INSA INSTITUT NATIONAL DES SCIENCES APPLIQUÉES TOULOUSE

Outils Mathématiques pour l'Ingénieur II

2ème année FAS Construction

Florent Chazel, Aude Rondepierre, Adeline Rouchon

florent.chazel@insa-toulouse.fr,
aude.rondepierre@insa-toulouse.fr,rouchon@insa-toulouse.fr

Table des matières

1 Algèbre bilinéaire **7**

1.1 Formes bilinéaires 7

1.1.1 Définition 7

1.1.2 Formes bilinéaires et matrices 9

1.1.3 Formes bilinéaires symétriques 12

1.2 Espaces euclidiens 13

1.2.1 Produit scalaire 13

1.2.2 Norme euclidienne 14

1.2.3 Orthogonalité 17

1.2.4 Introduction aux espaces vectoriels normés 22

2 Moindres carrés **27**

2.1 Présentation générale 27

2.2 La méthode des moindres carrés 29

2.2.1 Première méthode : interpolation des données 29

2.2.2 Seconde méthode : moindres carrés 30

2.2.3 Caractérisation d'une solution au sens des moindres carrés 31

2.2.4 Résolution des équations normales 33

2.2.5 Moindres carrés non-linéaires 34

3 Intégrales multiples **35**

3.1 Intégrales doubles 35

3.1.1 Formule d'additivité 35

3.1.2 Autres propriétés 36

3.1.3 Théorème de la moyenne 36

3.1.4 Calcul des intégrales doubles 36

3.1.5 Changement de variables dans les intégrales doubles 39

3.2 Intégrales triples et multiples 41

3.2.1 Calcul d'intégrales triples 41

3.2.2 Changement de variable dans une intégrale triple 41

4 Interpolation Numérique **47**
4.1 Trois méthodes d'interpolation . . . 47
4.1.1 Méthode directe . . . 48
4.1.2 Méthode de Lagrange . . . 50
4.1.3 Méthode de Newton . . . 53
4.2 Erreur d'interpolation d'une fonction . . . 56
4.3 Phénomène de Runge . . . 57

5 Intégration Numerique **59**
5.1 Introduction . . . 59
5.2 Méthodes des rectangles . . . 61
5.3 Méthodes du point milieu . . . 63
5.4 Méthode des trapèzes . . . 64
5.5 Méthode de Simpson . . . 66

6 Séries de fonctions **69**
6.1 Différents types de convergence . . . 69
6.1.1 Convergence simple et convergence absolue . . . 69
6.1.2 Convergence uniforme . . . 70
6.1.3 Convergence normale . . . 70
6.2 Convergence uniforme et régularité . . . 71

7 Les séries entières **73**
7.1 Définition et propriétés de convergence . . . 73
7.1.1 Domaine et rayon de convergence . . . 73
7.1.2 Opérations sur les séries entières . . . 74
7.2 Applications et calculs de la somme de quelques séries . . . 76
7.2.1 Autour de la série géométrique . . . 76
7.2.2 Retour sur l'exponentielle . . . 76
7.2.3 Calcul de $\sin(x)$ avec une précision arbitraire . . . 77
7.3 Résolution d'équations différentielles à coefficients polynomiaux . . . 77

8 EDO linéaires d'ordre 1 et 2 **79**
8.1 Equations différentielles ordinaires linéaires du premier ordre . . . 80
8.1.1 Solution générale . . . 80
8.1.2 Résolution de l'équation homogène . . . 81
8.1.3 Recherche d'une solution particulière . . . 82
8.1.4 Exercices . . . 83
8.2 EDO linéaires du second ordre . . . 83
8.2.1 Solution générale . . . 84
8.2.2 Résolution de l'équation homogène . . . 85
8.2.3 Recherche d'une solution particulière . . . 87

9 Systèmes différentiels linéaires à coefficients constants 89
9.1 Exponentielle de matrice . . . 89
9.2 Résolution du système : $X'(t) = AX(t) + F(t)$ . . . 91
9.2.1 Solution du système homogène $X'(t) = AX(t)$ . . . 91
9.2.2 Recherche d'une solution particulière . . . 91

10 EDO : résolution numérique des pbs aux conditions initiales 93
10.1 Réduction d'ordre . . . 93
10.2 Approximation numérique des solutions : principe généraux . . . 95
10.2.1 Définitions . . . 95
10.2.2 Méthodes d'Euler . . . 96
10.2.3 Deux premiers exemples d'applications . . . 96
10.2.4 Méthode d'Euler améliorée . . . 100
10.2.5 Définition d'un schéma à un pas . . . 101
10.2.6 Méthode de Runge-Kutta . . . 102
10.3 Etude générale des méthodes à un pas . . . 103
10.3.1 Contexte . . . 103
10.3.2 Erreur globale et erreurs locales . . . 104
10.3.3 Ordre de consistance d'un schéma . . . 107
10.3.4 Stabilité du problème . . . 108
10.3.5 Convergence et ordre de convergence d'un schéma . . . 110
10.3.6 Ordre des méthodes à un pas les plus courantes . . . 110
10.3.7 Seuil de stabilité . . . 111

Chapitre 1

Algèbre bilinéaire

1.1 Formes bilinéaires

1.1.1 Définition

Soit E un espace vectoriel sur $\mathbb{K}$ ($\mathbb{R}$ ou $\mathbb{C}$). Une **forme bilinéaire** sur E est une application :

$$f : E \times E \to \mathbb{K}$$

qui satisfait :

$$\forall \lambda \in \mathbb{K}\,,\ \forall (v, w, v', w') \in E^4\,, \quad \begin{cases} f(\lambda v + v', w) = \lambda f(v, w) + f(v', w) & (1) \\ f(v, \lambda w + w') = \lambda f(v, w) + f(v, w') & (2) \end{cases}$$

On exprime la condition (1) en disant que f est linéaire par rapport à la première variable et la condition (2) en disant que f est linéaire par rapport à la deuxième variable.

Exemple 1.1.1

1. *Montrer que l'application $g : \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow K$ définie par :*

$$g((x, y), (x', y')) = xx' + yy' - xy' - x'y$$

est bilinéaire.

2. *Soient ϕ et σ des applications linéaires de E dans $\mathbb{K}$. L'application $f : E \times E \longrightarrow K$ définie par $f(x, y) = \phi(x)\sigma(y)$ est bilinéaire. La vérification est immédiate à partir de la linéarité de ϕ et de σ.*

3. *Soit $f : \mathbb{R}^N \times \mathbb{R}^N \longrightarrow \mathbb{R}$ définie par :*

$$f(x, y) = x^T Ay = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} x_i y_j \tag{1.1}$$

où est une A matrice de $M_N(\mathbb{R})$. On vérifie là aussi directement en utilisant les règles de calcul matriciel que cette application est bilinéaire. D'une certaine façon, c'est le prototype des applications bilinéaires comme nous le verrons ci-dessous. En fait, souvent les formes bilinéaires sont données sous la forme :

$$f(x, y) = \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} x_i y_j \tag{1.2}$$

On vérifie directement en écrivant :

$$\begin{aligned} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{ij} x_i y_j &= \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} \sum_{j=1}^{N} a_{1j} y_j \\ \vdots \\ \sum_{j=1}^{N} a_{Nj} y_j \end{bmatrix} \\ &= \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} \end{aligned}$$

et en identifiant les vecteurs de $\mathbb{R}^N$ à des matrices colonne que (1.2) définit une forme bilinéaire sur $\mathbb{R}^N$.

Exemple : $E = \mathbb{R}^3$ *et* $A = \begin{pmatrix} 4 & 0 & 5 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$.

1.1.2 Formes bilinéaires et matrices

Supposons ici que E est de dimension finie. Soit f une forme bilinéaire sur E et $\{e_1, ...e_N\}$ une base de E.
Soient x et y dans E tels que $x = x_1e_1 + \cdots + x_Ne_N$ et $y = y_1e_1 + \cdots + y_Ne_N$. On peut alors écrire :

$$\begin{aligned} f(x,y) &= f(x_1e_1 + \cdots + x_Ne_N, y_1e_1 + \cdots + y_Ne_N) \\ &= x_1y_1f(e_1,e_1) + x_1y_2f(e_1,e_2) + \cdots + x_Ny_Nf(e_N,e_N) \\ &= \sum_{i=1}^{N}\sum_{j=1}^{N} x_iy_jf(e_i,e_j) \end{aligned}$$

Ainsi f est complètement déterminée par les N^2 valeurs $f(e_i, e_j)$.

Définition 1.1.1 *La matrice* $A = (a_{ij})_{1\leq i\leq n}$ *où* $a_{ij} = f(e_i, e_j)$ *est appelée matrice de* f *dans la base* $\{e_i\}_{1\leq$

Si X et Y représentent les vecteurs colonne formés des composantes des vecteurs x et y dans la base $\{e_i\}$ alors :

$$\boxed{f(x,y) = X^TAY}\,.$$

Dans le second exemple ci-dessus de forme bilinéaire sur $\mathbb{R}^N$, la matrice A de la formule (1.1) est exactement la matrice de la forme bilinéaire dans la base canonique.

Exemple 1.1.2 *1. Ecrire la matrice de l'application* $g : \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow K$ *définie par*

$$g((x, y), (x', y')) = xx' + yy' - xy' - x'y$$

2. On considère l'application f *définie de* $\mathbb{R}_2[x] \times \mathbb{R}_2[x]$ *dans* $\mathbb{R}$ *par :*

$$\forall (P, Q) \in \mathbb{R}_2[x] \times \mathbb{R}_2[x], \quad f(P, Q) = \frac{1}{2} \int_0^1 (P(x)Q"(x) + P"(x)Q(x))dx$$

où $P"$ *et* $Q"$ *désignent les dérivées secondes de* P *et* Q*. Montrer que* f *est une forme bilinéaire symétrique sur* $\mathbb{R}_2[x]$*. Soit* $\mathcal{E} = (P_0, P_1, P_2)$ *la base canonique de* $\mathbb{R}_2[x]$ *où* $P_i(x) = x^i$*. Déterminer la matrice de* f *dans cette base.*

Si on pose $P(x) = x^2 - 3x + 1$ et $Q(x) = -2x^2 - 1$, déterminer $f(P, Q)$ en utlisant la matrice.

Que se passe-t-il si on fait un changement de base dans E ? Le théorème suivant indique la correspondance entre les matrices assurant la représentation.

Théorème 1.1.1 *Soit P la matrice de passage d'une base à une autre. Si A est la matrice de f dans la base initiale alors :*

$$\boxed{B = P^T AP}$$

est la matrice de f dans la nouvelle base. Les matrices A et B sont dites alors congruentes.

Démonstration. Soit $f(x, y) = X^T AY$ dans la base initiale et $f(x, y) = X'^T BY'$ dans la nouvelle base ;
on a $X = PX'$ et $Y = PY'$ d'où $f(x, y) = (X'^T P^T)A(PY') = X'^T(P^T AP)Y'$. □

Proposition 1.1.1 *Deux matrices congruentes ont même rang.*

Définition 1.1.2 *Soit f une forme bilinéaire. On appelle rang de f, noté $\mathrm{rang}(f)$ ou $\mathrm{rg}(f)$, le rang de n'importe quelle matrice représentant f.*

1.1.3 Formes bilinéaires symétriques

Définition 1.1.3 *Une forme bilinéaire f est dite symétrique si :*

$$\boxed{\forall (v,w) \in E^2\,,\ f(v,w) = f(w,v)\ .}$$

Conséquences : Si E est de dimension finie et si A est une représentation matricielle de f, on a $f(e_i, e_j) = f(e_j, e_i)$; d'où $A^T = A$ et donc que A est symétrique. Réciproquement, la forme bilinéaire définie par :

$$f(x,y) = x^T A y, \quad \forall\ (x,y) \in \mathbb{R}^N \times \mathbb{R}^N,$$

associée à une matrice $A \in M_N(\mathbb{R})$ symétrique est une forme bilinéaire symétrique.

Cas où $\mathbb{K} = \mathbb{R}$, forme bilinéaire positive, forme bilinéaire définie

Définition 1.1.4 *On dit que f est positive si :*

$$\boxed{\forall v \in E, \quad F(v,v) \geq 0}$$

On dit que f est définie si :

$$\boxed{\forall v \in E, \quad f(v,v) = 0 \Longleftrightarrow v = 0.}$$

Exemple 1.1.3 *Soit $g : \mathbb{R}^2 \times \mathbb{R}^2 \longrightarrow K$ la forme bilinéaire définie par :*

$$g((x,y),(x',y')) = xx' + yy' - xy' - x'y.$$

g est elle positive ? définie ?

1.2 Espaces euclidiens

Nous nous limiterons ici au cas des espaces vectoriels réels.

1.2.1 Produit scalaire

Définition 1.2.1 *Soit E un espace vectoriel sur $\mathbb{R}$ et f une forme bilinéaire sur E. On dit que f est un* **produit scalaire** *sur E si et seulement si f est symétrique définie et positive.*

Notation 1.2.1 *On utilise souvent la notation $\langle x, y\rangle$ ou $x \cdot y$ (au lieu de $f(x, y)$) pour désigner le produit scalaire de deux vecteurs x et y de E.*

L'équivalence suivante découle directement de la définition précédente :

$$\langle \cdot, \cdot \rangle \text{ est un produit scalaire sur } E$$
$$\Updownarrow$$
$$\begin{cases} \langle x, y\rangle \in \mathbb{R}\,, & \text{pour tout } x \text{ et } y \text{ dans } E, \\ \langle x + x', y\rangle = \langle x, y\rangle + \langle x', y\rangle\,, & \text{pour tout } x,\ x' \text{ et } y \text{ dans } E, \\ \langle \lambda x, y\rangle = \lambda \langle x, y\rangle\,, & \text{pour tout } x \text{ et } y \text{ dans } E \text{ et tout } \lambda \text{ dans } \mathbb{R}, \\ \langle x, y\rangle = \langle y, x\rangle\,, & \text{pour tout } x \text{ et } y \text{ dans } E, \\ \langle x, x\rangle \geq 0\,, & \text{pour tout } x \text{ dans } E, \\ \langle x, x\rangle = 0 \text{ si et seulement si } x = 0. \end{cases}$$

Remarque 1.2.1 *Dans l'équivalence précédente, on remarque que l'on montre uniquement la linéarité par rapport à la première variable : en effet, la propriété de symétrie permet d'en déduire directement la linéarité par rapport à la seconde variable, et donc la bilinéarité.*

Remarque 1.2.2 *Si F est un sous-espace vectoriel de E alors le produit scalaire de E induit un produit scalaire sur F.*

Définition 1.2.2 *Un espace vectoriel E muni d'un produit scalaire $\langle \cdot, \cdot \rangle$ est appelé espace préhilbertien réel.*

Si E est en plus de dimension finie, il est alors appelé **espace euclidien**.

1.2.2 Norme euclidienne

Définition

Définition 1.2.3 *Soit E est un espace préhilbertien réel et $\langle ., . \rangle$ le produit scalaire associé.*

On appelle **norme euclidienne** *sur E l'application associée au produit scalaire $\langle \cdot, \cdot \rangle$ par :*

$$||x|| = \sqrt{\langle x, x \rangle}, \quad \text{pour tout } x \in E.$$

Exemple 1.2.1 *Les exemples suivants décrivent des situations souvent rencontrées en pratique.*

1. *L'exemple suivant est particulièrement fondamental. Montrer que l'espace $\mathbb{R}^2$ est un espace euclidien pour le produit scalaire*

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2.$$

En identifiant les éléments de $\mathbb{R}^2$ à des vecteurs colonne (ce qu'on fera de façon implicite dans toute la suite)

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix},$$

le produit scalaire peut être défini à l'aide du produit matriciel

$$\langle x, y \rangle = x^T y.$$

le produit scalaire ci-dessus est le produit scalaire défini géométriquement lorsqu'on identifie les vecteurs du plan ou de l'espace au vecteur colonne formé par leurs composantes dans un repère orthonormé.

Dans le cas où $N \geq 2$, $\mathbb{R}^N$ est un espace euclidien pour le produit scalaire

$$\langle x, y \rangle = x_1 y_1 + \cdots + x_N y_N.$$

En identifiant les éléments de $\mathbb{R}^N$ à des vecteurs colonne (ce qu'on fera de façon implicite dans toute la suite)

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix},$$

le produit scalaire peut être défini à l'aide du produit matriciel

$$\langle x, y \rangle = x^T y.$$

2. *Montrer que l'ensemble E des fonctions à valeurs réelles, continues sur [a,b], est un espace préhilbertien réel pour le produit scalaire*

$$\langle f, g \rangle = \int_a^b f(x)g(x)dx.$$

3. *L'espace $\mathbb{R}_N[x]$ des polynômes à coefficients réels, de degré $\leq N$, est un sous-espace de l'espace préhilbertien E ci-dessus. Il est donc un espace euclidien pour le produit scalaire induit.*
   Si on pose $P_0(x) = 1$, $P_1(x) = x$ et $P_2(x) = x^2$, déterminer par exemple $||P_0||, ||P_1||$, $< P_0, P_1 >, < P_1, P_2 >$.

Propriétés de la norme euclidienne

L'inégalité suivante est fondamentale.

Théorème 1.2.1 (Inégalité de Cauchy-Schwarz) *Soit E un espace préhilbertien. On a alors :*

$$|\langle x, y\rangle| \leq ||x||\,||y||, \quad \textit{pour tout } x \textit{ et } y \textit{ dans } E.$$

Démonstration. Si $x = 0$ ou $y = 0$, il n'y a rien à démontrer puisque les deux membres sont nuls.
Dans le cas contraire, calculer : $\left\| \frac{x}{\|x\|} + \frac{y}{\|y\|} \right\|^2 =$

Puis $\left\| \frac{x}{\|x\|} - \frac{y}{\|y\|} \right\|^2 =$

On en déduit alors que :

$$-||x||\,||y|| \leq \langle x, y\rangle \leq ||x||\,||y||$$

d'où le résultat.

□

1.2.3 Orthogonalité

Définition

L'importance des espaces préhilbertiens par rapport aux espaces normés quelconques vient du fait qu'on peut y définir une notion d'orthogonalité.

Définition 1.2.4 *Deux vecteurs x et y de E sont dits* **orthogonaux** *si leur produit scalaire est nul :*

$$\langle x, y\rangle = 0.$$

Théorème 1.2.2 *Le seul vecteur orthogonal à tous les autres d'un espace préhilbertien E est le vecteur nul 0_E.*

Démonstration. Si x est orthogonal à tout y il est orthogonal à lui même ; et donc $\langle x, x\rangle = \|x\|^2 = 0$. Ce qui équivaut à $x = 0$. □

Théorème 1.2.3 (Théorème de Pythagore) *Soit E un espace préhilbertien réel et $\langle \cdot, \cdot \rangle$ le produit scalaire associé. On a alors :*

$$\boxed{\langle x, y\rangle = 0 \text{ si et seulement si } ||x+y||^2 = ||x||^2 + ||y||^2}$$

Démonstration. Montrons que $||x+y||^2 = ||x||^2 + 2\langle x, y\rangle + ||y||^2$.

□

Systèmes orthonormés

Définition 1.2.5 *Un système de vecteurs $\{e_i\}_{i\in I}$ de E est dit orthogonal si deux vecteurs différents quelconques de ce système sont orthogonaux, c'est à dire si :*

$$\langle e_i, e_j\rangle = 0 \text{ si } e_i \neq e_j.$$

Ce système est dit orthonormé si il est orthogonal et si chaque vecteur est unitaire (ou normé), c'est à dire si :

$$\langle e_i, e_j\rangle = 0 \text{ si } e_i \neq e_j \quad \textbf{et} \quad \|e_i\| = 1, \quad \forall i \in I.$$

Théorème 1.2.4 *Tout système orthogonal $\{e_i\}_{i\in I}$ qui ne contient pas 0_E est libre.*

Démonstration. Faisons la preuve pour 3 vecteurs :

Cas général : Soient $\{e_{i_1}, \ldots, e_{i_m}\}$ un sous système fini de $\{e_i\}_{i\in I}$ et soient λ_1, ..., λ_m, m nombres réels tels que :

$$\lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m} = 0_E.$$

On a d'abord :

$$\langle \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}, e_{i_k} \rangle = \lambda_{i_1} \langle e_{i_1}, e_{i_k} \rangle + \cdots + \lambda_{i_m} \langle e_{i_m}, e_{i_k} \rangle .$$

Dans le second membre seul le terme $\langle e_{i_k}, e_{i_k} \rangle = \|e_{i_k}\|^2 \neq 0$ car $0_E \neq e_i$, $\forall i \in I$. Il vient donc :

$$\langle \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}, e_{i_k} \rangle = \lambda_k \|e_{i_k}\|^2 .$$

Comme d'autre part :

$$\langle \lambda_1 e_{i_1} + \cdots + \lambda_m e_{i_m}, e_{i_k} \rangle = \langle 0_E, e_{i_k} \rangle = 0,$$

on obtient $\lambda_k \|e_{i_k}\|^2 = 0$. Ce qui entraine que $\lambda_k = 0$.

□

Produit scalaire et norme dans une base orthonormée

On se limite dans la suite pour simplifier à un espace euclidien E qu'on suppose être de dimension N.

Soit $\{e_i\}_{i=1}^{i=N}$ une famille orthornormée de E. Le théorème précédent montre qu'elle est libre : c'est donc **une base** de E. On dit que c'est **une base orthornormée** de E.

L'intérêt des bases orthonormées vient du fait que l'on peut exprimer le produit scalaire de deux vecteurs matriciellement : si on prend deux vecteurs x et y de E et qu'on les décompose sur une base orthonormée $\{e_i\}_{i=1}^{i=N}$ de E, on a alors :

$$x = \sum_{i=1}^{N} x_i e_i, \qquad y = \sum_{i=1}^{N} y_i e_i$$

et en utilisant l'expression des formes bilinéaires vue dans la section précédente, on déduit que :

$$\begin{aligned} \langle x, y \rangle &= \sum_{i=1}^{N} \sum_{j=1}^{N} x_i y_j \langle e_i, e_j \rangle \\ &= x_1 y_1 + \cdots + x_N y_N \\ &= \begin{bmatrix} x_1 & \cdots & x_N \end{bmatrix} \begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}, \end{aligned}$$

où on a utilisé le fait que $\langle e_i, e_j \rangle = 0$ si $e_i \neq e_j$ et $\|e_i\| = 1$ puisque la base est orthonormée.

Remarque 1.2.3 *La dernière égalité aurait pu être obtenue directement en montrant que la matrice du produit scalaire $\langle \cdot, \cdot \rangle$ dans la base orthonormée $\{e_i\}_{i=1}^{i=N}$ est égale à la matrice identité I, puis en utilisant l'expression $\langle x, y \rangle = X^T A Y$ vue à la section précédente, où X est la représentation colonne de x, Y celle de y, et A la matrice de $\langle \cdot, \cdot \rangle$ dans la base orthonormée $\{e_i\}_{i=1}^{i=N}$.*

La norme euclidienne dans E s'exprime alors à partir de la norme euclidienne de $\mathbb{R}^N$:

$$||x|| = \sqrt{\sum_{i=1}^{N} x_i^2}$$

Exemple 1.2.2 *Dans $\mathbb{R}^2$, les ensembles $\{u, v\}$ suivants sont ils des bases orthonormées ?*

1. $u = (1, 0)$, $v = (0, 1)$.
2. $u = (1, 1)$, $v = (-1, 1)$.

Procédé de Gram-Schmidt.

A partir d'une base quelconque $\{u_i\}_{i=1}^{i=N}$, on peut toujours construire une base orthonormée. Le procédé qui permet de le faire est appelé **procédé de Gram-Schmidt**.

Description du procédé de Gram-Schmidt :

1. On pose $e_1 = u_1 / \|u_1\|$.
2. On suppose que e_1, e_2,..., e_{k-1} ont été construits et engendrent u_1, u_2,..., u_{k-1},
   (a) on définit :
   $$\tilde{e}_k = u_k - \sum_{j=1}^{k-1} \langle u_k, e_j \rangle e_j.$$
   (b) On vérifie que $\tilde{e}_k$ est orthogonal à e_1, e_2,..., e_{k-1} :
   $$\begin{aligned} \langle \tilde{e}_k, e_i \rangle &= \left\langle u_k - \sum_{j=1}^{k-1} \langle u_k, e_j \rangle e_j, e_i \right\rangle \\ &= \langle u_k, e_i \rangle - \sum_{j=1}^{k-1} \langle u_k, e_j \rangle \langle e_j, e_i \rangle \\ &= \langle u_k, e_i \rangle - \langle u_k, e_i \rangle = 0. \end{aligned}$$
   (c) On vérifie que $\tilde{e}_k \neq 0$ et on le normalise. Si $\tilde{e}_k = 0$, $u_k = \sum_{j=1}^{k-1} \langle u_k, e_j \rangle e_j$ et donc u_k serait engendré par e_1, e_2,..., e_{k-1} et, par suite, par u_1, u_2,..., u_{k-1}. Ceci entraine une contradiction avec le fait que $\{u_i\}_{i=1}^{i=N}$ est une base. On peut donc poser :
   $$e_k = \tilde{e}_k / \|\tilde{e}_k\| .$$
   (d) On vérifie que e_1, e_2,..., e_k engendrent u_1, u_2,..., u_k. Il suffit de vérifier que e_1, e_2,..., e_k engendrent u_k. Or de l'égalité ci-dessus définissant $\tilde{e}_k$, on déduit :
   $$u_k = \sum_{j=1}^{k-1} \langle u_k, e_j \rangle e_j + \|\tilde{e}_k\| e_k$$
   ce qui exprime exactement que e_1, e_2,..., e_k engendrent u_k.

Exemple 1.2.3 *1. Appliquer ce procédé sur la famille $u = (1, 2)$, $v = (1, 0)$ de $\mathbb{R}^2$.*

2. *On considère dans* $\mathbb{R}^4$ *muni de sa structure euclidienne canonique, les sous espace vectoriel* $F = Vect(u_1, u_2, u_3)$ *où* $u_1 = (1, 1, 1, 1)$, $u_2 = (1, 0, 1, 0)$ *et* $u_3 = (1, 1, 2, 0)$ *dans la base canonique. Donner, par ce procédé, un base orthonormée de* F.

3. *On considère* $\mathbb{R}_2[X]$, *muni du produit scalaire :*

$$\langle P, Q\rangle = \frac{1}{2}\int_{-1}^{1} P(x)Q(x)dx.$$

On utilise Gram-Schmidt pour orthonormaliser la base canonique :

$$P_0 = 1,\ P_1 = X,\ P_2 = X^2.$$

1.2.4 Introduction aux espaces vectoriels normés

Définitions

Définition 1.2.6 *Soit E un $K-$espace vectoriel avec $K = \mathbb{R}$ ou $\mathbb{C}$. Une norme sur E est une application notée $\|.\|_E$ telle que :*

$$\begin{array}{rcl} \|.\|_E : E & \longrightarrow & \mathbb{R}^+ \\ x & \longmapsto & \|x\|_E \end{array}$$

et qui vérifie les trois propriétés suivantes :

1. $\|x\|_E = 0 \Leftrightarrow x = 0_E$
2. $\forall\ \lambda \in K,\ \|\lambda x\|_E = |\lambda|\,\|x\|_E$

3. $\forall\ (x,y) \in E \times E,\ \|x+y\|_E \leq \|x\|_E + \|y\|_E$.

On dit que $(E, \|.\|_E)$ *est un espace vectoriel normé (en abrégé evn).*

Remarque 1.2.4 $\|0_E\|_E = 0_K$.

Définition 1.2.7 *Soit* $(E,+,.)$ *un espace vectoriel normé par* $\|.\|_E$ *et muni d'une multiplication notée* $\times$ *qui soit distributive par rapport à l'addition. La norme* $\|.\|_E$ *est dite multiplicative si*

$$\forall (x,y) \in E \times E,\ \ \|x \times y\|_E \leq \|x\|_E \, \|y\|_E.$$

Le théorème suivant nous permet de montrer qu'une norme euclidienne est **une norme** (ouf !).

Théorème 1.2.5 *Soit* E *un espace préhilbertien réel. Sa norme euclidienne est une norme.*

Démonstration. Les propriétés 1. 2. sont la conséquence directe de la définition d'une norme euclidienne.

En utilisant la définition d'une norme euclidienne, on a :

$$\|x+y\|^2 = \|x\|^2 + 2\langle x,y\rangle + \|y\|^2 \leq \|x\|^2 + 2\,|\langle x,y\rangle| + \|y\|^2$$

L'inégalité de Cauchy-Schwarz donne alors :

$$\|x+y\|^2 \leq \|x\|^2 + 2\,\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2\ ;$$

d'où le résultat en passant aux racines carrées. □

Exemples

- Si $E = \mathbb{R}$, alors $\forall x \in \mathbb{R}$, $\|x\| = |x|$ est une norme multiplicative sur $\mathbb{R}$.
- Soit $E = \mathbb{R}^n$. Notons $x = (x_1, x_2, \cdots, x_n)$ un élément de $\mathbb{R}^n$. Les trois applications suivantes sont des normes sur $\mathbb{R}^n$:

$$\begin{aligned} ||x||_1 &= \sum_{i=1}^{n} |x_i|, \\ ||x||_2 &= \sqrt{\sum_{i=1}^{n} |x_i|^2}, \\ ||x||_\infty &= \sup_{1\leq i\leq n}(|x_i|) = \max_{1\leq i\leq n}(|x_i|)\,. \end{aligned}$$

- Soit $E = \mathcal{C}([a,b],\mathbb{R})$, l'espace vectoriel des fonctions réelles continues sur $[a,b]$. Les trois applications suivantes sont des normes sur $\mathcal{C}([a,b],\mathbb{R})$:

$$\begin{aligned} \forall f \in E, \quad ||f||_1 &= \int_a^b |f(t)|dt, \\ ||f||_p &= \left(\int_a^b |f(t)|^p dt\right)^{\frac{1}{p}}, \forall p > 0, \\ ||f||_\infty &= \sup_{t\in[a,b]} \left(|f(t)|\right). \end{aligned}$$

— Soit $E = \mathcal{M}_n(K)$, l'espace vectoriel des matrices carrées d'ordre n. Soit $A \in E$. On note a_{ij} les cœfficients de A. L'application suivante est une norme sur $\mathcal{M}_n(K)$:

$$||A||_\infty = \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}| \, .$$

Exemple 1.2.4

1. *Dans $\mathbb{R}^3$, soit v le vecteur de coordonnées $(1, 2, -1)$, calculer sa norme 1, sa norme 2 et sa norme infinie.*

2. *Soit $f \in \mathcal{C}([0, 1], \mathbb{R})$ la fonction définie par $f : x \mapsto x^2 - 1$. Calculer sa norme 1, sa norme 2 et sa norme infinie.*

3. *Dans $\mathcal{M}_2(\mathbb{R})$, soit $D = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$,calculer sa norme 1 et sa norme infinie.*

Chapitre 2

Moindres carrés

2.1 Présentation générale

La méthode des **moindres carrés** permet de comparer des **données expérimentales**, généralement entachées d'incertitudes ou d'aléa, à un **modèle mathématique** sensé décrire ces données. Quelques exemples typiques : la calibration de modèle pour la prédiction météo ou la bourse, la prédiction d'une trajectoire, ou bien encore la prédiction de l'évolution de la température du globe ou d'une pandémie.

J.C. Gauss (mathématicien allemand, 1777-1855) et A.-M. Legendre (mathématicien français, 1752-1833) ont été les premiers à définir et calculer des solutions au sens des moindres carrés. J.C. Gauss, âgé de 24 ans, a réussi l'exploit de prévoir de manière précise la trajectoire de l'asteroïde Céres sur la base de quelques observations (et sur l'hypothèse à l'époque originale d'une trajectoire elliptique et non circulaire).

Pour présenter la méthode des moindres carrés, considérons un jeu de données brutes qui peuvent être des **mesures** ou des **observations**. Elles sont représentées par m paires $(x_i, y_i)_{i=1,\dots,m}$, où x_i représente la variable (position, instant, ...) dont dépend la mesure et y_i la mesure elle-même. Un exemple de jeu de données est présenté sur la figure 2.1 dans le cas $m = 10$:

Dans la suite, on va chercher à **déterminer les n paramètres $\theta_1, \dots, \theta_n$ d'un modèle mathématique donné** $f(x, \theta_1, \dots, \theta_n)$ pour que ce modèle **représente au mieux ces observations**. On parle parfois d'apprentissage. Une fois le modèle recalé sur les observations, il pourra être utilisé pour **estimer des données manquantes**, soit entre les données existantes (interpolation), soit au delà des données (**prédiction, extrapolation**).

Essayons d'utiliser nos connaissances sur l'interpolation pour construire une interpolation de ces données. On peut effectuer l'une des interpolations suivantes :

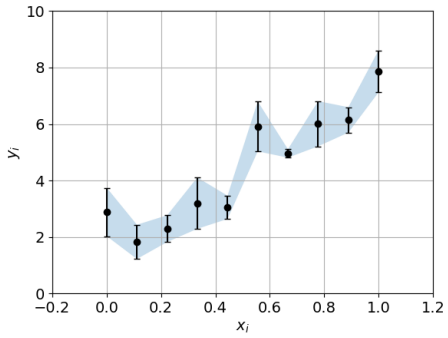


Figure 2.1 – Données brutes et estimation de leurs incertitudes

- **une interpolation polynomiale globale** : sachant qu'il y a m données, on peut choisir un polynôme d'interpolation de degré $m-1$. Avec la méthode de Van der Monde ou celle de Lagrange, on aboutit à un système linéaire de taille $m \times m$ qui admet une unique solution, pour peu que son rang soit plein : $rg(A_{VM}) = m$ (le rang peut ne pas être plein si deux mesures ont été prises pour la même variable x). Le résultat est présenté sur la figure 2.2 (gauche). Par définition, cette interpolation passe par tous les points mais les défauts sont nombreux. On n'a pas le choix du modèle mathématique (polynôme de degré m) et ce modèle dépend du nombre de points. Par ailleurs, dès lors que le nombre de points est important, l'extrapolation et l'interpolation n'ont plus beaucoup de sens puisque le phénomène de Runge produit des oscillations importantes entre et au delà des valeurs.

- **une interpolation polynomiale partielle** : à la vue des données, on peut être tenté de construire une interpolation linéaire (par un polynôme de degré 1). Pour le construire à partir du chapitre sur l'interpolation polynomiale, il faut deux points (degré du polynôme $+1$). On peut choisir par exemple, les premier et dernier points. Le résultat est présenté sur la figure 2.2 (au milieu). Dans ce cas, l'extrapolation est un peu plus robuste (si l'interpolation linéaire est physiquement réaliste), mais on accorde beaucoup d'importante à deux mesures et on n'exploite pas du tout toutes les autres. Cela peut être particulièrement préjudiciable dans le cas où les incertitudes (précision des mesures) sur ces deux points sont importantes.

- **une interpolation polynomiale par morceaux** : on peut construire une interpolation linéaire avec chaque paire de points, comme présenté sur la figure 2.2 (à droite). On aboutit à une interpolation par morceaux qui n'est pas régulière mais qui a l'avantage de passer par les points. Ici l'extrapolation en dehors du domaine est très hasardeuse car elle ne dépend que des dernières mesures.

On voit qu'aucune de ces techniques ne répond vraiment, de façon satisfaisante, à la question de déterminer les paramètres $\theta_1, \ldots, \theta_n$ d'un modèle mathématique donné $f(x, \theta_1, \ldots, \theta_n)$ pour

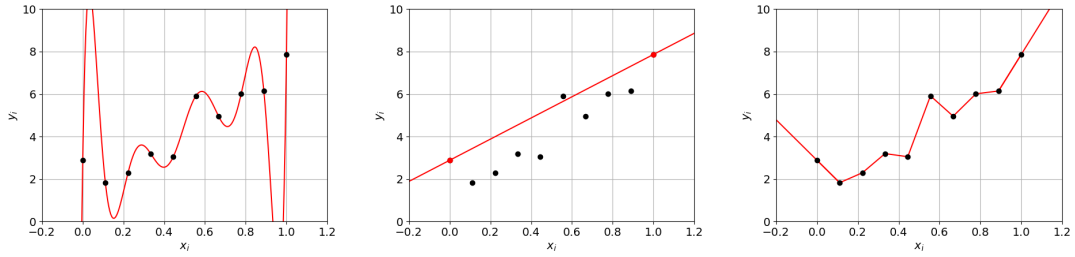
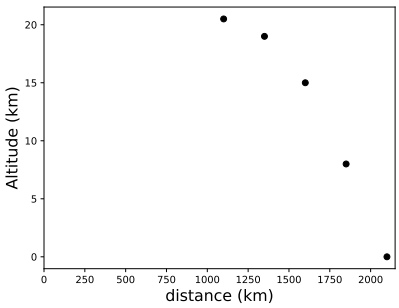


Figure 2.2 – Données brutes et interpolation

que celui-ci représente le mieux possible les données. C'est le but de la **méthode des moindres carrés**.

2.2 La méthode des moindres carrés

Afin d'introduire la méthode des moindres carrés, prenons l'exemple suivant : le défenseur d'une région a pour mission d'intercepter d'éventuels missiles. Il dispose de radars relevant les positions de missiles agresseurs, et doit prévoir leur trajectoire pour les intercepter.



Ses radars relèvent les positions suivantes d'un missile :

distance x_i (km)	2100	1850	1600	1350	1100
altitude y_i (km)	0	8	15	19	20

Les x_i désignent la distance à vol d'oiseau et les y_i l'altitude du missile, tous deux exprimés en kilomètres.

2.2.1 Première méthode : interpolation des données

Une première idée ici consiste à construire une interpolation polynomiale des données. Étant donné que nous avons les mesures à 5 instants diffèrents, on peut choisir une interpolation polynomiale de degré 4 :

$$p(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

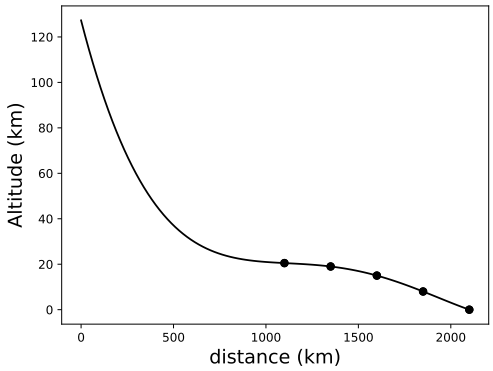
où les paramètres $\theta_0, \ldots, \theta_4$ doivent être déterminés en fonction des mesures existantes. Ces coefficients inconnus doivent vérifier chacun :

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \theta_3 x_i^3 + \theta_4 x_i^4 \,, \qquad \forall i \in \{1, ..., 5\}$$

Avec la méthode de Van der Monde, ceci nous donne un système de 5 équations à 5 inconnues :

$$\begin{pmatrix} 1 & x_1 & x_1^2 & x_1^3 & x_1^4 \\ 1 & x_2 & x_2^2 & x_2^3 & x_2^4 \\ 1 & x_3 & x_3^2 & x_3^3 & x_3^4 \\ 1 & x_4 & x_4^2 & x_4^3 & x_4^4 \\ 1 & x_5 & x_5^2 & x_5^3 & x_5^4 \end{pmatrix} \begin{pmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix}$$

où les (x_i, y_i) sont les couples du tableau de données provenant des radars. Ce système est carré et inversible pour peu que tous les points x_i soient différents deux à deux. On peut donc le résoudre afin de déterminer les coefficients du polynôme.



Sur la figure ci-dessus, on a tracé le polynôme obtenu après obtention des paramètres $\theta_0, \ldots, \theta_4$. On peut constater qu'il est **très délicat d'extrapoler une interpolation polynomiale en dehors du domaine** qui a servi à sa définition. Même si les points mesurés du missile sont interpolés par un polynôme de degré raisonnable (4 ici), on peut voir que la trajectoire obtenue semble correcte proche des mesures, mais complètement non-physique en dehors de l'intervalle de mesure (le missile s'envole dans la stratosphère).

2.2.2 Seconde méthode : moindres carrés

On suppose maintenant que ce défenseur possède de très bons ingénieurs qui lui affirment que la trajectoire des missiles est parabolique. Elle satisfait donc une équation de la forme :

$$y(x) = a_0 + a_1 x + a_2 x^2$$

Ceci est un modèle de trajectoire dont les paramètres a_0, a_1, a_2 doivent être déterminés en fonction des mesures existantes. Ces coefficients inconnus doivent ainsi chacun vérifier :

$$y_i = a_0 + a_1 x_i + a_2 x_i^2 \ , \qquad \forall i \in \{1, ..., 5\}$$

où les (x_i, y_i) sont les couples du tableau de données provenant des radars. En choisissant l'unité des mesures égale à la centaine de km (au lieu du km) afin de simplifier les écritures, on obtient alors le nouveau tableau de données suivant :

distance x_i (km)	21	18.5	16	13.50	11
altitude y_i (km)	0	0.08	0.15	0.19	0.2

ce qui nous donne le système linéaire suivant :

$$\left\{\begin{array}{ccc} + & + & = \\ + & + & = \\ + & + & = \\ + & + & = \\ + & + & = \end{array}\right. ,$$

que l'on peut réécrire sous forme matricielle $\mathbf{M}\,\mathbf{a} = \mathbf{y}$, où la matrice $\mathbf{M}$ est donnée par :

$$M = \left(\begin{array}{ccc} & & \\ & & \\ & & \\ & & \\ & & \end{array}\right) ,$$

et les vecteurs a et y sont donnés par :

$$a = (a_0, a_1, a_2)^T \quad \text{et} \quad y = (\quad , \quad , \quad , \quad , \quad)^T .$$

La matrice $\mathbf{M}$ est de taille 5×3 et dépend uniquement des données. Le second membre $\mathbf{y}$ dépend également uniquement des données. Nos inconnues ici sont les paramètres a_0, a_1, a_2 de la trajectoire, regroupés dans le vecteur $\mathbf{a}$. Il y a cependant un problème de taille ici : le système linéaire précédent est **sur-déterminé** : il y a plus d'équations que d'inconnues.

Ce système n'est pas inversible et on ne peut donc pas raisonnablement espérer en calculer une solution. On peut cependant chercher un vecteur $a \in \mathbb{R}^3$ tel que les vecteurs Aa et y (de $\mathbb{R}^5$) soient les plus proches possibles : on va donc chercher à rendre la quantité $||\mathbf{Aa} - \mathbf{y}||$ aussi petite que possible, ou de manière équivalente à minimiser $||\mathbf{Aa} - \mathbf{y}||^2$. C'est ce que l'on appelle un **problème de moindres carrés**, et plus précisément le problème de moindres carrés linéaire associé à la matrice A et au vecteur y.

2.2.3 Caractérisation d'une solution au sens des moindres carrés

On note ici $\langle \cdot, \cdot \rangle$ le **produit scalaire** canonique de $\mathbb{R}^5$ (mais les calculs suivants restent valables pour n'importe quel produit scalaire de $\mathbb{R}^5$).

Supposons que le problème de moindres carrés précédent (**minimiser** $||\mathbf{Aa} - \mathbf{y}||^2$) possède une solution : il existe donc un vecteur $\alpha \in \mathbb{R}^5$ tel que :

$$\forall a \in \mathbb{R}^5 , \quad ||Aa - y||^2 \geq ||A\alpha - y||^2 . \tag{2.1}$$

En notant $h = a - \alpha$, on peut écrire pour tout $a \in \mathbb{R}^5$ que :

$$\begin{aligned} ||Aa - y||^2 &= ||A(\alpha + h) - y||^2 \\ &= \\ &= \\ &= \\ &\qquad \text{(bilinéarité du produit scalaire)} \\ &= ||A\alpha - y||^2 + 2\langle A\alpha - y, Ah\rangle + ||Ah||^2 \qquad \text{(symétrie du produit scalaire)} \end{aligned}$$

Comme $||Ah||^2 \geq 0$, on voit que si on a la relation d'orthogonalité suivante :

$$\forall h \in \mathbb{R}^3 \quad \langle A\alpha - y\,,\, Ah\rangle = 0\,, \tag{2.2}$$

alors l'inégalité (2.1) $||Aa - y||^2 \geq ||A\alpha - y||^2$ sera vérifiée pour tout $h \in \mathbb{R}^3$, c'est à dire pour tout $a \in \mathbb{R}^3$, et donc que α est une solution du problème de moindres carrés que l'on cherche à résoudre.

En utilisant les connaissances vues dans le chapitre sur les espaces euclidiens, puisque la base canonique de R^3 est une base orthonormée pour le produit scalaire canonique, on peut écrire que :

$$\begin{aligned} \langle A\alpha - y\,,\, Ah\rangle &= \langle Ah\,,\, A\alpha - y\rangle \\ &= \\ &= \\ &= \\ &= \left\langle h\,,\, A^TA\alpha - A^Ty\right\rangle \end{aligned}$$

La relation d'orthogonalité (2.2) peut donc se réécrire :

$$\forall h \in \mathbb{R}^3 \quad \left\langle h\,,\, A^TA\alpha - A^Ty\right\rangle = 0\,,$$

ce qui entraîne que le vecteur $A^TA\alpha - A^Ty$ est orthogonal à tout vecteur $h \in \mathbb{R}^3$. Or d'après nos connaissances sur les espaces euclidiens, le seul vecteur orthogonal à tout les vecteurs de l'espace est le vecteur nul ! Ceci entraîne donc au final que :

$$A^TA\alpha - A^Ty = 0\,,$$

ou de manière équivalente :

$$\boxed{A^TA\alpha = A^Ty}\,.$$

Dans notre exemple, la matrice A^TA est une matrice carrée de taille 3, ce qui veut dire que le système précédent est un système à 3 équations et 3 inconnues : on l'appelle le système des **équations normales** et sa résolution permet de déterminer, lorsqu'elle existe, la solution du problème des moindres carrés associé à la matrice A et au vecteur y.

2.2.4 Résolution des équations normales

Si l'on calcule A^TA et A^Ty, le système des équations normales précédent s'écrit :

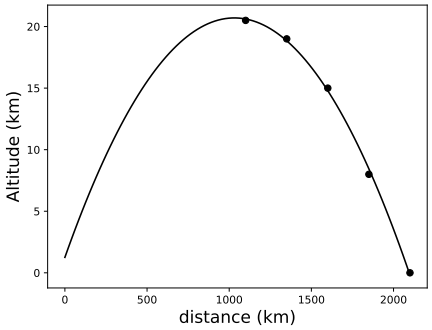
$$\begin{pmatrix} 5 & 25 & 187.5 \\ 25 & 187.5 & 1562.5 \\ 187.5 & 1562.5 & 13828.125 \end{pmatrix} \begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} 0.62 \\ 4.375 \\ 34.9375 \end{pmatrix} .$$

La résolution de système (par factorisation LU par exemple) nous donne la solution suivante :

$$\alpha = \begin{pmatrix} -0.00228571 \\ 0.03982857 \\ -0.00194286 \end{pmatrix}$$

On peut donc maintenant tracer la trajectoire complète du missile (cf figure 2.2.4), dont l'équation est :

$$y(x) = -0.00228571x^2 + 0.03982857x - 0.00194286 ,$$



Étant donné que la trajectoire obtenue est basée sur un modèle qui a du sens physique (trajectoire parabolique), son extrapolation en dehors des données est souvent assez fiable. Mais **le caractère prédictif dépend fortement du choix du modèle**. Et ce choix nécessite une certaine expertise !

De manière plus générale, on a le théorème suivant sur la résolution d'un problème de moindres carrés.

Théorème 2.2.1 *Soit $A \in \mathcal{M}_m, n(\mathbb{R})$ avec $m > n$ et $b \in \mathbb{R}^m$. Si le rang de A est maximal, autrement dit si $rg(A) = n$, alors les équations normales associées $A^TAx = A^Tb$ possèdent une unique solution $x^* \in \mathbb{R}^n$. Cette solution vérifie :*

$$x^* = \arg\min_{x \in \mathbb{R}^n} ||Ax - b||^2$$

et constitue donc l'unique solution du problème de moindres carrés linéaire associé à la matrice A et au vecteur b.

2.2.5 Moindres carrés non-linéaires

Dans l'exemple précédent, nous avons résolu un problème de moindres carrés dit **linéaire**, car le modèle mathématique choisi au départ **dépend linéairement des paramètres** a_0, a_1 et a_2. Mais ce n'est pas toujours le cas !

De manière générale, si l'on considère des données $(x_i, y_i)_{1 \leq i \leq n}$ et que l'on essaye de déterminer des paramètres $\theta_1, \ldots, \theta_n$ afin que le modèle mathématique $f(x, \theta_1, \ldots, \theta_n)$ représente au mieux les données, le problème de moindres carrés associé consiste alors en la recherche du vecteur $\theta = (\theta_1, \ldots, \theta_n)$ tel que :

$$\theta = \underset{t \in \mathbb{R}^n}{\arg\min} \sum_{i=1}^{n} (f(x_i, t_1, \ldots, t_n) - y_i)^2$$

Si la fonction f n'est pas linéaire en les paramètres $t_1, \ldots, t_n$, alors ce problème de moindres carrés est dit **non-linéaire**. Sa résolution ne peut dès lors pas s'effectuer à l'aide des équations normales. Dans certains cas, il est possible de linéariser le modèle mathématique et de se ramener à la résolution d'un problème de moindres carrés linéaires. Ceci sera vu lors de la séance de TP associée à ce chapitre.

Chapitre 3

Intégrales multiples

3.1 Intégrales doubles

Soit f une fonction de $\mathbb{R}^2$ dans $\mathbb{R}$ continue et bornée sur le domaine D, f est bornée signifie :

$$\exists M \in \mathbb{R}, \quad \forall (x,y) \in D, |f(x,y)| \leq M.$$

Pour définir l'intégrale de f sur D, comme en dimension 1, il faut construire une somme de Riemann. Cela peut s'effectuer sur des domaines "quadrables". On admet ici cette construction, on se concentre dans ce cours sur le calcul de ces intégrales.

3.1.1 Formule d'additivité

Elle généralise la relation de Chasles pour les intégrales simples

$$\int_a^b f(x)\,dx = \int_a^c f(x)\,dx + \int_c^b f(x)\,dx, \quad \text{pour } a < c < b.$$

Théorème 3.1.1 *Soient D un domaine de $\mathbb{R}^2$ et A et B deux domaines vérifiant*

$$D = A \cup B \text{ et aire de } A \cap B = \gamma.$$

où γ désigne une courbe. Alors, f est intégrable sur D si et seulement si f est intégrable sur A et sur B et

$$\int_D f(x,y)\,dxdy = \int_A f(x,y)\,dxdy + \int_B f(x,y)\,dxdy$$

Remarque 3.1.1 *Le théorème a deux conséquences importantes en pratique.*

1. *Il montre que les valeurs d'une fonction f sur une courbe n'a pas d'influence sur l'intégrabilité et la valeur de l'intégrale. Ceci est à comparer avec les valeurs en un point d'une fonction pour l'intégrale simple.*
2. *On peut simplifier le domaine d'intégration en le subdivisant en sous-domaines, ayant au plus une courbe en commun, et de forme plus simple.*

3.1.2 Autres propriétés

Les propriétés suivantes sont analogues à celles des intégrales simples. On suppose que f et g sont deux fonctions données intégrables sur D (ces propositions sont admises).

Proposition 3.1.1 (Linéarité de l'intégrale) *Pour λ et μ réels, on a*

$$\int_D (\lambda f + \mu g)(x,y)\,dxdy = \lambda \int_D f(x,y)\,dxdy + \mu \int_D g(x,y)\,dxdy.$$

Proposition 3.1.2 (Positivité de l'intégrale) *Si $f \geq 0$, alors*

$$\int_D f(x,y)\,dxdy \geq 0.$$

Corollaire 3.1.1 (Comparaison des intégrales) *Si $f \geq g$, alors*

$$\int_D f(x,y)\,dxdy \geq \int_D g(x,y)\,dxdy.$$

Corollaire 3.1.2 (Inégalité triangulaire) *On a si les intégrales sont définies*

$$\left|\int_D f(x,y)\,dxdy\right| \leq \int_D |f(x,y)|\,dxdy.$$

3.1.3 Théorème de la moyenne

Ce théorème sert à évaluer rapidement la valeur d'une intégrale sans avoir à la calculer exactement. Il est par exemple utilisé pour les dimensionnements dans les pré-études techniques.

Théorème 3.1.2 *Notons par Δ l'aire de D. La moyenne de f sur D vérifie*

$$\Delta \inf_{(x,y)\in D} f(x,y) \leq \int_D f(x,y)\,dxdy \leq \Delta \sup_{(x,y)\in D} f(x,y).$$

3.1.4 Calcul des intégrales doubles

En pratique, pour calculer

$$\int_S f(x,y)\,dxdy$$

on note

$$A_x = \{x \in \mathbb{R}, \exists y \in \mathbb{R}, (x,y) \in S\}$$

puis, pour $x \in A_x$ fixé, on pose

$$A_y(x) = \{y \in \mathbb{R}, (x,y) \in S\}$$

Alors

$$\int_S f(x,y)\,dxdy = \int_{A_x} \left(\int_{A_y(x)} f(x,y)\,dy \right) dx.$$

Ceci signifie que pour intégrer une fonction $f : \mathbb{R}^2 \to \mathbb{R}$, avec des hypotèses sur S, on peut intégrer séparément une variable puis l'autre.

Exemple 3.1.1 *En physique, on définit l'aire de S par*

$$\mathcal{A}_S = \int_S 1\,dxdy.$$

Effectuons alors les calculs de l'aire d'un carré de coté a et du triangle isocèle de base b et de hauteur h.

Aire du carré :

$$\int_S 1\,dxdy = \int_0^a \left(\int_0^a 1\,dy \right) dx =$$

Aire du triangle :

Ainsi, lorsque $x \in \left[0, \frac{1}{2}\right]$, y varie de 0 à $2x$, en effet

et lorsque $x \in \left[\frac{1}{2}, 1\right]$, y varie de

L'aire du triangle est :

$$\int_S 1\, dxdy =$$

Exemple 3.1.2 *On considère la fonction*

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

et le domaine

$$D = \{(x, y) \in \mathbb{R}^2 : 0 \leq x \leq 1, 0 \leq y \leq 1\} = [0, 1] \times [0, 1].$$

Alors, on a :

$$\int_D f(x, y)dxdy =$$

Exemple 3.1.3 *On considère maintenant le domaine suivant :*

$$D = \left\{(x,y) \in \mathbb{R}^2\,;\ 0 \leq y \leq 1,\ 0 \leq x \leq 1,\ \textit{et}\ x+y \leq 1\right\}.$$

et la fonction :

$$\begin{array}{rcl} f : \mathbb{R}^2 & \to & \mathbb{R} \\ (x,y) & \mapsto & 24xy. \end{array}$$

On a :
$\int_D f(x,y)dxdy =$

3.1.5 Changement de variables dans les intégrales doubles

Définition d'un changement de variable

Soient D et H deux domaines de $\mathbb{R}^2$. Une application $\Psi : D \to H \subset \mathbb{R}^2$ est caractérisée à l'aide de ses fonctions coordonnées

$$\Psi(x,y) = (u = \Psi_1(x,y), v = \Psi_2(x,y)).$$

On dit que $\Psi : D \to H$ définit un **changement de variable** de D sur H si les conditions suivantes sont vérifiées

1. Ψ est une bijection différentiable de D sur H,
2. Ψ est continue,
3. $\det J_\Psi(x,y) \neq 0,\ \forall (x,y) \in D$.

Exemple 3.1.4 (Coordonnées polaires) *Un exemple fondamental est celui des coordonnés polaires :*

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

$$\Delta = \{(r,\theta) \in]0,R] \times [0, 2\pi[\}$$

avec $\Psi : \Delta \to D$ *et* $x = r\cos\theta$ *et* $y = r\sin\theta$.

Vérifier que $\det J_\Psi(x,y) \neq 0$

Changement de variable dans une intégrale double

La formule permettant d'effectuer un changement de variable dans une intégrale double, est donnée par le théorème suivant (admis).

Théorème 3.1.3 *Soient* f, *une fonction intégrable sur* H, *et* $\Psi : D \to H$ *un changement de variable ; alors,* $v \mapsto f(\Psi(v))\,|\det J_\Psi(v)|$ *est intégrable sur* D *et l'on a*

$$\boxed{\int_H f(w)\,dw = \int_D f(\Psi(v))\,|\det J_\Psi(v)|\;dv.}$$

Exemple 3.1.5 *Calcul de l'aire du disque unité :*

3.2 Intégrales triples et multiples

Les intégrales triples et multiples se définissent et se calculent de manière analogue aux intégrales doubles. Nous donnerons quelques indications sur les intégrales triples en laissant le soin au lecteur d'effectuer les adaptations requises pour les intégrales de fonctions de plus de trois variables.

3.2.1 Calcul d'intégrales triples

Le calcul d'une intégrale triple sur le domaine d'intégration

$$D = \left\{(x, y, z) \in \mathbb{R}^3 ;\ g(x, y) \leq z \leq h(x, y),\ (x, y) \in A\right\}$$

se ramène par une intégrale simple au calcul d'une intégrale double

$$\int_D f(x, y, z)\, dxdydz = \int_A \int_{g(x,y)}^{h(x,y)} f(x, y, z)\, dz.$$

3.2.2 Changement de variable dans une intégrale triple

On a la même formule de changement de variable que pour les intégrales doubles où maintenant $\Psi : D \to H$ est un changement de variable entre deux domaines de $\mathbb{R}^3$

$$\boxed{\int_H f(w)\, dw = \int_D f(\Psi(v))\ |\det J_\Psi(v)|\ dv.}$$

Exemple 3.2.1 (Coordonnées cylindriques et sphériques) *Dans $\mathbb{R}^3$, deux examples très utilisés en physiques sont les coordonnés cylindriques et sphériques.*

— *Coordonnés cylindriques :*

Expression des coordonnées cartésiennes

$$\left[\begin{array}{c} x = r\cos\theta \\ y = r\sin\theta \\ z = z \end{array}\right]$$

Déterminer J_Ψ ainsi que son déterminant :

— *Coordonnées sphériques*

Expression des coordonnées cartésiennes

$$\begin{bmatrix} x = r\cos\theta\cos\varphi \\ y = r\sin\theta\cos\varphi \\ z = r\sin\varphi \end{bmatrix}$$

Déterminer J_Ψ ainsi que son déterminant :

Application. *Calcul du volume de la sphère de rayon* R

$$D = \left\{(x, y, z) \in \mathbb{R}^3 \,;\; x^2 + y^2 + z^2 \leq R^2\right\}$$

$$\Delta = \left\{(r, \theta, \varphi) \in]0, R] \times [0, 2\pi[\times \left]\frac{-\pi}{2}, \frac{\pi}{2}\right[\right\}$$

$\int_D dxdydz =$.

Exemple 3.2.2 *Une question se pose si le domaine est infini. Que ce passe-t-il si on veut intégrer une fonction sur* $\mathbb{R} \times \mathbb{R}$ *? Comme dans le cas réel des inTégrales généralisées, il ne suffit plus d'être continu, et on a besoin d'autres propriétés. Considérons la fonction :*

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

Nous admettons ici que l'intégrale sur $\mathbb{R} \times \mathbb{R}$ *de* f *est bien définie. Nous cherchons à la calculer explicitement par un changement en coordonnés polaires. Dans un premier temps, on remarque que par symétrie, on a :*

$$\int_{\mathbb{R}\times\mathbb{R}} f(x, y)dxdy = 4\int_0^{+\infty}\int_0^{+\infty} e^{-x^2-y^2}dxdy.$$

L'intégrale porte maintenant sur le quart de plan :

$$D = \{(x, y) \in \mathbb{R}^2 : x \geq 0, y \geq 0\}.$$

Ce domaine s'exprime à l'aide des coordonnés polaires :

$$D = \left\{(r, \theta) \in \mathbb{R}_+ \times \left[0, \frac{\pi}{2}\right]\right\}.$$

Comme $x^2 + y^2 = r^2$; il vient :

$$\begin{aligned} \int_{\mathbb{R}\times\mathbb{R}} f(x, y)dxdy &= 4\int_0^{+\infty}\int_0^{+\infty} e^{-(x^2+y^2)}dxdy \\ &= 4\int_0^{\frac{\pi}{2}}\int_0^{+\infty} e^{-r^2}rdrd\theta. \end{aligned}$$

Peut-être savez vous que la fonction $x \mapsto e^{-x^2}$ n'admet pas de primitives s'exprimant à l'aide des fonctions usuelles ? Mais il est possible de calculer la valeur de son intégrale sur $\mathbb{R}$ à l'aide du calcul précédent. Il suffit de remarquer que comme $e^{-x^2-y^2} = e^{-x^2}e^{-y^2}$, on a :

Remarque 3.2.1 *Bien qu'il ne soit pas possible de calculer explicitement une primitive de* $x \mapsto e^{-x^2}$*, on arrive, par des moyens détournés à calculer la valeur de son intégrale.*

Chapitre 4

Interpolation Numérique

Dans ce chapitre, nous allons considérer le problème de l'interpolation polynomiale puis l'intégration numérique, qui en découle. En effet, une fois une fonction f interpolée par un certain polynôme p, le calcul de l'intégrale du polynôme se fait immédiatement et fournit donc une approximation de l'intégrale de f.

4.1 Trois méthodes d'interpolation

Le principe de l'interpolation polynomial de Lagrange est simple : étant donnés une suite de points $(x_i, y_i)_{i\in I} \subset \mathbb{R}^2$, où I est un ensemble fini de $\mathbb{N}$, on cherche un polynôme p qui vérifie, $\forall i \in I$,

$$p(x_i) = y_i.$$

Géométriquement parlant, on veut que la courbe de $(x, p(x))_{x\in\mathbb{R}}$ passe par l'ensemble des points $(x_i, y_i)_{i\in I}$ dans le plan. Ces points peuvent provenir par exemple

- d'une fonction f compliqué dont l'évaluation est très coûteuse numériquement et, donc, que l'on a uniquement évalué sur un ensemble fini de points,
- du résultat d'un ensemble de mesures que l'on a effectué lors d'une expérience.

Dans les deux cas cités précédemment, on va en fait tenter de construire une approximation régulière (les polynômes étant de classe $\mathcal{C}^\infty$) d'une fonction inconnue. Étant donné qu'un polynôme, de degrés n, s'exprime complètement grâce à ces $n+1$ coefficients, il est préférable de stocker un polynôme d'interpolation dans un ordinateur là où un grand ensemble de points

ou une fonction quelconque ne peuvent être stockés de manière économique. Bien évidemment, la qualité de l'approximation va dépendre du nombre de coefficients (et donc du degré du polynôme) et il est ainsi nécessaire de trouver un bon compromis entre l'approximation et le degré du polynôme. Enfin, il est aussi désirable d'avoir une évaluation de l'erreur commise.

4.1.1 Méthode directe

On commence par rappeler la définition d'un polynôme.

Définition 4.1.1 *Soit $n \geq 0$. On note $\mathcal{P}_n$ l'ensemble des polynômes de degré n de $\mathbb{R}$ dans $\mathbb{R}$. Ainsi, $\forall p \in \mathcal{P}_n$, il existe un ensemble de coefficients réels $(a_k)_{0\leq k\leq n}$ tels que, $\forall x \in \mathbb{R}$,*

$$p(x) = \sum_{k=0}^{n} a_k x^k.$$

On se donne un ensemble D de $n+1$ points $D = \{(x_i, y_i)_{0\leq i\leq n}\}$ dans le plan $\mathbb{R}^2$ tels que

$$\forall j \neq k, \quad x_j \neq x_k, \tag{4.1}$$

et on cherche un polynôme p tel que, $\forall 0 \leq i \leq n$,

$$p(x_i) = y_i. \tag{4.2}$$

Il se trouve que, pour un nombre $n+1$ de points $(x_i, y_i)_{0\leq i\leq n}$ et un polynôme p de degrés m, le problème de l'interpolation polynomial (4.2) s'écrit comme un système de $n+1$ équations sur les $m+1$ coefficients puisque

$$\left\{\begin{array}{l} P(x_0) = y_0, \\ P(x_1) = y_1, \\ \vdots \\ P(x_n) = y_n. \end{array}\right. \Leftrightarrow \left\{\begin{array}{l} a_m x_0^m + a_{m-1} x_0^{m-1} + \ldots + a_0 = y_0, \\ a_m x_1^n + a_{m-1} x_1^{m-1} + \ldots + a_0 = y_1, \\ \vdots \\ a_m x_n^m + a_{m-1} x_n^{m-1} + \ldots + a_0 = y_n. \end{array}\right. , \tag{4.3}$$

ce que l'on peut donc écrire sous forme matricielle

$$\begin{bmatrix} 1 & x_0 & \cdots & x_0^{m-1} & x_0^m \\ 1 & x_1 & \cdots & x_1^{m-1} & x_1^m \\ \vdots & \vdots & & & \vdots \\ 1 & x_n & \cdots & x_n^{m-1} & x_n^m \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \Leftrightarrow Ma = y. \tag{4.4}$$

Remarque 4.1.1 *On voit que la matrice M du système* (4.4) *possède une structure bien particulière puisque, étant donné un vecteur $(x_0, \ldots, x_n)$ et $m \geq 0$, on peut l'identifier complètement. Ce type de matrice s'appelle* **matrice de Vandermonde**.

Ainsi, l'existence (et l'unicité) d'un polynôme d'interpolation de degré m pour un ensemble de $n+1$ points va dépendre de l'existence (et l'unicité) de solutions au système linéaire (4.4). On en déduit donc que

- si $n > m$, le problème est sur-déterminé et, si $y \notin \mathrm{Im}(M)$, il n'existe pas de solution,
- si $n < m$, le problème est sous-déterminé et il n'y a pas une unique solution.

Remarque 4.1.2 *En fait, on l'hypothèse* (4.1) *sur les points entraine que* $rang(M) = \min(n, m)$.

On va donc se focaliser plus particulièrement sur le cas $n = m$, c'est-à-dire lorsque le nombre de coefficients du polynôme est exactement le nombre de points que l'on veut interpoler. Le résultat suivant nous permet d'affirmer l'existence et l'unicité d'un polynôme p de degré n et passant par les $n + 1$ points D.

Lemme 4.1.1 *La matrice* M *du système* (4.4) *est inversible.*
Ce résultat est admis.

Grâce au résultat précédent, on sait qu'il est possible d'interpoler par un unique polynôme de degré n l'ensemble de points $D = \{(x_i, y_i)_{0 \leq i \leq n}\}$.

Exemple 4.1.1 *On se propose d'identifier le polynôme interpolant les points* D *donnés par*

$$D = \{(0, 7), (2, -3), (4, 2)\}.$$

On obtient donc le système suivant

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 2 & 4 \\ 1 & 4 & 16 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 7 \\ -3 \\ 2 \end{bmatrix}.$$

On applique la méthode du pivot de Gauss pour déduire le système équivalent

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 4 \\ 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 7 \\ -10 \\ 15 \end{bmatrix},$$

et trouver que

$$a_2 = 15/8, \quad a_1 = -35/4, \quad et \quad a_0 = 7.$$

On illustre cette interpolation dans le figure 4.1.

Malheureusement, comme on le sait, lorsque n est grand, résoudre un tel système peut se révéler coûteux. Par l'utilisation d'une factorisation LU, son coût est à peu près

$$C(n) \approx \frac{2}{3} n^3.$$

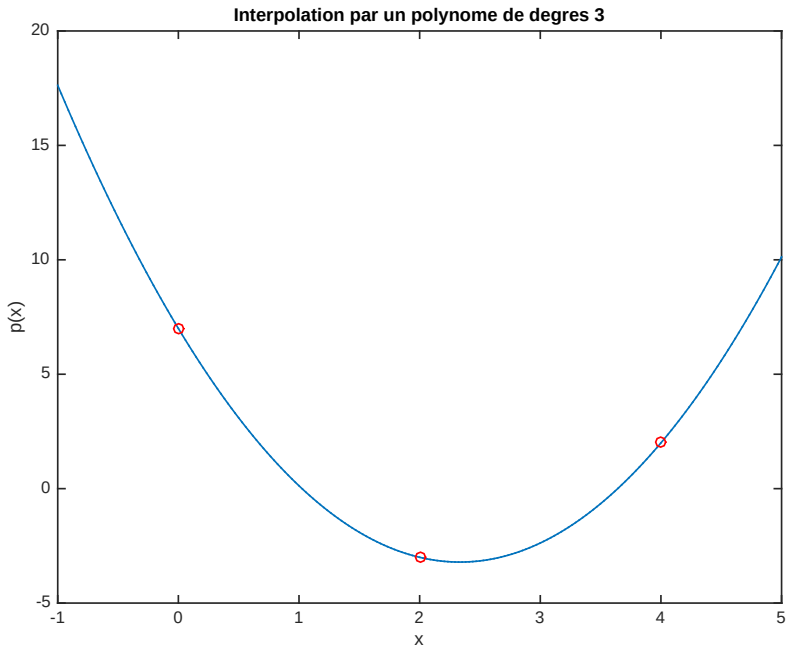


Figure 4.1 – Exemple d'interpolation polynomiale de 3 points par un polynôme de degré 3

4.1.2 Méthode de Lagrange

Heureusement, il existe une approche plus astucieuse. Le prochain résultat nous donne explicitement ce polynôme, que l'on appelle **polynôme de Lagrange**. L'interpolation d'un ensemble de points par un tel polynôme s'appelle **l'interpolation polynomiale de Lagrange**.

Théorème 4.1.1 *Soit $n \geq 0$ et $D = \{(x_i, y_i)_{0 \leq i \leq n}\}$ un ensemble de points. Alors il existe un ensemble de polynômes $(L_k)_{0 \leq k \leq n} \in \mathcal{P}_n$ tels que*

$$L_k(x_i) = \begin{cases} 1 \text{ si } i = k, \\ 0 \text{ sinon.} \end{cases} \tag{4.5}$$

Ces polynômes sont explicitement donnés, $\forall 0 \leq k \leq n$, par la formule, $\forall x \in \mathbb{R}$,

$$L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}. \tag{4.6}$$

Enfin, le polynôme d'interpolation p passant par l'ensemble D est alors donné par

$$p(x) = \sum_{j=0}^{n} y_j L_j(x). \tag{4.7}$$

Démonstration. On commence par construire les polynômes $(L_k)_{0\leq k\leq n}$ à partir de la relation (4.5). On voit que, $\forall 0 \leq k \leq n$, le polynôme L_k admet n racines qui sont les $(x_i)_{\substack{0\leq i\leq n \\ i\neq k}}$. Ainsi, il prend la forme factorisée

$$L_k(x) = C_k \prod_{\substack{j=0 \\ j\neq k}}^{n} (x - x_j), \tag{4.8}$$

où $C_k \in \mathbb{R}$ est une constante. Pour déterminer C_k, on utilise la relation

$$\begin{aligned} & L_k(x_k) = 1 \\ \Leftrightarrow\ & C_k \prod_{\substack{j=0 \\ j\neq k}}^{n} (x_k - x_j) = 1 \\ \Leftrightarrow\ & C_k = \left(\prod_{\substack{j=0 \\ j\neq k}}^{n} (x_k - x_j) \right)^{-1} = \prod_{\substack{j=1 \\ j\neq k}}^{n} (x_k - x_j)^{-1}. \end{aligned}$$

On en déduit ainsi la formule (4.6) en remplaçant C_k par l'expression précédente. Il reste donc à vérifier que le polynôme p donné par (4.10) interpole bien l'ensemble D. Pour cela, il suffit de voir que, $\forall 0 \leq i \leq n$,

$$p(x_i) = \sum_{j=0}^{n} y_j L_j(x_i) = y_i L_i(x_i) = y_i.$$

□

Le résultat précédent donne en fait la solution explicite du système (4.4) dans le cas $n = m$. Lorsque l'on veut faire de l'interpolation polynomiale de Lagrange d'une fonction f avec un ensemble de points d'abscisses

$$D_1 = \{x_0, x_1, \ldots, x_n\},$$

on pose, $\forall 0 \leq i \leq n$,

$$y_i = f(x_i),$$

et le polynôme de Lagrange p est alors donné par

$$p(x) = \sum_{j=0}^{n} f(x_j) L_j(x).$$

Exemple 4.1.2 *On veut interpoler la fonction $f(x) = cos(x)$ sur l'ensemble de points d'abscisses*

$$D_1 = \{-\pi/2, 0, 3\pi/4\}.$$

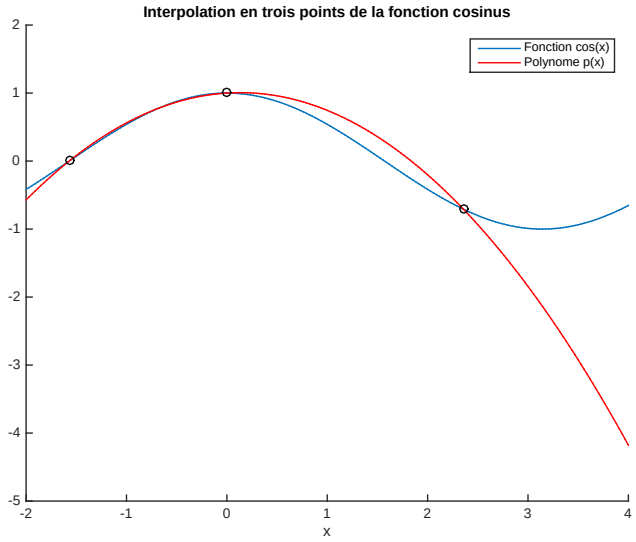


Figure 4.2 – Exemple d'interpolation polynomiale de Lagrange de la fonction cosinus en 3 points

On a donc

$$f(-\pi/2) = 0, \quad f(0) = 1 \quad \textit{et} \quad f(3\pi/4) = -\frac{\sqrt{2}}{2}.$$

Ceci nous permet de calculer les 3 polynômes

$$\begin{aligned} L_0(x) &= \frac{x(x-3\pi/4)}{-\pi/2(-\pi/2-3\pi/4)} = \frac{8}{5\pi^2}x(x-3\pi/4) \\ L_1(x) &= \frac{(x+\pi/2)(x-3\pi/4)}{\pi/2(-3\pi/4)} = -\frac{8}{3\pi^2}(x+\pi/2)(x-3\pi/4) \\ L_2(x) &= \frac{x(x+\pi/2)}{(3\pi/4)(3\pi/4+\pi/2)} = \frac{16}{15\pi^2}x(x+\pi/2), \end{aligned}$$

et, ainsi, de former le polynôme d'interpolation de Lagrange

$$\begin{aligned} p(x) &= L_1(x) + \frac{\sqrt{2}}{2}L_2(x) \\ &= -\frac{8}{3\pi^2}(x+\pi/2)(x-3\pi/4) - \frac{8\sqrt{2}}{15\pi^2}x(x+\pi/2) \\ &= \frac{1}{15\pi^2}(\pi+2x)(15\pi-4x(\sqrt{2}+5)) \end{aligned}$$

ou sous forme développée $p(x) = \frac{1}{15\pi^2}(\pi+2x)(15\pi-4x(\sqrt{2}+5))$. *L'interpolation obtenue est donnée dans la figure 4.2.*

Exemple 4.1.3 *Déterminer le polynôme d'interpolation vérifiant* $f(-1) = 2$, $f(0) = 0$ *et* $f(1) = 4$ *avec la méthode de Lagrange.*

Le théorème 4.1.1 a répondu de satisfaisante à notre problème : pour un ensemble de points D, on construit explicitement le polynôme d'interpolation de Lagrange p. On établit un premier algorithme qui permet l'évaluation d'un polynôme de la base $(L_k)_{0\leq k\leq n}$ dans la table 4.1.

En suivant cet algorithme, on peut déterminer le coût de l'évaluation du polynôme de Lagrange en un point est

$$C(n) \approx 2n^2.$$

Remarque 4.1.3 *Grâce à la formule* (4.6)*, on peut se rendre compte que l'évaluation d'un polynôme de* $(L_k)_{0\leq k\leq n}$ *ou le calcul des coefficients du polynôme de Lagrange risquent d'entraîner des erreurs d'élimination totale (que l'on a vu au premier semestre).*

4.1.3 Méthode de Newton

Dans la section précédente, on a vu que, étant donné l'ensemble de points d'abscisses $D_1 = \{x_0, x_1, \ldots, x_n\}$ sur lesquels on fait l'interpolation, on a construit le polynôme $p \in \mathcal{P}_n$ comme une combinaison linéaire des polynômes $(L_k)_{0\leq k\leq n}$ qui forme une base pour cette interpolation. Ce choix de polynômes n'est pas unique et il est possible de se placer sous une autre base. On a le résultat suivant qui est une alternative intéressante au théorème 4.1.1 dans une base de polynômes différente, que l'on appelle les **polynômes de Newton**, et qui donne lieu à **l'interpolation de Newton**.

```
Pour k allant de 0 jusqu'à n
    ‖   Pour i allant de 0 jusqu'à k − 1
    ‖       ‖   α_{k,i} ← 1/(x_k − x_i)
    ‖       ‖   α_{i,k} ← −α_{i,k}
Fonction w = P(x, y, α):
    |   w ← 0
    |   Pour k allant de 0 jusqu'à n
    |       ‖   β_k ← (x − x_k)
    |   Pour k allant de 0 jusqu'à n
    |       ‖   L_k ← y_k
    |       ‖   Pour i allant de 0 jusqu'à n
    |       ‖       ‖   Si i ≠ k
    |       ‖       ‖       ‖   L_k ← L_k * β_k * α_{k,i}
    |       ‖   w ← w + L_k
```

Table 4.1 – Algorithme permettant l'évaluation du polynôme de Lagrange en un point x *via* les polynômes $(L_k)_{0\leq k\leq n}$

Théorème 4.1.2 *Soit $n \geq 0$ et $D = \{(x_i, y_i)_{0\leq i\leq n}\}$ un ensemble de points. Alors il existe un ensemble de polynômes unitaires $(N_k)_{0\leq k\leq n} \in \mathcal{P}_n$ tels que, $\forall 0 \leq k \leq n$,*

$$\begin{cases} N_k \in \mathcal{P}_k, \\ N_k(x_i) = 0 \quad \textit{si} \quad i < k, \end{cases} \tag{4.9}$$

Ces polynômes sont explicitement donnés, pour $k = 0$ et $\forall x \in \mathbb{R}$, par

$$N_0(x) = 1,$$

et, $\forall 1 \leq k \leq n$ et $\forall x \in \mathbb{R}$, par la formule

$$N_k(x) = \prod_{j=0}^{k-1}(x - x_j).$$

Enfin, le polynôme d'interpolation p passant par l'ensemble D est alors donné par

$$p(x) = \sum_{j=0}^{n} z_j N_j(x). \tag{4.10}$$

où les coefficients $z = (z_i)_{0\leq i\leq n}$ sont solutions du système triangulaire inférieur

$$\begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & (x_1 - x_0) & 0 & \cdots & 0 \\ 1 & (x_2 - x_0) & (x_2 - x_0)(x_2 - x_1) & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & 0 \\ 1 & (x_n - x_0) & (x_n - x_0)(x_n - x_1) & \cdots & \prod_{j=0}^{n-1}(x_n - x_j) \end{bmatrix} \begin{bmatrix} z_0 \\ z_1 \\ \vdots \\ z_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}. \tag{4.11}$$

Démonstration. La construction des polynômes $(N_k)_{0\leq k\leq n}$ à partir de la relation (4.9) est immédiate en reprenant les mêmes arguments que la preuve du théorème 4.1.1. Pour obtenir le système linéaire dont z est la solution, il suffit d'utiliser les conditions d'interpolation et la seconde relation de (4.9) pour obtenir que

$$\begin{cases} p(x_0) = z_0 N_0(x_0) = y_0, \\ p(x_1) = z_0 N_0(x_1) + z_1 N_1(x_1) = y_1, \\ \vdots \\ p(x_n) = z_0 N_0(x_n) + z_1 N_1(x_n) + \cdots + z_n N_n(x_n) = y_n, \end{cases},$$

ce qui est bien le système triangulaire attendu.

□

Exemple 4.1.4 *Déterminer le polynôme d'interpolation vérifiant* $f(-1) = 2$, $f(0) = 0$ *et* $f(1) = 4$ *avec la méthode de Newton.*

On a l'algorithme d'évaluation des polynômes $(N_k)_{0\leq k\leq}$ en un point x donné par la table 4.2.

Étant donné un ensemble de points D, on suppose que l'on calcule les polynômes $(N_k)_{0\leq k\leq n}$ et le vecteurs $z \in \mathbb{R}^{n+1}$. L'intérêt de l'interpolation de Newton réside tout d'abord dans le fait que, si l'on veut ajouter un nouveau point d'interpolation, il suffit simplement de calculer le nouveau polynôme N_{n+1} et le nouveau point z_{n+1} puis les ajouter à ceux déjà trouvés sans avoir à tout recalculer.

```
Pour k allant de 0 jusqu'à n
    ‖  Pour i allant de 0 jusqu'à k
    ‖      ‖  Si i = 0
    ‖      ‖      ‖  T_{k,i} ← 1
    ‖      ‖  Si i ≠ 0
    ‖      ‖      ‖  T_{k,i} ← T_{k,i-1} * (x_k − x_{i-1})
Résoudre Tz = y
Fonction w = P(x, y, α):
    |  w ← 0
    |  Pour k allant de 0 jusqu'à n
    |      ‖  N_k ← z_k
    |      ‖  Si k ≠ 0
    |      ‖      ‖  Pour i allant de 0 jusqu'à k − 1
    |      ‖      ‖      ‖  N_k ← N_k * (x − x_i)
    |      ‖  w ← w + N_k
```

Table 4.2 – Algorithme permettant l'évaluation du polynôme de Newton en un point x *via* les polynômes $(N_k)_{0\leq k\leq n}$

On en déduit le coût pour l'évaluation du polynôme de Lagrange en un point est :

$$C(n) = \frac{n^2}{2}.$$

On voit donc que le coût de l'évaluation même est inférieur à celui de l'algorithme 4.1.

4.2 Erreur d'interpolation d'une fonction

Pour une fonction f et un ensemble de points d'abscisses D_1 donnés, même si f et son polynôme d'interpolation coïncident exactement aux points d'interpolation, ils peuvent être relativement éloignés lorsque l'on s'éloigne de ces points. En notant

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

où p est le polynôme d'interpolation de Lagrange de f, on peut se demander quel est l'ordre de grandeur de $|e(x)|$ lorsque $x \notin D_1$. Le prochain résultat nous permet d'obtenir borne sur cette valeur.

Théorème 4.2.1 *Soit $n \geq 0$, f une fonction de classe $\mathcal{C}^{n+1}$ définie sur un intervalle $[a, b]$ et $D_1 = \{x_0, x_1, \ldots, x_n\} \subset [a, b]$. Alors, $\forall x \in [a, b]$, il existe $\eta(x) \in]a, b[$ tel que*

$$e(x) = f(x) - p(x) = \frac{f^{(n+1)}(\eta(x))}{(n+1)!} \prod_{j=0}^{n} (x - x_j), \tag{4.12}$$

où p est le polynôme d'interpolation de Lagrange de f sur D_1. Ce résultat est admis.

Le résultat précédent nous donne donc que, si la fonction f est suffisamment dérivable, on peut espérer avoir une erreur assez petit. En effet, pour $f \in \mathcal{C}^\infty$, on a, lorsque $n \to \infty$, une erreur de plus en plus petite puisque

$$\frac{\sup_{y\in[a,b]} \prod_{j=0}^{n} |y - x_j|}{(n+1)!} \leq \frac{(b-a)^{n+1}}{(n+1)!} \underset{n\to\infty}{\to} 0.$$

4.3 Phénomène de Runge

En pratique, le terme $\sup_{y\in[a,b]} |f^{(n+1)}(y)|$ joue énormément dans l'erreur d'interpolation. Ceci donne lieu au **phénomène de Runge** que l'on met en avant dans l'exemple 4.3.1. Ce phénomène se caractérise par une erreur d'interpolation qui, à partir d'un certain nombre, augmente avec les points d'interpolation.

Exemple 4.3.1 *On considère la fonction*

$$f(x) = \frac{1}{1+x^2}.$$

Lorsque l'on veut faire de l'interpolation polynomiale sur l'intervalle $[-3, 3]$ *avec un nombre* $n+1$ *de points uniformément répartis*

$$D_1^n = \left\{ -3 + 6\frac{j}{n} \right\}_{0\leq j\leq n},$$

on obtient un phénomène, illustré par la figure 4.3 où l'on a tracé la fonction f *ainsi que différents polynômes d'interpolation de degré de croissant. On peut constater que l'erreur se situe près des bords de l'intervalle où se forment des oscillations de grandes amplitudes.*

L'explication du phénomène de Runge tient en deux facteurs

1) une fonction $\mathcal{C}^\infty$ n'est pas nécessairement une fonction dont les dérivées sont bornées de manière uniforme par rapport à n, autrement dit, elles peuvent croître très vite en fonction de n,
2) un mauvais choix de points d'interpolation peut rendre le terme

$$\frac{\sup_{y\in[a,b]} \prod_{j=0}^{n} |y - x_j|}{(n+1)!},$$

relativement grand.

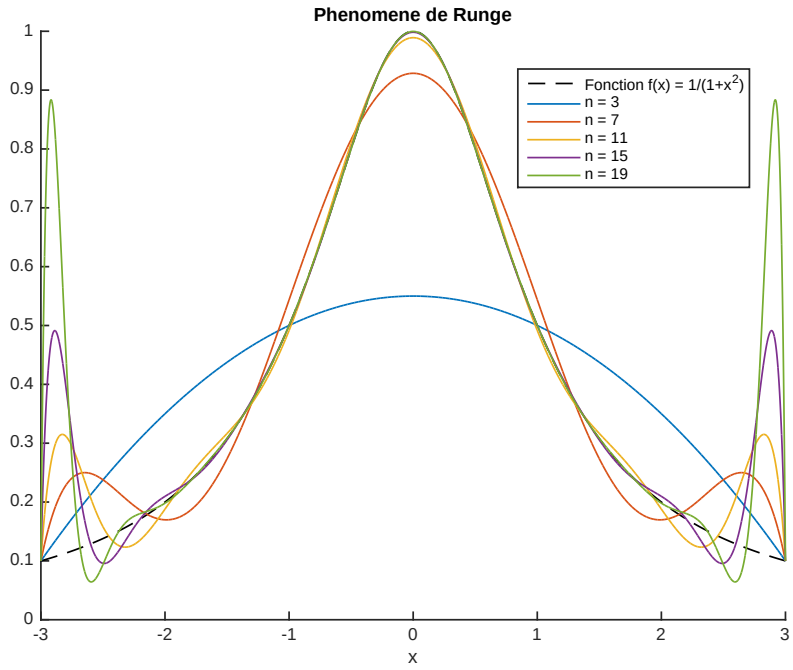


Figure 4.3 – Illustration du phénomène de Runge

Chapitre 5

Intégration Numerique

5.1 Introduction

L'intégration numérique consiste à trouver une approximation de l'intégrale d'une fonction f sur un intervalle $[a, b]$. On rappelle que l'intégrale de Riemann est définie comme suit.

Définition 5.1.1 *Soit $[a, b]$ un intervalle de $\mathbb{R}$ et f une fonction continue sur $[a, b]$. En notant, $\forall 0 \leq j \leq n$,*

$$x_j = a + (b - a)\frac{j}{n}, \tag{5.1}$$

l'intégrale de Riemann de f sur l'intervalle $[a, b]$ est définie comme

$$\int_a^b f(x)dx = \lim_{n\to\infty} \sum_{j=1}^{n} f(x_{j-1})\frac{b-a}{n}.$$

Remarque 5.1.1 *Dans cette définition, on aurait aussi pu prendre, dans le membre de droite,*

$$\lim_{n\to\infty} \sum_{j=1}^{n} f(x_j)\frac{b-a}{n}.$$

L'idée de l'intégrale de Riemann est la suivante : l'intégrale est l'aire sous la courbe de f dans l'intervalle $[a, b]$. Des fonctions dont l'intégrale est élémentaire sont les fonctions constantes par morceaux χ_I, pour I un intervalle de $\mathbb{R}$, définies telles que

$$\chi_I(x) = \begin{cases} 1, & \text{si} \quad x \in I, \\ 0, & \text{si} \quad x \notin I. \end{cases} \tag{5.2}$$

puisque l'intégrale correspond alors à l'aire du rectangle dont la largeur est la taille de l'intervalle et la hauteur 1. Autrement dit, pour $I = [x, y[$, pour $a \leq x < y \leq b$ on a

$$\int_a^b \chi_{[x,y[}(z)dz = y - x.$$

Si l'on approxime maintenant la fonction f sur l'intervalle $[a, b]$ par une fonction constante par morceaux f_n contruite à partir de fonction χ de la manière suivante, $\forall n \geq 1$, en utilisant (5.1),

$$f_n(x) = \sum_{j=1}^{n} f(x_{j-1})\chi_{[x_{j-1},x_j[}(x),$$

alors on obtient
$\int_a^b f_n(x)dx =$

(5.3)

La théorie de l'intégrale de Riemann nous donne que, $\forall x \in [a, b]$,

$$f_n(x) \underset{n\to\infty}{\to} f(x),$$

et il est donc naturel de définir l'intégrale comme la limite, lorsque n tend vers l'infini, de (5.3). On rappelle que l'intégrale de Riemann possède les propriétés suivantes

-(*Linéarité*) pour toutes fonctions f, g continues sur $[a, b]$ et $\forall \lambda \in \mathbb{R}$

$$\int_a^b (f(x) + \lambda g(x))\, dx = \int_a^b f(x)dx + \lambda \int_a^b g(x)dx,$$

-(*Relation de Chasles*) pour toute fonction f continue sur $[a, b]$ et pour tout $a < r < b$

$$\int_a^b f(x)dx = \int_a^r f(x)dx + \int_r^b f(x)dx,$$

-(*Inégalité triangulaire*) pour toute fonction f continue sur $[a, b]$

$$\left| \int_a^b f(x)dx \right| \leq \int_a^b |f(x)|dx.$$

5.2 Méthodes des rectangles

L'approche la plus simple pour construire une approximation de l'intégrale de Riemann consiste à revenir à sa définition et considérer la somme

$$I^n_{1,\mathrm{Riemann}}(f) = \sum_{j=1}^{n} f\left(x_{j-1}\right) \frac{b-a}{n},$$

pour un certain $n \geq 1$. On a alors le résultat suivant.

Théorème 5.2.1 *Soit f une fonction $\mathcal{C}^1$ sur l'intervalle $[a,b]$. Alors on a que*

$$\left| \int_a^b f(x)dx - I^n_{1,Riemann}(f) \right| \leq \frac{(b-a)^2 \sup_{x\in[a,b]} |f'(x)|}{2n}. \tag{5.4}$$

Démonstration. On a, par la relation de Chasles,

$\int_a^b f(x)dx - I^n_{1,\mathrm{Riemann}}(f) =$

et, d'autre part, grâce à la formule des accroissements finis,

Ainsi, on obtient que

$\left| \int_a^b f(x)dx - I^n_{1,\mathrm{Riemann}}(f) \right|$

□

On appelle **méthode des rectangles à gauche** cette approximation. Cette méthode est la plus basique des méthodes d'intégration numérique. On illustre sa mise en oeuvre dans la figure 5.1.

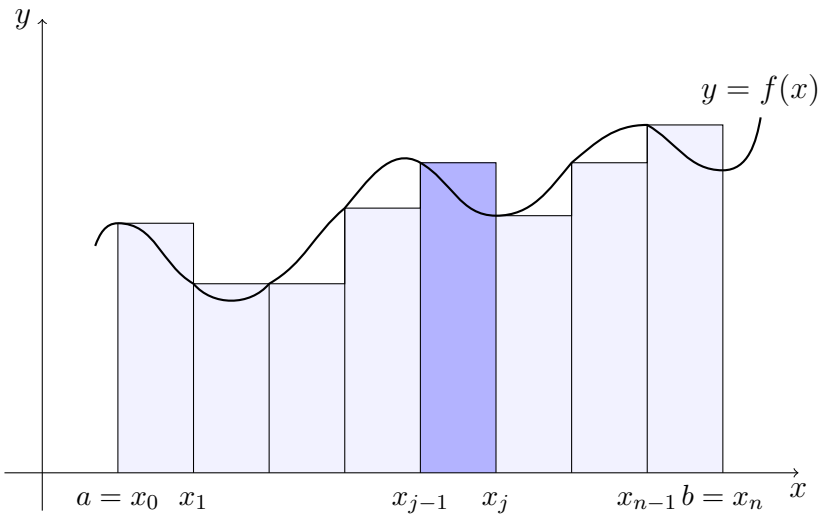


Figure 5.1 – Illustration de la méthode des rectangles à gauche

Remarque 5.2.1 *Dans la définition de l'intégrale discrétisée $I^n_{1,Riemann}(f)$, nous avons choisi de découper l'intervalle $[a, b]$ en des sous-intervalles de longueur uniforme $(b - a)/n$. Ce choix simplifie grandement l'approche mais il est aussi possible de prendre des sous-intervalles de longueurs différentes. Dans la suite, nous considérons uniquement des découpages uniformes.*

On introduit la définition suivante :

Définition 5.2.1 *Soit $n \geq 1$, $[a, b]$ un intervalle de $\mathbb{R}$ et f une fonction $\mathcal{C}^\infty$. Alors, si l'on considère une méthode d'intégration numérique donnant une approximation $I_n(f)$ de l'intégrale de f sur $[a, b]$, on dira que la méthode est d'ordre $p \in \mathbb{N}^*$ si p est le plus grand entier tel que l'on a*

$$\left| \int_a^b f(t)dt - I_n(f) \right| \leq C \left(\frac{b-a}{n} \right)^p,$$

pour une certaine constante $C > 0$ qui peut dépendre de f, a et b.

L'expression (5.4) nous permet en fait d'établir que la **méthode des rectangles à gauche** qui est aussi d'ordre 1 en effet

Remarque 5.2.2 *En prenant l'approximation*

$$I^n_{2,Riemann}(f) = \sum_{j=1}^{n} f(x_j) \left(\frac{b-a}{n} \right),$$

on obtient la méthode des rectangles à droite. Elle est aussi d'ordre 1.

5.3 Méthodes du point milieu

Une deuxième approche consiste à prendre, au lieu du point à gauche ou à droite, le point milieu. On appelle naturellement cette méthode la **méthode du point milieu**. On obtient alors l'approximation suivante

$$I^n_{\text{Milieu}}(f) = \sum_{j=1}^{n} f(x_{j-1/2}) \left(\frac{b-a}{n} \right),$$

où l'on note

$$x_{j-1/2} = \frac{x_{j-1} + x_j}{2} = a + (j - 1/2)\frac{b-a}{n}.$$

Le résultat suivant fournit une estimation de l'erreur commise avec cette méthode.

Théorème 5.3.1 *Soit $f \in \mathcal{C}^2([a,b])$ et $n \geq 1$. Alors, on a l'estimation suivante*

$$\left| \int_a^b f(x)dx - I^n_{Milieu}(f) \right| \leq \frac{b-a}{24} \sup_{x\in[a,b]} |f''(x)| \left(\frac{b-a}{n} \right)^2.$$

Démonstration. Par la relation de Chasles, on a que

$$\int_a^b f(x)dx - I^n_{\text{Milieu}}(f) = \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} \left(f(x) - f(x_{j-1/2}) \right) dx.$$

On utilise ensuite la formule de Taylor-Lagrange qui nous donne

$$f(x) = f(x_{j-1/2}) + f'(x_{j-1/2})(x - x_{j-1/2}) + \frac{1}{2} f''(\eta(x))(x - x_{j-1/2})^2,$$

avec $\eta(x) \in [x_{j-1}, x_j]$.
Sachant que

$$\int_{x_{j-1}}^{x_j} (x - x_{j-1/2})dx$$

on obtient
$$\left| \int_{x_{j-1}}^{x_j} \left(f(x) - f(x_{j-1/2}) \right) dx \right|$$

Après sommation, on en conclut que

$$\begin{aligned} \left| \int_a^b f(x)dx - I^n_{\text{Milieu}}(f) \right| &\leq \frac{\sup_{x\in[a,b]} |f''(x)|}{24} \sum_{j=1}^{n} \left(\frac{b-a}{n} \right)^3 \\ &\leq \frac{(b-a)\sup_{x\in[a,b]} |f''(x)|}{24} \left(\frac{b-a}{n} \right)^2 . \end{aligned}$$

□

On en déduit donc que la méthode du point milieu est d'ordre 2. On illustre cette méthode dans la figure 5.2.

5.4 Méthode des trapèzes

Elle consiste à interpeler par morceaux la fonction avec un polynôme de degré 1 (voir chapitre précédent) Ainsi, on obtient l'approximation de l'intégrale de f sur le segment $[x_{j-1}, x_j]$

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx \sum_{m=0}^{1} f(x_{m,j})\ell_{m,j} = \frac{1}{2}(x_j - x_{j-1}) \left(f(x_{j-1}) + f(x_j) \right),$$

que l'on appelle **la méthode des trapèzes**. On a finalement l'approximation de l'intégrale de f sur $[a, b]$ qui va être donnée par

$$I^n_{Trapeze}(f) = \sum_{j=1}^{n} \frac{1}{2}(x_j - x_{j-1}) \left(f(x_{j-1}) + f(x_j) \right) = \frac{b-a}{2n} \sum_{j=1}^{n} \left(f(x_{j-1}) + f(x_j) \right).$$

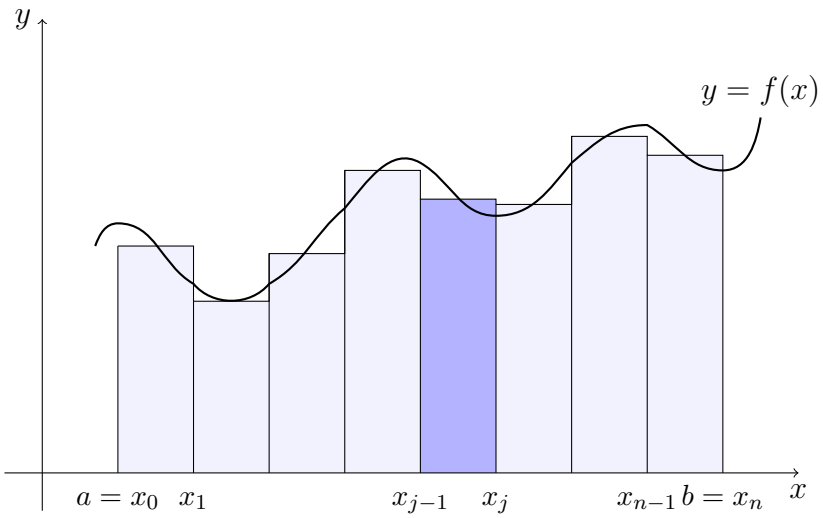


Figure 5.2 – Illustration de la méthode du point milieu

La méthode est illustrée dans la figure 5.3. On peut voir que l'interpolation par un polynôme de degré 1 donne en fait, dans chaque intervalle $[x_{j-1}, x_j]$, une fonction affine qui relie $f(x_{j-1})$ à $f(x_j)$.

Remarque 5.4.1 *On peut voir la méthode des trapèzes comme une moyenne des méthodes des rectangles à droite et à gauche. En effet, on a*

$$\begin{aligned} I^n_{Trapeze}(f) &= \frac{b-a}{2n}\sum_{j=1}^{n}\left(f(x_{j-1})+f(x_j)\right) \\ &= \frac{b-a}{2n}\sum_{j=1}^{n} f(x_{j-1}) + \frac{b-a}{2n}\sum_{j=1}^{n} f(x_j) \\ &= \frac{1}{2}\left(I^n_{1,Riemann}(f) + I^n_{2,Riemann}(f)\right). \end{aligned}$$

Remarque 5.4.2 *On a l'estimation d'erreur suivante pour la formule des trapèzes*

$$\left|\int_a^b f(x)dx - I^n_{Trapeze}(f)\right| \leq \frac{(b-a)\sup_{\eta\in[a,b]}|f^{(2)}(\eta)|}{12}\left(\frac{b-a}{n}\right)^2.$$

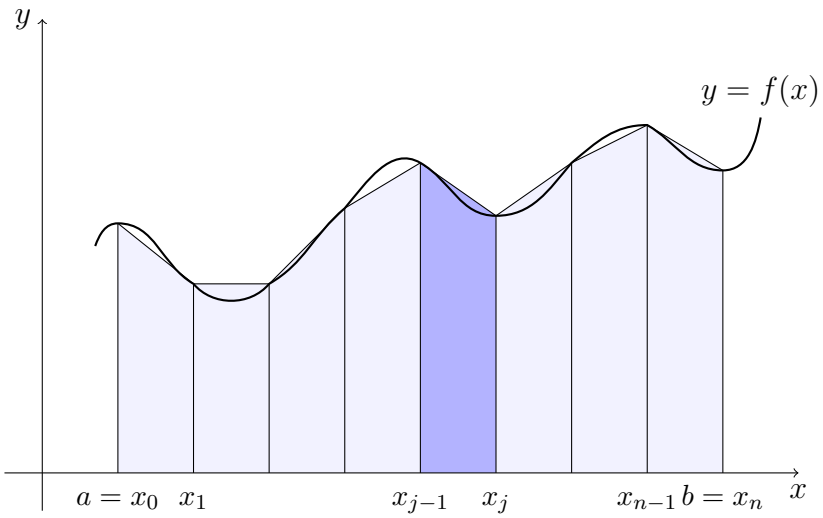


Figure 5.3 – Illustration de la méthode des trapèzes

5.5 Méthode de Simpson

La **méthode de Simpson** est une méthode d'ordre plus élevé, on approche la fonction par morceaux avec des polynômes de degré 2. On obtient l'approximation de l'intégrale de f sur l'intervalle $[x_{j-1}, x_j]$ suivante

$$\int_{x_{j-1}}^{x_j} f(x)dx \approx \sum_{m=0}^{2} f(x_{m,j})\ell_{m,j} = \frac{b-a}{6n}\left(f(x_{j-1}) + 4f(x_{j-1/2}) + f(x_j)\right),$$

ce qui nous amène à l'intégrale discrète de f sur l'intervalle $[a, b]$

$$I^n_{Simpson}(f) = \frac{b-a}{6n}\sum_{j=1}^{n}\left(f(x_{j-1}) + 4f(x_{j-1/2}) + f(x_j)\right).$$

La méthode de Simpson est illustrée dans la figure 5.4. On peut voir que la courbe de f est interpolée par des fonctions paraboliques sur chacun des sous-intervalles.

Remarque 5.5.1 *On a l'expression suivante de l'approximation de l'intégrale de f sur $[a, b]$ à l'aide de la méthode de Simpson en fonction des approximations obtenues par la méthode du point milieu et les méthodes des rectangles*

$$I^n_{Simpson}(f)^n(f) = \frac{1}{6}\left(I^n_{1,Riemann}(f) + 4I^n_{Milieu}(f) + I^n_{2,Riemann}(f)\right).$$

Remarque 5.5.2 *On a l'estimation d'erreur suivante pour la formule de Simpson*

$$\left|\int_a^b f(x)dx - I^n_{Simpson}(f)^n(f)\right| \leq \frac{(b-a)\sup_{\eta\in[a,b]}|f^{(3)}(\eta)|}{90}\left(\frac{b-a}{n}\right)^3,$$

mais la méthode est en réalité d'ordre 4.

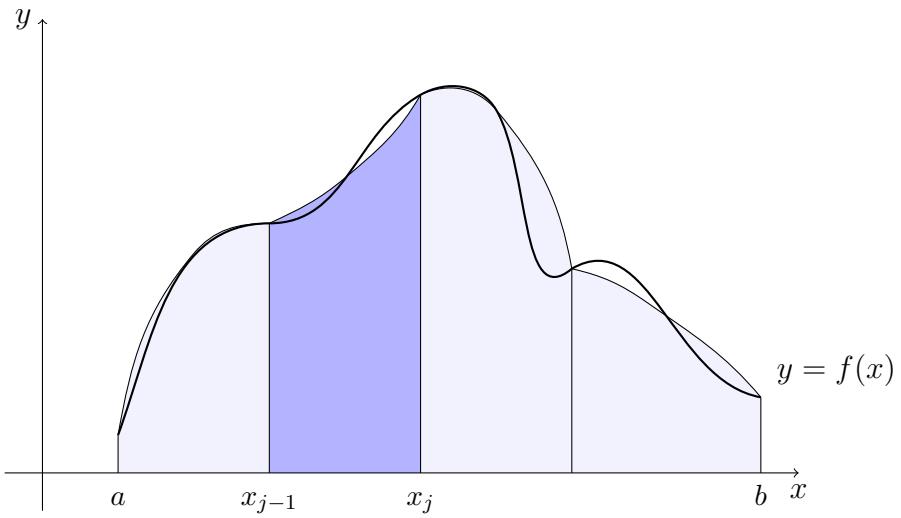


Figure 5.4 – Illustration de la méthode de Simpson

Exemple 5.5.1 *On se propose de comparer les différentes méthodes d'intégration numérique que nous avons vues sur une intégrale dont on connait la valeur exacte. L'intégrale que l'on va chercher à approximer est la suivante*

$$I = \int_0^1 x \ln(1+x) dx = 2\ln(2) - 1 \approx 0.38629436111.$$

On indique dans la figure 5.5 l'erreur absolue commise entre l'approximation de l'intégrale et la valeur I en fonction du nombre de points n pour 3 méthodes : la méthode des rectangles à gauche, la méthode des trapèzes et enfin la méthode de Simpson. On voit tout l'intérêt de prendre des méthodes d'ordre élevé puisque la méthode de Simpson, avec seulement 5 points, donne une erreur de l'ordre de 10^{-6} !

n	Rectangles	Trapèzes	Simpson
5	0.070977543607286	0.001662825551292	$0.957315084870736 \times 10^{-6}$
10	0.035073783402134	0.000416424374137	$0.060525534029132 \times 10^{-6}$
15	0.023290043260145	0.000185137241481	$0.011981725278609 \times 10^{-6}$
20	0.017432831001683	0.000104151487685	$0.003793996394830 \times 10^{-6}$
25	0.013929604058789	0.000066660447590	$0.001554572581774 \times 10^{-6}$
30	0.011598746305996	0.000046293296664	$0.000749842077319 \times 10^{-6}$
35	0.009936114565592	0.000034011986165	$0.000404793198960 \times 10^{-6}$
40	0.008690380474418	0.000026040717418	$0.000237300235018 \times 10^{-6}$
45	0.007722210878607	0.000020575539052	$0.000148152934365 \times 10^{-6}$
50	0.006948138083426	0.000016666277827	$0.000097206798166 \times 10^{-6}$

Figure 5.5 – Erreur sur l'intégration numérique de la fonction $f(x) = x\ln(1+x)$ sur $[0,1]$

Chapitre 6

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 :

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

6.1 Différents types de convergence

6.1.1 Convergence simple et convergence absolue

Définition 6.1.1 *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 vus 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 6.1.1 *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 6.1.2 *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 6.1.1 *$\sum f_n$ CVA sur $\Delta \Rightarrow \sum f_n$ CVS sur Δ.*

6.1.2 Convergence uniforme

Définition 6.1.3 *Soit $\Delta \subset \mathbb{R}$ et $(f_n)_{n\in\mathbb{N}}$ une suite de fonctions de Δ dans $\mathbb{R}$. La série $\sum f_n$ On dit que la série $\sum f_n$ converge uniformément sur Δ vers f 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.

6.1.3 Convergence normale

Définition 6.1.4 *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*

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

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

Théorème 6.1.2 (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 6.1.1 *Etudier 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* $\Delta = [0,\pi]$.

6.2 Convergence uniforme et régularité

Théorème 6.2.1 (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 6.2.2 (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 6.2.3 (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 : Etudier 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 7

Les séries entières

7.1 Définition et propriétés de convergence

On appelle série entière une "somme infinie" de fonctions de la forme

$$S(z) = \sum_{n=0}^{+\infty} a_n \, z^n, \tag{7.1}$$

où $(a_n)_{n\geq 0}$ est une suite de nombres complexes et $z \in \mathbb{C}$. Les a_n sont appelés *coefficients* de la série entière.

La question que l'on se pose dans un premier temps est de savoir si ces sommes sont bien définies et prennent des valeurs finies.

Définition 7.1.1 (Nature d'une série entière)

- *On dit que la série $\sum a_n z^n$, $n \in \mathbb{N}$, converge si et seulement si la limite de $\sum_{n=0}^{N} a_n z^n$ quand N tend vers $+\infty$ existe et est finie.*
- *Sinon on dit que la série $\sum a_n z^n$, $n \in \mathbb{N}$, diverge.*

Proposition 7.1.1

1. *Condition nécessaire de convergence. Si la série $\sum a_n z^n$ converge alors : $\lim_{n\to+\infty} |a_n z^n| = 0$.*
2. *Convergence absolue. Si la série $\sum |a_n z^n|$ converge, alors la série $\sum a_n z^n$ converge.*

7.1.1 Domaine et rayon de convergence

Proposition 7.1.2 *Il existe un unique $R \in [0, +\infty]$, appelé rayon de convergence de la série (7.1), tel que la série (7.1) converge pour $|z| < R$ et diverge pour $|z| > R$.*

Quelques remarques :

- Si $R = 0$, la série $\sum a_n z^n$ converge pour $z = 0$ (sa somme vaut a_0) et diverge pour $z \neq 0$.
- Si $R = +\infty$, la série $\sum a_n z^n$ converge pour tout $z \in \mathbb{C}$.

- Si $0 < R < +\infty$, les séries $\sum a_n z^n$ prises en $|z| = R$ doivent être étudiées à part (pas de résultats généraux).

Le résultat suivant est utile pour le calcul du rayon de convergence :

Proposition 7.1.3 (Calcul pratique du rayon de convergence)

Règle de D'Alembert. *On calcule :* $l = \lim_{n\to+\infty} |a_{n+1}/a_n|$. *Alors :*

$$R(\sum a_n z^n) = 1/l.$$

Règle de Cauchy. *On calcule :* $l = \lim_{n\to+\infty} |a_n|^{\frac{1}{n}}$. *Alors :*

$$R(\sum a_n z^n) = 1/l.$$

Exercice 7.1.1 *Calculer les rayons de convergence des séries entières suivantes :*

- $\sum x^n$, $n \in \mathbb{N}$, *et donner la valeur de sa somme.*
- $\sum x^{2n+1}$, $n \in \mathbb{N}$, *et donner la valeur de sa somme.*
- $\sum \dfrac{x^n}{n}$, $n \in \mathbb{N}^\star$.
- $\sum n! x^n$, $n \in \mathbb{N}$.
- $\sum \dfrac{x^n}{n!}$, $n \in \mathbb{N}$.
- $\sum \left(\dfrac{n}{n+3}\right)^{n^2} x^n$, $n \in \mathbb{N}$.
- $\sum \dfrac{(-1)^n}{8^n} x^{3n+1}$, $n \in \mathbb{N}$.

7.1.2 Opérations sur les séries entières

a. Somme et produits de séries entières

Soient $\sum a_n z^n$ et $\sum b_n z^n$ deux séries entières de rayon de convergence respectifs R_a et R_b.

1. La série entière $\sum (a_n + b_n) z^n$ a pour rayon de convergence R vérifiant :

$$\begin{array}{rcll} R & = & \min(R_a, R_b) & \text{si } R_a \neq R_b \\ & \geq & R_a & \text{si } R_a = R_b \end{array}$$

On a alors :

$$\forall |z| < R, \sum_{n=0}^{+\infty} (a_n + b_n) z^n = \sum_{n=0}^{+\infty} a_n z^n + \sum_{n=0}^{+\infty} b_n z^n.$$

2. Soit $\lambda \in \mathbb{R}$. La série $\sum (\lambda a_n) z^n$ a pour rayon de convergence R_a et :

$$\forall |z| < R_a, \sum_{n=0}^{+\infty} (\lambda a_n) z^n = \lambda \sum_{n=0}^{+\infty} a_n z^n.$$

3. La série produit, notée $\sum c_n z^n$, s'écrit :

$$\sum_{n=0}^{+\infty} c_n z^n = \left(\sum_{n=0}^{+\infty} a_n z^n\right)\left(\sum_{n=0}^{+\infty} b_n z^n\right),$$

avec $c_n = \sum_{k=0}^{n} a_k b_{n-k}$, et a pour rayon de convergence $R \geq \min(R_a, R_b)$.

Exercice 7.1.2 *Etudier la convergence des séries* $\sum ch(na)z^n$, $a > 0$, *et* $\sum \frac{n(n+1)}{2}x^n$.

b. Continuité, dérivabilité et intégrabilité des séries entières

Soient $\sum a_n x^n$ une série entière de rayon de convergence $R > 0$, $x \in \mathbb{R}$. On note :

$$f(x) = \sum_{n=0}^{+\infty} a_n x^n$$

la somme de cette série, définie pour tout $x \in]-R, R[$.

1. la fonction somme f est continue sur $]-R, R[$, et même C^∞ sur $]-R, R[$.
2. Les séries dérivées se calculent terme à terme et ont même rayon de convergence R :

$$\begin{aligned}
f(x) &= \sum_{n=0}^{+\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \dots \\
f'(x) &= \sum_{n=1}^{+\infty} n a_n z^{n-1} = a_1 + 2a_2 z + 3a_3 z^2 + \dots \\
f''(x) &= \sum_{n=2}^{+\infty} n(n-1) a_n z^{n-2} = 2a_2 z + 6a_3 z + \dots \\
&\vdots \\
f^{(p)}(x) &= \sum_{n=p}^{+\infty} n(n-1)\dots(n-p+1) a_n z^{n-p}
\end{aligned}$$

3. **Série de Taylor.** Par récurrence, on démontre que :

$$f^{(p)}(0) = p(p-1)\dots 2.1 a_p = p! a_p, \text{ soit } a_p = \frac{f^{(p)}(0)}{p!}.$$

D'où par conséquent :

$$f(x) = \sum_{p=0}^{+\infty} a_p z^p = \sum_{p=0}^{+\infty} \frac{f^{(p)}(0)}{p!} z^p, \quad x \in]-R, R[.$$

4. La série $\sum \frac{a_n}{n+1} x^{n+1}$ est une primitive de $\sum a_n x^n$ (intégrée terme à terme) et a même rayon de convergence R.

7.2 Applications et calculs de la somme de quelques séries

7.2.1 Autour de la série géométrique

Dans ce paragraphe, nous nous intéressons à la série entière $\sum x^n$ (i.e. : $\forall n \in \mathbb{N},\ a_n = 1$), appelée *série géométrique*.

En utilisant la règle de Cauchy, on voit immédiatement que le rayon de convergence est égal à 1 et on démontre alors :

$$\forall t \in]-1,1[,\ \sum_{n=0}^{+\infty} t^n = \frac{1}{1-t}.$$

On en déduit en intégrant entre 0 et x :

$$\forall x \in]-1,1[,\ \sum_{n=0}^{+\infty} \frac{x^{n+1}}{n+1} = -\ln(1-x).$$

De même en dérivant :

$$\forall t \in]-1,1[,\ \sum_{n=1}^{+\infty} nt^{n-1} = \frac{1}{(1-t)^2}.$$

7.2.2 Retour sur l'exponentielle

Dans ce paragraphe, nous allons utiliser la théorie des EDO et des séries entières pour démontrer :

$$\forall x \in \mathbb{R}, e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}.$$

Considérons la série $\sum \frac{x^n}{n!}$. La règle de D'Alembert nous permet de calculer son rayon de convergence :

$$R(\sum \frac{x^n}{n!}) = +\infty,$$

ce qui signifie que la somme infinie $S(x) = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$ admet des valeurs finies en tout point $x \in \mathbb{R}$.

D'après les propriétés de dérivation des séries entières, on sait que la série $\sum \frac{x^n}{n!}$ est infiniment dérivable sur $\mathbb{R}$ et ses dérivées se calculent terme à terme :

$$\begin{aligned} S'(x) &= \sum_{n=1}^{+\infty} \frac{x^{n-1}}{(n-1)!} = \sum_{p=0}^{+\infty} \frac{x^p}{p!} \text{ en posant : } p = n-1, \\ &= S(x) \end{aligned}$$

La somme S de la série est donc solution de l'équation différentielle ordinaire suivante : $y' - y = 0$. L'idée est maintenant de rajouter une condition initiale de sorte que S soit l'unique solution

de l'EDO précédente. Comme $S(0) = 1$, on va donc résoudre :

$$\begin{cases} y' - y = 0 \\ y(0) = 1 \end{cases}$$

dont la solution est : $y(x) = e^x$. On en déduit donc : $S(x) = e^x$ quel que soit $x \in \mathbb{R}$ soit :

$$\forall x \in \mathbb{R},\ \sum_{n=0}^{+\infty} \frac{x^n}{n!} = e^x.$$

7.2.3 Calcul de $\sin(x)$ avec une précision arbitraire

On a vu qu'une fonction est égale à sa série de Taylor à condition que le rayon de convergence de celle-ci soit non nul. La fonction $\sin$ est infiniment dérivable et :

$$\begin{cases} \sin^{(2n)}(x) &= (-1)^n \sin(x) \\ \sin^{(2n+1)}(x) &= (-1)^n \cos(x) \end{cases} \Rightarrow \begin{cases} \sin^{(2n)}(0) &= 0 \\ \sin^{(2n+1)}(0) &= (-1)^n \end{cases}$$

Sa série de Taylor s'écrit alors :

$$\sum_{n=0}^{+\infty} \frac{\sin^{(n)}(0)}{n!} x^n, \qquad \text{soit :} \qquad \sum_{n=0}^{+\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

En utilisant la règle de D'Alembert, on montre que son rayon de convergence est infini, donc non nul ce qui nous permet de montrer que $\sin$ est égal à sa série de Taylor :

$$\forall x \in \mathbb{R},\ \sin(x) = \sum_{n=0}^{+\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1}.$$

La somme finie $\displaystyle\sum_{n=0}^{N} \frac{(-1)^n}{(2n+1)!} x^{2n+1}$ est une approximation de $\sin(x)$ à $\dfrac{|x|^{2N+1}}{(2N+1)!}$ près. En effet regardons d'un peu plus près l'erreur commise :

$$\begin{aligned} \left| \sin(x) - \sum_{n=0}^{N} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \right| &= \left| \sum_{n=N+1}^{+\infty} \frac{(-1)^n}{(2n+1)!} x^{2n+1} \right| \\ &\leq \sum_{n=N+1}^{+\infty} \frac{1}{(2n+1)!} |x|^{2n+1} \leq \frac{|x|^{2N+1}}{(2N+1)!} \end{aligned}$$

7.3 Résolution d'équations différentielles à coefficients polynomiaux

Les séries entières permettent de trouver des solutions de certaines classes d'équations différentielles à coefficients polynomiaux, comme le garantit le résultat suivant dans le cas des équations du second ordre.

Proposition 7.3.1 *Soient a, b, c des polynômes avec $a(0) \neq 0$. Toutes les solutions du problème*

$$a(x)\, y'' + b(x)\, y' + c(x)\, y = 0 \tag{7.2}$$

s'écrivent sous la forme de séries entières $y(x) = \sum_{n=0}^{+\infty} a_n\, x^n$ de rayon de convergence $R > 0$. Par ailleurs, R est supérieur ou égal au plus petit des modules des zéros de a dans $\mathbb{C}$. En particulier, si a est constant alors $R = +\infty$.

On peut dire également que sous les hypothèses de la proposition 7.3.1, les solutions de (8.1) s'écrivent comme combinaisons linéaires de deux solutions linéairement indépendantes y_1 et y_2, de rayons de convergence $R_1, R_2 > 0$ (pas forcément égaux).

Dans le cas où $a(0) = 0$ il peut très bien n'exister aucune solution sous la forme de série entière, ou seulement une droite de telles solutions $y(x) = Cy_1(x)$, avec une solution linéairement indépendante y_2 qui ne s'écrit pas sous la forme d'une série entière.

Nous verrons en travaux dirigés différents exemples d'utilisation des séries entières pour résoudre des équations de la forme (8.1), ainsi que des exemples dégénérés pour lesquels $a(0) = 0$.

Exercice 7.3.1 *On cherche une solution de l'EDO ci-après sous forme de série entière :*

$$(E)\quad xy'' - (2x-1)y' + (x-1)y = 0.$$

1. *Déterminer la relation de récurrence entre les coefficients a_n de la série entière.*
2. *Calculer a_1, a_2, a_3 et a_4. En déduire l'expression générale des coefficients a_n, ainsi que les solutions de l'équation (E) obtenues par ce procédé.*
3. *Obtient-on toutes les solutions de l'équation (E) ? Donner l'expression de la solution générale de l'équation (E).*

Chapitre 8

Equations différentielles ordinaires linéaires d'ordre 1 et 2

On appelle Equation Différentielle Ordinaire (EDO) toute équation dont l'inconnue est une fonction y de la variable indépendante x et faisant intervenir les dérivées successives de y :

$$F(x, y(x), y'(x), \ldots, y^{(n)}(x)) = 0.$$

Terminologie :

- La fonction y est l'inconnue de l'EDO.
- L'EDO est dite **d'ordre** n si la dérivée de y d'ordre le plus élevé dans l'équation, est d'ordre n
- L'EDO est dite **homogène** si elle ne contient que des termes en y et ses dérivées.
- L'EDO est dite **linéaire** si $F(\cdot)$ est linéaire par rapport à y et à toutes ses dérivées.
- L'EDO est dite **linéaire à coefficients constants** si $F(\cdot)$ est linéaire par rapport à y et à toutes ses dérivées et les coefficients ne dépendent pas de x.

Exercice 8.0.1 *Donner le type des EDO suivantes :*

1. $y''(x) + xy(x)y'^2(x) = \sin(x)$.

2. $y''(x) + xy(x)y'(x) = 0$.

3. $(x+1)y^{(3)}(x) + 2xy''(x) + xy'(x) = \sin(x)$.

4. $y^{(3)}(x) + 2y''(x) + y'(x) + y(x) = \sin(x)$.

Définition 8.0.1 *Soit :*

$$F(x, y(x), y'(x), \cdots, y^{(n)}(x)) = 0 \tag{8.1}$$

une EDO d'inconnue y et définie sur l'intervalle I. Une solution de l'EDO est une fonction y de classe C^n définie sur un intervalle $I \subset \mathbb{R}$ et vérifiant l'EDO (8.1) *en tout point $x \in I$.*

Dans ce chapitre, nous nous intéressons aux EDO linéaires d'ordre 1 à coefficients variables, ainsi qu'aux EDO linéaires d'ordre 2, d'abord à coefficients constants puis à coefficients variables.

8.1 Equations différentielles ordinaires linéaires du premier ordre

Dans ce paragraphe, nous nous intéresserons à la résolution des EDO linéaires du premier ordre i.e. à la résolution d'équations du type :

$$(E) \qquad a(x)y'(x) + b(x)y(x) = f(x)$$

où les fonctions a et b sont données et s'appellent les coefficients de l'EDO et la fonction f est donnée et s'appelle le second membre. Reprenons la terminologie vue précédemment :

- (E) est dite **homogène** si le second membre f est identiquement nul :$f \equiv 0$.
- (E) est dite **à coefficients constants** si les fonctions a et b sont constantes.

8.1.1 Solution générale

Considérons l'EDO linéaire d'ordre 1 suivante :

$$(E) \qquad a(x)y'(x) + b(x)y(x) = f(x)$$

Rappelons le principe général de résolution des EDO linéaires d'ordre 1 :

Théorème 8.1.1 *Supposons que les fonctions a, b et f sont continues sur l'intervalle $I \subset \mathbb{R}$. Alors la solution générale $y(x)$ de (E) est donnée par :*

$$y(x) = y_h(x) + y_p(x)$$

où

- *y_h est une solution générale de l'EDO homogène (E_h) définie par :*

$$(E_h) \qquad a(x)y'(x) + b(x)y(x) = 0.$$

- *y_p est une solution particulière de l'EDO complète (E) i.e. :*

$$a(x)y_p'(x) + b(x)y_p(x) = f(x).$$

Il reste donc maintenant à savoir calculer la solution générale de l'EDO homogène associée à (E) ainsi qu'une solution particulière. Avant cela remarquons que nous disposons pour les EDO linéaires d'un principe de superposition très utile en pratique :

Proposition 8.1.1 (Principe de superposition) *Sous les hypothèses du Théorème 8.1.1, si y_{p1} et y_{p2} sont respectivement solution particulière des EDO :*

$$\begin{aligned} a(x)y'_{p1}(x) + b(x)y_{p1}(x) &= f_1(x), \\ a(x)y'_{p2}(x) + b(x)y_{p2}(x) &= f_2(x), \end{aligned}$$

alors, pour tous $\lambda_1, \lambda_2 \in \mathbb{R}$, la fonction $y_p = \lambda_1 y_{p1} + \lambda_2 y_{p2}$ est une solution particulière de l'EDO :

$$a(x)y'_p(x) + b(x)y_p(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x).$$

8.1.2 Résolution de l'équation homogène

Intéressons nous dans un premier temps à la résolution de l'équation homogène associée à l'EDO (E) :

$$(E_h) \qquad a(x)y'(x) + b(x)y(x) = 0.$$

On note I un intervalle où les fonctions a et b sont définies et continues et tel que :

$$\forall x \in I,\ a(x) \neq 0.$$

La solution de l'EDO homogène (E_h) s'écrit :

$$y_h(x) = Ce^{u(x)},$$

avec $u'(x) = -\frac{b(x)}{a(x)}$ et C est une constante arbitraire. Autrement dit, l'ensemble des solutions de l'EDO homogène (E_h) est un espace vectoriel de dimension 1 : cela signifie en pratique que les solutions de (E_h) font intervenir une (et une seule !) constante arbitraire (ici, notée C).

Exemple 8.1.1 *Résoudre les EDO suivantes :*

1. $4y'(x) + 3y(x) = 0.$

2. $y' + 2xy = 0$.

3. $xy' + 2y = 0$.

8.1.3 Recherche d'une solution particulière

On commence toujours par regarder s'il n'y a pas de solution évidente, sinon on peut appliquer l'une des méthodes suivantes

Cas d'une EDO à <u>coefficients constants</u> et second membre de la forme : $f(x) = e^{\lambda x} P_n(x)$ **où** P_n **est un polynôme de degré** n **:**

1^{er} **cas : si** $\lambda \neq r = -\frac{b}{a}$**,** alors on cherche une solution sous la forme :

$$y_p(x) = e^{\lambda x} Q_n(x),$$

où Q_n est un polynôme de degré n.

2^{eme} **cas : si** $\lambda = r = -\frac{b}{a}$**,** on cherche une solution sous la forme :

$$y_p(x) = e^{\lambda x} x Q_n(x)$$

où Q_n est un polynôme de degré n.

Méthode de variation de la constante

La méthode est la suivante : on sait que la solution générale de l'équation homogène (E) est de la forme : $y_h(x) = Ce^{u(x)}$. On va alors :

1. Chercher une solution particulière de (E) sous la forme :
$$y_p(x) = C(x)e^{u(x)}.$$
2. Injecter cette solution particulière dans l'équation (E) en calculant au préalable $y_p'(x)$.
3. Après simplification des termes en $C(x)$, montrer que la fonction C vérifie :
$$C'(x) = \frac{f(x)}{a(x)}e^{-u(x)}.$$
4. Chercher une primitive quelconque de $x \mapsto \frac{f(x)}{a(x)}e^{-u(x)}$ afin de trouver $C(x)$ et obtenir $y_p(x)$.

8.1.4 Exercices

Résoudre les EDO suivantes en précisant leur type :

1. $y'(x) + 2y(x) = 2\cos x$.
2. $xy'(x) + y(x) = x$, pour $x > 0$.
3. $\sin(x)y' - \cos(x)y = x$, pour $x \in]0, \frac{\pi}{2}[$.
4. $\sin(x)y' - \cos(x)y = 2\cos(x) + 3x$, pour $x \in]0, \frac{\pi}{2}[$.

8.2 Equations différentielles ordinaires linéaires du second ordre

Dans ce paragraphe, nous nous intéresserons à la résolution des EDO linéaires du second ordre i.e. à la résolution d'équations du type :

$$(E) \qquad a(x)y''(x) + b(x)y'(x) + c(x)y(x) = f(x),$$

où les fonctions a, b et c sont données et s'appellent coefficients de l'EDO et la fonction f est donnée et appelée second membre de l'EDO. Une solution de (E) est une fonction y de classe C^2 sur un intervalle I vérifiant (E) pour tout $x \in I$.

- L'EDO (E) est dite **homogène** si le second membre f est identiquement nul : $f \equiv 0$.
- L'EDO (E) est dite à coefficients constants si les fonctions a, b et c sont constantes.

On peut démontrer (mais on ne le fera pas ici) l'existence et l'unicité de la solution d'une EDO linéaire d'ordre 2 à laquelle on impose exactement deux conditions initiales :

Théorème 8.2.1 *Soient I un intervalle de $\mathbb{R}$, $x_0 \in I$ et $(c_0, c_1) \in \mathbb{R}^2$. Si a, b, c et f sont des fonctions continues sur un intervalle I de $\mathbb{R}$ et si : $\forall x \in I,\ a(x) \neq 0$, alors le problème :*

$$\begin{aligned} &a(x)y''(x) + b(x)y'(x) + c(x)y(x) = f(x) \\ &y(x_0) = c_0,\ y'(x_0) = c_1 \end{aligned}$$

admet une unique solution y définie sur I.

Par conséquent, si jamais l'EDO (E) est homogène (i.e. si $f = 0$) alors toute solution du problème :

$$a(x)y''(x) + b(x)y'(x) + c(x)y(x) = 0,\ x \in I,$$

avec comme conditions initiales :

$$y(x_0) = 0,\ y'(x_0) = 0,$$

est nécessairement la solution nulle : $\forall x \in I,\ y(x) = 0$.

8.2.1 Solution générale

Intéressons nous maintenant au calcul de la solution générale d'une EDO linéaire d'ordre 2 : le théorème ci-après nous donne les étapes pour cette solution :

Théorème 8.2.2 *Considérons l'EDO linéaire d'ordre 2 suivante :*

$$(E) \qquad a(x)y''(x) + b(x)y'(x) + c(x)y(x) = f(x)$$

Supposons que les fonctions a, b, c et f sont continues sur l'intervalle $I \subset \mathbb{R}$. Alors la solution générale $y(x)$ de (E) est donnée par :

$$y(x) = y_h(x) + y_p(x)$$

où

- *y_h est solution de l'EDO homogène (E_h) définie par :*

$$(E_h) \qquad a(x)y''(x) + b(x)y'(x) + c(x)y(x) = 0.$$

- *y_p est <u>une</u> solution particulière de l'EDO complète (E) i.e. :*

$$a(x)y_p''(x) + b(x)y_p'(x) + c(x)y_p(x) = f(x).$$

On dispose à nouveau du principe de superposition :

Proposition 8.2.1 (Principe de superposition) *Sous les hypothèses du Théorème 8.2.2, si y_{p1} et y_{p2} sont respectivement solution particulière des EDO :*

$$\begin{aligned} a(x)y_{p1}''(x) + b(x)y_{p1}'(x) + c(x)y_{p1}(x) &= f_1(x), \\ a(x)y_{p2}''(x) + b(x)y_{p2}'(x) + c(x)y_{p2}(x) &= f_2(x), \end{aligned}$$

alors, pour tous $\lambda_1, \lambda_2 \in \mathbb{R}$, la fonction $y_p = \lambda_1 y_{p1} + \lambda_2 y_{p2}$ est une solution particulière de l'EDO :

$$a(x)y_p''(x) + b(x)y_p'(x) + c(x)y_p(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x).$$

8.2.2 Résolution de l'équation homogène

Soit :

$$(E_h) \qquad a(x)y''(x) + b(x)y'(x) + c(x)y(x) = 0.$$

Contrairement aux EDO linéaires homogènes du premier ordre, on n'a pas d'expression explicite des solutions lorsque les coefficients sont non constants. Commençons par donner la structure des solutions d'une EDO linéaire homogène du second ordre :

Proposition 8.2.2 *Soit I un intervalle où les fonctions a, b et c sont définies et continues et tel que $a(x) \neq 0$, pour tout $x \in I$. Les solutions de l'équation homogène (E_h) sont de la forme :*

$$y(x) = \lambda y_1(x) + \mu y_2(x),$$

où λ et μ sont des constantes arbitraires et y_1, y_2 deux solutions linéairement indépendantes (non proportionnelles). Autrement dit, l'ensemble des solutions de (E_h) est un espace vectoriel de dimension 2.

La solution de l'équation (E_h) dépend ainsi de deux constantes arbitraires, soit 2 degrés de liberté. Il suffit donc de connaître deux solutions linéairement indépendantes de (E_h) pour les connaître toutes. En réalité, nous verrons grâce à la méthode de réduction d'ordre décrite ci-après, qu'il suffit de connaître une solution de l'équation homogène pour pouvoir en construire une deuxième, linéairement indépendante, et ainsi les obtenir toutes.

Caractérisation de l'indépendance linéaire.

Pour caractériser l'indépendance linéaire, il sera utile (surtout si on cherche à résoudre des EDO d'ordre $n \geq 2$) d'avoir un critère systématique simple pour tester l'indépendance linéaire de n solutions données.

Définition 8.2.1 *Soit $y_1, \ldots, y_n$ n fonctions réelles de classe C^{n-1} sur $I = [a, b]$. On appelle Wronskien de ces n fonctions le déterminant suivant :*

$$W(y_1, y_2, \cdots, y_n) = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \vdots & \vdots & & \vdots \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix}$$

Proposition 8.2.3 (Caractérisation de l'indépendance linéaire)

1. *S'il existe $x_0 \in I$ tel que $W(y_1(x_0), \cdots, y_n(x_0)) = 0$ alors y_1, $\cdots$, y_n ne sont pas des fonctions linéairement indépendantes sur I.*
2. *Si, pour tout $x \in I$, $W(y_1(x), \cdots, y_n(x)) \neq 0$ alors y_1, $\cdots$, y_n sont des fonctions linéairement indépendantes sur I.*

Dans le cas particulier des EDO linéaires homogènes du second ordre, il suffit donc de calculer :

$$W(y_1(x), y_2(x)) = \begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix}$$

et de montrer que ce déterminant est non nul en tout point de l'intervalle I sur lequel sont définies les solutions de (E_h) pour en déduire que les solutions y_1 et y_2 sont linéairement indépendantes.

Exemple 8.2.1 *Montrer que les fonctions $y_1(x) = \cos(x)$ et $y_2(x) = \sin(x)$ sont deux solutions linéairement indépendantes de l'équation :*

$$y''(x) + y(x) = 0.$$

Intéressons nous maintenant au calcul de ces solutions dans le cas d'une EDO linéaire du second ordre générale. L'idée est d'essayer de se ramener à la résolution d'une EDO linéaire du premier ordre pour laquelle on pourra calculer les solutions explicites. C'est le principe de la méthode de réduction d'ordre (qui est une variante de la méthode de la variation de la constante).

Méthode de réduction d'ordre.

Étape 1. "Deviner" une solution y_1 de l'équation (E_h) : on pourra pour cela regarder la forme de l'équation et essayer par exemple des polynômes : x, x^2, e^x (si $a(x)+b(x)+c(x) = 0$, $y(x) = e^x$ est solution !), etc.

Étape 2. On cherche une solution y_2 de (E_h) sous la forme : $y_2(x) = C(x)y_1(x)$ avec C fonction non constante.

Étape 3. La solution générale de (E_h) s'écrit alors : $y(x) = \lambda y_1(x) + \mu y_2(x)$ avec $\lambda, \mu \in \mathbb{R}$ (ou $\mathbb{C}$ si on cherche des solutions complexes).

Exemple Résoudre : $x^2y'' + xy' - 4y = 0$.

Application au cas particulier des EDO linéaires du second ordre à coefficients constants

Soit :

$$(E_c) \qquad ay''(x) + by'(x) + cy(x) = 0 \tag{8.2}$$

où les réels $a,\ b$ et c sont donnés dans $\mathbb{R}^* \times \mathbb{R} \times \mathbb{R}$.

1. Soit $r \in \mathbb{C}$. Montrer que $y_1(x) = e^{rx}$ soit solution de l'EDO (E_c) si et seulement si :

$$ar^2 + br + c = 0 \qquad \text{(Equation caractéristique)} \tag{8.3}$$

2. On note $\Delta = b^2 - 4ac$ le discriminant de l'équation caractéristique. Montrer que l'on retrouve les résultats connus suivants :

Si $\Delta > 0$ alors l'équation caractéristique admet deux solutions $r_1 \neq r_2$ réelles. La solution générale de (E_c) s'écrit :

$$y(x) = C_1 e^{r_1 x} + C_2 e^{r_2 x}, \text{ avec : } C_1,\, C_2 \in \mathbb{R} \text{ constantes arbitraires.}$$

Si $\Delta < 0$ alors l'équation caractéristique admet deux solutions complexes conjuguées $r = \alpha + i\beta$ et $\bar{r} = \alpha - i\beta$. Comme précédemment on a donc :

$$y(x) = C_1 e^{rx} + C_2 e^{\bar{r}x}, \quad \text{avec } C_1, C_2 \in \mathbb{C}, \qquad \textbf{solutions complexes}$$

ou bien si on veut les solutions réelles de (E_c) :

$$y(x) = e^{\alpha x}\left(A\cos(\beta x) + B\sin(\beta x)\right), \quad \text{avec } A, B \in \mathbb{R}, \qquad \textbf{solutions réelles}$$

Si $\Delta = 0$ alors l'équation caractéristique admet une unique solution $r = -\frac{b}{2a}$ (racine double) et les solutions de (E_c) s'écrivent :

$$y(x) = (A + Bx)e^{rx}, \quad \text{où } A \text{ et } B \text{ sont deux constantes arbitraires réelles.}$$

8.2.3 Recherche d'une solution particulière

Revenons à l'EDO complète à coefficients constants :

$$(E) \qquad ay''(x) + by'(x) + cy(x) = f(x)$$

dont on cherche une solution particulière. On commence toujours par regarder s'il n'y a pas de solution évidente, sinon on peut appliquer l'une des méthodes suivantes :

- Si le second membre est de la forme $f(x) = \alpha\cos(x) + \beta\sin(x)$ alors on peut chercher une solution sous la forme : $y_0(x) = c_1\cos(x) + c_2\sin(x)$.

- Si le second membre est de la forme $f(x) = e^{\lambda x}P_n(x)$ avec P_n polynôme de degré n :

  1^{er} **cas** si $a\lambda^2 + b\lambda + c \neq 0$ i.e. $\lambda \neq r_1$ et $\lambda \neq r_2$, on cherche une solution sous la forme $y_0(x) = e^{\lambda x}Q_n(x)$ où Q_n est un polynôme de degré n.

  2^{eme} **cas** si $a\lambda^2 + b\lambda + c = 0$ et $2a\lambda + b \neq 0$ i.e. si $\lambda = r_1$ ou $\lambda = r_2$ avec $r_1 \neq r_2$, on cherche une solution sous la forme $y_0(x) = e^{\lambda x}xQ_n(x)$ où Q_n est un polynôme de degré n.

  3^{eme} **cas** si $\lambda = r_1 = r_2$, on cherche une solution sous la forme $y_0(x) = e^{\lambda x}x^2Q_n(x)$ où Q_n est un polynôme de degré n.

Cas général - Méthode de variation des constantes Le principe est le suivant : soit $y_h(x) = Ay_1(x) + By_2(x)$ la solution générale de l'EDO homogène associée. On cherche une solution particulière de l'EDO complète sous la forme :

$$y_p(x) = A(x)y_1(x) + B(x)y_2(x).$$

Pour cela :

1. On calcule :

$$y_p'(x) = A'(x)y_1(x) + B'(x)y_2(x) + A(x)y_1'(x) + B(x)y_2'(x),$$

et on impose arbitrairement :

$$A'(x)y_1(x) + B'(x)y_2(x) = 0.$$

2. On calcule alors $y_p''(x)$ et on injecte dans (E) les expressions de $y_p(x)$, $y_p'(x)$ et $y_p''(x)$. On obtient alors :

$$(S) \quad \begin{cases} y_1(x)A'(x) + y_2(x)B'(x) = 0 \\ y_1'(x)A'(x) + y_2'(x)B'(x) = \dfrac{f(x)}{a(x)}. \end{cases}$$

qui est un système linéaire à résoudre en les variables $(A'(x), B'(x))$.

Remarquons ici que le déterminant du système (S) s'écrit :

$$W(y_1(x), y_2(x)) = \begin{vmatrix} y_1(x) & y_2(x) \\ y_1'(x) & y_2'(x) \end{vmatrix}$$

et ce n'est autre que le *Wronskien* des solutions y_1 et y_2 de l'EDO homogène. Le système (S) a donc des solutions si et seulement si son déterminant $W(x)$ est non nul, i.e. si et seulement si les solutions y_1 et y_2 sont linéairement indépendants (ce qui est bien le cas ici par construction).

3. On intègre les expressions de $A'(x)$ et $B'(x)$ pour trouver $A(x)$ et $B(x)$ et en déduire :

$$y_p(x) = A(x)y_1(x) + B(x)y_2(x).$$

Chapitre 9

Systèmes différentiels linéaires à coefficients constants

Dans ce chapitre, nous nous intéressons à la résolution des systèmes différentiels linéaires à coefficients constants i.e. des systèmes d'équations différentielles ordinaires linéaires à coefficients a_{ij} constants :

$$(S) \qquad \begin{cases} \dfrac{dx_1}{dt} &= a_{11}\,x_1 + \cdots + a_{1n}\,x_n + f_1(t), \\ \vdots & \vdots \qquad\qquad\qquad \vdots \\ \dfrac{dx_n}{dt} &= a_{n1}\,x_1 + \cdots + a_{nn}\,x_n + f_n(t). \end{cases}$$

Sous forme matricielle, le système (S) s'écrit :

$$(S) \qquad \dot{X}(t) = AX(t) + F(t)$$

en posant : $A = (a_{ij}) \in M_n(\mathbb{R})$ est la matrice $n \times n$ dont le coefficient de la $i$ème ligne et $j$ème colonne est a_{ij}.

$$X(t) = \begin{pmatrix} x_1(t) \\ \vdots \\ x_n(t) \end{pmatrix}, \quad F(t) = \begin{pmatrix} f_1(t) \\ \vdots \\ f_n(t) \end{pmatrix} \in \mathbb{R}^n$$

De façon tout à fait analogue à ce qui a été vu au chapitre 1 pour la résolution des EDO linéaires du premier ordre, on aimerait pouvoir écrire e^{tA}. Pour cela, il nous faut définir l'exponentielle de matrice.

9.1 Exponentielle de matrice

Définition 9.1.1 (Exponentielle de matrice) *Soit $M \in \mathbb{M}_n(\mathbb{R})$ une matrice carrée de taille n.*

$$e^M = \sum_{n=0}^{+\infty} \frac{M^n}{n!} = 1 + M + \frac{M^2}{2} + \frac{M^3}{3!} + \dots.$$

Cette définition est inspirée du fait que :

$$\forall x \in \mathbb{R}, e^x = \sum_{n=0}^{+\infty} \frac{x^n}{n!}.$$

Cette somme infinie est une *série entière*, et est bien définie et a des valeurs finies sur $\mathbb{R}$.

L'exponentielle d'une matrice diagonalisable se calcule facilement : soit

$$M = P \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix} P^{-1}.$$

On obtient alors :

$$e^M = P \begin{pmatrix} e^{\lambda_1} & 0 & \cdots & 0 \\ 0 & e^{\lambda_2} & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & e^{\lambda_n} \end{pmatrix} P^{-1}.$$

On peut également obtenir de manière assez similaire une formule donnant l'exponentielle d'une matrice non diagonalisable. Pour cela il faut mettre la matrice sous sa *forme de Jordan*, mais cette notion d'algèbre n'est pas à notre programme... ! Retenez cependant que l'exponentielle d'une matrice (a priori définie comme une somme infinie) peut toujours être calculée explicitement.

L'exponentielle de matrice possède les propriétés suivantes :

Lemme 9.1.1 (Propriétés de l'exponentielle de matrice)

- $e^0 = I$.
- *Si les matrices A et B commutent ($AB = BA$), alors :*

$$e^{A+B} = e^A\, e^B = e^B\, e^A.$$

- $(e^A)^{-1} = e^{-A}$.
- $e^A\, A = A\, e^A$.
- $\dfrac{d}{dt} e^{At} = A\, e^{At}$.

Exercice 9.1.1 *Calculer les exponentielles des matrices suivantes :*

- $A_1 = aI_n$ *où* I_n *désigne la matrice identité d'ordre* n.
- $A_2 = \begin{pmatrix} 1 & 4 \\ 2 & 3 \end{pmatrix}$
- $A_3 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$
- *Généralisation :* $A = I_n + N$ *où* N *est une matrice nilpotente d'ordre* p *c'est-à-dire :*

$$N^p = 0 \quad \textit{mais} \quad N^{p-1} \neq 0.$$

9.2 Résolution du système : $X'(t) = AX(t) + F(t)$

Les solutions du système (S) $X'(t) = AX(t) + F(t)$ s'écrivent :

$$X(t) = X_h(t) + X_0(t),$$

où :

- X_h est solution du système homogène : $X'(t) = AX(t)$.
- X_0 est une solution particulière du système (S).

9.2.1 Solution du système homogène $X'(t) = AX(t)$

Proposition 9.2.1 *Étant donné $A \in M_n(\mathbb{R})$, la solution du système : $\dot{X}(t) = AX(t)$ s'écrit :*

$$X(t) = e^{At}C,$$

où $C \in \mathbb{R}^n$ un vecteur de n constantes arbitraires.

Le résultat précédent montre que les solutions du système homogène $X'(t) = AX(t)$ forment un espace vectoriel de dimension n (n degrés de liberté). Il suffit donc de trouver n solutions linéairement indépendantes pour toutes les avoir. On a en particulier :

Proposition 9.2.2 *Si A est diagonalisable, la solution générale de $X'(t) = AX(t)$ est de la forme :*

$$X(t) = c_1 e^{\lambda_1 t} V_1 + \cdots + c_n e^{\lambda_n t} V_n = \sum_{i=1}^{n} c_i \, e^{\lambda_i t} \, V_i$$

où :

- *λ_i sont les valeurs propres de A,*
- *V_i des vecteurs propres associés ($(V_1, \ldots, V_n)$ forme une base de vecteurs propres),*
- *c_i des constantes arbitraires réelles quelconques (ou complexes dans le cas de valeurs propres complexes).*

Exercice 9.2.1 *Ecrire les systèmes différentiels suivants sous forme matricielle puis les résoudre :*

$$(S_1) \begin{cases} x_1'(t) &= x_1 + 4x_2 \\ x_2'(t) &= 2x_1 + 3x_2 \end{cases} \qquad (S_2) \begin{cases} x_1'(t) &= 2x_1 - 3x_2 \\ x_2'(t) &= x_1 - 2x_2 \end{cases}$$

9.2.2 Recherche d'une solution particulière

Revenons au système avec second membre :

$$X'(t) = AX(t) + F(t)$$

On commence toujours par regarder s'il n'existe pas une solution évidente. Sinon dans le cas général, on utilise une fois encore la méthode de variation de la constante.

On cherche une solution du système (S) sous la forme :

$$\begin{array}{rcl} X(t) & = & e^{At}X_0(t) \\ \dot{X}(t) & = & Ae^{At}X_0(t) + e^{At}X_0'(t) \end{array}$$

On réinjecte dans le système différentiel (S) ce qui donne :

$$X_0'(t) = F(t)e^{-At}$$

dont on cherche une primitive pour enfin obtenir une solution particulière $X(t) = e^{At}X_0(t)$.

Exercice 9.2.2 *Chercher les solutions des systèmes différentiels suivants :*

$$(S_1)\begin{cases} x_1'(t) &=& 2x_1 - 3x_2 + \cos t \\ x_2'(t) &=& x_1 - 2x_2 + \sin t \end{cases} \qquad (S_2)\begin{cases} x' &=& x & & +1 \\ y' &=& x + 3y - z - 2 \\ z' &=& x + 2y \end{cases}$$

Chapitre 10

Equation différentielle ordinaire : Résolution numérique des problème aux conditions initiales

Nous allons dans ce chapitre présenter la résolution numérique de problème aux conditions initiales. Il est à noter qu'avant de résoudre un problème numérique, il faut s'assurer de l'existence et l'inicité de la solution. Cette étude sera faite dans le cours de mathématique de ce semestre pour les problèmes linéaires. Ce n'est cependant pas toujours le cas. Exemple $x'(t) = \sqrt{x(t)}$

Nous supposerons dans la suite du cours que les problèmes étudiés ont une unique solution.

10.1 Réduction d'ordre

On rappelle qu'un *problème à valeur initiale* est un problème de la forme

$$\left\{ \begin{array}{rcl} y'(t) & = & f(t, y(t)) \,, \quad t \in I, \quad y(t) \in \mathbb{R}^n, \\ y(a) & = & \alpha \in \mathbb{R}^n. \end{array} \right. \tag{10.1}$$

Les problèmes du second ordre, ou d'ordre supérieur d'inconnue $y : I \to \mathbb{R}$ peuvent être également des problèmes aux valeurs initiales ; on les transforme alors en problèmes d'ordre 1 dont la nouvelle fonction inconnue Y est à valeur dans $\mathbb{R}^p$, p étant l'ordre de l'équation différentielle d'origine. Partant d'une équation différentielle, on forme un **système différentiel** ; c'est le cas pour le problème suivant, avec $a \in I$:

$$\left\{ \begin{array}{l} y''(x) = f(x) \,, \, x \in I, \\ y(a) = \alpha \,, \, y'(a) = \beta. \end{array} \right. \tag{10.2}$$

que l'on écrira sous la forme :

$$\left\{ \begin{array}{rcl} Y'(x) & = & F(x, Y(x)) \; ; \; x \in I, \\ Y(a) & = & Y_a, \end{array} \right.$$

avec $Y = \left[\begin{array}{c} y(x) \\ y'(x) \end{array} \right]$ et $Y_a = \left[\begin{array}{c} \alpha \\ \beta \end{array} \right]$. On obtient :

$$Y'(x) = \left[\begin{array}{cc} 0 & 1 \\ 0 & 0 \end{array} \right] Y(x) + \left[\begin{array}{c} 0 \\ f(x) \end{array} \right].$$

On dit que l'**on a réduit l'ordre du problème** originel.

Exercice 10.1.1 *Les problèmes différentiels suivants sont-ils des EDO aux conditions initiales ?*

$$\left\{ \begin{array}{l} y^{(3)}(t) + (t^2 - 1)y(t) = 0, \; , t \in \mathbb{R}_+, y(t) \in \mathbb{R}, \\ y(0) = 1, \; y'(1) = 0, \; y''(0) = 0, \end{array} \right. \tag{10.3}$$

$$\left\{ \begin{array}{l} y^{(3)}(t) + (t^2 - 1)y(t) = 0, \; , t \in \mathbb{R}_+, y(t) \in \mathbb{R}, \\ y(0) = 1, \; y'(0) = 0, \; y''(0) = 0, \end{array} \right. \tag{10.4}$$

et

$$\left\{ \begin{array}{l} y^{(3)}(t) + (t^2 - 1)y(t) = 0, \; , t \in \mathbb{R}_+, y(t) \in \mathbb{R}, \\ y(0) = 1, \; y''(0) = 0. \end{array} \right. \tag{10.5}$$

On peut ainsi, lorsqu'on a un problème différentiel d'ordre p bien posé aux conditions initiales, toujours se ramener après réduction d'ordre, à un problème du type

$$\left\{ \begin{array}{rcl} Y'(t) & = & f(t, Y(t)), \quad t \in I, \quad Y(t) \in \mathbb{R}^n, \\ Y(a) & = & \alpha \in \mathbb{R}^p. \end{array} \right. \tag{10.6}$$

Les méthodes numériques que nous allons présenter dans ce chapitre s'appliqueront une fois que l'ordre du problème différentiel a été réduit et que l'on s'est ramené à un problème aux conditions initiales de la forme ci-dessus.

Exemple du pendule :

Autre exemple :

10.2 Approximation numérique des solutions : principe généraux

10.2.1 Définitions

Soit le problème différentiel à condition initiale suivant :

$$\left\{ \begin{array}{rcl} y'(t) & = & f(t, Y(t)), \quad t \in I, \quad y(t) \in \mathbb{R}, \\ y(a) & = & \alpha \in \mathbb{R}. \end{array} \right. \tag{10.7}$$

Il est bien évident que les méthodes numériques que nous allons présenter s'appliqueront au cas où $y(t) \in \mathbb{R}^p$, la résolution du schéma numérique étant bien entendu liés à la dimension p de l'espace.

Pour simplifier les notations, on supposera désormais que $a = 0$, et on notera $b = T$.

L'objectif est de trouver des valeurs approchées de la solution $y(t)$ sur $[0,\, T]$. Pour cela, on se donne une partition $0 = t_0 < t_1 < ... < t_N = T$ et on cherche des y_n aussi voisins que possible des $y(t_n)$.

Définition 10.2.1 *On appelle* **pas de discrétisation** *les réels* $h_n = t_{n+1} - t_n$, $\forall n = 0..N-1$.

Lorsque $\forall n = 0..N-1$, $h_n = h = \dfrac{T}{N}$, *on dit que la subdivision est à* **pas constant**. *Dans le cas contraire, on parlera de* **pas variable**.

Remarque 10.2.1 *En analyse numérique, pour désigner une sudvision de l'intervalle* $[0, T]$, *on utilise également le terme de* **grille de discrétisation**. *Lorsque le pas est constant, on parle de* **grille uniforme**.

Comment trouver les $y_n \approx y(t_n)$?

L'idée première repose sur l'intégration de l'équation différentielle 10.7 sur $[t_n, t_{n+1}]$. On obtient :

$$y(t_{n+1}) = y(t_n) + \int_{t_n}^{t_{n+1}} y'(s)\, ds = y(t_n) + \int_{t_n}^{t_{n+1}} f(s, y(s))\, ds. \tag{10.8}$$

On remplacera alors l'intégrale $\int_{t_n}^{t_{n+1}} f(s, y(s))\, ds$ par une formule approchée. Le choix de la méthode d'intégration numérique utilisée pour approximer cette intégrale conduit à différents schémas numériques de résolution.

10.2.2 Méthodes d'Euler

Méthodes d'Euler explicites et implicites

Si on choisit la formule des rectangles à gauche :

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\, ds \simeq h\, f(t_n,\, y(t_n)),$$

on obtient la formule d'**Euler explicite** :

$$y_{n+1} = y_n + h\, f(t_n, y_n)\ ,\ 0 \leq n \leq N-1.$$

Remarque 10.2.2 *On parle de méthode explicite car y_{n+1} est déterminé par un calcul direct.*

On peut aussi choisir une formule des rectangles à droite, et dans ce cas, c'est la méthode d'**Euler implicite** que l'on obtient :

$$y_{n+1} = y_n + h\, f(t_{n+1}, y_{n+1})\ ,\ 0 \leq n \leq N-1.$$

Remarque 10.2.3 *Cette méthode est dite implicite car y_{n+1} ne s'obtient qu'après la résolution d'une équation.*

Définition 10.2.2 (Schéma à un pas) *Un schéma à un pas est une méthode numérique permettant de calculer y_{n+1} en fonction de y_n, t_n et h_n. Dans le cas d'un pas constant, il est de la forme*

$$y_{n+1} = y_n + h\phi(t_n, y_n, h)$$

où $\phi : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$.

10.2.3 Deux premiers exemples d'applications

Une mise en œuvre réussie

Considérons le problème suivant :

$$\left\{ \begin{array}{rcl} y'(t) & = & y(t)\ \ 0 \leq t \leq 1, \\ y(0) & = & 1. \end{array} \right. \tag{10.9}$$

Nous allons résoudre (10.9) à l'aide des méthodes d'Euler explicite et implicite.

Déterminer la solution exacte de (10.9)

Exercice 10.2.1 *o*

Soit $N > 0$, $h = 1/N$.

Montrer que la méthode d'Euler explicite consiste à calculer $y_{n+1,e} = (1+h)^{n+1}$ *avec* $y_{0,e} = 1$.

o

On retrouve donc pour y_N une valeur approchée de $y(1) = e$ puisque $\lim_{N\to\infty} y_{N,e} = \lim_{N\to\infty} \left(1 + \frac{1}{N}\right)^N = e$. En effet :

Exercice 10.2.2 *o*

Soit $N > 0$, $h = 1/N$.

Montrer que la méthode d'Euler implicite consiste à calculer $y_{n+1,imp} = \dfrac{1}{\left(1 - \frac{1}{n+1}\right)^{n+1}}$ *avec* $y_{0,imp} = 1$.

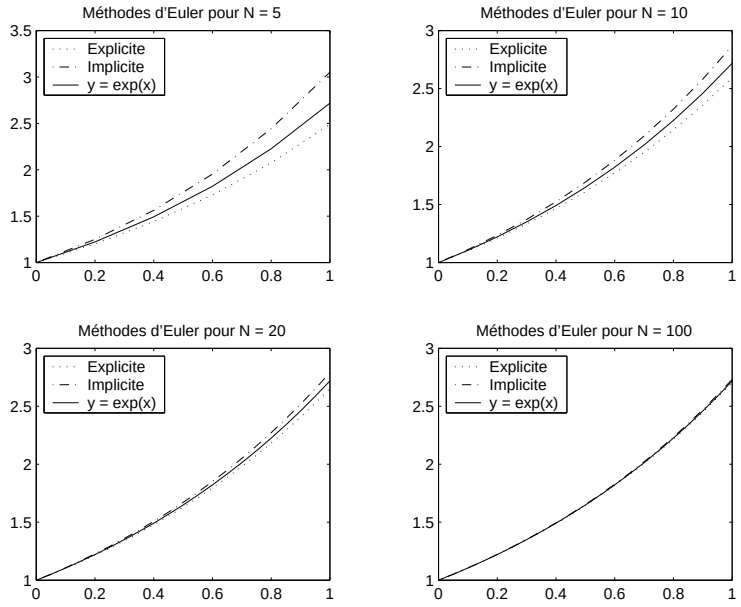


Figure 10.1 – Les méthodes d'Euler pour le problème (10.9)

o

Comme pour le cas explicite, on a $\lim_{N \longrightarrow +\infty} y_{N,imp} = e$.

La figure 10.1 illustre cette convergence des deux méthodes ; on y constate que la méthode explicite calcule des valeurs légèrement inférieures aux $y(t_n)$, alors que la méthode implicite calcule des valeurs supérieures.

En fait, ces comportements s'expliquent par un développement limité :

$$\left\{ \begin{array}{rcccl} y_{N,e} & = & e^{N \ln\left(1+\frac{1}{N}\right)} & = & e\left(1 - \dfrac{1}{2\,N} + \mathcal{O}(\dfrac{1}{N^2})\right), \\ y_{N,imp} & = & e^{-N \ln\left(1-\frac{1}{N}\right)} & = & e\left(1 + \dfrac{1}{2\,N} + \mathcal{O}(\dfrac{1}{N^2})\right). \end{array} \right. \tag{10.10}$$

Cette estimation est confortée par les valeurs données dans le tableau suivant, où e_N désigne l'erreur absolue au point $t = 1$ i.e. $n = N$ pour chacune des méthodes.

	Euler explicite		**Euler implicite**	
N	ϵ_N	$\frac{\epsilon_N}{h}$	ϵ_N	$\frac{\epsilon_N}{h}$
10	0.1245	1.2454	0.1497	1.4969
25	0.0524	1.3111	0.0564	1.4110
50	0.0267	1.3347	0.0277	1.3845
100	0.0135	1.3468	0.0137	1.3717
250	0.0054	1.3542	0.0055	1.3641
500	0.0027	1.3567	0.0027	1.3616
1000	0.0014	1.3579	0.0014	1.3604

Les valeurs du rapport $\frac{e_N}{h}$ encadrent de mieux en mieux $\frac{e}{2} \simeq 1.3591$.

Une mise en œuvre défaillante

1. **Un exemple surprenant**

   Considérons le problème suivant :

$$\begin{cases} y'(t) &= 3y(t) - 3t \ ; \ 0 \leq t \leq T, \\ y(0) &= \frac{1}{3}, \end{cases} \tag{10.11}$$

   dont la solution est

   Appliquons la méthode d'Euler explicite pour en chercher la solution approchée à l'instant T, pour différentes valeurs de T et différents choix du pas h. Les résultats obtenus peuvent surprendre. Le tableau suivant donne les erreurs dans le calcul de y_N :

h	$T = 5$	$T = 10$	$T = 20$
1	$1.8651\ 10^{-14}$	$1.9403\ 10^{-11}$	$2.0345\ 10^{-05}$
0.5	$4.9560\ 10^{-13}$	$4.7278\ 10^{-09}$	$4.2999\ 10^{-01}$
0.1	$5.2402\ 10^{-14}$	$2.6318\ 10^{-08}$	$6.5252\ 10^{+03}$
0.05	$1.5525\ 10^{-11}$	$1.8232\ 10^{-05}$	$2.5142\ 10^{+07}$

On constate que l'erreur semble croître avec le pas, contrairement à ce que l'on espérait. Par ailleurs, pour $T = 20$, cette erreur atteint des valeurs astronomiques !

2. **Analyse de l'exemple**

Calculons la solution générale de l'équation différentielle :

$$y'(t) = 3y(t) - 3t$$

La solution de l'équation homogène est $y(t) = Ke^{3t}$. La méthode de variation de la constante fournit pour notre équation

$$\begin{aligned} K'(t) &= -3t\, e^{-3t}, \\ K(t) &= \left(t + \frac{1}{3}\right) e^{-3t} + k, \\ y(t) &= t + \frac{1}{3} + k\, e^{3t}. \end{aligned}$$

La constante k peut être évaluée à l'aide de la valeur de y en 0, avec $k = 0$ si $y(0) = \dfrac{1}{3}$. Mais si $y(0) = \dfrac{1}{3} + \epsilon$, on obtient $k = \epsilon$ de sorte que la solution réelle du problème devient $y_\epsilon(t) = \dfrac{1}{3} + t + \epsilon\, e^{3t}$.

Les valeurs de e^{3T} sont respectivement :
- pour $T = 5$, $e^{3T} \simeq 3.269\ 10^6$,
- pour $T = 10$, $e^{3T} \simeq 1.068\ 10^{13}$,
- pour $T = 20$, $e^{3T} \simeq 1.142\ 10^{26}$.

Avec une unité d'arrondi de l'ordre de $1.1102\ 10^{-16}$, on retrouve l'ordre de grandeur des erreurs constatées ! Le problème (10.11) est typiquement un problème instable ; **une petite perturbation de la donnée initiale entraine une grande perturbation de la solution.**

10.2.4 Méthode d'Euler améliorée

L'intégrale $\displaystyle\int_{t_n}^{t_{n+1}} f(s, y(s))\, ds$ peut-être approchée par la formule des trapèzes. On obtient alors un schéma implicite défini par :

$$y_{n+1} = y_n + \frac{h}{2}\left(f(t_n,\, y_n) + f(t_{n+1},\, y_{n+1})\right).$$

Si l'on souhaite éviter le caractère implicite de ce schéma, on remplace y_{n+1} par l'estimation qu'en donne le schéma d'Euler explicite. On obtient alors la méthode du trapèze explicite aussi appelée méthode d'Euler améliorée :

$$y_{n+1} = y_n + \frac{h}{2}\big[f(t_n,\, y_n) + f(t_{n+1},\, y_n + h\, f(t_n,\, y_n))\big]. \tag{10.12}$$

10.2.5 Définition d'un schéma à un pas

Définition 10.2.3 (Schéma à un pas) *Un schéma à un pas est une méthode numérique permettant de calculer y_{n+1} en fonction de y_n, t_n et h_n. Dans le cas d'un pas constant, il est de la forme*

$$y_{n+1} = y_n + h\phi(t_n, y_n, h)$$

où $\phi : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^+$.

Remarque 10.2.4 *1) Dans le cas plus général où $y \in \mathbb{R}^p$, on a $\phi : \mathbb{R} \times \mathbb{R}^p \times \mathbb{R}^+$.*
2) Par opposition aux méthodes à un pas, les méthodes multipas utilisent aussi des valeurs calculées aux étapes précédentes.

Exercice 10.2.3 *Les méthodes d'Euler explicite et d'Euler amélioré sont des méthodes à un pas. Donner pour ces méthodes, les fonctions Φ associée.*

10.2.6 Méthode de Runge-Kutta

Nous allons tout d'abord présenter la méthode de Runge-Kutta à 2 étages, puis nous expliquerons la construction dans le cas général i.e. les méthodes de Runge-Kutta à q étages.

Schéma de Runge-Kutta à 2 étages ou méthode d'Euler modifiée

On peut obtenir une valeur plus précise de l'intégrale (10.8) en utilisant la méthode du point milieu. On a

$$\int_{t_n}^{t_{n+1}} f(s, y(s))\, ds \simeq h\, f(t_n + \frac{h}{2},\, y_{n,1})$$

où $y_{n,1}$ est une valeur approchée de $y(t_n + \frac{h}{2})$. On a alors

$$y_{n+1} = y_n + h\, f(t_n + h/2, y_{n,1})\,.$$

Nous n'avons pas encore ici "la forme" d'un schéma à un pas. Nous allons donc essayer de trouver une valeur approchée de $y_{n,1}$ en fonction de y_n, h et t_n. Pour ce faire, on peut utiliser une méthode d'Euler explicite et poser

$$y_{n,1} = y_n + \frac{h}{2} f(t_n, y_n).$$

Ainsi, on obtient le schéma à un pas suivant appelé méthode d'Euler modifée :

$$y_{n+1} = y_n + h\, f(t_n + \frac{h}{2}, y_n + \frac{h}{2} f(t_n, y_n))\,.$$

Son écriture est très compacte et difficilement lisible. Cela risque d'empirer si l'on souhaite raffiner ce type de construction.

On préfère alors décrire ce schéma sous une forme algorithmique :

$$\left\{ \begin{array}{lcl} k_1 & = & f(t_n,\, y_n), \\ k_2 & = & f(t_n + \dfrac{h}{2},\, y_n + \dfrac{h}{2}\, k_1), \\ y_{n+1} & = & y_n + h k_2. \end{array} \right. \tag{10.13}$$

Ce schéma est aussi appelé **méthode de Runge-Kutta à 2 étages**.

Exercice 10.2.4 *o*

Donner la fonction Φ associée à la méthode de Runge-Kutta à 2 étages.

o

Remarque 10.2.5 *Dans le même esprit, la méthode (10.12) s'écrira :*

$$\left\{ \begin{array}{lcl} k_1 & = & f(t_n,\, y_n), \\ k_2 & = & f(t_{n+1}, y_n + hk_1), \\ y_{n+1} & = & y_n + \dfrac{h}{2}k_1 + \dfrac{h}{2}k_2. \end{array} \right. \tag{10.14}$$

Quelle est la fonction Φ associée ?

La méthode "classique" de Runge-Kutta à 4 étages

C'est la méthode habituellement désignée par le nom de méthode de Runge-Kutta. On verra plus tard pourquoi elle est aussi appelée *méthode de Runge-Kutta d'ordre* 4. C'est une excellente méthode pour la plupart des problèmes de Cauchy. C'est en tous cas souvent la première à essayer.

Elle est à 4 étages, définie par la fonction $\Phi(t, y, h) = \dfrac{1}{6}\,(k_1 + 2k_2 + 2k_3 + k_4)$ avec :

$$\begin{array}{lcl} k_1 & = & f(t, y), \\ k_2 & = & f(t + \dfrac{h}{2},\, y + \dfrac{h}{2}\, k_1), \\ k_3 & = & f(t + \dfrac{h}{2},\, y + \dfrac{h}{2}\, k_2), \\ k_4 & = & f(t + h,\, y + h\, k_3)). \end{array}$$

Cette méthode combine les formules de Simpson, du trapèze, et les évaluations de $y(s)$ par la méthode d'Euler.

10.3 Etude générale des méthodes à un pas

Nous avons expliciter les méthodes à un pas les plus courantes. Nous allons dans cette section, les étudier de manière plus approfondie.

10.3.1 Contexte

On rappelle que nous nous intéressons au problème :

$$\left\{ \begin{array}{lcl} y'(t) & = & f(t,\, y(t))\ ; \quad 0 \leq t \leq T, \\ y(0) & = & \alpha. \end{array} \right. \tag{10.15}$$

et que l'on suppose que les hypothèses du théoréme de Cauchy-Lipschitz **??** sont vérifiées.

On souhaite résoudre numériquement le problème (10.15) à l'aide d'une méthode à un pas $(y_n)_{1\leq n\leq N}$, le pas h étant constant.

On a vu au paragraphe 10.2.1 définition 10.2.3 que l'on parle de méthode à un pas lorsque le calcul de la valeur approchée y_{n+1} de $y(t_{n+1})$ ne fait intervenir que y_n, t_n et le pas h_n. Dans le cas d'un pas constant $h = \dfrac{T}{N}$, la forme générale des schémas à un pas est donc la suivante :

$$\left\{ \begin{array}{lcl} y_0 & = & \alpha, \\ y_{n+1} & = & y_n + h\,\Phi(t_n,\, y_n,\, h), \quad 0 \leq n \leq N-1. \end{array} \right. \tag{10.16}$$

C'est la fonction Φ qui caractérise le schéma. Par exemple, pour la méthode d'Euler explicite, on a $\Phi(t,\, y,\, h) = f(t,\, y)$.

Remarque 10.3.1 *La forme (10.16) correspond à des méthodes* **explicites** *à un pas. Pour un schéma implicite, on a en général*

$$y_{n+1} = y_n + h\,\Phi(t_n,\, y_n,\, y_{n+1},\, h).$$

Dans la suite, nous supposons disposer de l'écriture (10.16).

10.3.2 Erreur globale et erreurs locales

Le schéma (10.16) sera convergent si les valeurs y_n qu'il calcule se rapprochent des valeurs $y(t_n)$ de la solution de (10.15) aux instants t_n, pour n assez grand.

Pour caractériser cette convergence, on introduit la notion d' **erreurs locales.**

Définition 10.3.1 *Pour $n = 0..N$, on appelle* **erreur locale** *que l'on note e_n la différence*

$$e_n = y(t_n) - y_n. \tag{10.17}$$

Ces erreurs locales sont dominées par l'erreur globale E_N définie par

$$E_N = \max_{0\leq n\leq N} |e_n| \tag{10.18}$$

Le problème va être d'évaluer ces erreurs locales. En effet, bien que l'on sache que la solution $y(t)$ existe et est unique, on ne sait pas la calculer analytiquement. Dans la définition de e_n seul y_n est connu, la valeur $y(t_n)$ n'étant pas accessible.

Cependant, ces erreurs résultent du cumul d'une contribution "locale" δ_n qui correspond aux calculs pour le pas courant, et à la propagation des erreurs précédentes.

Une façon de définir la contribution locale consiste considérer les **erreurs locales de discrétisation**.

Définition 10.3.2 *Pour $n = 0..N-1$, on appelle* **erreur locale de discrétisation** *la quantité δ_n définie par*

$$\delta_{n+1} = z_n(t_{n+1}) - y_{n+1} \tag{10.19}$$

où z_n est la solution du problème

$$\left\{ \begin{array}{lcl} z_n'(t) & = & f(t,\, z_n(t))\ , \quad t_n \leq t \leq T, \\ z_n(t_n) & = & y_n. \end{array} \right. \tag{10.20}$$

Remarque 10.3.2 *On note que $z_n(t)$ satisfait la même équation différentielle que $y(t)$, mais elle ne vérifie pas le même système différentiel, les conditions initiales étant différentes. Pour n fixé, z_n "passe" par y_n à l'instant t_n (alors que y passe par $y(t_n)$ à l'instant t_n). On obtient alors*

$$\delta_{n+1} = \left\{ \begin{array}{l} z_n(t_{n+1}) - y_n - h\,\Phi(t_n,\, y_n,\, h), \\ z_n(t_{n+1}) - z_n(t_n) - h\,\Phi(t_n,\, z_n(t_n),\, h). \end{array} \right. \tag{10.21}$$

La contribution δ_n ajoutée à l'erreur lors de l'étape n pourra être amplifiée ou réduite dans les calculs ultérieurs selon que le problème aura tendance à amplifier ou à atténuer les perturbations. Cette tendance pourra varier dans le cas de problèmes non linéaires (par exemple, au cours du temps, on pourra avoir une atténuation suivie d'une augmentation etc...).

Elle caractérise la **stabilité** du problème.

Pour les problèmes linéaires, cette tendance ne varie pas. Soit le problème amplifie les perturbations soit il les atténue. On n'aura pas de changement de comportement au cours de la résolution.

Exercice 10.3.1 *Intéressons nous au problème :*

$$\left\{ \begin{array}{lcl} y'(t) & = & y - t^2 + \dfrac{7\,t - 3}{2}\ , \quad 0 \leq t \leq 1.6, \\ y(0) & = & 0. \end{array} \right. \tag{10.22}$$

C'est un problème linéaire du premier ordre à coefficients constants dont la solution est

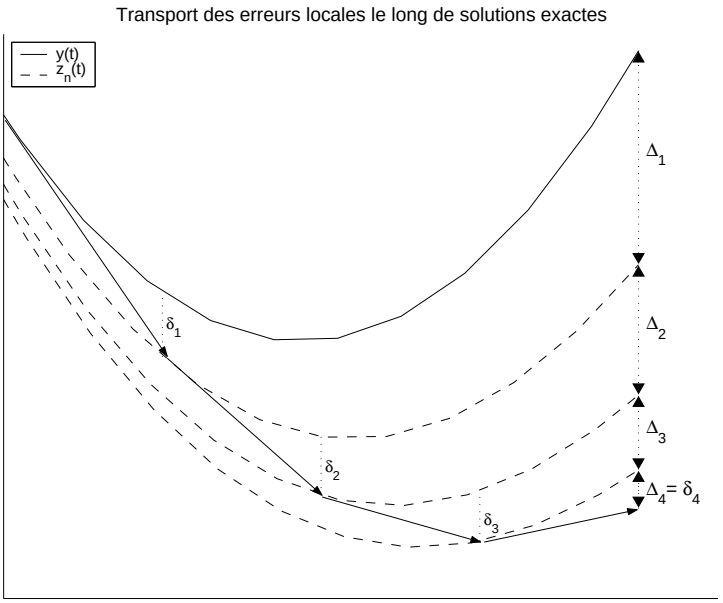


Figure 10.2 – Les z_n transportent les erreurs locales de discrétisation

La figure 10.2 montre 4 étapes de la méthode d'Euler explicite appliquée à ce problème, avec un pas $h = 0,4$. *Les valeurs* y_n, $1 \leq n \leq 4$ *sont pointées par les flèches qui illustrent la trajectoire de la méthode. Cette figure fait également apparaitre les trajectoires des différentes solutions exactes* $z_n(t),\ 0 \leq n \leq 3$ *aux problèmes (10.20), avec bien sur* $z_0(t) = y(t)$.

Les erreurs locales (10.19) ne sont pas simplement additionnées pour contribuer à l'erreur globale : elles sont d'abord "transportées" par les solutions z_n *pour fournir des contributions* Δ_n *à l'erreur finale* e_4 *qui vérifie :*

$$e_4 = \sum_{n=1}^{4} \Delta_n.$$

Sur notre exemple, ce "transport" a plutôt enrichi les erreurs locales.

Exercice 10.3.2 *o On applique la méthode d'Euler explicite au problème (10.9). Montrez que les erreurs locales de discrétisation vérifient*

$$\delta_{n+1} \simeq y_n \frac{h^2}{2} \tag{10.23}$$

Montrez que si l'on applique la méthode du trapèze (10.12), cette erreur vérifie

$$\delta_{n+1} \simeq y_n \frac{h^3}{6} \tag{10.24}$$

o

10.3.3 Ordre de consistance d'un schéma

En fait, (10.21) montre que δ_{n+1} peut-être considérée comme une erreur locale de consistance : elle mesure avec quelle précision le schéma approche localement une solution de l'équation différentielle. L'exercice 10.3.2 montre que leur comportement dépend du schéma utilisé.

Définition 10.3.3 *On dit que le schéma (10.16) est consistant à l'ordre p s'il existe une constante $K > 0$ telle que*

$$|\delta_n| \leq K\, h^{p+1}. \tag{10.25}$$

Lorsque $p = 1$, on dit simplement que la méthode est **consistante**.

Exercice 10.3.3 *o Montrer que pour une méthode à un pas consistante (à l'ordre 1), on aura*

$$\lim_{h\to 0} \sum_{n=1}^{N} |\delta_n| = 0. \tag{10.26}$$

o

10.3.4 Stabilité du problème

Le transport des erreurs locales de discrétisation par les solutions $z_n(t)$ ne les enrichit pas nécessairement, et cela ne dépend pas de la solution $y(t)$.

Exercice 10.3.4 *o*

Déterminer les solutions des deux problèmes suivants :

$$\begin{cases} y'(t) &= \frac{4}{3}(y - t^2) + 4t - \frac{3}{2}\,; \\ y(0) &= 0. \end{cases} \tag{10.27}$$

$$\begin{cases} y'(t) &= \frac{4}{3}(t^2 - y) - \frac{3}{2}\,; \\ y(0) &= 0. \end{cases} \tag{10.28}$$

o

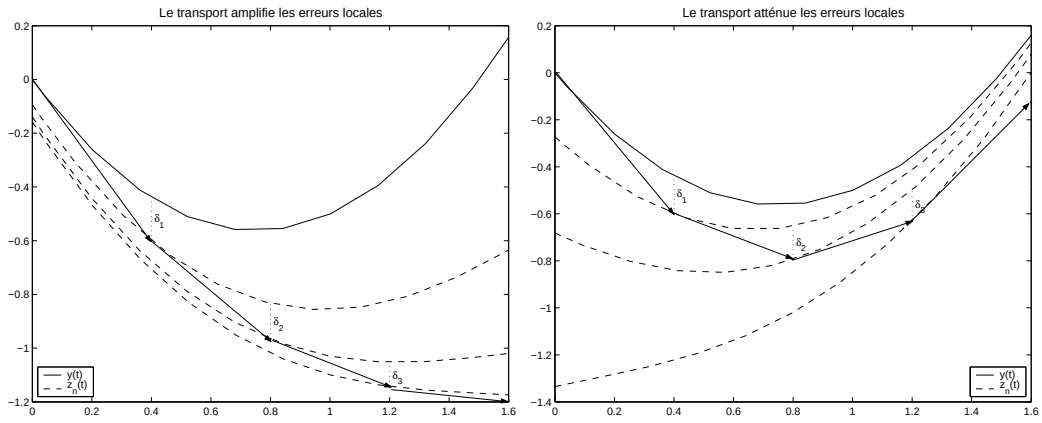


Figure 10.3 – Les erreurs locales de discrétisation peuvent être amplifiée ou atténuées

En appliquant à nouveau la méthode d'Euler implicite avec un pas $h = 0.4$ à ces deux problèmes, on obtient les résultats représentés par la figure 10.3. On constate que dans un cas les erreurs locales de discrétisation sont clairement amplifiées, et dans l'autre clairement atténuées.

Ces deux problèmes sont définis par des équations différentielles linéaires.
Ecrire les solutions des équations homogènes associées. Que remarquez vous :

La dérivée partielle $\dfrac{\partial f}{\partial y}$ est constante, et c'est elle qui permet de savoir si les erreurs locales seront amplifiées ou atténuées. Dans le cas "stable", elle est égale à $-\dfrac{4}{3}$, et dans le cas "instable"

à $+\frac{4}{3}$.
Dans le cas général, ce comportement va varier au cours du calcul. On peut montrer qui si $\frac{\partial f}{\partial y}(t,\, y) < 0$, alors on aura $|\Delta_n| \leq |\delta_n|$ pour tout n.

On ne peut pas se contenter de résoudre les seuls problèmes "stables".

La caractérisation (10.21) permet d'obtenir une estimation de l'erreur locale de discrétisation en cours de calcul, car elle s'appuie sur les valeurs y_n que l'on a calculées.

Remarque : Cette notion est très utile pour mettre en place des stratégies de contrôle du pas, le pas ne sera alors pas constant mais on choisira alors le pas en fonction de cette estimation.

10.3.5 Convergence et ordre de convergence d'un schéma

On rappelle que l'erreur globale E_N est donnée par $E_N = max\{|e_1|, |e_2|, ..., |e_{n-1}|, |e_n|\}$.

Définition 10.3.4 *On dit que la méthode (10.16) est convergente à l'ordre p s'il existe une constante $C > 0$ telle que*

$$(0 \leq)\, E_N \leq C\, h^p. \tag{10.29}$$

La consistance à l'ordre p ne suffit pas à garantir cette convergence. Il faut ajouter une condition de stabilité. A priori, c'est la méthode qui doit être stable, mais une méthode à un pas consistante à l'ordre p sera toujours stable si le problème satisfait les hypothèse du théorème de Cauchy-Lipchitz.

Théorème 10.3.1 *Si la fonction f est telle que*

$$\forall t \in [0,\ T],\ \forall (y,\, z) \in \mathbb{R}^2,\quad |f(t,y) - f(t,z)| \leq L\, |y - z|,$$

la méthode (10.16) utilisée pour calculer une solution approchée du problème (10.15) est convergente à l'ordre p.

Ce théorème est admis.

10.3.6 Ordre des méthodes à un pas les plus courantes

Nous avons supposé que le problème à résoudre vérifie les hypothèses de Cauchy-Lipschitz. Ainsi, si la méthode utilisée est consistante d'ordre p, alors elle sera convergente à l'ordre p. On parlera alors d'ordre sans préciser de consistance ou de convergence.

Nous allons ici donner les ordres des méthodes présentées section **??**.

Méthodes d'Euler explicites et implicites

On rappelle que la méthode d'Euler explicite est consistante. En général, le schéma d'Euler explicite n'est que d'ordre 1.

Il en est de même pour la méthode d'Euler implicite.

Méthodes d'Euler améliorée et d'Euler modifiée

On peut montrer que les méthodeq d'Euler améliorée et d'Euler modifée sont d'ordre 2.

Méthode de Runge-Kutta à 4 étages.

La méthode de Runge-Kutta à 4 étages est d'ordre 4.

10.3.7 Seuil de stabilité

Considérons par exemple le problème

$$\left\{ \begin{array}{rcl} y'(t) & = & -a\ (y(t) - \sin(t)) + \cos(t), \quad 0 < t < 2\,\pi, \\ y(0) & = & 1. \end{array} \right. \tag{10.30}$$

La solution générale de l'équation différentielle qui le définit est donnée en fonction du paramètre a par

Si le paramètre a est positif, et par exemple pour $a = 10$, c'est un problème linéaire à coefficient constant avec $\dfrac{\partial f}{\partial y}(t, y) = -10 < 0$: il est stable.

Appliquons la méthode d'Euler explicite, l'alogrithme est le suivant :

Pour le pas $h = 0.19$, la solution calculée commence par osciller autour des valeurs de la solution exacte avant de se stabiliser.

Pour le pas $h = 0.20$, la solution oscille autour de la solution exacte sans se stabiliser, et finalement, pour le pas $h = 0.21$, les oscillations initiales sont constament amplifiées au cours du calcul ! La valeur 0.20 semble bien être un seuil de stabilité pour ce calcul.

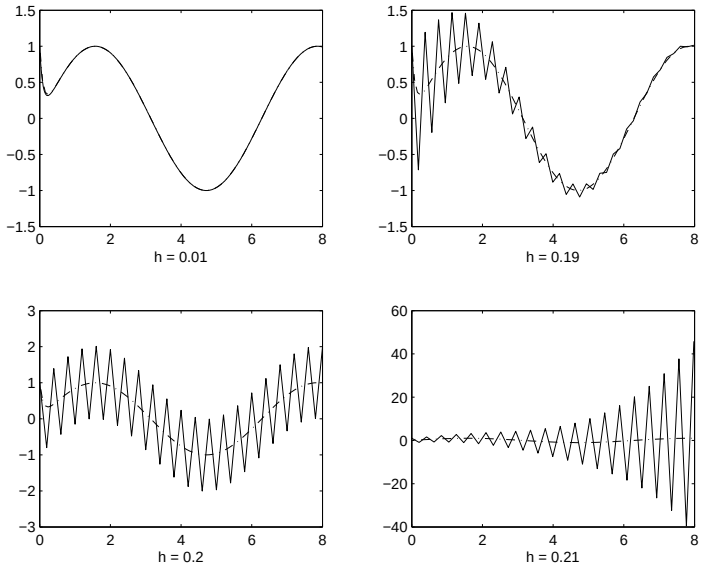


Figure 10.4 – Apparition d'instabilités pour certaines valeurs du pas.

Pour comprendre ce phénomène, on peut d'abord remarquer que la solution du problème est une combinaison linéaire de deux fonctions :

— $y_R(t) = \exp(-a\,t)$ est la solution générale de l'équation homogène ; lorsque le paramètre a devient grand, elle décrit un phénomène rapide ; son effet sur la solution a une durée de vie de l'ordre de $\dfrac{1}{a}$.

— $y_L(t) = \sin(t)$ qui dépend ici directement du second membre décrit un phenomène plus lent.

La figure 10.5 ilustre ce comportement ; on y constate que la contribution de la fonction y_R à la solution devient vite négligeable.

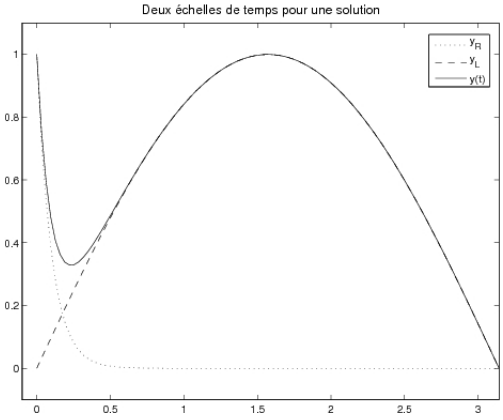


Figure 10.5 – Le problème (10.30) est raide si a est grand.

On va donc essayer de comprendre ce qui se passe en s'intéressant au calcul d'une solution de l'équation homogène associée au problème (10.30) :

$$y'(t) = -10\,y(t). \tag{10.31}$$

Sa solution générale est $y(t) = y(0)\,e^{-10t}$ et tend vers 0 lorsque t tend vers l'infini. Il faut que le schéma qu'on utilise pour calculer la solution de (10.30) respecte ce comportement.

Après avoir choisi $y_0 = y(0)$, le schéma d'Euler calcule

On constate que y_n ne tendra vers 0 lorsque n tend vers l'infini qu'à la condition que $|1 - 10\,h| < 1$, et comme le pas h est positif, il devra vérifier

Cette notion est généralisée par le théorème suivant.

Théorème 10.3.2 *Lorsqu'on applique une méthode à un pas à l'équation différentielle $y'(t) = -a\,y(t)$, en utilisant le pas h, cette méthode calcule des approximations y_n de $y(tn)$ sous la forme*

$$y_n = R(a\,h)^n\,y_0,\ n > 0. \tag{10.32}$$

La fonction $R(x)$ s'appelle **fonction de stabilité** *de la méthode.*

La méthode est inconditionellement stable si $|R(x)| < 1$ pour tout $x > 0$. Sinon, on appelle **rayon de stabilité** *de cette méthode le nombre $r > 0$ tel que $0 < x < r$ entraine $|R(x)| < 1$.*

Le seuil de stabilité de la méthode pour un problème donné sera donc $s = \dfrac{r}{a}$.

Exemple 10.3.1 *Montrer que la fonction de stabilité de la méthode RK2 (10.13) est*

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

En déduire le rayon de stabilité :

Exemple 10.3.2 *Montrer que pour la méthode d'Euler implicite, la fonction $R(x)$ est donnée par*

$$R(x) = \frac{1}{1+x}. \tag{10.33}$$

Cette fonction est telle que $|R(x)| < 1$ *pour tout* $x > 0$. *Il n'y a donc pas de restriction à la stabilité de cette méthode : elle est* **inconditionnellement stable**.