouse.fr

Table des matières

1 Rappels d'algèbre linéaire **7**
- 1.1 Propriétés du déterminant . . . 7
- 1.2 Systèmes linéaires . . . 8
  - 1.2.1 Position du problème . . . 8
  - 1.2.2 Pivot de Gauss . . . 9
  - 1.2.3 Systèmes de Cramer . . . 10
- 1.3 Réduction des endomorphismes . . . 11
  - 1.3.1 Valeurs propres, vecteurs propres . . . 11
  - 1.3.2 Recherche des valeurs propres et des vecteurs propres . . . 12
- 1.4 Caractérisation des endomorphismes diagonalisables . . . 12
- 1.5 Applications . . . 13
  - 1.5.1 Calcul de A^m . . . 13
  - 1.5.2 Étude des suites récurrentes . . . 14
  - 1.5.3 Résolution de systèmes linéaires différentiels du premier ordre . . . 15

2 Séries numériques **17**
- 2.1 Définitions et généralités . . . 17
  - 2.1.1 Définition . . . 17
  - 2.1.2 Quelques exemples à connaitre : . . . 18
  - 2.1.3 Opérations sur les séries . . . 18
  - 2.1.4 Condition nécessaire de convergence . . . 19
- 2.2 Séries numériques à termes positifs . . . 20
  - 2.2.1 Théorèmes de comparaison . . . 20
  - 2.2.2 Règles de Cauchy et de d'Alembert . . . 21
  - 2.2.3 Comparaison d'une série avec une intégrale . . . 21
  - 2.2.4 Exercices . . . 21
- 2.3 Séries à termes réels de signe quelconque . . . 23
  - 2.3.1 Séries absolument convergentes . . . 23
  - 2.3.2 Séries alternées . . . 24

3 Séries de fonctions **27**
- 3.1 Différents types de convergence . . . 27
  - 3.1.1 Convergence simple et convergence absolue . . . 27

3.1.2 Convergence uniforme 28
3.1.3 Convergence normale 28
3.2 Convergence uniforme et régularité 29

4 Fonctions de plusieurs variables **31**
4.1 Notion de fonction de plusieurs variables 31
4.2 Continuité 31
4.3 Dérivées partielles et différentielle 33
4.3.1 Définitions 33
4.3.2 Théorèmes 34
4.4 Dérivées d'ordre supérieur 36
4.4.1 Définition 36
4.4.2 Théorème de Schwarz 37
4.5 Matrice jacobienne 38
4.6 Opérateurs différentiels 40
4.7 Règle de la chaîne 42
4.8 Étude des extremums 46

5 Erreurs numériques **49**
5.1 Erreurs 49
5.1.1 Contexte 49
5.1.2 Erreur absolue et erreur relative 49
5.2 Représentation des nombres réels 50
5.2.1 Représentation traditionnelle des nombres réels en base 10 50
5.2.2 Représentation des nombres réels en base b 51
5.2.3 Représentation des nombres réels en base 2 52
5.2.4 Représentation des nombres réels en machine 53
5.2.5 Précision machine et mécanismes d'absorption et d'élimination . . 54
5.3 Conditionnement et stabilité 55
5.3.1 Objectif 55
5.3.2 Stabilité : exemple 56

6 Systèmes linéaires **59**
6.1 Résolution de systèmes linéaires simples 60
6.1.1 Systèmes linéaires diagonaux 60
6.1.2 Systèmes linéaires triangulaires 60
6.2 Résolution de systèmes linéaires quelconques 63
6.2.1 Pivot de Gauss 63
6.2.2 Interprétation matricielle 64
6.2.3 Factorisation LU 66
6.2.4 Cas particulier des matrices symétriques définies positives 70

7 Équations non linéaires : Résolution numérique **73**
7.1 Héron d'Alexandrie et les racines carrées 74
7.1.1 La méthode de Héron 74
7.2 Résolution d'une équation non linéaire 77
7.2.1 Méthode de dichotomie 79
7.2.2 Méthode du point fixe 81
7.2.3 La méthode de Newton 84
7.2.4 Combiner deux méthodes 90

Chapitre 1

Rappels d'algèbre linéaire

1.1 Propriétés du déterminant

Soit A une matrice de $\mathcal{M}_n(K)$.

1. $\det\,(AB) = \det\ A\ \det\ B$
2. A est inversible si et seulement si $\det\ A\ \neq 0$
   En effet, nous savons que si $f \in End(E)$ et E de dimension finie n alors

   $$f \text{ bijective} \Longleftrightarrow f \text{ injective} \Longleftrightarrow f \text{ surjective} \Longleftrightarrow \text{rg } f = n$$

   En transposant ceci à la matrice A, interprétée comme la matrice d'un endomorphisme on obtient en particulier

   $$\text{rg } A = n \Longleftrightarrow A \text{ inversible}$$

   et

   $$\text{rg } A = n \Longleftrightarrow \text{ Les vecteurs colonnes sont libres } \Longleftrightarrow \det_{\mathcal{E}}(C_1, C_2, ..., C_n) \neq 0$$

   De plus, puisque $AA^{-1} = I_n$ nous avons la formule

   $$1 = \det\ I_n = \det\,(AA^{-1}) = \det\ A\ \det\ A^{-1}$$

   et

   $$\det\ A^{-1} \ = \ \frac{1}{\det\ A}$$

3. Si A et A' sont deux matrices représentant le même endomorphisme f alors il existe une matrice P telle que $A' = P^{-1}AP$. On a

   $$\det\ A' = \det\,(P^{-1}AP) = \det\,(P^{-1}A)\ \det\ P = \det\ P^{-1} \det\ A\ \det\ P = \det\ A$$

   Ainsi le déterminant d'une application linéaire ne dépend pas de la base dans laquelle on écrit sa matrice et on a

   $$\det\ f = \det\ A \text{ où } A \text{ est la matrice de } f \text{ dans n'importe quelle base de } E$$

1.2 Systèmes linéaires

1.2.1 Position du problème

Soit à résoudre le système d'équations linéaires de m équations à n inconnues suivant :

$$(S) \quad \left\{ \begin{array}{ccccccccc} a_{11}x_1 & + & a_{12}x_2 & + & ... & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & ... & + & a_{2n}x_n & = & b_2 \\ a_{31}x_1 & + & a_{32}x_2 & + & ... & + & a_{3n}x_n & = & b_3 \\ & & & & & & & & \\ ... & & ... & & ... & & ... & & ... \\ a_{i1}x_1 & + & a_{i2}x_2 & + & ... & + & a_{in}x_n & = & b_i \\ ... & & ... & & ... & & ... & & ... \\ & & & & & & & & \\ a_{m-11}x_1 & + & a_{m-12}x_2 & + & ... & + & a_{m-1n}x_n & = & b_{m-1} \\ a_{m1}x_1 & + & a_{m2}x_2 & + & ... & + & a_{mn}x_n & = & b_m \end{array} \right.$$

Les a_{ij} ainsi que les b_i sont des données, éléments de K.

Les x_i sont les n inconnues, éléments de K.

1. **Écriture matricielle** : On peut écrire ce système de la façon suivante :

$$AX = B$$

où

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} \quad ; \quad B = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{pmatrix}$$

et

$$A = \begin{pmatrix} a_{11} & a_{12} & ... & a_{1n} \\ a_{21} & a_{22} & ... & a_{2n} \\ & & & \\ ... & ... & ... & ... \\ a_{i1} & a_{i2} & ... & a_{in} \\ ... & ... & ... & ... \\ & & & \\ a_{m-11} & a_{m-12} & ... & a_{m-1n} \\ a_{m1} & a_{m2} & ... & a_{mn} \end{pmatrix}$$

2. **Écriture vectorielle** : Si on note $A_1, A_2, \cdots, A_n$ les colonnes de la matrice A alors le système s'écrit

$$x_1A_1 + x_2A_2 + \cdots + x_nA_n = B$$

3. **Écriture fonctionnelle** : Si la matrice A représente l'application linéaire f de $L(E,F)$ et si X et B sont les matrices colonnes représentant les vecteurs x et b alors le système s'écrit aussi
$$f(x) = b$$

Résoudre ce système, c'est trouver tous les n-uplets $(x_1, x_2, \cdots, x_n)$ qui satisfont aux m équations. C'est définir l'ensemble des antécédents de b par f.
Si $b = 0_F$, l'équation $f(x) = 0_F$ est appelée équation homogène, elle représente le système homogène. Résoudre ce système revient à déterminer le noyau Ker f. On a toujours $0_E \in$ Ker f donc 0_E est toujours solution du système homogène, c'est la solution triviale.
Soit x_0 une solution particulière de (S) alors on a $f(x_0) = b$ et le système à résoudre devient
$$f(x) = f(x_0) \text{ soit } f(x - x_0) = 0_F \text{ et } x - x_0 \in \text{Ker } f$$
Les solutions de (S) sont de la forme $z + x_0$ où $z \in$ Ker f et x_0 est une solution particulière.

1.2.2 Pivot de Gauss

Nous nous contenterons de présenter cette méthode sur trois exemples. Cette méthode consiste à transformer le système en un système triangulaire (c'est-à-dire tel que la matrice du système soit triangulaire), que l'on résout facilement en partant du bas.
Cette transformation se fait en remplaçant une ligne par cette ligne plus une constante fois une ligne fixée à l'avance de façon à faire apparaître un zéro.

Exemples : Écrire sous forme matricielle, calculer le déterminant de la matrice puis résoudre les systèmes d'équations linéaires suivants :

$$(I) \quad \begin{cases} 5x_1 + 2x_2 + x_3 = 12 \\ 5x_1 - 6x_2 + 2x_3 = -1 \\ -4x_1 + 2x_2 + x_3 = 3 \end{cases}$$

$$(II) \quad \begin{cases} 5x_1 & + & 2x_2 & + & x_3 & = & 12 \\ 5x_1 & + & 2x_2 & + & 2x_3 & = & -1 \\ -4x_1 & - & \frac{8}{5}x_2 & + & x_3 & = & 3 \end{cases}$$

$$(III) \quad \begin{cases} 5x_1 & + & 2x_2 & + & x_3 & = & 12 \\ 5x_1 & - & x_2 & + & 2x_3 & = & 2 \\ -5x_1 & + & 4x_2 & - & 3x_3 & = & 8 \end{cases}$$

1.2.3 Systèmes de Cramer

On se place dans le cas particulier où $m = n$. On peut alors caractériser l'existence, ou non, d'une solution unique. Le résultat est le suivant :

Théorème 1.2.1 *Le système (S) a une unique solution si et seulement si le déterminant de A n'est pas nul.*
Dans ce cas le système (S) est qualifié de Cramer.

1.3 Réduction des endomorphismes

On considère un endomorphisme f d'un espace vectoriel E. K désigne $\mathbb{R}$ ou $\mathbb{C}$.

1.3.1 Valeurs propres, vecteurs propres

On appelle valeur propre de f un scalaire λ de K tel qu'il existe un vecteur v de E non nul tel que $f(v) = \lambda v$.
Le vecteur v est appelé vecteur propre de f associé à la valeur propre λ. On appelle spectre de f l'ensemble des valeurs propres de f. On note cet ensemble $Spec(f)$.

Remarque : Cette définition reste valable même lorsque E est de dimension infinie.
Propriété : Soit $\lambda \in K$. On note $E_\lambda = \{v \in E, f(v) = \lambda v\} =$ Ker $(f - \lambda id_E)$.
E_λ est un sous-espace vectoriel de E.
Si $\lambda \in Spec(f)$, alors $E_\lambda \neq \{0_E\}$ et dans ce cas, on l'appelle le sous-espace propre de E associé à λ.
Si $\lambda \notin Spec(f)$, alors $E_\lambda = \{0_E\}$.

Théorème 1.3.1
Soit λ_1 et λ_2 deux valeurs propres distinctes de f. Alors $E_{\lambda_1} \cap E_{\lambda_2} = \{0_E\}$.

Théorème 1.3.2 *(Admis) Soit $f \in L(E)$ avec $Spec(f) = \{\lambda_1, \lambda_2, \cdots, \lambda_p\}$.*

$$\begin{array}{rcl} f \text{ est diagonalisable} & \iff & \text{il existe une base de } E \text{ formée de vecteurs propres de } f \\ & \iff & E = E_{\lambda_1} \bigoplus E_{\lambda_2} \bigoplus \cdots \bigoplus E_{\lambda_p}. \end{array}$$

Théorème 1.3.3 *(Admis) Considérons $\{v_1, \cdots, v_p\}$, p vecteurs propres associés à p valeurs propres $\lambda_1, \cdots, \lambda_p$, distinctes deux à deux. Alors la famille $\mathcal{V} = (v_1, \cdots, v_p)$ est libre.*

Conséquences immédiates :

1. f possède au maximum n valeurs propres.
2. Si f possède n valeurs propres distinctes avec $n =$ dim E, alors f est diagonalisable.

1.3.2 Recherche des valeurs propres et des vecteurs propres

Théorème 1.3.4 $\lambda \in Spec(f) \iff \det(f - \lambda id_E) = 0$.

Théorème 1.3.5 *(Admis)*
Soit $f \in L(E)$. Soit $A = (a_{ij})_{(i,j)\in\{1,2,...,n\}^2} = Mat\,(f, \mathcal{E})$. On définit la fonction $P_f(x)$ par

$$P_f(x) = \det(f - x id_E) = \det(A - xI_n).$$

Cette fonction est une fonction polynomiale de degré n (où dim $E = n$), son terme de plus haut degré a pour coefficient $(-1)^n$ et son terme constant est $det(A)$. Elle s'appelle le polynôme caractéristique de A.

Théorème 1.3.6 *Deux matrices semblables ont le même polynôme caractéristique. Par conséquent, si $f \in L(E)$, on définit le polynôme caractéristique de f par $P_f = P_A$, où A est la matrice de f dans une base quelconque de E.*

En résumé, on a :

Théorème 1.3.7 *(Admis)*

1. *Les valeurs propres de f sont les racines dans K de P_f.*
2. *Si P_f est scindé sur K (c'est-à-dire que toutes ses racines sont dans K) alors f a exactement $n = dim\ E$ valeurs propres, distinctes ou non.*

Soit $\lambda \in Spec(f)$. On appelle multiplicité de la valeur propre λ l'ordre de multiplicité de λ en tant que racine de P_f. On la note m_λ.

Théorème 1.3.8 *(Admis)*
Si λ est une valeur propre de f de multiplicité m_λ, alors $1 \leq dim\ E_\lambda \leq m_\lambda$.

En conséquence nous avons la propriété suivante :
Si λ est une valeur propre simple de f (c'est-à-dire $m_\lambda = 1$), alors dim $E_\lambda = 1$.

1.4 Caractérisation des endomorphismes diagonalisables

Théorème 1.4.1 *(Admis)*
Soit $f \in L(E)$. On note m_λ la multiplicité de la valeur propre λ.

$$f \text{ est diagonalisable } \iff \begin{cases} P_f \text{ est scindé dans } K \\ \forall\, \lambda \in Spec(f), dim\ E_\lambda = m_\lambda. \end{cases}$$

On a donc aussi, si on note $Spec(f) = \{\lambda_1, \lambda_2, ..., \lambda_p\}$ l'ensemble des valeurs propres distinctes de f :

$$f \text{ est diagonalisable } \iff E = E_{\lambda_1} \oplus E_{\lambda_2} \oplus ... \oplus E_{\lambda_p}.$$

En conséquence : Si P_f possède n racines distinctes, alors f est diagonalisable.

1.5 Applications

1.5.1 Calcul de A^m

Exemple : Calculer A^m où $A = \begin{pmatrix} -7 & -9 \\ 6 & 8 \end{pmatrix}$ pour tout $m \in \mathbb{N}$.

1.5.2 Étude des suites récurrentes

Exemple : Soit les suites réelles $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ telles que $u_0 = 1$, $v_0 = 2$ et pour tout $n \in \mathbb{N}$,

$$(S)\begin{cases} u_{n+1} &= -7u_n - 9v_n \\ v_{n+1} &= 6u_n + 8v_n \end{cases}$$

Déterminer pour tout entier n l'expression de u_n et v_n en fonction de n.

1.5.3 Résolution de systèmes linéaires différentiels du premier ordre

Exemple : Soit le système différentiel suivant :

$$(S)\begin{cases} x'(t) &= -7x(t) - 9y(t) \\ y'(t) &= 6x(t) + 8y(t) \end{cases}$$

où x et y sont deux fonctions dérivables définies sur $\mathbb{R}$ vérifiant $x(0) = 0$ et $y(0) = 1$. Déterminer les fonctions x et y.

Chapitre 2

Séries numériques

2.1 Définitions et généralités

2.1.1 Définition

Soit $(u_n)_{n\in\mathbb{N}}$ une suite de nombres réels (ou complexes). On appelle *suite des sommes partielles des* u_n la suite $(S_N)_{N\in\mathbb{N}}$ où pour tout $N \in \mathbb{N}$,

$$S_N = u_0 + u_1 + \cdots + u_N = \sum_{k=0}^{N} u_k.$$

Définition 2.1.1 *Soit $(u_n)_{n\in\mathbb{N}}$ une suite de nombres réels (ou complexes). On appelle série de terme général $u_n, n \in \mathbb{N}$, notée $\sum u_n$, la suite $(S_N)_{N\in\mathbb{N}}$ des sommes partielles de u_n.*

Étudier la nature d'une série, c'est étudier la convergence de la suite des sommes partielles, c'est-à-dire se demander si la somme *infinie* des u_n existe et a une valeur finie :

- Si la suite $(S_N)_{N\in\mathbb{N}}$ admet une limite finie S ($\lim\limits_{N\to\ +\infty} S_N = S$), on dit que la série $\sum u_n$ converge ou est convergente et on écrit alors :

  $$S = \sum_{n=0}^{+\infty} u_n.$$

  et S s'appelle la somme de la série $\sum u_n$.
- Si la suite $(S_N)_{N\in\mathbb{N}}$ n'admet pas de limite, on dit que la série $\sum u_n$ diverge ou est divergente.

Exercice 2.1.1 *Pour chacune des suites suivantes, calculer les sommes partielles, déterminer si les séries associées convergent et dans ce cas calculer leur somme :*

1. $u_n = 1$, $n \in \mathbb{N}$.
2. $u_n = n$, $n \in \mathbb{N}$.
3. $u_n = q^n$, $n \in \mathbb{N}$, $q \in \mathbb{R}$.

2.1.2 Quelques exemples à connaitre :

- **Série harmonique :** $\sum \frac{1}{n}, n \in \mathbb{N}^*$
  Étudions la série de terme général $\frac{1}{n}$. On peut démontrer que

$$\forall k \in \mathbb{N}^*, \quad \int_k^{k+1} \frac{dx}{x} \leq \frac{1}{k} \leq \int_{k-1}^{k} \frac{dx}{x}$$

  alors, en sommant pour k allant de 1 à n on obtient

$$\ln(n+1) \leq S_n \leq 1 + \ln(n).$$

  Ainsi la suite des sommes partielles S_n tend vers $+\infty$ quand n tend vers $+\infty$. Donc la série $\sum \frac{1}{n}$ est divergente. Cependant, on a : $\lim\limits_{n\to+\infty} \frac{1}{n} = 0$.
- **Série géométrique :** Soit $q \in \mathbb{R}$, la série de terme général q^n converge si et seulement si $|q| < 1$. De plus, si $|q| < 1$, alors

$$\sum_{n=p}^{+\infty} q^n = \frac{q^p}{1-q}.$$

- **Série télescopique :** soit $(v_n)_{n\in\mathbb{N}}$ une suite réelle quelconque et $(u_n)_{n\in\mathbb{N}}$ une suite de la forme :

$$u_n = v_n - v_{n+1}, n \in \mathbb{N}.$$

  La série $\sum u_n$ converge si et seulement si la suite $(v_n)_{n\in\mathbb{N}}$ converge et a une limite finie.

2.1.3 Opérations sur les séries

On peut démontrer facilement que l'ensemble des séries numériques convergentes est un espace vectoriel sur $\mathbb{R}$:

Théorème 2.1.1 *Soient $(a, b) \in \mathbb{R}^2$. Si les séries $\sum u_n$ et $\sum v_n$ convergent et ont pour sommes respectives U et V alors la série $\sum(au_n + bv_n)$ converge vers $aU + bV$.*

Théorème 2.1.2 *On peut sans changer la nature d'une série, supprimer ou ajouter un nombre fini de termes.*

Démonstrations en exercice.

De façon générale, la définition 2.1.1 n'est pas commode à utiliser en pratique car la somme partielle n'est pas toujours calculable explicitement. Nous aurons donc besoin de critères de convergence permettant de déterminer la nature d'une série SANS avoir à calculer ses sommes partielles.

2.1.4 Condition nécessaire de convergence

Proposition 2.1.1 (Condition nécessaire de convergence)

$$\textit{La série } \sum u_n \textit{ converge} \Longrightarrow \lim_{n\to+\infty} u_n = 0.$$

La condition $\lim_{n\to+\infty} u_n = 0$ est nécessaire à la convergence de la série mais elle n'est **pas suffisante.** Autrement dit, si la condition $\lim_{n\to+\infty} u_n = 0$ est vérifiée, on ne peut pas conclure quant à la nature de la série $\sum u_n$. En revanche, la contraposée nous sera très utile :

$$\lim_{n\to+\infty} u_n \neq 0 \Longrightarrow \text{La série } \sum u_n \text{ diverge}$$

Exercices :

1. Étudier la convergence des séries de terme général :

$$u_n = \left(\frac{1}{2}\right)^n, \quad v_n = 2^n, \quad w_n = \frac{1}{n(n+1)}.$$

2. Démontrer que $0,999\ldots = 1$.

2.2 Séries numériques à termes positifs

Définition 2.2.1 *On dit qu'une série $\sum u_n$ est à termes positifs à partir d'un certain rang s'il existe $n_0 \in \mathbb{N}$ tel que : $\forall n \geqslant n_0,\ u_n \geq 0$.*

Dans tout ce paragraphe, les séries que nous considèrerons seront à termes tous positifs ($n_0 = 0$) pour simplifier la présentation des résultats. Mais ces résultats restent valables pour les séries numériques à termes positifs à partir d'un certain rang ($n_0 > 0$). En effet, la suppression des termes $u_0, u_1, \ldots, u_{n_0-1}$ (en nombre fini) ne modifie pas la nature de la série.

Théorème 2.2.1 *Si la série $\sum u_n$ est à termes positifs alors*
la série $\sum u_n$ converge $\Longleftrightarrow (S_N)_{N\in\mathbb{N}}$ est majorée.

En pratique nous utiliserons les 4 critères de convergence suivants : comparaison, règle de Cauchy, règle de d'Alembert, équivalents, détaillés ci-après.

2.2.1 Théorèmes de comparaison

Théorème 2.2.2 (Comparaison de séries à termes positifs)
Soient $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites réelles à termes positifs telles que :

$$\forall\ n \in \mathbb{N},\ 0 \leq u_n \leq v_n.$$

Alors :

- *Si la série $\sum v_n$ converge alors la série $\sum u_n$ converge et*

$$\sum_{n=0}^{+\infty} u_n \leq \sum_{n=0}^{+\infty} v_n.$$

- *Si la série $\sum u_n$ diverge alors la série $\sum v_n$ diverge.*

Théorème 2.2.3 (Séries de termes généraux équivalents)
Soient $(u_n)_{n\in\mathbb{N}}$ et $(v_n)_{n\in\mathbb{N}}$ deux suites réelles à termes positifs dont les termes généraux sont équivalents quand n tend vers $+\infty$, i.e.

$$\lim_{n\to+\infty} \frac{u_n}{v_n} = 1.$$

Alors les deux séries $\sum u_n$ et $\sum v_n$ sont de même nature.

2.2.2 Règles de Cauchy et de d'Alembert

Théorème 2.2.4 (Règle de Cauchy) *Soit $(u_n)_{n\in\mathbb{N}}$ une suite à termes positifs. S'il existe un réel positif ou nul l tel que*

$$\lim_{n\to+\infty} \sqrt[n]{u_n} = l,$$

alors

- *Si $0 \leq l < 1$, la série $\sum u_n$ converge.*
- *Si $l > 1$, la série $\sum u_n$ diverge.*
- *Si $l = 1$, la règle de Cauchy ne permet pas de conclure.*

Théorème 2.2.5 (Règle de d'Alembert) *Soit $(u_n)_{n\in\mathbb{N}}$ une suite à termes* **strictement** *positifs. S'il existe un réel positif ou nul l tel que*

$$\lim_{n\to+\infty} \frac{u_{n+1}}{u_n} = l,$$

alors

- *Si $0 \leq l < 1$, la série $\sum u_n$ converge.*
- *Si $l > 1$, la série $\sum u_n$ diverge.*
- *Si $l = 1$, la règle de d'Alembert ne permet pas de conclure.*

2.2.3 Comparaison d'une série avec une intégrale

Théorème 2.2.6 *Soit $f : \mathbb{R} \to \mathbb{R}$ une fonction réelle, positive, continue et décroissante sur $[n_0, +\infty[$ pour un certain $n_0 \in \mathbb{N}$. Alors la série $\sum f(n)$, $n \geqslant n_0$, et l'intégrale $\int_{n_0}^{+\infty} f(x)dx$ sont de même nature.*

Grâce à ce théorème on peut démontrer le résultat suivant (à connaitre !) :

Séries de Riemann : soit $\alpha \in \mathbb{R}$.
La série $\sum \frac{1}{n^\alpha}$, $n \in \mathbb{N}^*$ converge si et seulement si $\alpha > 1$.

2.2.4 Exercices

Étudier la nature des séries de terme général :

1. $u_n = \dfrac{\ln n}{n}$. *(Comparer u_n avec la suite définie par : $v_n = \dfrac{1}{n}$, $n \in \mathbb{N}^*$).*
2. $u_n = \left(\dfrac{n}{n+a}\right)^{n^2}$ avec $a \in \mathbb{R}$.
3. $u_n = \dfrac{a^n}{n!}$ avec $a \in \mathbb{R}_+^*$.

4. $u_n = nq^n \ln(1 + \frac{1}{n})$, $q > 0$, $n \in \mathbb{N}^*$.
5. $u_n = \frac{1}{n^2-n+1}$, $n \in \mathbb{N}$.

Cas des séries à termes négatifs. Soit $\sum u_n$ une série à termes négatifs i.e. telle qu'il existe $n_0 \in \mathbb{N}$ tel que :

$$\forall n \geqslant n_0, u_n \leqslant 0.$$

Alors les séries $\sum u_n$ et $\sum(-u_n)$ sont de même nature. En particulier, la série $\sum(-u_n)$ est à termes positifs et on peut utiliser les règles précédentes.

2.3 Séries à termes réels de signe quelconque

Dans le cas de séries numériques dont le terme général est de signe quelconque, on se pose deux questions :

1. La série $\sum |u_n|$ est-elle convergente ? La série $\sum u_n$ est-elle convergente ?
2. Quel est le lien entre la convergence en valeur absolue de la série et sa convergence ?

2.3.1 Séries absolument convergentes

Définition 2.3.1 *Une série $\sum u_n$ est dite absolument convergente lorsque la série à termes positifs $\sum |u_n|$ converge.*

Exemples : étudier la convergence absolue des séries numériques dont les termes généraux sont définis par $u_n = \dfrac{(-1)^n}{n^2}$ et $v_n = \dfrac{(-1)^n}{n}$.

On admet le théorème suivant faisant le lien entre la CVA et la CV d'une série numérique.

Théorème 2.3.1 *Soit $(u_n)_{n\in\mathbb{N}}$ une suite numérique à termes réels. Si la série $\sum u_n$ est absolument convergente alors la série $\sum u_n$ converge et :*

$$\left|\sum_{n=0}^{+\infty} u_n\right| \leq \sum_{n=0}^{+\infty} |u_n|.$$

Ainsi l'étude de la convergence absolue se ramène à l'étude d'une série à termes positifs et donc à appliquer les techniques du paragraphe précédent. Attention la réciproque est fausse !

Définition 2.3.2 *Une série $\sum u_n$ est dite semi-convergente (SCV) si elle n'est pas absolument convergente mais qu'elle converge.*

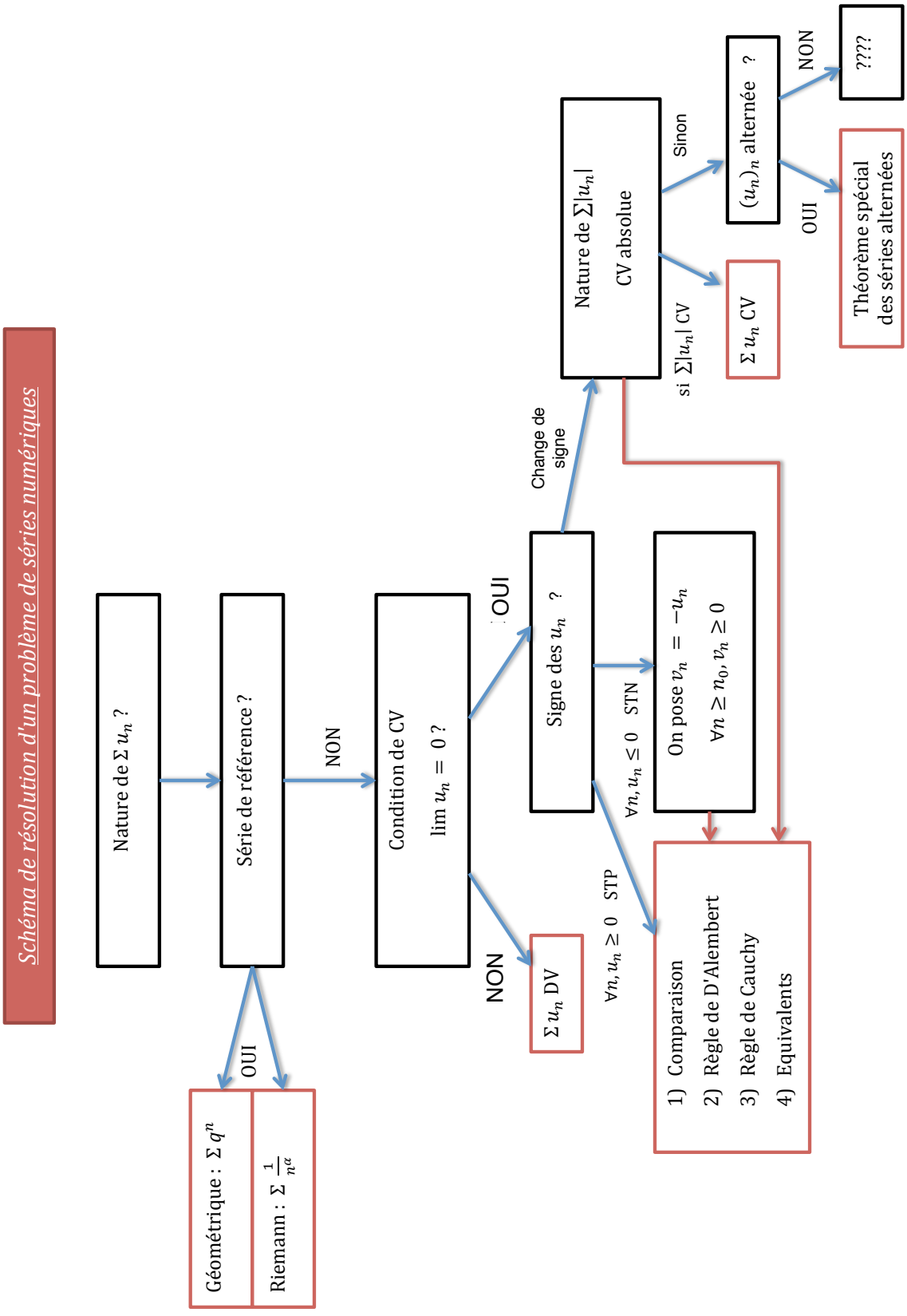
Exercice : Montrer que la série de terme général $u_n = \frac{(-1)^{n-1}}{n}$, $n \in \mathbb{N}^*$ est convergente mais pas absolument convergente.

2.3.2 Séries alternées

Définition 2.3.3 *La série* $\sum u_n$ *est dite alternée si* $u_n u_{n+1} \leqslant 0$*. Autrement dit* u_n *peut s'écrire sous la forme :* $u_n = (-1)^n |u_n|$ *pour tout* $n \in \mathbb{N}$ *ou bien sous la forme :* $u_n = (-1)^{n+1} |u_n|$ *pour tout* $n \in \mathbb{N}$

Théorème 2.3.2 (Théorème spécial des séries alternées) *Soit la série alternée* $\sum u_n$*. On pose :* $v_n = |u_n|$*. Si la suite* $(v_n)_{n\in\mathbb{N}}$ *est décroissante et tend vers* 0 *quand* n *tend vers* $+\infty$ *alors la série* $\sum u_n$ *converge.*

Exemple : étudier la nature des séries $\sum \dfrac{(-1)^n}{\sqrt{n}}$ et $\sum \dfrac{(-1)^n}{n},\ n \in \mathbb{N}$.


Schéma de résolution d'un problème de séries numériques
Nature de Σu_n ?
Série de référence ?
OUI
Géométrique : Σq^n
Riemann : $\Sigma \frac{1}{n^\alpha}$
NON
Condition de CV
$\lim u_n = 0$?
NON
Σu_n DV
OUI
Signe des u_n ?
$\forall n, u_n \geq 0$ STP
$\forall n, u_n \leq 0$ STN
On pose $v_n = -u_n$
$\forall n \geq n_0, v_n \geq 0$
1) Comparaison
2) Règle de D'Alembert
3) Règle de Cauchy
4) Equivalents
Change de signe
Nature de $\sum |u_n|$
CV absolue
si $\sum |u_n|$ CV
Σu_n CV
Sinon
$(u_n)_n$ alternée ?
OUI
Théorème spécial des séries alternées
NON
????

Chapitre 3

Séries de fonctions

Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions de Δ dans $\mathbb{R}$. On appelle série de fonctions et on note $\sum f_n$, la suite de fonctions $(\mathcal{S}_N)_{N\in\mathbb{N}}$ définie pour tout $N \in \mathbb{N}$ et pour tout $x \in \Delta$ par

$$\mathcal{S}_N(x) = \sum_{k=0}^{N} f_k(x).$$

On notera que :

- À $x \in \Delta$ fixé, $\sum f_n(x)$, $n \in \mathbb{N}$, est une série numérique de terme général $f_n(x)$.
- À $N \in \mathbb{N}$ fixé, $\mathcal{S}_N$ est la fonction qui à un $x \in \Delta$ associe la somme partielle des $N+1$ premiers termes des $f_n(x)$.

3.1 Différents types de convergence

3.1.1 Convergence simple et convergence absolue

Définition 3.1.1 *Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions de Δ dans $\mathbb{R}$. La série $\sum f_n$ converge simplement sur Δ vers la fonction S si*

$$(\forall t \in \Delta)(\lim_{N\to+\infty} \sum_{k=0}^{N} f_k(t) = S(t)).$$

Autrement dit, étudier la convergence simple d'une série de fonctions $\sum f_n$, c'est montrer que pour n'importe quel $x \in \Delta$, la série numérique $\sum f_n(x)$ converge en faisant appel aux techniques vues dans le chapitre des séries numériques.

On notera : $\sum f_n$ CVS vers S sur Δ où S désigne la somme de la série de fonctions. Ainsi on a :

$$\forall x \in \Delta,\ S(x) = \sum_{n=0}^{+\infty} f_n(x).$$

En particulier, la condition nécessaire de convergence s'applique aux séries de fonctions de la façon suivante :

Théorème 3.1.1 *Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions de Δ dans $\mathbb{R}$. Si la série de fonctions $\sum f_n$ converge simplement sur Δ alors nécessairement la suite de fonctions $(f_n)_{n\in\mathbb{N}}$ converge simplement sur Δ vers la fonction nulle i.e. :*

$$\forall x \in \Delta, \ \lim_{n\to+\infty} f_n(x) = 0.$$

Exemple : étudier la convergence simple de la série $\sum x^n, n \in \mathbb{N}$ sur $\Delta = [0, 1]$ et sur $\Delta = [0, 1[$.

Pour étudier la convergence simple d'une série de fonctions, il peut être utile de passer par la convergence absolue comme cela a été fait dans le chapitre sur les séries numériques :

Définition 3.1.2 *Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions de Δ dans $\mathbb{R}$. La série de fonctions $\sum f_n$ converge absolument sur Δ si la série de fonctions $\sum |f_n|$ converge simplement sur Δ.*

Proposition 3.1.1 *$\sum f_n$ CVA sur $\Delta \Rightarrow \sum f_n$ CVS sur Δ.*

3.1.2 Convergence uniforme

Définition 3.1.3 *Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions de Δ dans $\mathbb{R}$. On dit que la série $\sum f_n$ converge uniformément sur Δ vers S si*

$$\lim_{n\to+\infty} \sup_{t\in\Delta} \left| \sum_{k=0}^{n} f_k(t) - S(t) \right| = 0.$$

Exemple : étudier la convergence uniforme de la série $\sum x^n, n \in \mathbb{N}$ sur $\Delta = [0, a]$ avec $a \in [0, 1[$.

En 2ème année FAS Construction, nous ne travaillerons pas directement avec la convergence uniforme qui est une conséquence de la convergence normale définie ci-après.

3.1.3 Convergence normale

Définition 3.1.4 *Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions définies sur Δ. La série de fonctions $\sum f_n$ converge normalement s'il existe une série numérique $\sum u_n$ convergente telle que*

$$\forall n \in \mathbb{N}, \ \ \forall t \in \Delta \ \ |f_n(t)| \leq u_n.$$

Exemple : Étudier la convergence normale de la série $\sum \frac{\sin x\, e^{-nx}}{n^2}, n \in \mathbb{N}^*$ sur $\Delta = [0, \pi]$.

Théorème 3.1.2 (Comparaison des différents modes de convergence)

$$\sum_{n=0}^{+\infty} f_n \overset{CVN}{=} S \implies \sum_{n=0}^{+\infty} f_n \overset{CVU}{=} S \Longrightarrow \sum_{n=0}^{+\infty} f_n \overset{CVS}{=} S.$$
$$\sum_{n=0}^{+\infty} f_n \overset{CVN}{=} S \implies \sum_{n=0}^{+\infty} f_n \overset{CVA}{=} S \Longrightarrow \sum_{n=0}^{+\infty} f_n \overset{CVS}{=} S.$$

La CVN d'une série de fonctions a un double intérêt. Le premier est que l'étude de la CVN revient à l'étude d'une série numérique à termes positifs, donc à l'application des résultats du chapitre précédent. Le second, selon le théorème précédent, est que ce mode de convergence implique tous les autres modes de convergence. Cela permet plus facilement de montrer la CVU d'une série afin d'appliquer les divers théorèmes d'interversion et de propriétés de la somme d'une série de fonctions détaillées dans le paragraphe suivant.

Exercice 3.1.1 *Étudier la nature des séries de fonctions suivantes :*

1. $f_n(x) = x^n$, $n \in \mathbb{N}$, *sur* $\Delta = [0, 1]$, *puis* $\Delta = [0, 1[$, *puis sur* $\Delta = [0, a]$, $a > 0$.
2. $f_n(x) = \frac{1}{n^2+x^2}$, $n \in \mathbb{N}$, *sur* $\Delta = [0, 1]$.
3. $f_n(x) = \frac{\sin(x)}{n^2} e^{-nx}$, $n \in \mathbb{N}$, *sur* $\Delta = [0, 1]$, *puis* $\Delta = [0, \pi]$.

3.2 Convergence uniforme et régularité

Théorème 3.2.1 (Limite et continuité) *Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions continues sur Δ. Si la série $\sum f_n$ converge uniformément sur Δ vers S alors S est continue sur Δ.*

Théorème 3.2.2 (Limite et intégration) *Nous supposerons ici que $\Delta = [a, b]$. Si la série $\sum f_n$ converge uniformément vers S et si pour tout $n \in \mathbb{N}$, $f_n \in \mathcal{C}_\Delta$ alors*
$\forall [\alpha, \beta] \subset [a, b]$,

$$\sum_{n=0}^{+\infty} \left(\int_\alpha^\beta f_n(t)dt \right) = \int_\alpha^\beta \left(\sum_{n=0}^{+\infty} f_n(t) \right) dt = \int_\alpha^\beta S(t)dt.$$

Théorème 3.2.3 (Limite et dérivation) *Nous supposerons ici que $\Delta = [a, b]$. Soit $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions continues et dérivables sur Δ. Si $\exists t_0 \in \Delta$ tel que $\sum f_n(t_0)$ converge et si la série $\sum f'_n$ converge uniformément vers g alors la série $\sum f_n$ converge uniformément vers S avec $S' = g$:*

$$\forall t \in \Delta \quad \left(\sum_{n=0}^{+\infty} f_n(t)\right)' = \sum_{n=0}^{+\infty} f'_n(t).$$

Exemple : Étudier le domaine de définition, de continuité et de dérivabilité de la fonction f définie par

$$f(x) = \sum_{n=1}^{+\infty} \frac{\cos nx}{n^3}$$

Chapitre 4

Fonctions de plusieurs variables

4.1 Notion de fonction de plusieurs variables

Plusieurs formules en physique ou en ingénierie s'expriment à l'aide de deux ou de plusieurs variables réelles :

— aire A d'un triangle : $A = \frac{1}{2}bh$, b : base, h : hauteur,

— volume V d'une boite parallélépipédique : $V = L\ell h$, L : longueur, ℓ : largeur, h : hauteur,

— moyenne $\overline{x}$ de N variables x_1, x_2,..., x_N : $\overline{x} = \frac{1}{N}(x_1 + \cdots + x_N)$.

On dit que

— A est une fonction des 2 variables b et h,

— V est une fonction des 3 variables L, ℓ et h,

— $\overline{x}$ est une fonction des N variables x_1, x_2,..., x_N.

On est amené ainsi à considérer la notion suivante.

Définition 4.1.1 ***Une fonction de*** N ***variables réelles*** x_1, x_2,..., x_N, *qu'on note*

$$f : D \subset \mathbb{R}^N \to \mathbb{R}$$

est une règle qui permet d'associer à tout système de N *nombres réels contenu dans un sous-ensemble* D *de* $\mathbb{R}^N$ ***un nombre réel et un seul***

$$(x_1, \ldots, x_N) \mapsto f(x_1, \ldots, x_N).$$

Dans la suite, nous pourrons utiliser un seul symbole $x = (x_1, x_2)$ pour désigner deux variables ou $x = (x_1, \ldots, x_N)$ pour désigner N variables afin d'alléger les notations.

4.2 Continuité

On munit $\mathbb{R}^N$ d'une norme, on sait alors définir les limites, et on a :

Définition 4.2.1 *La fonction $f : D \subset \mathbb{R}^N \to \mathbb{R}$ tend vers L lorsque $x = (x_1, \ldots, x_N)$ tend vers $a = (a_1, \ldots, a_N)$ si la limite de $f(x)$ est égale à L lorsque $\|x - a\|$ tend vers 0, autrement dit lorsque les composantes x_i tendent vers a_i* ***de façon indépendante*** *:*

$$\forall \varepsilon > 0, \exists \eta > 0, \forall x \in D \; ; \|x - a\| \leq \eta \Rightarrow |f(x) - L| \leq \varepsilon$$

Proposition 4.2.1 *La limite de $f(x)$ lorsque x tend vers a, lorsqu'elle existe, est unique.*

Définition 4.2.2 *On dit que f est* ***continue au point*** *a si $\lim\limits_{x \to a} f(x) = f(a)$.*

Les théorèmes sur les limites d'une somme, d'un produit, d'un quotient, etc,... restent valables pour les fonctions de plusieurs variables. Ils permettent ainsi de s'assurer que des fonctions d'expressions compliquées sont continues à partir de la continuité de fonctions plus simples. De plus, les polynômes en plusieurs variables restent continus.

Exemple 4.2.1 *On a ainsi*

— *$f(x, y) = x^2 + y^2 + 1$ est continue en tout point (x, y) de $\mathbb{R}^2$, comme polynôme en les variables x, y.*

— *$f(x, y) = \sqrt{x + y}$ est continue en tout point de l'ensemble*

$$D := \left\{(x, y) \in \mathbb{R}^2 \; ; \; x + y \geq 0\right\}$$

qui est l'ensemble des points sur et au-dessus de la seconde bissectrice du plan cartésien, comme composé de fonctions continues.

— *$f(x, y, z) = xyz/(x^2 + y^2 + z^2)$ est continue en tout point de l'espace $\mathbb{R}^3$ où elle est définie, c'est-à-dire en tout point (x, y, z) différent de l'origine $(0, 0, 0)$.*

Remarque 4.2.1 *Il y a un point dans la définition précédente qui peut passer inaperçu de prime abord mais qui généralement* ***complique considérablement le passage de l'étude des fonctions d'une variable à celles à plusieurs variables*** *: les composantes x_i doivent tendre vers a_i* ***de façon indépendante****.*

Exemple 4.2.2 *On considère la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :*

$$f(x, y) = \begin{cases} \dfrac{xy}{x^2 + y^2} & \text{si } (x, y) \neq (0, 0), \\ 0 & \text{si } (x, y) = (0, 0). \end{cases}$$

Alors, en choisissant comment (x, y) tend vers 0, il est possible d'obtenir deux limites différentes.

— *si (x, y) tend vers $(0, 0)$ en* ***restant sur la parabole*** *$y = x^2$, alors :*

$f(x, y) = f(x, x^2) =$

— *si* (x, y) *tend vers* $(0, 0)$ *en* ***restant sur la droite*** *d'équation* $y = 2x$*, alors :*

$f(x, y) = f(x, 2x) =$

Ainsi, la fonction f *n'admet pas de limite en* $(0, 0)$ *(les limites obtenues selon des chemins différents ne coïncident pas), et ne peut donc pas être continue en ce point.*

4.3 Dérivées partielles et différentielle

4.3.1 Définitions

Les définitions suivantes s'étendent sans problèmes à des fonctions de N variables. Pour la clarté de l'exposé, nous les écrivons pour deux variables.

Définition 4.3.1 *Soit* $f : \mathbb{R}^2 \to \mathbb{R}$ *une fonction de 2 variables. Fixons* $y = y_0$*. Si la fonction d'une variable :*

$$\begin{array}{rccc} f(\cdot, y_0) : \mathbb{R} & \to & \mathbb{R} \\ x & \mapsto & f(x, y_0) \end{array}$$

est dérivable au point x_0 *on dit que* f *admet une dérivée partielle par rapport à la première variable au point* (x_0, y_0)*. De façon classique, cette dérivée partielle est notée*

$$\boxed{\frac{\partial f}{\partial x}(x_0, y_0) = \lim_{x \to x_0} \frac{f(x, y_0) - f(x_0, y_0)}{x - x_0}}$$

Dans certains cours, on utilise la notation $\partial_x f(x_0, y_0)$ *pour désigner* $\dfrac{\partial f}{\partial x}(x_0, y_0)$*.*

On définit de même lorsque cela est possible $\dfrac{\partial f}{\partial y}(x_0, y_0)$*.*

Les dérivées partielles se calculent par les règles usuelles de dérivation des fonctions à une variable.

Exemple 4.3.1 *Soit la fonction* f *définie par* $f(x, y) = \sin(x^2 + y^2)$*. On a*

$$\frac{\partial f}{\partial x}(x, y) =$$

$$\frac{\partial f}{\partial y}(x, y) =$$

Définition 4.3.2 *On dit que la fonction $f : \mathbb{R}^2 \to \mathbb{R}$ est différentiable (ou admet une différentielle) au point (x_0, y_0) si elle admet des dérivées partielles $\frac{\partial f}{\partial x}(x_0, y_0)$ et $\frac{\partial f}{\partial y}(x_0, y_0)$ et vérifie :*

$$f(x,y) = f(x_0,y_0) + \frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) \\ + \|(x,y)-(x_0,y_0)\|\, R(x,y,x_0,y_0).$$

avec $R(x, y, x_0, y_0)$ tendant vers 0 lorsque (x, y) tend vers (x_0, y_0).

Remarque 4.3.1 *Remarquez qu'on peut réécrire le terme de dérivées partielles à l'aide d'un produit de matrices :*

$$\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) = \begin{bmatrix} \frac{\partial f}{\partial x}(x_0,y_0) & \frac{\partial f}{\partial y}(x_0,y_0) \end{bmatrix} \begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix}$$

On appelle le vecteur ligne $\begin{bmatrix} \frac{\partial f}{\partial x}(x_0,y_0) & \frac{\partial f}{\partial y}(x_0,y_0) \end{bmatrix}$ la différentielle de f au point (x_0, y_0).

Certains ouvrages utilisent la notation $df_{(x_0,y_0)}$, ou encore $f'(x_0, y_0)$.
Enfin, notez qu'on peut également utiliser le produit scalaire usuel de $\mathbb{R}^2$:

$$\frac{\partial f}{\partial x}(x_0,y_0)(x-x_0) + \frac{\partial f}{\partial y}(x_0,y_0)(y-y_0) = \left\langle \begin{bmatrix} \frac{\partial f}{\partial x}(x_0,y_0) \\ \\ \frac{\partial f}{\partial y}(x_0,y_0) \end{bmatrix}, \begin{bmatrix} x-x_0 \\ y-y_0 \end{bmatrix} \right\rangle.$$

Dans ce cas, le vecteur colonne $\begin{bmatrix} \frac{\partial f}{\partial x}(x_0,y_0) \\ \\ \frac{\partial f}{\partial y}(x_0,y_0) \end{bmatrix}$ est appelé le **gradient** *de f en (x_0, y_0), et on le note souvent :*

$$\nabla f(x_0,y_0) = \begin{bmatrix} \frac{\partial f}{\partial x}(x_0,y_0) \\ \\ \frac{\partial f}{\partial y}(x_0,y_0) \end{bmatrix}$$

4.3.2 Théorèmes

On a immédiatement le résultat important suivant.

Théorème 4.3.1 *Si f est différentiable au point (x_0, y_0), elle est continue au point (x_0, y_0).*

Démonstration. Immédiate car

$$(x-x_0)\frac{\partial f}{\partial x}(x_0,y_0)+(y-y_0)\frac{\partial f}{\partial y}(x_0,y_0)+\|(x,y)-(x_0,y_0)\|\,R(x,y,x_0,y_0)$$

tend vers 0 lorsque (x,y) tend vers (x_0,y_0).

□

Notez que dans ce résultat, on a bien besoin de f différentiable et pas juste de l'existence de ses dérivées partielles ; ce qui est une différence majeure par rapport à l'analyse des fonctions d'une variable (pour lesquelles dérivable implique continue). On peut par ailleurs donner des contre-exemples pour lesquels les dérivées partielles existent, mais où la fonction n'est pas continue (et donc a fortiori, pas différentiable) :

Remarque 4.3.2 *On considère de nouveau*

$$f(x,y)=\begin{cases}\dfrac{xy}{x^2+y^2} & \textit{si } (x,y)\neq(0,0),\\ 0 & \textit{si } (x,y)=(0,0).\end{cases}$$

Écrire ∇f*.*

Alors, on a :

$$\frac{\partial f}{\partial x}(0,0)=\lim_{x\to 0}\frac{f(x,0)-f(0,0)}{x-0}=$$

$$\frac{\partial f}{\partial y}(0,0)=\lim_{y\to 0}\frac{f(0,y)-f(0,0)}{y-0}=$$

En d'autres termes, les dérivées partielles en $(0,0)$ *existent. Cependant, elle ne peut pas être différentiable en* $(0,0)$ *car on a vu plus haut qu'elle n'était pas continue en ce point.*

Ce théorème nous montre donc une différence importante entre les fonctions d'une variable pour lesquelles la différentiabilité se confond avec la dérivabilité et les fonctions de deux variables ou plus qui peuvent avoir des dérivées partielles et ne pas être différentiables. On a cependant le résultat suivant qu'on admettra.

Théorème 4.3.2 *Soit* $f : D \subset \mathbb{R}^2 \to \mathbb{R}$ *une fonction de deux variables. Si* $\dfrac{\partial f}{\partial x}(x,y)$ *et* $\dfrac{\partial f}{\partial y}(x,y)$ *existent au voisinage de* (x_0,y_0) *et sont continues au point* (x_0,y_0)*,* f *est différentiable au point* (x_0,y_0)*.*

4.4 Dérivées d'ordre supérieur

4.4.1 Définition

Si $f : \mathbb{R}^2 \to \mathbb{R}$ admet une dérivée partielle $\dfrac{\partial f}{\partial x}(x, y)$ en tout point $(x, y) \in D$, on obtient de cette façon une nouvelle fonction de deux variables :

$$\begin{array}{rcl} g : \mathbb{R}^2 & \to & \mathbb{R} \\ (x, y) & \mapsto & \dfrac{\partial f}{\partial x}(x, y), \end{array}$$

et on peut maintenant étudier la dérivabilité de g. Lorsque g est dérivable, ses dérivées partielles sont les dérivées d'ordre 2 de f :

$$\frac{\partial^2 f}{\partial x^2}(x, y) = \frac{\partial}{\partial x}\frac{\partial f}{\partial x}(x, y) = \frac{\partial g}{\partial x}(x, y).$$

De même par rapport à la variable y :

$$\frac{\partial^2 f}{\partial y \partial x}(x, y) = \frac{\partial}{\partial y}\frac{\partial f}{\partial x}(x, y) = \frac{\partial g}{\partial y}(x, y).$$

Maintenant, les mêmes considérations peuvent être faites pour

$$\begin{array}{rcl} h : \mathbb{R}^2 & \to & \mathbb{R} \\ (x, y) & \mapsto & \dfrac{\partial f}{\partial y}(x, y), \end{array}$$

ce qui donne les dérivées :

$$\frac{\partial^2 f}{\partial x \partial y}(x, y) = \frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x, y) = \frac{\partial h}{\partial x}(x, y).$$

De même par rapport à la variable y :

$$\frac{\partial^2 f}{\partial y^2}(x, y) = \frac{\partial}{\partial y}\frac{\partial f}{\partial y}(x, y) = \frac{\partial h}{\partial y}(x, y).$$

Ces dérivées d'ordre 2 sont notées

$$\frac{\partial^2 f}{\partial x^2}, \quad \frac{\partial^2 f}{\partial x \partial y}, \quad \frac{\partial^2 f}{\partial y \partial x}, \quad \frac{\partial^2 f}{\partial y^2}.$$

Exemple 4.4.1 *Considérons la fonction :*

$$\begin{array}{rcl} f : \mathbb{R}^2 & \to & \mathbb{R} \\ (x, y) & \mapsto & e^{y^2} \sin(x). \end{array}$$

Déterminer les dérivées partielles d'ordre 1 et 2 :

4.4.2 Théorème de Schwarz

On remarque que le résultat de la dérivation est indépendant de l'ordre des dérivées partielles. Cette propriété est en fait générale. Elle est établie dans le théorème suivant dont on admettra le résultat.

Théorème 4.4.1 *(Théorème de Schwarz). Soit f une fonction de deux variables. Si $\frac{\partial^2 f}{\partial x \partial y}$ et $\frac{\partial^2 f}{\partial y \partial x}$ existent et sont continues, alors*

$$\boxed{\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.}$$

Exemple 4.4.2 *Dans le Théorème de Schwarz, il est crucial que les dérivées secondes soient continues. Considérons le contre-exemple suivant :*

$$\begin{array}{rcl} f : \mathbb{R}^2 & \to & \mathbb{R} \\ (x,y) & \mapsto & \begin{cases} \dfrac{xy(x^2-y^2)}{x^2+y^2} & \text{si } (x,y) \neq (0,0), \\ 0 & \text{si } (x,y) = (0,0). \end{cases} \end{array}$$

Alors, on a $\forall y \in \mathbb{R}$, on calcule la dérivée partielle de f par rapport à x en 0 :

$$\frac{\partial f}{\partial x}(0,y) = \lim_{x \to 0} \frac{f(x,y) - f(0,y)}{x - 0} =$$

et donc on trouve :

$$\frac{\partial}{\partial y}\frac{\partial f}{\partial x}(0,y) = -1, \text{ donc en particulier : } \frac{\partial^2 f}{\partial y \partial x}(0,0) = -1.$$

De même, on calcule maintenant pour tous $x \in \mathbb{R}$, la dérivée partielle de f par rapport à y en 0 :

$$\frac{\partial f}{\partial y}(x,0) = \lim_{y\to 0} \frac{f(x,y) - f(x,0)}{y-0} =$$

Il vient alors :

$$\frac{\partial}{\partial x}\frac{\partial f}{\partial y}(x,0) = \quad , \text{ donc en particulier : } \frac{\partial^2 f}{\partial x \partial y}(0,0) = \quad .$$

Donc ici, on a :

$$\frac{\partial^2 f}{\partial y \partial x}(0,0) = -1 \neq \quad = \frac{\partial^2 f}{\partial x \partial y}(0,0).$$

Notons $a = (a_1, \ldots, a_N)$, la matrice des dérivées secondes est appelée **matrice hessienne** de f au point a. On note :

$$H_f(a) = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f(a) & \cdots & \frac{\partial^2}{\partial x_1 \partial x_N} f(a) \\ \vdots & & \vdots \\ \frac{\partial^2}{\partial x_N \partial x_1} f(a) & \cdots & \frac{\partial^2}{\partial x_N^2} f(a) \end{bmatrix}$$

Exemple : Déterminer le gradient et la matrice hessienne de

$$\begin{array}{rcl} f : \mathbb{R}^2 & \to & \mathbb{R} \\ (x,y) & \mapsto & x^3 + y^3 \end{array}$$

4.5 Matrice jacobienne

On considère dans ce paragraphe, non plus des fonctions définies de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}$ mais des fonctions définies de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^p$.

Définition 4.5.1 *Soit ψ une fonction définie dans U un ouvert de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^p$ qui à $x = (x_1, \cdots, x_n)$ associe $(\psi_1(x_1, \cdots, x_n), \cdots, \psi_p(x_1, \cdots, x_n))$. On dit que ψ est différentiable si les fonctions $\psi_1, \cdots, \psi_p$ le sont. Dans ce cas, on lui associe une matrice à p lignes et n colonnes J_ψ, appelée matrice jacobienne de ψ, définie par :*

$$J_\psi(x) = \left[\frac{\partial \psi_i}{\partial x_j}(x)\right]_{\substack{1 \leq j \leq n \\ 1 \leq i \leq p}} \quad \textit{pour tout } x \in U.$$

Par exemple, si on considère

$$\begin{array}{rcl} \psi : \mathbb{R}^2 & \to & \mathbb{R}^2 \\ \begin{bmatrix} x \\ y \end{bmatrix} & \mapsto & \begin{bmatrix} \psi_1(x,y) \\ \psi_2(x,y) \end{bmatrix} \end{array}$$

la matrice jacobienne de ψ est donnée par :

$$J_\psi\left(\begin{bmatrix} x \\ y \end{bmatrix}\right) = \begin{bmatrix} \dfrac{\partial \psi_1}{\partial x}(x,y) & \dfrac{\partial \psi_1}{\partial y}(x,y) \\ \dfrac{\partial \psi_2}{\partial x}(x,y) & \dfrac{\partial \psi_2}{\partial y}(x,y) \end{bmatrix}.$$

Exemple : Déterminer la matrice jacobienne de

$$\begin{array}{rcl} \psi : \mathbb{R}^2 & \to & \mathbb{R}^2 \\ \begin{bmatrix} x \\ y \end{bmatrix} & \mapsto & \begin{bmatrix} x^2 + y \\ y^3 + xy \end{bmatrix} \end{array}$$

4.6 Opérateurs différentiels

Nous présentons rapidement dans ce paragraphe les opérateurs différentiels que vous rencontrerez dans la suite de votre scolarité.

Pour simplifier les notations, on peut introduire « Nabla » : $\nabla = \begin{bmatrix} \frac{\partial}{\partial x_1} \\ \frac{\partial}{\partial x_2} \\ \frac{\partial}{\partial x_3} \end{bmatrix}$.

Le **gradient** a déjà été défini précédemment :

Soit la fonction $f : \mathbb{R}^3 \to \mathbb{R}$ suffisamment régulière.

$\nabla f = \begin{bmatrix} \frac{\partial f}{\partial x_1} \\ \frac{\partial f}{\partial x_2} \\ \frac{\partial f}{\partial x_3} \end{bmatrix}$.

Laplacien

Soit la fonction $f : \mathbb{R}^3 \to \mathbb{R}$ suffisamment régulière.

$\Delta f = \sum_{k=1}^{3} \frac{\partial^2 f}{\partial x_k^2} = \frac{\partial^2 f}{\partial x_1^2} + \frac{\partial^2 f}{\partial x_2^2} + \frac{\partial^2 f}{\partial x_3^2}$.

Divergence

Soit la fonction $f : \mathbb{R}^3 \to \mathbb{R}^3$ suffisamment régulière.

$div(f) = \sum_{k=1}^{3} \frac{\partial f_k}{\partial x_k} = \frac{\partial f_1}{\partial x_1} + \frac{\partial f_2}{\partial x_2} + \frac{\partial f_3}{\partial x_3}$

On peut noter que $div(f) = \nabla^T . f$ le produit scalaire entre l'opérateur ∇ et le vecteur $f = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}$.

Rotationnel

Soit la fonction $f : \mathbb{R}^3 \to \mathbb{R}^3$ suffisamment régulière.

$$rot(f) = \begin{bmatrix} \dfrac{\partial f_3}{\partial x_2} - \dfrac{\partial f_2}{\partial x_3} \\ \dfrac{\partial f_1}{\partial x_3} - \dfrac{\partial f_3}{\partial x_1} \\ \dfrac{\partial f_2}{\partial x_1} - \dfrac{\partial f_1}{\partial x_2} \end{bmatrix}.$$

On peut noter que $rot(f) = \nabla \wedge f$ où $f = \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix}$.

Exemple 4.6.1 *Soit la fonction $f : \mathbb{R}^3 \to \mathbb{R}$ suffisamment régulière. Montrer que $div(\nabla f) = \Delta f$.*

4.7 Règle de la chaîne

Nous allons introduire une formule importante connue sous le nom de **règle de la chaîne** qui permet de dériver la fonction composée $g(t) = f(x(t), y(t))$ obtenue lorsque les deux variables d'une fonction à deux variables dépendent elles-mêmes d'une même troisième variable.

Théorème 4.7.1 *Si $x = u(t)$ et $y = v(t)$ sont deux fonctions dérivables de la variable réelle t et si $f : \mathbb{R}^2 \to \mathbb{R}$ est une fonction différentiable au point $(u(t_0), v(t_0))$, alors la fonction :*

$$\begin{array}{rcl} \varphi : \mathbb{R} & \to & \mathbb{R} \\ t & \mapsto & f\big(u(t), v(t)\big) \end{array}$$

est dérivable au point t_0 et

$$\boxed{\varphi'(t_0) = \frac{\partial f}{\partial x}(u(t_0), v(t_0))u'(t_0) + \frac{\partial f}{\partial y}(u(t_0), v(t_0))v'(t_0).}$$

Démonstration. La différentiabilité de f au point $(u(t_0), v(t_0))$ donne

$$\begin{aligned} \varphi(t) = f(u(t), v(t)) = f(u(t_0), v(t_0))+ \\ (u(t) - u(t_0))\frac{\partial f}{\partial x}(u(t_0), v(t_0)) + (v(t) - v(t_0))\frac{\partial f}{\partial y}(u(t_0), v(t_0))+ \\ \|(u(t) - u(t_0), v(t) - v(t_0))\| \, R(u(t), v(t), u(t_0), v(t_0)). \end{aligned}$$

avec

$$\lim_{t \to t_0} R(u(t), v(t), u(t_0), v(t_0)) = 0$$

car $\lim_{t \to t_0} u(t) = u(t_0)$ et $\lim_{t \to t_0} v(t) = v(t_0)$. On écrit alors que u et v sont dérivables au point t_0

$$u(t) - u(t_0) = (t - t_0)\left(u'(t_0) + \varepsilon_1(t)\right), \quad v(t) - v(t_0) = (t - t_0)\left(v'(t_0) + \varepsilon_2(t)\right)$$

avec $\lim_{t \to t_0} \varepsilon_i(t) = 0$ $(i = 1, 2)$. La formule ci-dessus donnant $\varphi(t)$ se réécrit

$$\begin{aligned} \varphi(t) = \varphi(t_0)+ \\ (t - t_0)\left(\frac{\partial f}{\partial x}(u(t_0), v(t_0))u'(t_0) + \frac{\partial f}{\partial y}(u(t_0), v(t_0))v'(t_0)\right) + \\ (t - t_0)\,\varepsilon(t) \end{aligned}$$

avec $\lim_{t \to t_0} \varepsilon(t) = 0$. Ceci établit le théorème. □

Remarque 4.7.1 *Si $f : \mathbb{R}^2 \to \mathbb{R}$ admet des dérivées partielles continues sur tout D (on dira alors que f est de classe $\mathcal{C}^1$ sur D et on notera ceci par $f \in \mathcal{C}^1(D)$), et si $x(t)$ et $y(t)$ sont deux fonctions à valeurs dans D continument dérivables, alors la fonction $z(t) = f(x(t), y(t))$ est dérivable en chaque point t. La règle de la chaîne s'écrit alors de façon mnémotechnique en utilisant la notation classique pour les dérivées partielles*

$$\frac{dz}{dt} = \frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt}.$$

La formule de la chaîne se généralise directement aux fonctions de trois variables ou plus

$$z = f(x_1(t), \ldots, x_N(t)),$$
$$\frac{dz}{dt} = \frac{\partial f}{\partial x_1}\frac{dx_1}{dt} + \cdots + \frac{\partial f}{\partial x_N}\frac{dx_N}{dt}.$$

Exemple 4.7.1 *On considère $f : \mathbb{R}^2 \to \mathbb{R}$ définie par :*

$$f(x, y) = x^2 + y^2.$$

On prend $x(t) = \cos(t)$ et $y(t) = \sin(t)$, et on pose

$$z(t) = f(x(t), y(t)).$$

Calculer $z'(t)$.

Ainsi, la fonction $z(t)$ est constante. Comme on a en particulier que $z(0) = 1$, on retrouve l'identité trigonométrique :

$$\forall t \in \mathbb{R},\ \cos(t)^2 + \sin(t)^2 = 1.$$

Considérons maintenant des fonctions définies de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^p$. On rappelle la définition suivante :

Définition 4.7.1 *Soit ψ une fonction définie dans U un ouvert de $\mathbb{R}^n$ à valeurs dans $\mathbb{R}^p$ qui à $x = (x_1, \cdots, x_n)$ associe $(\psi_1(x_1, \cdots, x_n), \cdots, \psi_p(x_1, \cdots, x_n))$. On dit que ψ est différentiable si les fonctions $\psi_1, \cdots, \psi_p$ le sont. Dans ce cas, on lui associe une matrice à p lignes et n colonnes J_ψ, appelée matrice jacobienne de ψ, définie par :*

$$J_\psi(x) = \left[\frac{\partial \psi_i}{\partial x_j}(x)\right]_{\substack{1 \le j \le n \\ 1 \le i \le p}} \quad \textit{pour tout } x \in U.$$

Proposition 4.7.1 *Soient f une fonction de $U \subset \mathbb{R}^n$ dans $W \subset \mathbb{R}^p$ et g une fonction de $W \subset \mathbb{R}^p$ dans $\mathbb{R}^q$ toutes deux différentiables. Dans ce cas, la matrice jacobienne de la composée $g \circ f$ s'obtient pour tout a dans $\mathbb{R}^n$ en considérant* ***le produit de matrices*** *:*

$$J_{(g\circ f)}(a) = J_g(f(a)) \times J_f(a).$$

Démonstration. Posons $F = g \circ f$ qui est ainsi la fonction de $U \subset \mathbb{R}^n$ dans $\mathbb{R}^q$ définie par

$$\begin{aligned} F(x) &= (F_1(x), \cdots, F_q(x)) \\ &= g(f(x)) = (g_1(f(x)), \cdots, g_q(f(x)) \end{aligned}$$

Sachant que $x = (x_1, \cdots, x_n)$ et en posant $y = f(x)$ ou encore suivant les composantes

$$\begin{aligned} y &= (y_1, \cdots, y_p) \\ &= (f_1(x_1, \cdots, x_n), \cdots, f_p(x_1, \cdots, x_n)), \end{aligned}$$

en utilisant la règle de la chaîne précédente :

$$\frac{\partial F_i}{\partial x_j}(a) = \sum_{k=1}^{p} \frac{\partial g_i}{\partial y_k}(f(a)) \frac{\partial f_k}{\partial x_j}(a),$$

avec $j = 1, \cdots, n$ et $i = 1, \cdots, q$, la proposition est une conséquence directe de la formule donnant le produit des matrices. □

Exemple 4.7.2 *Soient f une fonction de deux variables x et y dans $\mathbb{R}$. Appliquons cette formule au passage des coordonnées cartésiennes aux coordonnées polaires. La relation de changement de variables est définie de façon plus aisée en exprimant les coordonnées cartésiennes à l'aide des coordonnées polaires*

$$\begin{aligned} x &= r\cos(\theta) \\ y &= r\sin(\theta) \end{aligned}$$

avec $r > 0$ et (par exemple) $0 \leq \theta < 2\pi$. La relation suivante permet d'exprimer la fonction f des coordonnées cartésiennes (x, y) à partir de la fonction g des coordonnées polaires (r, θ)

$$f(x, y) = g(r, \theta).$$

On veut calculer les dérivées partielles $\dfrac{\partial f}{\partial x}(x, y)$ et $\dfrac{\partial f}{\partial y}(x, y)$ à partir des dérivées partielles $\dfrac{\partial g}{\partial r}(r, \theta)$ et $\dfrac{\partial g}{\partial \theta}(r, \theta)$. La règle de la chaîne donne

$$\frac{\partial f}{\partial x}(x, y) =$$

$$\frac{\partial f}{\partial y}(x, y) =$$

Il est facile d'exprimer $r = \sqrt{x^2 + y^2}$ *à partir de* x *et* y*. On a alors directement*

$$\frac{\partial r}{\partial x}(x, y) =$$

$$\frac{\partial r}{\partial y}(x, y) =$$

Ce n'est pas le cas pour l'argument θ *qui ne peut être donné par une formule à l'aide de* x *et* y *que si par exemple* $0 < \theta < \pi$ *:* $\theta = \mathrm{Arccos}(x/\sqrt{x^2 + y^2})$*. Il n'y a aucune difficulté en revanche à calculer*

$$\frac{\partial x}{\partial r}(r, \theta) =$$

$$\frac{\partial x}{\partial \theta}(r, \theta) =$$

$$\frac{\partial y}{\partial r}(r, \theta) =$$

$$\frac{\partial y}{\partial \theta}(r, \theta) =$$

La formule de la chaîne donne alors :

$$\frac{\partial g}{\partial r}(r, \theta) =$$
$$\frac{\partial g}{\partial \theta}(r, \theta) =$$

qu'on peut écrire sous forme matricielle

$$\begin{bmatrix} \dfrac{\partial g}{\partial r}(r, \theta) \\ \dfrac{1}{r}\dfrac{\partial g}{\partial \theta}(r, \theta) \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} \dfrac{\partial f}{\partial x}(x, y) \\ \dfrac{\partial f}{\partial y}(x, y) \end{bmatrix}$$

Vérifier que la matrice $P = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$ *est une matrice inversible et calculer son inverse :*

On obtient alors $P^{-1} =$ *, et on en déduit donc*

$$\frac{\partial f}{\partial x}(x,y) =$$

$$\frac{\partial f}{\partial y}(x,y) =$$

Remarque 4.7.2 *Le théorème de Schwarz s'étend de façon évidente aux fonctions de plus de deux variables et on a, si* $z = f(x_1, \dots x_N)$*,*

$$\boxed{\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}, \quad 1 \leq i, j \leq N.}$$

4.8 Étude des extremums

Le but est d'étudier les extremums d'une fonction de $\mathbb{R}^N$ dans $\mathbb{R}$, s'ils existent. Pour une fonction $f : \mathbb{R} \to \mathbb{R}$, on sait qu'on peut localiser les extremums en regardant les points où la dérivée s'annule. Cette notion s'étend au cas multidimensionnel.

Définition 4.8.1 *Un point* $a \in \mathbb{R}^N$ *est un* **point critique** *pour* f *si pour tous* $i = 1, \dots, N$*,*

$$\frac{\partial}{\partial x_i} f(a) = 0.$$

C'est un point où la différentielle s'annule.

Pour une fonction de la variable réelle, une fois les zéros de la dérivée localisés, pour savoir si le point est un minimum ou un maximum, on peut regarder le signe de la dérivée seconde. Ici encore, cette notion s'étend au cas multidimensionnel. Il faut toutefois être vigilant sur la notion de « signe » pour la dérivée seconde : dans le cas multidimensionnel, la dérivée seconde est une matrice (matrice hessienne). Commençons par donner la définition d'un maximum ou minimum local :

Définition 4.8.2 *On dit que* a *est un minimum (resp. maximum) local pour* f *s'il existe un voisinage* V *de* a *tel que* $\forall x \in V$*,*

$$f(a) \leq f(x) \text{ (resp. } f(a) \geq f(x)\text{)}.$$

Lorsque f est de classe $\mathcal{C}^2$, le théorème de Schwarz s'applique et $H_f(a)$ est une matrice symétrique (elle est égale à sa transposée).

On a le théorème suivant :

Théorème 4.8.1 *Soit $f : D \subset \mathbb{R}^N \to \mathbb{R}$ de classe $\mathcal{C}^3$, et $a \in D$ un **point critique** pour f. Soit $H_f(a)$ la matrice hessienne de f au point a.*

Si $H_f(a)$ est diagonalisable et que ses valeurs propres sont strictement positives (resp. strictement négatives), alors f admet un minimum (resp. maximum) local en a.

On admet ce résultat, il se montre avec la formule de Taylor et utilise le cours d'algèbre bilinéaire (S4).

Exercice 4.8.1 *On considère pour tout $(x, y) \in \mathbb{R}^2$ la fonction :*

$$g(x, y) = \exp\left(-\frac{1}{2}x^2 - \frac{1}{4}(y-1)^2\right)$$

1. *Déterminer le seul point critique de g.*
2. *Montrer qu'il s'agit d'un maximum local.*

Chapitre 5

Erreurs numériques

5.1 Erreurs

5.1.1 Contexte

Les méthodes de l'analyse numérique visent à traiter numériquement un certain nombre de problèmes mathématiques, provenant pour la plupart de la modélisation de phénomènes ou comportements issus du monde réel, au sein de disciplines aussi variées que la physique, la biologie, ou encore l'économie. Ce traitement se fait à l'aide d'un ou plusieurs ordinateurs. Évidemment, tous les nombres ne peuvent pas être représentés en machine puisqu'en général ils possèdent une infinité (ou un trop grand nombre) de décimales, ce qui induit des erreurs numériques.

5.1.2 Erreur absolue et erreur relative

Notons plus précisément x la solution théorique d'un problème donné et $\tilde{x}$ la solution approchée lorsqu'on résout le problème en question numériquement. Il existe deux manières de quantifier l'erreur commise entre la valeur théorique et la valeur numérique :

- l'**erreur absolue** définie par :

$$\boxed{\Delta x = |x - \tilde{x}|}\,,$$

- l'**erreur relative** telle que :

$$\boxed{\delta x = \frac{|x - \tilde{x}|}{|x|}}\,.$$

Il est à noter que l'erreur absolue possède une unité physique (mètres, grammes...) qui est la même que celle de x et $\tilde{x}$, tandis que l'erreur relative est sans unité. Aussi, si l'erreur relative est de 10^{-8}, on peut penser que les 8 premiers chiffres du nombre $\tilde{x}$ seront corrects ; en revanche, on ne peut rien dire de tel au vu d'une erreur absolue. En

pratique, on préférera ainsi utiliser l'erreur relative pour estimer correctement la qualité de l'approximation numérique.

Exemple 5.1.1 *Calculer les erreurs absolues et relatives entre* $x = 10^{-8}$ *et* $\tilde{x} = 2.10^{-8}$*.*

5.2 Représentation des nombres réels

Cette section introduit la manière dont les nombres sont représentés et stockés en machine, autrement dit dans la mémoire d'un ordinateur. Pour commencer, on rappelle la façon traditionnelle d'écrire les nombres avant de s'intéresser, à proprement parler, à leur représentation (et leur approximation) en machine.

5.2.1 Représentation traditionnelle des nombres réels en base 10

Nombres décimaux

Définition 5.2.1 *Soit* x *un nombre* **décimal***, il existe deux entiers* N *et* P*, et* $N + P + 1$ *entiers* $\alpha_0, \alpha_1, \ldots, \alpha_{N+P}$ *compris entre* 0 *et* 9*, tels que :*

$$x = \sum_{i=0}^{N+P} \alpha_i 10^{N-i} \, .$$

Le nombre x *s'écrit alors en base* 10 *:*

$$x = \alpha_0 \, \alpha_1 \, \cdots \, \alpha_N \, , \, \alpha_{N+1} \, \cdots \, \alpha_{N+P} \, .$$

Remarque 5.2.1 N *correspond à l'unique entier tel que* $10^N \leqslant x < 10^{N+1}$*.*

Exemple 5.2.1

$305 =$

$345,87 =$

$23704,135 =$

Remarque 5.2.2 *La langue française impose l'utilisation de la virgule comme séparateur pour les nombres décimaux, mais dans les pays anglo-saxons c'est le point qui est utilisé. Ainsi,* $345,87$ *s'écrira* 345.87*. En Analyse Numérique, on utilise principalement la notation anglo-saxonne avec un point car les normes des architectures informatiques sont pour la plupart écrites en anglais. À partir de maintenant, nous utiliserons donc le point comme séparateur pour les nombres décimaux.*

Nombres réels

Définition 5.2.2 *Soit $x \in \mathbb{R}$, il existe un entier N et une suite $(\alpha_i)_{i\in\mathbb{N}}$ d'entiers compris entre 0 et 9 tels que*

$$x = \sum_{i=0}^{+\infty} \alpha_i 10^{N-i} .$$

Le nombre x s'écrit alors en base 10 :

$$x = \alpha_0\, \alpha_1 \cdots \alpha_N \,.\, \alpha_{N+1}\, \alpha_{N+2} \cdots .$$

Remarque 5.2.3 *La définition correspond à celle donnée pour les nombres décimaux avec cette fois-ci $p = +\infty$.*

Proposition 5.2.1 *Notons deux cas particuliers :*

- *x est décimal si et seulement s'il existe un entier p tel que :*

$$\forall i \in \mathbb{N},\ \alpha_{N+p+i} = 0.$$

- *x est rationnel si et seulement si la suite $(\alpha_i)_{i\in\mathbb{N}}$ est périodique à partir d'un certain rang.*

Exemple 5.2.2 $x = 28.73041504150415... = 28.73\overline{0415}$ *est rationnel, mais* $\pi = 3.141592653589...$ *ne l'est pas.*

5.2.2 Représentation des nombres réels en base b

La représentation en base b généralise la représentation traditionnelle utilisant la base 10. Le principe est simplement de remplacer 10 par l'entier positif b :

Définition 5.2.3 *Soit $x \in \mathbb{R}$, il existe $N \in \mathbb{N}$ et une suite $(\alpha_i)_{i\in\mathbb{N}}$ d'entiers compris entre 0 et $b-1$ tels que :*

$$x = \sum_{i=0}^{+\infty} \alpha_i b^{N-i}.$$

Le nombre x s'écrit alors en base b :

$$x = \alpha_0\, \alpha_1 \cdots \alpha_N \,.\, \alpha_{N+1}\, \alpha_{N+2} \cdots .$$

Remarque 5.2.4 *En dehors de la base $b = 10$, les bases les plus fréquemment utilisées sont la base $b = 2$ (cf paragraphe suivant) et la base $b = 16$ (représentation hexadécimale des couleurs par exemple).*

5.2.3 Représentation des nombres réels en base 2

Pour le cas particulier $b = 2$, la représentation n'utilise que les deux chiffres 0 et 1 pour les entiers $(\alpha_i)_{i\in\mathbb{N}}$.

Définition 5.2.4 *Soit $x \in \mathbb{R}$, il existe $N \in \mathbb{N}$ et une suite $(\alpha_i)_{i\in\mathbb{N}}$ d'entiers égaux à 0 ou 1 tels que :*

$$x = \sum_{i=0}^{+\infty} \alpha_i 2^{N-i}.$$

Le nombre x s'écrit alors en base 2 :

$$x = \alpha_0\, \alpha_1 \cdots \alpha_N \,.\, \alpha_{N+1}\, \alpha_{N+2} \cdots ,$$

où chaque α_i est égal à 0 ou 1.

Elle est par conséquent intéressante pour l'informatique, car pour stocker ou transmettre des nombres, il suffit de disposer d'un mécanisme physique à deux états. Les tables d'opérations sont simples, et peuvent être simulées par des circuits électroniques simples. En pratique, c'est donc à partir de la représentation en base 2 que les nombres seront stockés en machine.

Exemple 5.2.3 *Écrire les nombres suivants en base 2.*

- $10_{10} =$
- $71_{10} =$
- $10.25_{10} =$

Remarque 5.2.5 *Le nombre écrit 1010 désigne à la fois le nombre en base 10 "mille dix" et le nombre en base 2 "dix". Pour lever l'ambiguité, on écrira plutôt 1010_{10} pour désigner le nombre "mille dix" écrit en base 10 et 1010_2 pour désigner le nombre "dix" écrit en base 2. Par exemple,*

$$10_{10} = 1010_2.$$

Ci-après la table des puissances de 2, très utile pour l'écriture d'un nombre en base 2.

Les méthodes pratiques pour écrire les réels en base b (et plus spécifiquement en base 2) sont abordées dans une fiche annexe disponible sur la page Moodle du cours.

2^{10}	1024	2^{2}	4
2^{9}	512	2^{1}	2
2^{8}	256	2^{0}	1
2^{7}	128	2^{-1}	0.5
2^{6}	64	2^{-2}	0.25
2^{5}	32	2^{-3}	0.125
2^{4}	16	2^{-4}	0.0625
2^{3}	8	2^{-5}	0.03125

TABLE 5.1 – Puissances de 2.

5.2.4 Représentation des nombres réels en machine

Le principal mode de représentation des nombres réels en machine est la représentation en virgule flottante normalisée. Cette représentation est utilisée aussi bien dans les calculatrices de poche que dans les stations de travail et les supercalculateurs. Plus précisément, $\forall x \in \mathbb{R}$, on associe sa représentation en machine notée $\mathrm{fl}(x)$ par l'expression suivante :

$$\boxed{\mathrm{fl}(x) \;=\; \pm\, m \;\cdot\; b^{e}}\,,$$

où :

- $\pm$ caractérise le **signe**,
- b est la **base** de représentation, en général $b = 2$,
- m s'appelle la **mantisse** : elle est normalisée, c'est-à-dire que :

  $\boxed{1 \leq m < b}$ et elle s'écrit avec p chiffres dans la base b,

- e désigne l'**exposant** ; il est compris entre les deux valeurs $e_{\min}$ et $e_{\max}$.

Avec l'expression de $\mathrm{fl}(x)$ ci-dessus, on voit bien qu'il est impossible de représenter tous les nombres en machine. En effet, **la mantisse ne peut avoir plus de p chiffres en base b**, p étant un entier fixé par le format utilisé. De plus, l'exposant e est compris entre deux valeurs minimum et maximum. Du fait de la mémoire finie des machines, la plupart des nombres réels ne peuvent être représentés qu'approximativement : cette représentation passe par le stockage des écritures en base 2 (ou b plus généralement) de sa mantisse m et de son exposant e. En ce qui concerne les valeurs de p et de e, il existe principalement deux formats de représentation, correspondant chacun à une précision spécifique : la **simple précision** et la **double précision**.

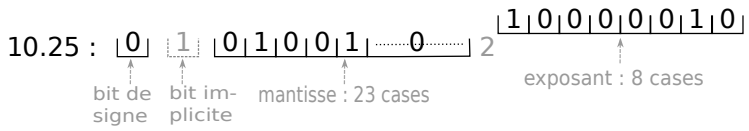
précision	taille	mantisse+signe	exposant	$e_{\min}$	$e_{\max}$
simple	32 bits	23+1 bits	8 bits	-126	127
double	64 bits	52+1 bits	11 bits	-1022	+1023

TABLE 5.2 – Tailles de stockage en simple et double précision

La base ici étant 2, le premier bit de la mantisse est implicite [1], c'est-à-dire qu'on ne le stockera pas. L'ensemble des réels représentables de manière exacte sera donc plus grand en double précision qu'en simple précision.

Attention, il n'y a pas de bit de signe pour l'exposant (qui peut pourtant être négatif) : on **rajoutera donc systématiquement $e_{\max}$ à l'exposant e** avant de le stocker. Par exemple, si l'on est en simple précision et que $e = 1$, alors on stockera l'écriture en base 2 de $e + e_{\max} = 128$ en machine.

Exemple 5.2.4 *Voici les 32 bits de la représentation machine binaire de 10.25 en simple précision. On rappelle que 10.25 s'écrit dans la base 2 tel que 1010.01 ou encore 1.01001×2^3. Son bit de signe est 0 car le nombre est positif. Ensuite, il y a le bit 1 implicite puis la mantisse. Enfin, comme l'exposant est positif, il faut lui ajouter $e_{\max} = 127$ d'où on doit coder 130 en binaire dans l'exposant.*



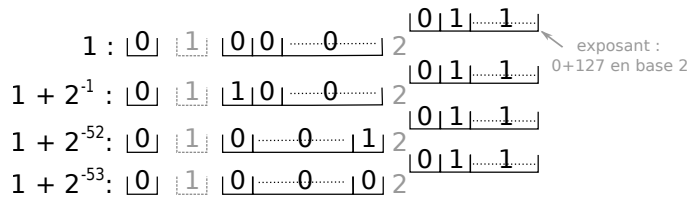
5.2.5 Précision machine et mécanismes d'absorption et d'élimination

Pour quantifier la précision d'un appareil de calcul, on introduit la **précision machine** ε. Plus précisément, ε est le plus grand réel strictement positif qui satisfait :

$$\boxed{\varepsilon > 0 \ \text{ tel que } \ \mathrm{fl}(1+\varepsilon) = 1.}$$

Si l'on connait la représentation des nombres en machine, cette précision peut être déterminée théoriquement. Par exemple, celle-ci sera de $2^{-24} \approx 10^{-7}$ en simple précision et de $2^{-53} \approx 10^{-16}$ en double précision. Une illustration est donnée ci-dessous pour le cas de la simple précision.

1. Ce bit est dit "implicite" car il vaut toujours 1. Le problème avec cette convention est qu'on ne peut alors pas représenter le zéro (et le NaN). En fait, le zéro et le NaN sont représentés, respectivement, avec l'exposant $e = (00000000)$ et $e = (11111111)$ si bien qu'en réalité l'exposant parcourt uniquement $[[-126, 127]]$.



La précision machine entraîne des erreurs d'arrondi qui peuvent déboucher sur des erreurs de calcul catastrophiques. Deux mécanismes néfastes peuvent en effet survenir :

- **Le mécanisme d'absorption :** lorsque l'on additionne des nombres d'ordres de grandeur très différents (comme 1 et ε par exemple),
- **Le mécanisme d'élimination :** lorsque l'on soustrait des nombres très proches l'un de l'autre.

Pour se rendre compte du problème, prenons l'exemple suivant : on cherche à évaluer numériquement pour x proche de 0 :

$$f(x) = \frac{1}{\sqrt{1+x}-1}.$$

Pour $x \leq \varepsilon$, vous obtiendrez une division par zéro du fait de l'arrondi de x à 1 sous la racine ! On évitera donc autant que possible d'utiliser des algorithmes de calcul susceptibles de faire apparaître des mécanismes d'absorption et/ou d'élimination.

5.3 Conditionnement et stabilité

5.3.1 Objectif

Le conditionnement et la stabilité ont un but commun : **analyser la sensibilité de la solution d'un problème aux différentes erreurs** pouvant interférer avec le calcul de celle-ci. Néanmoins, ces deux notions s'inscrivent dans deux contextes différents :

- **Le conditionnement** est une notion qui **s'applique à un problème théorique** (en dehors de tout contexte numérique), et qui mesurera principalement l'influence des erreurs de données sur la précision de la solution finale.
- **La stabilité** est une notion qui **s'applique à un algorithme de résolution** numérique d'un problème donné, où l'on essaiera de quantifier la sensibilité de l'ensemble des opérations de l'algorithme aux erreurs d'arrondi principalement.

Nous reviendrons sur le conditionnement plus tard dans le cours, notamment dans le cadre de la résolution d'un système linéaire de la forme $Ax = b$ où A est une matrice carrée quelconque et b un vecteur de même taille.

5.3.2 Stabilité : exemple

Certaines méthodes peuvent se révéler très instables, pour calculer la solution de problèmes qui pourtant sont très stables. Il faut par exemple se méfier de certaines relations de récurrence. L'exemple suivant illustre clairement la problématique. Ce dernier sera aussi traité en TP.

Le problème : on souhaite calculer les valeurs des intégrales suivantes :

$$I_k = \int_0^1 x^k e^{-x} dx, \qquad \forall k,\ 1 \leq k \leq 50.$$

Une première analyse du problème permet d'établir :

$$\begin{aligned} I_0 &= 1 - \tfrac{1}{e}, \\ I_{k+1} &= (k+1)I_k - \frac{1}{e}, \quad k \geq 0. \end{aligned}$$

Il s'agit donc d'une formule de récurrence. On propose immédiatement l'algorithme suivant :

$I_0 \leftarrow 1 - \dfrac{1}{e}$
Pour k **de** 0 **à** 49
$\quad \| \quad I_{k+1} \leftarrow (k+1)I_k - \dfrac{1}{e}$
Fin pour k.

Une fois implémenté, cet algorithme nous donne n'importe quoi. On obtient par exemple

$$I_{18} = -0.1959\ 248 \ \text{ et } \ I_{50} = -1.0275\ 336 \times 10^{48}.$$

C'est extravagant quand on sait que les valeurs doivent décroître en restant positives !

Analyse de sensibilité : On va perturber la donnée initiale et observer l'évolution de l'algorithme. On a

$$I_0 = 1 - \frac{1}{e},$$

mais la valeur calculée par la machine est légèrement différente. Supposons donc que l'on initialise notre récurrence par :

$$\tilde{I}_0 = I_0 + \varepsilon.$$

Les premiers itérés que l'on va calculer seront :

$$\begin{aligned}
\tilde{I}_1 &= \tilde{I}_0 - \frac{1}{e} &&= I_1 + \varepsilon, \\
\tilde{I}_2 &= 2\,\tilde{I}_1 - \frac{1}{e} &&= I_2 + 2\,\varepsilon, \\
\tilde{I}_3 &= 3\,\tilde{I}_2 - \frac{1}{e} &&= I_3 + 3!\,\varepsilon,
\end{aligned}$$

et si l'on suppose que :

$$\tilde{I}_k = I_k + k!\,\varepsilon,$$

on vérifie que $\tilde{I}_{k+1} = I_{k+1} + (k+1)!\,\varepsilon$. On voit ainsi que la perturbation de départ est amplifiée par un facteur qui croît très vite.

Cette analyse de sensibilité permet de mettre en évidence **l'instabilité de l'algorithme** proposé. En double précision, on a vu que $\varepsilon \approx 1.11 \times 10^{-16}$ et comme $50! \simeq 3.04 \times 10^{64}$, on comprend mieux les résultats précédents. La valeur calculée pour I_{50} vérifie alors :

$$(I_{50})_c \simeq 0.3 \times \varepsilon \times 50!$$

On a surtout calculé, et sans doute assez correctement, la propagation de l'erreur.

La bonne méthode : la récurrence que l'on a établie :

$$I_{k+1} = (k+1)\,I_k - \frac{1}{e},$$

se retourne selon :

$$I_k = \frac{I_{k+1} + \frac{1}{e}}{k+1}.$$

Cette idée sera exploitée pour implémenter une récurrence descendante : lors du passage de I_{k+1} à I_k, les erreurs sur le calcul de I_{k+1} seront divisées par $k+1$. Supposons que l'on parte, pour un réel ν arbitraire, de

$$\tilde{I}_{60} = I_{60} + \nu.$$

Ici, on peut par exemple imaginer que l'on a calculé l'intégrale de manière approchée avec une méthode de discrétisation (l'erreur correspondant donc à ν ici). La récurrence descendante va nous fournir :

$$\begin{aligned}
\tilde{I}_{59} &= \frac{\tilde{I}_{60} + \frac{1}{e}}{60} &&= I_{59} + \frac{\nu}{60}, \\
\tilde{I}_{58} &= \frac{\tilde{I}_{59} + \frac{1}{e}}{59} &&= I_{58} + \frac{\nu}{60 \times 59}, \\
\dots &\quad \dots &&\quad \dots \\
\tilde{I}_{50} &= \frac{\tilde{I}_{51} + \frac{1}{e}}{51} &&= I_{50} + \frac{\nu}{60 \times 59 \dots \times 51}.
\end{aligned}$$

En choisissant arbitrairement $\nu = 1$, on obtient une valeur de I_{50} à peu près exacte à la double précision près :

$$\frac{50!}{60!} \simeq 3.6551\ 10^{-18}.$$

Chapitre 6

Systèmes linéaires

Dans ce chapitre, on s'intéressera à la résolution des systèmes linéaires de type $\mathbf{Ax = b}$, où A est une matrice carrée de taille n et b est un vecteur de taille n, A et b étant à coefficients réels ou complexes.

L'écriture $Ax = b$ correspond à l'écriture matricielle du système à n équations et n inconnues suivant :

$$\begin{cases} a_{11}\,x_1 + a_{12}\,x_2 + ... + a_{1n}\,x_n = b_1 \\ a_{21}\,x_1 + a_{22}\,x_2 + ... + a_{2n}\,x_n = b_2 \\ ... \\ a_{n1}\,x_1 + a_{n2}\,x_2 + ... + a_{nn}\,x_n = b_n \end{cases}$$

où les coefficients $(a_{ij})_{1\leq i,j\leq n}$ correspondent aux coefficients de la matrice A, et $(b_j)_{1\leq j\leq n}$ à ceux du vecteur b.

Ces systèmes apparaissent dans la résolution de nombreux problèmes de la physique ou de la mécanique (systèmes masses-ressorts, éléments finis, physique quantique, etc.).

On ne s'intéressera dans ce chapitre qu'au cas où la matrice A est inversible afin de s'assurer de l'existence et de l'unicité d'une solution au système $Ax = b$, qui est dans ce cas un système de Cramer. Le cas des systèmes linéaires $Ax = b$ où A est une matrice rectangulaire sera traité lors du chapitre sur les moindres carrés.

6.1 Résolution de systèmes linéaires simples

6.1.1 Systèmes linéaires diagonaux

Soit A une matrice diagonale, c'est-à-dire telle que $a_{ij} = 0$ pour tout $i \neq j$ avec $(i,j) \in [\![1,n]\!]^2$:

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ 0 & a_{22} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & a_{nn} \end{pmatrix}.$$

Alors le système $Ax = b$ est dit **diagonal** et se résout trivialement en posant pour tout $i \in [\![1,n]\!]$:

$$\boxed{x_i = \frac{b_i}{a_{ii}}}.$$

Remarque 6.1.1 *On a ici* $\det(A) = \prod_{i=1}^{n} a_{ii}$ *(les autres coefficients* $a_{i,j}$*,* $i \neq j$*, étant nuls) et* A *est supposée inversible, donc* $\det(A) \neq 0$*, ce qui entraîne que* $a_{ii} \neq 0$ *pour tout* $i \in [\![1,n]\!]$*. La solution précédente est donc bien définie.*

Définition 6.1.1 *On appelle* ***coût de résolution*** *(ou bien* ***complexité****) l'entier noté* $c(n)$ *correspondant au nombre d'opérations élémentaires (additions, soustractions, multiplications, divisions) nécessaires pour obtenir la solution d'un système* $Ax = b$ *où* A *est une matrice de taille* n*.*

Dans le cas présent d'une matrice diagonale, ce coût de résolution est :

$$\boxed{c(n) = n},$$

puisque l'on effectue une seule division pour le calcul de chaque composante de x, qui a n composantes.

6.1.2 Systèmes linéaires triangulaires

On appelle système triangulaire tout système linéaire $Ax = b$ où la matrice A est triangulaire. Les systèmes triangulaires sont également simples à résoudre et il en existe deux catégories :

- les systèmes triangulaires supérieurs, où $a_{ij} = 0$ pour tout $i > j$ avec $(i,j) \in [\![1,n]\!]^2$:

$$A = \begin{pmatrix} a_{11} & \cdots & \cdots & \cdots & a_{1n} \\ 0 & \ddots & & & \vdots \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 & a_{nn} \end{pmatrix}$$

- et les systèmes triangulaires inférieurs, où $a_{ij} = 0$ pour tout $i < j$ avec $(i,j) \in [\![1,n]\!]^2$:

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & \ddots & 0 \\ a_{n1} & \cdots & \cdots & \cdots & a_{nn} \end{pmatrix}$$

Examinons d'abord le cas des systèmes triangulaires supérieurs, qui s'écrivent de manière explicite :

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

Ce genre de système se résout par une méthode dite de **remontée** (ou *substitution rétrograde*) : on commence par calculer la dernière composante, puis on utilise sa valeur pour calculer l'avant-dernière, et on recommence selon le même principe jusqu'à la première. Ce procédé se formalise de la manière suivante :

- Étape 1 : en utilisant la dernière équation du système pour obtenir :

$$x_n = \frac{b_n}{a_{nn}}\ .$$

- Étape 2 : on réécrit l'avant-dernière équation pour exprimer x_{n-1} en fonction de x_n :

$$x_{n-1} = \frac{1}{a_{n-1,n-1}}\left(b_{n-1} - a_{n-1,n}\,x_n\right)\ ,$$

puis on utilise la valeur de x_n trouvée à l'étape 1 pour obtenir celle de x_{n-1}.

- Étape i : on réécrit l'équation i pour exprimer x_i en fonction de $x_n, x_{n-1}, \cdots, x_{i+1}$:

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij}\, x_j \right) ,$$

  puis on utilise les valeurs de $x_n, x_{n-1}, \cdots, x_{i+1}$ trouvées aux étapes précédentes pour obtenir x_i.
  N.B. : cette expression est valable pour tout $i \in [\![1, n]\!]$.

- Étape n : on réécrit l'équation 1 pour exprimer x_1 en fonction de $x_n, x_{n-1}, \cdots, x_2$:

$$x_1 = \frac{1}{a_{11}} \left(b_1 - \sum_{j=2}^{n} a_{1j}\, x_j \right) ,$$

  puis on utilise les valeurs de $x_n, x_{n-1}, \cdots, x_2$ trouvées aux étapes précédentes pour obtenir x_1.

Exemple 6.1.1 *Prenons le système suivant :*

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

qui s'écrit aussi :

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

Lorsque le système est triangulaire inférieur, on le résout par une méthode dite de **descente** : on commence par calculer la première composante grâce à la première équation, puis on utilise sa valeur pour calculer la deuxième, et ainsi de suite jusqu'à la dernière.

Calculons maintenant le coût de résolution d'un système triangulaire. Comme vu précédemment, l'expression utilisée à l'étape i pour calculer x_i pour tout $i \in [\![1, n]\!]$ est :

$$x_i = \frac{1}{a_{ii}} \left(b_i - \sum_{j=i+1}^{n} a_{ij}\, x_j \right) ,$$

À chaque étape, il y a donc $n-i$ multiplication(s), $n-i$ soustraction(s) et 1 division. Étant donné qu'il y a n étapes en tout, le coût de résolution s'obtient par la somme suivante :

$$c(n) = \sum_{i=1}^{n}((n-i)+(n-i)+1) = \sum_{i=1}^{n}(2n+1-2i) = (2n+1)n - 2\left(\frac{n(n+1)}{2}\right) ,$$

autrement dit :

$$\boxed{c(n) = n^2} .$$

6.2 Résolution de systèmes linéaires quelconques

6.2.1 Pivot de Gauss

La stratégie du **pivot de Gauss** consiste à transformer progressivement le système de départ $Ax = b$ en un système triangulaire supérieur, que l'on résout ensuite par remontée : on utilise la première ligne et notamment le coefficient a_{11} pour, grâce à des opérations sur les lignes du système, mettre à 0 les coefficients a_{i1} pour tout $i > 1$. On procède ensuite de la même façon en utilisant la deuxième ligne, et notamment le coefficient a_{22}, pour mettre à 0 les coefficients a_{i2} pour tout $i > 2$. Et ainsi de suite, en utilisant le coefficient a_{kk}, pour mettre à 0 les coefficients a_{ik} pour tout $i > k$. À l'étape k, le coefficient a_{kk} est appelé le **pivot**.

Exemple 6.2.1 *Considérons le système :*

$$\begin{cases} 2x_1 + x_2 + x_3 = 1 \\ 6x_1 + 2x_2 + x_3 = -1 \\ -2x_1 + 2x_2 + x_3 = 7 \end{cases}$$

Le premier coefficient de la première ligne est non nul : on va s'en servir comme « pivot ». On va conserver la première ligne et éliminer la variable x_1 des lignes 2 et 3 :

- *on remplace la ligne* L_2 *par* $L_2 - \frac{6}{2}L_1$*, c'est-à-dire* $L_2 \longleftarrow L_2 - 3L_1$*,*
- *puis on remplace la ligne* L_3 *par* $L_3 - \frac{-2}{2}L_1$*, c'est-à-dire* $L_3 \longleftarrow L_3 + L_1$*.*

Notons qu'à chaque étape, on a divisé par le pivot. On obtient alors le système équivalent suivant :

$$\begin{cases} 2x_1 + x_2 + x_3 = 1 \\ - x_2 - 2x_3 = -4 \\ 3x_2 + 2x_3 = 8 \end{cases}$$

Enfin, on élimine la variable x_2 *de la dernière ligne grâce au nouveau pivot de la deuxième ligne* (-1) *en remplaçant la ligne* L_3 *par* $L_3 - \frac{3}{(-1)}L_2$*, c'est-à-dire* $L_3 \longleftarrow L_3 + 3L_2$*. On obtient ainsi le système :*

$$\begin{cases} 2x_1 + x_2 + x_3 = 1 \\ - x_2 - 2x_3 = -4 \\ - 4x_3 = -4 \end{cases}$$

On obtient ainsi un système triangulaire supérieur, que l'on résout ensuite par remontée.

6.2.2 Interprétation matricielle

Nous venons de voir qu'afin de transformer un système linéaire quelconque en un système triangulaire supérieur, on effectuait un certain nombre d'opérations sur les lignes du système. Ces opérations ont une interprétation matricielle.

Introduisons pour tout $(i, j) \in [\![1, n]\!]^2$ tel que $i > j$ la matrice $\mathcal{L}_{ij}(\lambda)$ définie par :

$$\boxed{\mathcal{L}_{ij}(\lambda) = Id_n + \lambda E_{ij}},$$

où E_{ij} est la matrice dont tous les coefficients valent 0 sauf sur la ligne i et la colonne j où le coefficient correspondant vaut 1.

Ainsi, la matrice $\mathcal{L}_{ij}(\lambda)$ est une matrice triangulaire inférieure avec des 1 sur sa diagonale, un λ sur le coefficient de la ligne i et de la colonne j, et des 0 partout ailleurs.

Exemple 6.2.2 *On aura par exemple dans le cas* $n = 3$ *:*

$$\mathcal{L}_{21}(\lambda) = \begin{pmatrix} 1 & 0 & 0 \\ \lambda & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad , \quad \mathcal{L}_{31}(\lambda) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \lambda & 0 & 1 \end{pmatrix} \quad \textit{et} \quad \mathcal{L}_{32}(\lambda) = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda & 1 \end{pmatrix} .$$

Afin d'illustrer l'intérêt de ces matrices, restons dans le cadre $n = 3$. Prenons une matrice A quelconque :

$$A = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix} ,$$

et calculons maintenant le produit de la matrice $\mathcal{L}_{32}(\lambda)$ par A :

$$\begin{aligned}\mathcal{L}_{32}(\lambda)A &= \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & \lambda & 1\end{pmatrix}\begin{pmatrix}a & b & c\\ d & e & f\\ g & h & i\end{pmatrix}\\ &= \begin{pmatrix}a & b & c\\ d & e & f\\ g+\lambda d & h+\lambda e & i+\lambda f\end{pmatrix}.\end{aligned}$$

On remarque ici qu'en multipliant A à gauche par $\mathcal{L}_{32}(\lambda)$, on a obtenu une matrice presque identique à A mais dans laquelle on a remplacé la ligne L_3 par $L_3 + \lambda L_2$.

Ce résultat peut se généraliser (démonstration laissée en exercice) :

$\boxed{\mathcal{L}_{ij}(\lambda)A \text{ est la matrice } A \text{ dans laquelle on a remplacé la ligne } L_i \text{ par } L_i + \lambda L_j}$.

On voit ainsi que l'on peut écrire toutes les étapes du pivot de Gauss de manière matricielle !

Regardons cela sur le système de l'exemple 2.1. Dans ce cas, la matrice A est donnée par :

$$A = \begin{pmatrix}2 & 1 & 1\\ 6 & 2 & 1\\ -2 & 2 & 1\end{pmatrix}$$

La première étape ($L_2 \longleftarrow L_2 - 3L_1$ puis $L_3 \longleftarrow L_3 + L_1$) revient matriciellement à écrire :

$$\mathcal{L}_{31}(1)\mathcal{L}_{21}(-3)A = \begin{pmatrix}2 & 1 & 1\\ 0 & -1 & -2\\ 0 & 3 & 2\end{pmatrix}$$

Si l'on note $A^{(1)}$ la matrice précédente, la seconde étape ($L_3 \longleftarrow L_3 + 3L_2$) revient matriciellement à écrire :

$$\mathcal{L}_{32}(3)A^{(1)} = \begin{pmatrix}2 & 1 & 1\\ 0 & -1 & -2\\ 0 & 0 & -4\end{pmatrix}$$

Si l'on note U cette matrice triangulaire supérieure, on peut résumer les deux étapes précédentes par :

$$\mathcal{L}_{32}(3)\mathcal{L}_{31}(1)\mathcal{L}_{21}(-3)A = U\ , \tag{6.1}$$

On obtient ainsi une écriture matricielle des différentes étapes du pivot de Gauss.

6.2.3 Factorisation LU

Sur un exemple

Nous allons maintenant voir que cette écriture matricielle permet d'obtenir in fine une factorisation de la matrice A. Pour cela, multiplions l'égalité (6.1) à gauche par l'inverse de la matrice $\mathcal{L}_{32}(3)\mathcal{L}_{31}(1)\mathcal{L}_{21}(-3)$:

$$A = (\mathcal{L}_{32}(3)\mathcal{L}_{31}(1)\mathcal{L}_{21}(-3))^{-1}\,U = \mathcal{L}_{21}(-3)^{-1}\mathcal{L}_{31}(1)^{-1}\mathcal{L}_{32}(3)^{-1}U\ .$$

Énonçons maintenant une propriété permettant de calculer l'inverse de n'importe quelle matrice $\mathcal{L}_{ij}(\lambda)$.

Proposition 6.2.1 *Pour tout $(i,j) \in [\![1,n]\!]^2$ tel que $i > j$, on a :*

$$\boxed{\mathcal{L}_{ij}(\lambda)^{-1} = \mathcal{L}_{ij}(-\lambda)}\,,$$

En utilisant la proposition précédente, on obtient alors la factorisation finale :

$$A = \mathcal{L}_{21}(3)\mathcal{L}_{31}(-1)\mathcal{L}_{32}(-3)U\ .$$

En posant $L = \mathcal{L}_{21}(3)\mathcal{L}_{31}(-1)\mathcal{L}_{32}(-3)$, on voit que l'on a alors factorisé la matrice A sous la forme :

$$A = L\,U\ ,$$

où L est une matrice triangulaire inférieure et U une matrice triangulaire supérieure.

Conclusion : l'écriture matricielle du pivot de Gauss à l'aide des matrices $\mathcal{L}_{ij}(\lambda)$ permet non seulement de garder en mémoire les opérations réalisées mais surtout d'aboutir à la factorisation de A sous la forme :

$$\boxed{A = L\,U}\,,$$

où L est une matrice triangulaire inférieure (L pour « *lower* ») qui n'a que des 1 sur sa diagonale, et U une matrice triangulaire supérieure (U pour « *upper* ») obtenue à partir de A grâce à la méthode du pivot de Gauss.

Cette factorisation est appelée **factorisation LU** de la matrice A.

Nous verrons dans les exercices que la matrice L se détermine en réalité sans calcul : c'est une matrice triangulaire inférieure avec des 1 sur la diagonale, et sous la diagonale on retrouve l'opposé des coefficients λ des différentes étapes du pivot de Gauss :

si on a effectué l'opération $L_i \leftarrow L_i + \lambda L_j$ lors du calcul, alors on aura $l_{ij} = -\lambda$ dans la matrice L .

Si l'on revient à l'exemple précédent, les étapes du pivot de Gauss sont :

$$L_2 \longleftarrow L_2 - 3L_1 \quad , \quad L_3 \longleftarrow L_3 + L_1 \quad \text{puis} \quad L_3 \longleftarrow L_3 + 3L_2$$

et on aura donc :

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

Existence d'une factorisation LU pour une matrice A inversible

Une question se pose à ce stade : peut-on toujours effectuer la factorisation LU d'une matrice A inversible ? Pour répondre à cette question, revenons au pivot de Gauss : nous avons vu que la factorisation LU n'était rien d'autre qu'une réécriture matricielle des opérations sur les lignes du pivot de Gauss. Or à chaque étape k du calcul ($k \in [\![1, n-1]\!]$), ces opérations s'écrivent pour tout $i \in [\![k+1, n]\!]$:

$$L_i \leftarrow L_i + \lambda L_k \quad \text{où} \quad \lambda = -\frac{a_{ik}}{a_{kk}} .$$

On voit donc clairement que le seul obstacle qui pourrait bloquer le calcul serait le cas d'un pivot nul ($a_{kk} = 0$).
Deux conséquences à cela :

- la factorisation LU d'une matrice A inversible n'est possible que si l'on ne rencontre pas de pivot nul,
- elle redevient possible si l'on rajoute la possibilité de permuter des lignes dans le cas d'un pivot nul.

Les deux théorèmes suivants (admis) permettent de préciser de manière rigoureuse les observations précédentes.

Théorème 6.2.1 ***Factorisation LU sans permutation***

Soit $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$). Si pour tout $k \in [\![1, n]\!]$, la sous-matrice principale $A_k = (a_{ij})_{1 \leq i,j \leq k}$ de A est telle que $det(A_k) \neq 0$, alors il existe une unique matrice triangulaire inférieure L dont les éléments diagonaux sont égaux à 1, et une unique matrice triangulaire supérieure U telles que :

$$A = LU .$$

Avant d'énoncer le second théorème, définissons ce qu'est une *matrice de permutation*.

Définition 6.2.1 *Une matrice $P \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) est appelée **matrice de permutation** s'il existe une bijection σ de $[\![1, n]\!]$ dans $[\![1, n]\!]$ telle que pour tout $i \in [\![1, n]\!]$, on a $Pe_i = e_{\sigma(i)}$.*

Théorème 6.2.2 ***Factorisation LU avec permutation***

Soit $A \in \mathcal{M}_n(\mathbb{R})$ (ou $\mathcal{M}_n(\mathbb{C})$) une matrice inversible. Alors il existe une matrice de permutation P, une matrice triangulaire inférieure L dont les éléments diagonaux sont égaux à 1 et une matrice triangulaire supérieure inversible U telles que :

$$PA = LU \ .$$

Exemple 6.2.3 *Matrice de permutation 3×3 qui échange la deuxième et la troisième ligne :*

$$P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

Remarque 6.2.1 *L'inverse d'une matrice de permutation est égale à sa transposée : $P^{-1} = P^T$.*

Coût d'une factorisation LU

Intéressons-nous maintenant au coût de la factorisation LU. À chaque étape k du pivot de Gauss (où k correspond au numéro de la colonne dans laquelle on essaye de faire apparaître des 0 grâce au pivot), on doit :

- effectuer $n - k$ divisions par le pivot pour le calcul des coefficients de la k-ième colonne de L,
- effectuer $(n - k)^2$ multiplications et $(n - k)^2$ soustractions pour la mise à jour de la matrice A.

Le coût de calcul de la factorisation LU s'obtient donc en sommant pour k allant de 1 à $n - 1$ (et pas n car on arrête le pivot de Gauss à la colonne $n - 1$) :

$$c(n) = \sum_{k=1}^{n-1} (n - k) + 2(n - k)^2 = \frac{1}{6}(4n^3 - 3n^2 - n) = \frac{1}{6} n(n - 1)(4n + 1)$$

On obtient ainsi le coût de calcul approximatif suivant pour la factorisation LU d'une matrice de taille n :

$$\boxed{c(n) \sim \frac{2n^3}{3}} \ .$$

Application de la factorisation LU à la résolution des systèmes linéaires

Remarque 6.2.2 *Une première conséquence intéressante de la factorisation LU d'une matrice est le calcul effectif du déterminant de cette matrice. En effet, si on a* $A = LU$ *avec* U *triangulaire supérieure et* L *triangulaire inférieure avec des* 1 *sur la diagonale, alors on peut facilement calculer le déterminant de* A *:*

$$det(A) = det(L)\, det(U) = \prod_{k=1}^{n} U_{kk} \ .$$

Mais le principal intérêt de la factorisation LU réside dans son application à la résolution de systèmes linéaires quelconques : considérons pour cela un système $Ax = b$ où la matrice A vérifie les hypothèse du théorème de factorisation LU sans permutation. Il existe donc une matrice triangulaire inférieure L et une matrice triangulaire supérieure U telle que :

$$A = LU \ .$$

Le système $Ax = b$ peut donc également s'écrire :

$$LUx = b \ ,$$

ou bien encore :

$$L(Ux) = b \ .$$

En posant $y = Ux$, on obtient ainsi l'équivalence suivante :

$$\boxed{Ax = b \quad \Longleftrightarrow \quad \begin{cases} Ly = b \\ Ux = y \end{cases}} \ .$$

On voit donc que la résolution du système $Ax = b$ équivaut à la résolution successive de deux systèmes triangulaires.

Remarque 6.2.3 *Si la matrice* A *est inversible mais ne vérifie pas les hypothèses du théorème de factorisation LU sans permutation, alors on résoudra le système* $PAx = Pb$ *au lieu de* $Ax = b$ *et* x *sera alors la solution de :*

$$\begin{cases} Ly = Pb \\ Ux = y \end{cases} \ .$$

Sachant que la résolution d'un système triangulaire est peu coûteux, la factorisation LU fournit donc une méthode très intéressante et relativement peu coûteuse pour résoudre

un système linéaire.

Plus précisément, le coût de résolution d'un système $Ax = b$ (où A est inversible) via cette méthode sera égal à :

$$c(n) = \frac{1}{6}n(n-1)(4n+1) + 2n^2 \ ,$$

où le premier terme correspond au coût de la factorisation LU et le second à la résolution des deux systèmes triangulaires. On obtient donc un coût de résolution via cette méthode approximativement égal à :

$$\boxed{c(n) \sim \frac{2n^3}{3}} \ .$$

6.2.4 Cas particulier des matrices symétriques définies positives

Commençons par définir ce qu'est une matrice symétrique définie positive.

Définition 6.2.2 *Soit $A \in \mathcal{M}_n(\mathbb{R})$. La matrice A est dite symétrique définie positive si et seulement si :*

- $A = A^T$,
- $\forall x \in \mathbb{R}^n \backslash \{0_{\mathbb{R}^n}\}$, $x^T A x > 0$.

Dans ce cas, on peut appliquer une autre factorisation appelée factorisation de Cholesky.

Théorème 6.2.3 ***Factorisation de Cholesky***

Soit A une matrice symétrique définie positive. Alors il existe une matrice triangulaire inférieure L telle que :

$$A = LL^T \ .$$

*Cette factorisation est **unique** si on impose les coefficients diagonaux de L strictement positifs.*

On obtient la factorisation directement par identification :

$$A_{ij} = \sum_{k=1}^{\min(i,j)} L_{ik}L_{jk}$$

La matrice A étant symétrique, on peut supposer $i \leq j$. Si $i = j = 1$, on a $L_{11}^2 = A_{11}$. On pose donc $L_{11} = \sqrt{A_{11}}$. Ensuite, pour $i = 1$ et $j = 2, ..., n$, on a :

$$A_{1j} = L_{11}L_{j1} \Rightarrow L_{j1} = \frac{A_{1j}}{L_{11}}$$

Ceci achève la détermination de la première colonne de L en au plus n opérations.

On réitère ensuite le procédé d'identification pour les autres colonnes $i \in [\![2, n]\!]$ et on obtient :

$$L_{ii}^2 = A_{ii} - \sum_{k=1}^{i-1} L_{ik}^2 \, , \quad L_{ji} = \frac{1}{L_{ii}} \left(A_{ij} - \sum_{k=1}^{i-1} L_{ik} L_{jk} \right) .$$

Le coût total de la factorisation de Cholesky est de l'ordre de :

$$\boxed{c(n) = \frac{n^3}{3}} \, ,$$

soit la moitié du coût de la factorisation LU d'une matrice.

Chapitre 7

Équations non linéaires : Résolution numérique

Dans le cas général, on ne peut pas espérer résoudre une équation ou un système d'équations non linéaires par une formule directe : ça n'est possible que pour quelques équations algébriques, comme par exemple les équations du second degré.

En fait, la formule qui donne les deux racines des équations du second degré est connue depuis très longtemps. Elle a été retrouvée sur des tablettes babyloniennes datées des environs de 1500 avant notre ère. Pour les équations de degré 3 la solution a été établie par l'italien Tartaglia en 1535, mais publiée en 1545 par Jérôme Cardan, qui par ailleurs a laissé son nom à un système de transmission cher aux automobilistes. Pour le degré 4, c'est encore un italien, dont le nom sonne de façon plaisante aux oreilles des amateurs d'automobiles qui a établi la méthode (Ludovico Ferrari vers 1540). Ces formules sont exactes si l'on considère que l'extraction d'une racine est une opération exacte. Et on sait depuis Niels Abel et Évariste Galois que l'on ne peut pas étendre ce type de formules au calcul des racines d'un polynôme quelconque de degré supérieur.

Mais si les calculettes nous fournissent immédiatement la valeur $\alpha = \sqrt{a}$ pour un réel a positif, cette opération est en fait la résolution de l'équation non linéaire $x^2 - a = 0$, et s'obtient en calculant de proche en proche les itérés d'une suite qui tend vers α. De façon plus générale, on peut être amené à rechercher une racine réelle x d'une équation

$$F(x) = 0,$$

où F est une fonction quelconque. On devra construire une suite $(x_k)_{k\in\mathbb{N}}$ qui converge vers une solution du problème. Il est important de remarquer que l'on parle ici d'une solution : on n'a généralement pas unicité de solution et il arrive que l'on en cherche plusieurs, ou qu'on les cherche toutes.
Il faudra alors

1. localiser la ou les racines que l'on veut calculer,
2. s'assurer que la suite (x_k) converge vers une limite x^* telle que $F(x^*) = 0$,

3. prévoir un test d'arrêt des calculs, et estimer la précision qu'il garantit,
4. éventuellement, savoir comparer différentes méthodes.

7.1 Héron d'Alexandrie et les racines carrées

Également appelé Héron l'ancien, il vécut au premier siècle de notre ère et fut un des grands mécaniciens de l'antiquité ; il inventa diverses machines et instruments de mesure comme l'odomètre permettant de mesurer les distances parcourues en comptant le nombre de tours d'une roue, et dont le principe est encore utilisé par les compteurs kilométriques des bicyclettes. L'algorithme de calcul des racines carrées qu'on lui attribue était en fait déjà connu des Babyloniens.

7.1.1 La méthode de Héron

Approche pragmatique

Soit a un nombre positif dont on veut calculer la racine carrée

$$\alpha = \sqrt{a}.$$

Il se trouve que α correspond à la longueur du côté d'un carré dont l'aire est a. On va chercher à construire un tel carré en partant d'un rectangle, qu'on va peu à peu transformer pour le faire ressembler à un carré. Si l'on considère une valeur x_0 donnée comme une estimation grossière de α, on construit d'abord un rectangle $\mathcal{R}$ d'aire a ayant un côté de longueur

$$L = x_0,$$

et un second côté égal à

$$l = \frac{a}{L} = \frac{a}{x_0}.$$

Pour construire un nouveau rectangle qui ressemble plus à un carré, il est raisonnable de choisir pour longueur d'un côté la moyenne des longueurs des côtés du rectangle précédent. Cette longueur sera

$$L = x_1 = \frac{1}{2}\,(x_0 + \frac{a}{x_0}).$$

Le second côté aura pour longueur

$$l = \frac{a}{L} = \frac{a}{x_1}.$$

Et l'on va continuer jusqu'à ce que notre appareil de mesure ne nous permette plus de distinguer les longueurs des deux côtés. En réitérant cette démarche, on en déduit donc la formule

$$x_{n+1} = \frac{1}{2}\,(x_n + \frac{a}{x_n}). \tag{7.1}$$

Exemple 7.1.1 *La figure 7.1 nous montre trois étapes de ce procédé. Sur cet exemple,*

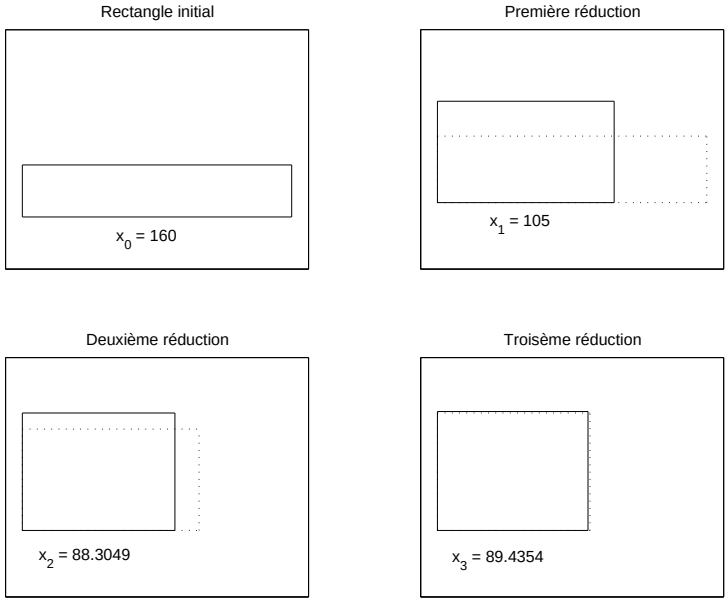


FIGURE 7.1 – Premiers itérés de la méthode de Héron.

nous cherchons à calculer $\alpha = \sqrt{8000}$*. La valeur initiale est* $x_0 = 160$*. Le rectangle initial a pour côtés* 160 *et* 50*. Pour construire un rectangle qui ressemble plus à un carré, on le remplace par le rectangle dont un côté est la moyenne des deux précédents, c'est-à-dire*

$$x_1 = \frac{1}{2}\,(x_0 + \frac{a}{x_0}) = 105,$$

de sorte que l'autre côté a pour longueur

$$\frac{8000}{105} = 76.1905.$$

x_n	$\dfrac{a}{x_n}$
160.00000000000000	50.00000000000000
105.00000000000000	76.19047619047619
90.59523809523810	88.30486202365309
$\underline{89.4}5005005944560$	$\underline{89.4}3538874135297$
$\underline{89.44271}940039928$	$\underline{89.44271}879958389$
$\underline{89.44271909999159}$	$\underline{89.44271909999158}$

On a obtenu la racine cherchée "à la précision machine". Les chiffres corrects sont soulignés, et l'on constate que dès que l'on a obtenu 3 chiffres corrects (pour x_3), la convergence devient très rapide : 7 chiffres pour x_4 et 15 pour x_5 !

Le principe des approximations successives

On peut regarder cet algorithme d'un point de vue légèrement différent. Partant de l'équation

$$F(x) = x^2 - a = 0,$$

qu'on écrit de façon équivalente $x^2 = a$ ou $x = \dfrac{a}{x}$, on obtient successivement

$$x^2 = a$$

On a remplacé l'équation $F(x) = 0$ par une équation équivalente

C'est un exemple de méthode de point fixe (que l'on verra plus précisément à la section 7.2.2).

Définition 7.1.1 *Soit f une fonction de $\mathbb{R}^d$ dans $\mathbb{R}^d$. Alors un point fixe de la fonction f est un $x \in \mathbb{R}^d$ tel que*

$$f(x) = x.$$

La méthode du point fixe tient son nom du fait que la limite $\alpha = \sqrt{a}$ de la suite définie par la récurrence ci-dessus est telle que

$$\alpha = f(\alpha).$$

Cette méthode reste néanmoins imprécise sur deux points fondamentaux

1) le choix de la donnée initiale x_0,
2) le critère à vérifier pour arrêter les itérations.

On va voir que pour notre algorithme, ces deux points peuvent être élucidés facilement.

Remarque 7.1.1 *Le critère qui va arrêter les itérations s'appelle* ***le critère d'arrêt****. Il doit être fixé sans quoi, en général, les itérations ne s'arrêtent jamais puisque les nombres stockés en machine n'ont qu'un nombre fini de décimales là où une racine est souvent un nombre irrationnel.*
Quels critères peut on fixer ?

7.2 Résolution d'une équation non linéaire

Notre but est de chercher les racines éventuelles de l'équation

$$F(x) = 0, \tag{7.2}$$

où la fonction F est définie de $\mathbb{R}$ dans $\mathbb{R}$.

La recherche d'une solution de l'équation (7.2) ne peut se faire, en général, que par une méthode itérative. On va maintenant s'intéresser à trois de ces méthodes : la **méthode de la dichotomie**, la **méthode du point fixe** et la **méthode de Newton**.

On verra que ces méthodes n'ont pas toutes la même vitesse de convergence. Cette vitesse va être donnée par l'**ordre de la méthode** qui est défini comme suit.

Définition 7.2.1 *Si la suite des itérés $(x_n)_{n\geq 0}$ d'une méthode itérative est telle qu'il existe une constante $c > 0$ telle que pour n suffisamment grand, les erreurs $e_n = |x_n - x|$ vérifient*

$$e_{n+1} \leq c\, e_n^p,$$

on dit que la méthode est d'ordre p.

Exemple 7.2.1 *Dans le cas de la méthode d'Héron, on a une méthode d'ordre 2.*

Enfin, ces méthodes doivent être alimentées par une ou deux valeurs initiales que l'on prendra, afin de gagner en temps de calcul, aussi proche que possible de la racine cherchée. Une connaissance *a priori* de la nature du problème vous permettra de situer grossièrement les racines qui vous intéressent. C'est d'ailleurs la première réflexion qu'il faut faire. Elle permet notamment de délimiter l'intervalle d'étude de la fonction.

Exemple 7.2.2 *On cherche les racines positives de la fonction*

$$F(x) = 0.01\ \exp x + 10\ \cos x - 3\,x. \tag{7.3}$$

Localisation graphique : *Une première façon de choisir une valeur initiale consiste à faire une représentation graphique. La figure 7.2 nous donne une représentation graphique de F sur l'intervalle $[-1,\ 10]$. Notre équation possède apparemment deux racines*

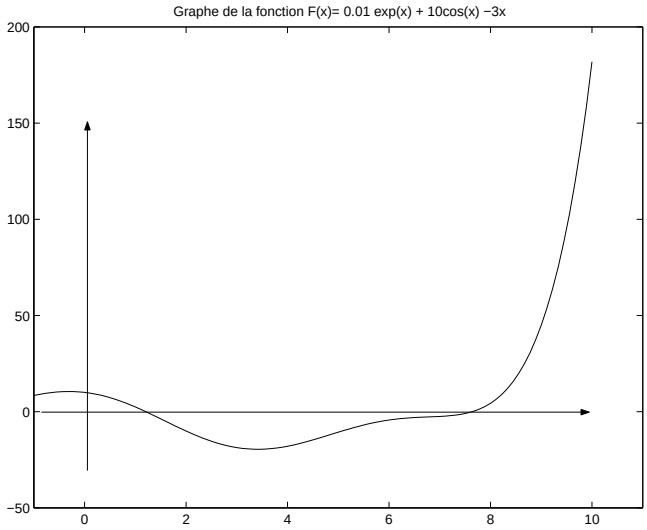


FIGURE 7.2 – Un graphique pour localiser les racines.

positives

- *la première est dans l'intervalle $[0,\ 2]$, plutôt vers le milieu,*
- *la seconde est dans l'intervalle $[7,\ 8]$, plus près de 8.*

Balayage systématique : *Faute de possibilité de tracer facilement un graphe, on peut balayer systématiquement l'intervalle qui nous intéresse pour rechercher d'éventuels changements de signe de F. Pour notre exemple, on peut calculer les valeurs de $F(x)$ pour les valeurs entières positives de x. Ces valeurs sont, dans l'ordre 10.01, 2.43, -10.08, -18.69, -17.99, -10.67, -4.36, -2.49, 4.35, 44.91, 181.87. On observe un changement de signe entre 1 et 2 ainsi qu'entre 7 et 8.*

7.2.1 Méthode de dichotomie

Principe de la dichotomie : Soit F une fonction continue de $\mathbb{R}$ dans $\mathbb{R}$ et qui change de signe sur un intervalle $[a,\ b]$, c'est-à-dire que l'on a $F(a)\,F(b) < 0$. Alors, par le théorème des valeurs intermédiaires, F s'annule au moins une fois sur cet intervalle. Cette remarque peut être utilisée pour former une suite d'intervalles de plus en plus petits qui contiennent une racine.

Mise en oeuvre : En notant a_0, b_0 les bornes de l'intervalle initial et $m = \dfrac{a_0 + b_0}{2}$ le milieu de cet intervalle, on aura deux possibilités :

- soit $F(a_0)\,F(m) \leq 0$, et dans ce cas nous aurons une racine dans $[a_0,\ m]$: on pose alors $a_1 = a_0$ et $b_1 = m$,
- soit $F(b_0)\,F(m) \leq 0$, et dans ce cas nous aurons une racine dans $[m,\ b_0]$: on pose alors $a_1 = m$ et $b_1 = b_0$.

Dans les deux cas, le nouvel intervalle $[a_1,\ b_1]$ est de longueur $b_1 - a_1 = \dfrac{b_0 - a_0}{2}$ et on a donc divisé par 2 la longueur de $[a_0, b_0]$. On recommence le même processus avec ce nouvel intervalle. On construit ainsi une suite d'intervalles $[a_n,\ b_n]$ qui sont emboîtés et contiennent une racine. Une fois l'intervalle $[a_n,\ b_n]$ suffisamment petit, on prendra pour racine la valeur

$$\tilde{x} = \frac{a_n + b_n}{2}.$$

Critère d'arrêt : On arrêtera ces itérations lorsque la longueur $|a_k - b_k|$ de l'intervalle sera inférieure à $2\,\varepsilon$, qui correspond à un seuil d'erreur absolue ε choisi à l'avance.
Algorithme : On en déduit ainsi l'algorithme de la table 7.1.

```
fa ← F(a)
fb ← F(b)
Erreur si  fa * fb > 0
Tant que |b − a| > 2ε
    ‖ c ← (a + b)/2
    ‖ fc ← F(c)
    ‖ Si   fa * fc < 0
    ‖     ‖ b ← c
    ‖ Sinon
    ‖     ‖ a ← c
    ‖     ‖ fa ← fc
```

TABLE 7.1 – Algorithme de la dichotomie

Exemple 7.2.3 *Voici les résultats obtenus pour la fonction F de l'exemple 7.2.2 en partant des intervalles $[1,\ 2]$ et $[7,\ 8]$.*

n	a_n	b_n	a_n	b_n
0	1.000000000	2.000000000	7.000000000	8.000000000
1	1.000000000	1.500000000	7.500000000	8.000000000
2	1.000000000	1.250000000	7.500000000	7.750000000
3	1.125000000	1.250000000	7.625000000	7.750000000
⋮	⋮	⋮	⋮	⋮
12	1.204589843	1.204833984	7.639648437	7.639892578
13	1.204589843	1.204711914	7.639770507	7.639892578
14	1.204589843	1.204650878	7.639831542	7.639892578

Les racines calculées sont alors respectivement $\tilde{r_1} = 1.204620361$ et $\tilde{r_2} = 7.639862060$.

Remarque 7.2.1 *Sur les deux racines calculées précédemment, la précision garantie est la même. En effet la longueur des derniers intervalles calculés est, dans un cas*

$$b_{14} - a_{14} = 6.1035 \times 10^{-5},$$

et dans l'autre

$$b_{14} - a_{14} = 6.1036 \times 10^{-5}.$$

Cette précision peut être prévue à l'avance.

À l'initialisation, si l'intervalle $[a_0,\ b_0]$ a pour longueur $\ell_0 = 1$ alors, à la première itération, sa longueur deviendra égale à

$$\ell_1 = \frac{1}{2},$$

et à la n-ième, elle sera égale à

$$\ell_{n-1} = \frac{1}{2^n}.$$

Comme on choisit le milieu de l'intervalle, l'erreur maximale sera donc

$$\nu = \frac{\ell_{n-1}}{2} = \frac{1}{2^{n+1}}.$$

Exemple 7.2.4 *Si l'on souhaite une précision de 10^{-4}, on cherchera n^* tel que*

$$\frac{1}{2^{n^*+1}} \leq 10^{-4},$$

soit encore

On a en fait le résultat suivant qui généralise les propos précédents.

Théorème 7.2.1 *Si r désigne une racine de l'équation $F(x) = 0$ contenue dans l'intervalle $[a_0,\ b_0]$. La méthode de dichotomie appliquée à la recherche de cette racine est telle que*

i) la racine approchée r_n calculée à l'itération n vérifie

$$|r_n - r| \leq \frac{|a_0 - b_0|}{2^{n+1}}.$$

ii) le nombre n_ε d'itérations nécessaire pour atteindre une précision ε est tel que

$$(n_\varepsilon + 1)\ln 2 \geq \ln \frac{|a_0 - b_0|}{\varepsilon}.$$

La borne de l'erreur est divisée par 2 à chaque itération.

Ordre de la dichotomie : grâce au théorème précédent et suivant la définition 7.2.1, la méthode de dichotomie est d'ordre 1. Ce type de vitesse de convergence est dit **linéaire**.

7.2.2 Méthode du point fixe

Principe de la méthode du point fixe : Lors de la recherche de la racine carrée par la méthode de Héron, nous avons construit une suite définie par $x_{n+1} = f(x_n)$ et un x_0 donné. Nous avons alors montré que cette suite converge vers une valeur qui n'est autre que $\sqrt{a}$. De façon générale, on peut chercher à suivre un principe d'approximations successives en construisant des suites de la forme

$$x_{n+1} = f(x_n),$$

pour $n \in \mathbb{N}$, avec $x_0 \in \mathbb{R}$ donné et nous nous intéresserons à la convergence de la suite $(x_n)_{n\geq 0}$ ainsi définie. Un résultat de convergence est donné par le théorème du point fixe qui nous fournit des conditions suffisantes pour la convergence de la suite $(x_n)_{n\geq 0}$.

Théorème 7.2.2 *Soit I un ensemble de $\mathbb{R}^d$ et f une application de I dans $\mathbb{R}^d$ telle que*

— *$f(I) \subset I$,*

— *il existe une constante $k \in [0, 1[$ telle que f est k-contractante, c'est-à-dire*

$$\forall (x, y) \in I^2,\ \|f(x) - f(y)\| \leq k\|x - y\|.$$

Alors la fonction f admet un unique point fixe $l \in I$ tel que

$$f(l) = l.$$

De plus, toute suite récurrente définie par

$$\begin{aligned} & x_0 \in I, \\ \textit{et } & x_{n+1} = f(x_n), \quad \forall n \in \mathbb{N}, \end{aligned}$$

converge vers l.

Exemple 7.2.5 *Pour la fonction F de (7.3), on a observé qu'on avait une racine dans l'intervalle $[1, 2]$ et une racine dans l'intervalle $[7, 8]$. Les performances de la méthode ne sont pas les mêmes dans les deux cas comme l'illustrent les itérations présentées dans le tableau 7.2, respectivement initialisées à 8 et 2.*

k	x_k	x_k
1	7.782271023383442	1.224032476578299
2	7.714930140950544	1.205607243906417
3	7.682084600165323	1.204670726723357
4	7.664303002069792	1.204620696293069
⋮	⋮	⋮
10	7.640956604817492	1.204617865207309
11	7.640524208045744	1.204617865207245
12	7.640265620540383	1.204617865207242
13	7.640110893202825	1.204617865207242
14	7.640018281330524	1.204617865207242

TABLE 7.2 – Deux comportements différents de la méthode du point fixe.

Exemple 7.2.6 *Suivant que f est croissante ou décroissante, la figure 7.3 montre les premières itérations d'une méthode de point fixe définie par f.*

Critère d'arrêt : On arrêtera ces itérations lorsque l'erreur relative entre 2 itérés successifs est suffisamment petite, c'est-à-dire que

$$|x_{n+1} - x_n| < \varepsilon |x_n|,$$

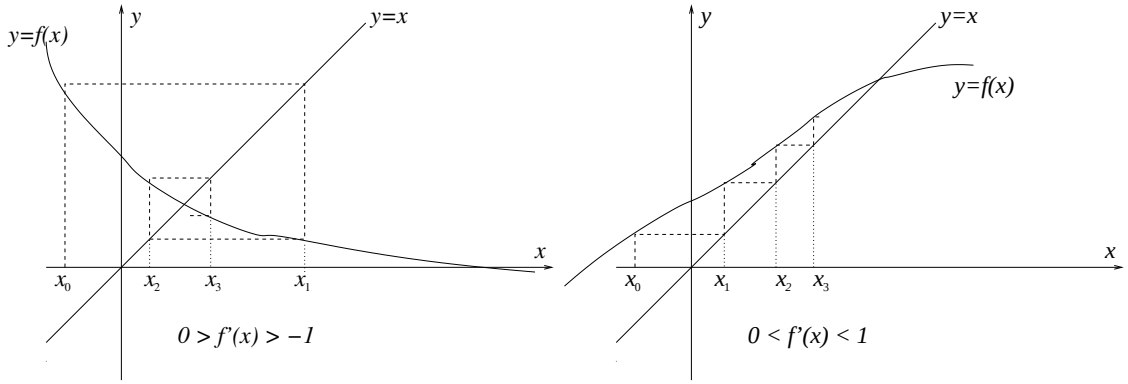


FIGURE 7.3 – Premiers itérés d'une méthode de point fixe.

$x = x_0$
Pour $k = 0, 1, \ldots,$itermax:
  $x_1 \leftarrow f(x)$
  Si $|x_1 - x| < \varepsilon|x|$:
    Arrêt de la boucle sur k
  $x \leftarrow x_1$

TABLE 7.3 – Algorithme de la méthode du point fixe

ou que le nombre maximal d'itérations est atteint. On remarque que la première condition est équivalente à

$$|F(x_n)/M| < \varepsilon|x_n|.$$

Algorithme : On a l'algorithme de la table 7.3 pour la méthode du point fixe.
Ordre de la méthode du point fixe : On peut évaluer l'ordre de la méthode du point fixe en introduisant l'erreur à l'étape n

$$e_n = |x_n - x|,$$

7.2.3 La méthode de Newton

Principe de la méthode de Newton : Supposons maintenant que, comme précédemment, une représentation graphique nous ait indiqué qu'une valeur x_0 est proche d'une racine x. Un développement de Taylor à l'ordre 1 au point x s'écrit

$$F(x) = F(x_0) + (x - x_0)\, F'(x_0) + \underset{|x-x_0|\to 0}{\mathcal{O}}(|x - x_0|^2). \tag{7.4}$$

On aura alors, puisque $F(x) = 0$,

$$F(x_0) + (x - x_0)\, F'(x_0) + \underset{|x-x_0|\to 0}{\mathcal{O}}(|x - x_0|^2) = 0 \tag{7.5}$$

Comme $|x - x_0|$ est assez petit, $\underset{|x-x_0|\to 0}{\mathcal{O}}(|x - x_0|^2)$ sera encore plus petit et on néglige donc ce terme. Cette approximation nécessite de remplacer $(x - x_0)$ par un certain $(x_1 - x_0) \neq (x - x_0)$ dans la formule et on obtient l'expression

$$F(x_0) + (x_1 - x_0)\, F'(x_0) = 0, \tag{7.6}$$

qui fournit ainsi

$$x_1 - x_0 = -\frac{F(x_0)}{F'(x_0)}.$$

On espère alors que x_1 sera plus proche de x que ne l'est x_0. On aura ainsi défini le premier itéré d'une nouvelle méthode en posant

$$x_1 = x_0 - \frac{F(x_0)}{F'(x_0)}.$$

Remarque 7.2.2 *Par ailleurs, on sait que la tangente à la courbe représentant $y = F(x)$ au point x_0 a pour équation*

$$y = F(x_0) + (x - x_0)\, F'(x_0). \tag{7.7}$$

On remarque alors que cette tangente coupe l'axe des x au point x_1 : le choix de x_1 est donc la solution d'un modèle linéaire de notre problème au voisinage de x_0 (voir la figure 7.4) !

Mise en oeuvre : La méthode de Newton va être définie en itérant cette démarche. L'algorithme consiste donc, partant de x_0 voisin d'une racine, à calculer pour $n \geq 0$:

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)}. \tag{7.8}$$

Critère d'arrêt : Comme on le verra par la suite (c.f. exemples 7.2.8 et 7.2.9), la méthode de Newton ne converge pas systématiquement. On pourra donc fixer

— un nombre maximum d'itérations,

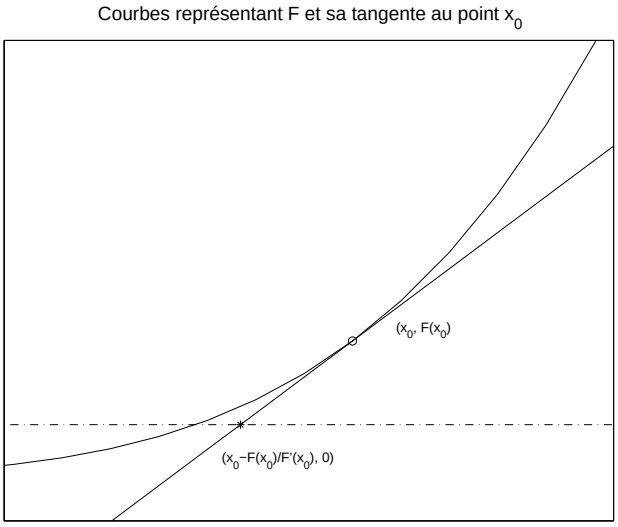


FIGURE 7.4 – La méthode de Newton choisit l'itéré suivant sur la tangente.

— un seuil de convergence, qui peut être
  — un test de convergence de la suite

$$|x_{n+1} - x_n| < \varepsilon_x |x_n|,$$

  — un test de réalisation de l'objectif

$$|F(x_n)| < \varepsilon_F,$$

  — une combinaison des précédents.

Algorithme : L'algorithme représenté dans la table 7.4 permet, à partir d'une valeur approchée x_0 de calculer une racine, à condition que la méthode converge vers cette racine. Attention, c'est une version très édulcorée (car en dimension 1).

$x = x_0$
Pour $k = 0, 1, \ldots,$itermax:
  $d \leftarrow -F(x)/F'(x)$
  $x \leftarrow x + d$
  Si $|d| < \varepsilon|x|$:
    Arrêt de la boucle sur k

TABLE 7.4 – Algorithme de la méthode de Newton en dimension 1

Exemple 7.2.7 *Nous l'avons appliqué à la recherche des zéros positifs de la fonction F de (7.3), partant respectivement des valeurs $x_0 = 2$ et $x_0 = 8$. Les valeurs calculées sont présentées au tableau 7.5. On voit que la suite converge à la précision de l'arithmétique en 4 ou 5 itérations.*

n	x_n	x_n
0	2.0000000000000000	8.0000000000000000
1	1.1607032574053444	7.7425762473069293
2	1.2049212788797627	7.6507460430283869
3	1.2046178784694039	7.6400156469715865
4	1.2046178652072419	7.6398801183259391
5	1.2046178652072419	7.6398800969514733
6	1.2046178652072419	7.6398800969514733

TABLE 7.5 – Deux convergences de la méthode de Newton.

Remarque 7.2.3 *La méthode de Newton est beaucoup plus rapide que la dichotomie. On peut aussi le voir en regardant les courbes représentant l'évolution de l'erreur des exemples 7.2.3 et 7.2.7 dans la recherche de la racine, sur la figure 7.5.*

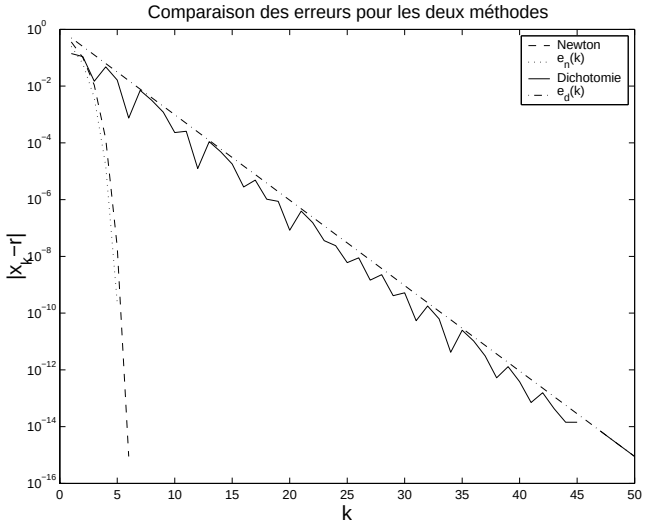


FIGURE 7.5 – Convergence comparée des méthodes de Newton et de dichotomie.

On y remarque que l'erreur pour chacune de ces méthodes est comparable à une courbe

- *pour la méthode de Newton, c'est la courbe représentant*

$$e_n(k) = \left(\frac{1}{2}\right)^{2^k},$$

- *pour la méthode de dichotomie, c'est la courbe représentant*

$$e_d(k) = \left(\frac{1}{2}\right)^k.$$

Pour la méthode de dichotomie, on retrouve le principe de la convergence linéaire et à chaque itération l'erreur est divisée par 2. Pour la méthode de Newton, c'est le principe observé pour la méthode de Héron que l'on retrouve. En effet, ici

$$|x_0 - r| \simeq \frac{1}{2},$$

et au-delà, pour $n \geq 0$*, la convergence d'ordre 2 de la méthode nous donne* $|x_{n+1} - r| \simeq |x_n - r|^2$*, ce qui entraine donc que*

$$|x_{n+1} - r| \simeq \left(|x_{n-1} - r|^2\right)^2.$$

Ainsi, en remontant jusqu'à $n = 0$*, on retrouve bien*

$$|x_n - r| \simeq \left(\frac{1}{2}\right)^{2^n}.$$

Ordre de la méthode de Newton : On peut montrer, que sous des hypothèses raisonnables, la méthode de Newton est, comme la méthode de Héron, d'ordre 2. C'est l'objet du théorème suivant.

Théorème 7.2.3 *Soit* f *une fonction définie sur* $\mathbb{R}$ *et à valeurs dans* $\mathbb{R}$ *et qui s'annule au moins en une racine* r*. On suppose qu'il existe un* $U =]a, b[$ *tel que*

- $r \in U$*,*
- *la fonction* f *est de classe* $\mathcal{C}^2$ *sur* U*,*
- *pour tout* $x \in U$*,* $f'(x) \neq 0$*,*
- *les itérés* $(x_n)_{n\geq 0}$ *de la méthode de Newton convergent vers* r*.*

Alors, pour n *suffisamment grand,*

$$|x_{n+1} - r| = \mathcal{O}(|x_n - r|^2).$$

Démonstration. Puisque la suite converge vers r, on sait que, pour un certain $N_U \in \mathbb{N}$, on a, $\forall n \geq N_U$,

$$x_n \in U.$$

Ainsi, pour tout $n \geq N_U$, la formule de Taylor-Lagrange permet d'écrire

$$0 = F(r) = F(x_n) + (r - x_n)\, F'(x_n) + \frac{1}{2}(r - x_n)^2\, F''(\eta_n),$$

avec $\eta_n \in U$, de sorte que

$$(x_n - r)\, F'(x_n) - F(x_n) = \frac{1}{2}(r - x_n)^2\, F''(\eta_n). \tag{7.9}$$

Par la définition (7.8) de la méthode, on a

$$x_{n+1} = x_n - \frac{F(x_n)}{F'(x_n)},$$

ce qui nous donne donc

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

Avec l'aide de (7.9), on obtient finalement

$$x_{n+1} - r = \frac{(r - x_n)^2\, F''(\eta_n))}{2\, F'(x_n)} = \mathcal{O}(|x_n - r|^2). \tag{7.10}$$

□

Remarque 7.2.4 *Il ne faut pas s'étonner de retrouver ici une propriété de la méthode de Héron. En effet, si l'on veut appliquer la méthode de Newton à la fonction* $F(x) = x^2 - 2$*,*

La vitesse de convergence observée ci-dessus peut laisser penser que la méthode de Newton est préférable à la méthode de dichotomie. Cependant, pour l'exemple traité, le choix de x_0 a permis la convergence, et nous avons cependant précisé, en présentant l'algorithme, qu'il devait être assez voisin de la racine cherchée afin d'assurer la convergence.

En fait, les choses sont un peu plus compliquées : on peut atteindre certaines racines en partant d'assez loin, et les manquer alors que l'on part d'assez près, soit pour converger vers une autre racine, soit pour diverger ! La méthode de Newton a donc une bonne vitesse de convergence mais elle n'est pas **robuste** puisqu'elle a tendance à avoir un comportement erratique.

Exemple 7.2.8 *Ainsi, en initialisant les itération à* $x_0 = 3$ *pour la fonction* F *de (7.3), la méthode va converger vers une racine indésirable, approximativement égale à* -2.35*.*

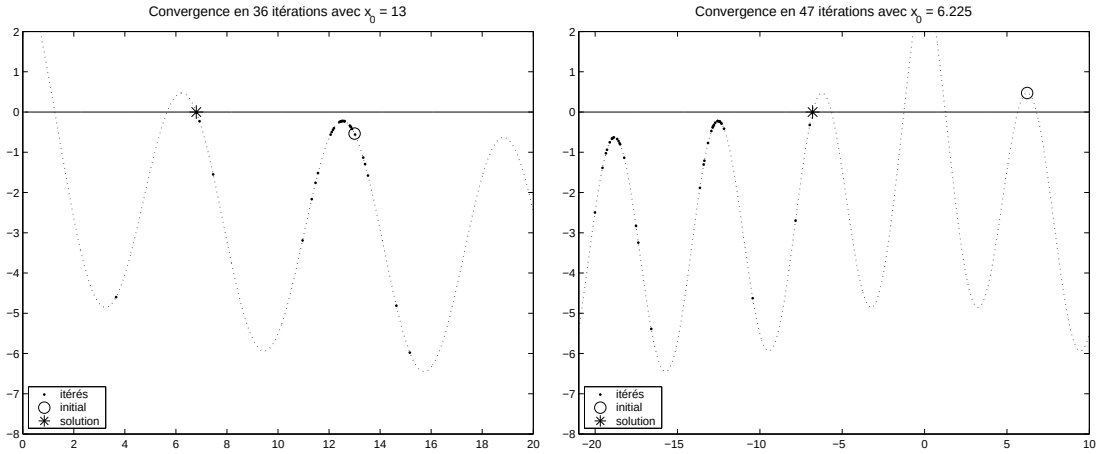


FIGURE 7.6 – On ne converge pas toujours vers la racine attendue !

Exemple 7.2.9 *Considérons la fonction*

$$F(x) = 3\,\cos(x) - \log|2\,x|\,. \tag{7.11}$$

La figure 7.6 montre une courbe représentant cette fonction ainsi les points $(x_n,\, F(x_n))$ *pour les itérés calculés avec deux choix différents de* x_0*.*

- *En partant de* $x_0 = 13$*, la méthode hésite un certain temps autour de* $x \simeq 12.53$ *en restant dans l'intervalle* $[0,\, 15]$ *jusqu'à l'itération* $n = 30$*, avant de finalement converger vers la solution approchée* $x \simeq 6.7988$ *en* 37 *itérations.*
- *En partant de* $x_0 = 6.225$*, la méthode commence par aller visiter l'intervalle* $[-20.5,\, -16.5]$*, hésitant jusqu'à l'itération* 19 *autour d'un "pic" de la fonction* $F(x)$ *approximativement situé au point* -18.8*. Puis, les* 20 *itérations suivantes vont osciller autour de la valeur* -12.5 *dans l'intervalle* $[-11.5,\, -13.5]$*, et finalement la convergence aura lieu vers* $x \simeq -6.7988$ *en* 47 *itérations.*

Dans les deux cas rapportés ci-dessus, la méthode a convergé, mais cela n'est pas garanti. Lorsque l'on part de $x_0 = 14$*, la méthode hésite jusqu'à l'itération* 18*, en restant d'abord autour du pic situé aux environs de* $x = 12.5$*, puis autour de celui que l'on trouve aux environs de* $x = 18.8$*, avec* $x_{18} \simeq 18.8307$*, avant de s'échapper jusqu'à* $x_{19} \simeq 198.744$*. Au-delà, si l'on suit l'algorithme présenté ci-dessus, les calculs ne s'interrompent pas, mais les itérations vont faire du "ping-pong" sur l'intervalle* $[198,\, 204]$*, autour du pic qu'il contient. La figure 7.7 montre les points* $(x_n,\, F(x_n))$ *pour* $k \leq 5000$ *! En fait, le point* $x_{18} \simeq 18.8307$ *est très proche d'un pic de* $F(x)$*, et* $F'(x_{18}) \simeq 0.0035$ *: la tangente à la courbe* $y = F(x)$ *en ce point est presque horizontale ! C'est ce qui explique que le point* x_{19} *est aussi éloigné de* x_{18}*.*

L'algorithme présenté ci-dessus n'est donc pas suffisant. Il faudra limiter le nombre d'itérations, et vérifier la convergence en sortie de boucle.

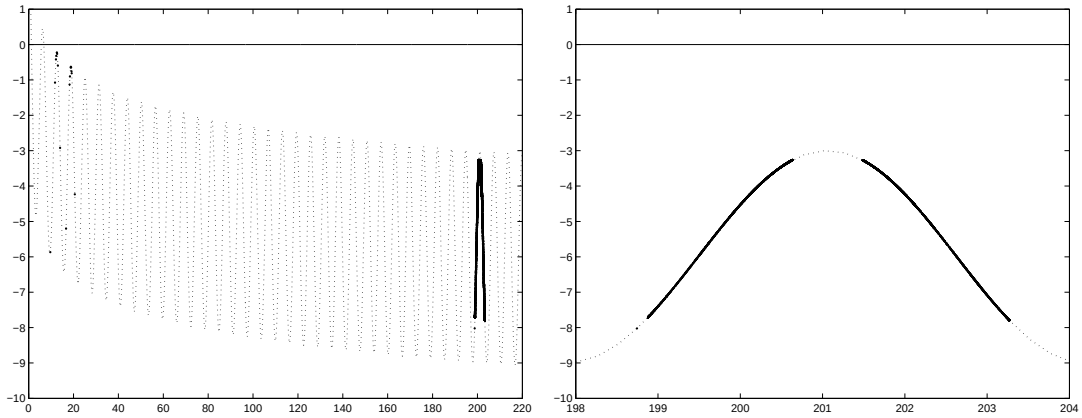


FIGURE 7.7 – À partir de la 19ème itération, les itérés restent autour d'un pic !

7.2.4 Combiner deux méthodes

Nous avons vu plusieurs méthodes et identifié deux qualités essentielles : la vitesse de convergence et la robustesse.

D'un côté, on a la méthode de Newton qui est incontestablement la meilleure en termes de vitesse lorsqu'on est assez proche de la racine recherchée. Par contre elle n'est pas très robuste, et sa convergence vers la racine cherchée dépend de la fonction et de la solution.

D'un autre côté, on a la méthode de dichotomie qui est beaucoup plus lente, mais elle est plus robuste : lorsque l'on a isolé une racine avec changement de signe, on ne risque plus rien.

On peut alors avoir l'idée de combiner les deux méthodes pour tirer profit des avantages de chacune.

Mise en oeuvre : La combinaison des 2 méthodes va se faire de la manière suivante.

1. À l'initialisation, on dispose d'un intervalle $[a_0,\ b_0]$ tel que $F(a_0)\,F(b_0) < 0$. On choisit une des bornes a_0 ou b_0 pour x_0. Puis on rentre dans une boucle sous le critère de convergence de la dichotomie (qui assure une certaine précision sur la solution calculée), c'est-à-dire sur la longueur de $[a_n, b_n]$.
2. À partir du point x_{n-1}, on fait une itération de Newton puis on vérifie si le nouveau point x_n n'est pas sorti de l'intervalle où est la racine. Si c'est le cas, on choisit de réinitialiser
$$x_n = \frac{a_{n-1} + b_{n-1}}{2}.$$
3. On calcule $F(x_n)$ et on procède par dichotomie pour localiser la racine.
   (a) Si $F(a_{n-1})F(x_n) < 0$ alors on pose
$$b_n = x_n \quad \text{et} \quad F(b_n) = F(x_n).$$

(b) Si $F(a_{n-1})F(x_n) > 0$ alors on pose

$$a_n = x_n \quad \text{et} \quad F(a_n) = F(x_n).$$

On voit donc que cet algorithme incorpore une étape de Newton dans la dichotomie. Cette étape vise à accélérer la méthode. D'autre part, on évite le problème lié à la robustesse de Newton en forçant la suite $(x_n)_{n\geq 0}$ à toujours être à l'intérieur de l'intervalle $[a_n, b_n]$ (où se situe la racine recherchée).

Algorithme : On en déduit donc l'algorithme de la table 7.6.

$fa \leftarrow F(a)$
$fb \leftarrow F(b)$
`Erreur si` $fa * fb > 0$
$x \leftarrow a$
$fx \leftarrow fa$
$fpx \leftarrow F'(x)$
`Tant que` $|b-a| > 2*\varepsilon*\max\{|a|,\ |b|\}$
  $x \leftarrow x - fx/fpx$
  `Si` $x < a$ `ou` $x > b$
    $x \leftarrow (a+b)/2$
  $fx \leftarrow F(x)$
  $fpx \leftarrow F'(x)$
  `Si` $fa * fx < 0$
    $b \longleftarrow x$
    $fb \longleftarrow fx$
  `Sinon`
    $b \longleftarrow x$
    $fb \longleftarrow fx$

TABLE 7.6 – Algorithme combinant la dichotomie et la méthode de Newton

Remarque 7.2.5 *Une autre idée de combinaison, plus élémentaire, consiste à suivre la mise en oeuvre suivante.*

1. *À l'initialisation, on dispose d'un intervalle* $[a_0,\ b_0]$ *tel que* $F(a_0)\,F(b_0) < 0$*. On lance la dichotomie jusqu'à obtenir une certaine précision (grossière)* $\varepsilon_1 > 0$ *sur la racine cherchée. On obtient*

$$x_1 = \frac{a_N + b_N}{2},$$

la solution approchée avec la dichotomie.

2. *Puis, on initialise la méthode de Newton avec x_1 qui est supposé être suffisamment proche de la racine. On lance alors la méthode avec une précision plus fine $\varepsilon_2 \ll \varepsilon_1$ et on obtient x_2 une solution très proche de la racine. Cette étape, qui aurait été longue avec la dichotomie, va être très rapide avec Newton.*

Dans l'idée précédente, la difficulté principale réside dans le choix de ε_1 : on doit le prendre suffisamment grand pour ne pas perdre trop de temps avec la dichotomie mais suffisamment petit pour que la méthode de Newton converge.